Towards Generalist Robot Learning from Internet Video: A Survey

Robert McCarthy[1], Daniel C.H. Tan[1], Dominik Schmidt[2], Fernando Acero[1], Nathan Herr[1], Yilun Du[3], Thomas G. Thuruthel[1], and Zhibin Li[1]

[1]University College London
[2]Weco AI
[3]Massachusetts Institute of Technology

Abstract

This survey presents an overview of methods for *learning from video* (LfV) in the context of reinforcement learning (RL) and robotics. We focus on methods capable of scaling to large internet video datasets and, in the process, extracting foundational knowledge about the world's dynamics and physical human behaviour. Such methods hold great promise for developing general-purpose robots.

We open with an overview of fundamental concepts relevant to the LfV-for-robotics setting. This includes a discussion of the exciting benefits LfV methods can offer (e.g., improved generalization beyond the available robot data) and commentary on key LfV challenges (e.g., challenges related to missing information in video and LfV distribution shifts). Our literature review begins with an analysis of video foundation model techniques that can extract knowledge from large, heterogeneous video datasets. Next, we review methods that specifically leverage video data for robot learning. Here, we categorise work according to which RL knowledge modality benefits from the use of video data. We additionally highlight techniques for mitigating LfV challenges, including reviewing action representations that address the issue of missing action labels in video.

Finally, we examine LfV datasets and benchmarks, before concluding the survey by discussing challenges and opportunities in LfV. Here, we advocate for scalable approaches that can leverage the full range of available data and that target the key benefits of LfV. Overall, we hope this survey will serve as a comprehensive reference for the emerging field of LfV, catalysing further research in the area, and ultimately facilitating progress towards obtaining general-purpose robots.

Correspondence to robert.mccarthy.22@ucl.ac.uk

Daniel C.H. Tan is supported by the Agency for Science, Technology, and Research (A*STAR), Singapore.

Fernando Acero and Nathan Herr are supported by the UKRI CDT in Foundational AI [EP/S021566/1] at the UCL Centre for Artificial Intelligence.

Contents

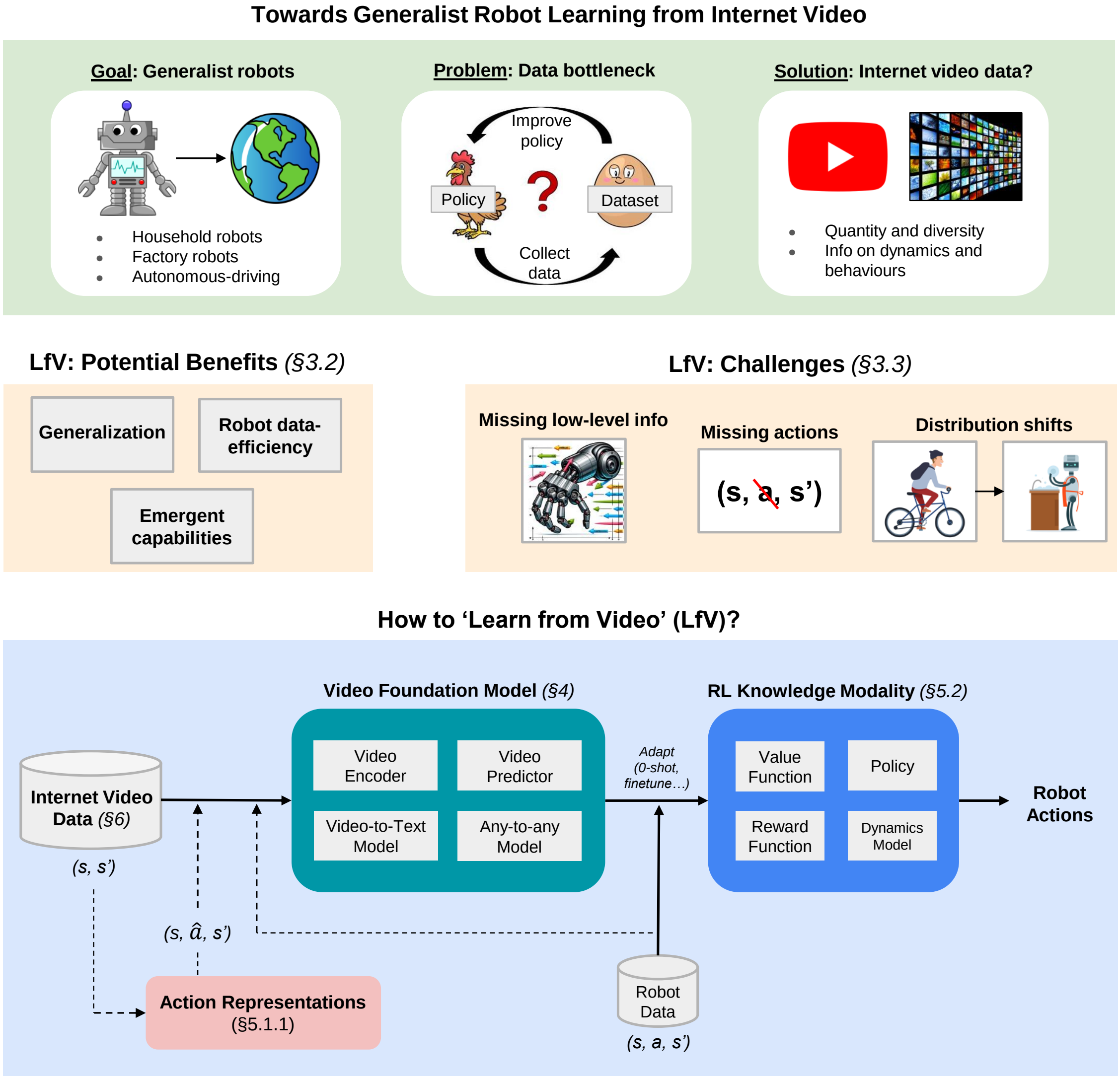


Figure 1: An overview of the key narratives, concepts, and taxonomies contained in this survey on Learning from Videos (LfV) for generalist robotics. The top green box presents the high-level motivation behind LfV approaches. The middle orange boxes highlight the potential benefits LfV can provide (Section 3.2) and the challenges in LfV (Section 3.3). The bottom blue box visualises possible components in a pipeline for learning from large-scale internet video, as per the taxonomies presented in the survey. Large internet video datasets (Section 6) can be used to help pretrain (video) foundation models (Section 4). These foundation models can further be adapted into reinforcement learning (RL) 'knowledge modalities' and used in the robot domain (Section 5.2). The diagram additionally highlights that action representations (Section 5.1.1) can be used to mitigate the issue of missing action labels in video.

1 Introduction

This survey is concerned with the goal of obtaining *generalist robots*. These are robots capable of performing a diverse range of physical tasks in unstructured real-world environments. Such robots would be highly useful and have a wide range of commercial applications (e.g., household robots, factory robots, or autonomous driving). Nevertheless, the generalist robot setting presents several challenges. First, a generalist robot must be highly competent. This includes maintaining capabilities ranging from the high-level (e.g., reasoning and planning) to the low-level (e.g., dexterity and skill). Second, to operate in unstructured settings, a generalist robot must rely on imperfect partial observations (e.g., visual and tactile sensing) to perceive the world.

How could we obtain such a robot? Classical robotics techniques are insufficient as they typically rely on hand-crafted idealisations of physics models and cannot usually handle unstructured and unseen scenarios [Krotkov et al., 2018]. In contrast, machine learning (ML) techniques are more promising and have been increasingly used in robotics, leading to the coining of the term *robot learning* [Argall et al., 2009, Peters et al., 2016, Kroemer et al., 2021, Ibarz et al., 2021]. Now, it is commonly argued that progress in ML is driven by improvements in data, algorithms, and computing power. Fortunately, the cost of compute is ever decreasing [Moore, 1998, Mack, 2011], and highly effective algorithms have recently been developed — including expressive deep learning architectures, such as transformers [Vaswani et al., 2017] and diffusion models [Ho et al., 2020] — whose performances consistently and predictably improve with increased compute and data [Kaplan et al., 2020]. Combining these algorithms with massive, diverse datasets scraped from the internet has led to remarkable improvements in language understanding and generation [OpenAI, 2023], image generation [Betker et al., 2023], and, most recently, video generation [Brooks et al., 2024].

Promisingly, these deep learning methods are transferable to robotics [Brohan et al., 2022, Team et al., 2023b]. However, unlike other domains, robotics is missing a key ingredient required for success: suitably large, diverse datasets. Indeed, robotics faces a *chicken-and-egg* problem. First, we cannot easily collect real-world robot data due to the limited capabilities of our robots. These limited capabilities mean deploying robots to collect data can be ineffective and dangerous. Subsequently, we cannot easily improve our robots due to the lack of data. Thus, arguably, data is currently *the* key bottleneck to progress in robotics.

How can we overcome this data bottleneck? To give insight into potential solutions, we now briefly discuss the primary sources of data for robotics. (1) *Real robot data:* This is the exact data we want. With high-quality real robot data, supervised learning or offline reinforcement learning (RL) can be used to train our robot control policy. However, collecting real-world robot data is expensive and difficult, regardless of whether collection is performed by human teleoperation or an automated policy. (2) *Simulated robot data:* Compared to real-world data collection, simulated collection is significantly faster and cheaper [Kaufmann et al., 2023a]. However, simulation comes with a number of issues. Simulated physics can be inaccurate. Additionally, creating a diversity of simulated environments and tasks suitable for training a generalist policy is not trivial. Moreover, we still often lack access to an automated policy capable of collecting the simulated data. (3) *Internet data:* The internet is a massive and diverse data source. It has laid the foundation for recent progress in deep learning [OpenAI, 2023, Betker et al., 2023]. Internet text, image, and video data contain vast amounts of information relevant to a generalist robot. However, internet data is not trivially or directly applicable to robotics. This is due to distribution shifts between internet data and the robot domain, and crucial missing information in internet data (e.g., text contains no visual information, whilst video contains no action labels).

Now, while any practical approach may utilise all three of the above data sources, in this paper we focus on learning from internet data. Specifically, our interest lies in *internet video data.* Our reasoning here is threefold.

1. *Information content of internet video.* The information content of video is highly relevant to generalist robotics. Compared to text or image data, video can uniquely offer information about the physics and dynamics of the world, and information about human behaviours and actions [Yang et al., 2024]. Like images, video can provide visual and spatial information. Meanwhile, video can ground knowledge and concepts learned from text data. Crucially, internet video has excellent coverage over many behaviours and tasks relevant to a generalist robot. For example, there are many videos of humans performing household chores, which would be relevant to a general-purpose household robot.

2. *Quantity and diversity of internet video.* There are *huge* quantities of video data freely available on the internet (e.g., YouTube alone contains over ten thousand years of footage [Sjöberg, 2023]). Importantly, this data is highly diverse. The largest open-source robot dataset [Padalkar et al., 2023] pales in comparison (see Figure 9), both in terms of quantity and diversity of the data.

3. *Internet video is relatively untapped.* The use of real and simulated robot data has been extensively explored [Brohan et al., 2022, Team et al., 2023b, Akkaya et al., 2019, Kaufmann et al., 2023a]. Meanwhile, internet text and image data have been heavily exploited to train foundation models [OpenAI, 2023, Betker et al., 2023], and leveraging these text and image foundation models for robotics has become increasingly common [Ahn et al., 2022, Liang et al., 2023, Brohan et al., 2023, Shah et al., 2023]. However, the use of internet video data is in more nascent stages, both in the general ML literature, and in robotics. As such, there

are opportunities for rapid progress to be obtained via research focused on utilising internet video data.

Given its abundant quantities and relevant content, internet video data can help mitigate the data bottleneck issue in robotics, and drive progress towards creating generalist robots. More specifically, we hope to obtain the following benefits from internet video: (1) improved generalization beyond the available robot data, (2) improved data-efficiency and performance in-distribution of the robot data, and, speculatively, (3) obtain emergent capabilities not attainable from robot data alone. Indeed, recent progress in the emerging field of LfV has been promising, demonstrating evidence of these benefits. This has included works that leverage large-scale video prediction models to act as robot dynamics models [Yang et al., 2023c, Bruce et al., 2024], or works that leverage both robot data and internet video to train foundational robot policies [Sohn et al., 2024].

Nevertheless, utilising internet video for robotics comes with a number of fundamental and practical challenges. First, in general, video is a challenging data modality. Video data is high-dimensional, noisy, stochastic, and poorly labelled. These issues have seen progress in video foundation models lag behind that of language and image models. Second, utilising video data specifically for robotics introduces its own set of issues. Video lacks information critical to robotics, including explicit action information and low-level information such as forces and proprioception. Moreover, there may be various distribution shifts between internet video and the downstream robot setting, including disparities in environments, embodiments, and viewpoints. Given these challenges, we outline two key questions for LfV research:

1. *How to extract relevant knowledge from internet video?*
2. *How to apply video-extracted knowledge to robotics?*

In this survey (see Figure 1), we review existing literature that attempts to answer these questions. For the first question, we survey video foundation modelling techniques promising for extracting knowledge from large-scale internet video. We note that improved video foundation models, and video foundation model pipelines customized for robotics, will likely be a key driver of future LfV progress. For the second question, we perform a thorough analysis of literature that leverages video data to aid robot learning. We taxonomise this literature according to which RL Knowledge Modality (KM) (i.e., which of representations, policy, dynamics model, reward function, or value function) directly benefits from the use of video data. Additionally, we review common techniques used to mitigate LfV challenges, such as the use of action representations to address missing action labels in video.

We conclude by discussing problems and opportunities for future LfV research. This includes advocating for scalable approaches that can best provide the promised benefits of LfV. Here, we recommend targeting the policy and dynamics model KMs. Additionally, we discuss directions for utilising video foundation model techniques for LfV, before touching on directions for overcoming key LfV challenges.

These promising opportunities, combined with encouraging recent advances in LfV [Yang et al., 2023c, Bruce et al., 2024], strongly suggest that the promised benefits of LfV are well within reach. We hope this comprehensive survey can encourage and inform future LfV research, ultimately serving to accelerate our progress towards the creation of generalist robots.

Survey structure. The remainder of the paper proceeds as follows.

- *Background (Section 2).* First, we introduce key RL concepts, including formalising the RL setting and providing more detail on RL KMs (Section 2.1). Second, a broad overview of relevant machine learning literature provides the reader with additional context to LfV (Section 2.2). Finally, details of related surveys and papers are discussed (Section 2.3).
- *LfV-for-robotics: Preliminaries (Section 3).* Here, we give a more detailed introduction to the LfV setting, providing useful preliminary information in advance of our literature review in Sections 4 and 5. We outline the potential benefits of LfV, the challenges in LfV, and metrics for evaluating LfV methods.
- *Towards Video Foundation Models (Section 4).* Improved foundational video models and techniques will be vital for future progress in LfV. As such, we dedicate this section to surveying literature regarding three relevant types of video foundation models: video encoders (Section 4.1), video prediction models (Section 4.2), and video-to-text models (Section 4.3). Associated limitations and promising future directions are discussed.

- *LfV-for-robotics: Methods (Section 5).* This section offers a review of literature that utilizes video data to enhance robot learning. First, we identify and taxonomise common strategies used to address LfV challenges (Section 5.1). This includes methods that use alternative representations of actions to tackle the missing action label problem (Section 5.1.1), and methods that aim to explicitly address LfV distribution shifts (Section 5.1.2). Second, we present our main taxonomy of the LfV-for-robotics literature (Section 5.2), categorising methods based on which component of the RL algorithm benefits from the use of video. For each component, we detail the different strategies in the literature and discuss their corresponding advantages, disadvantages, and promising directions.
- *Datasets (Section 6).* High quality video datasets are essential for LfV progress. Here, we outline the desiderata for LfV video datasets (Section 6.1), provide in-depth details on techniques for curating video data (Section 6.2), and review and critique relevant existing video datasets (Section 6.3).
- *Benchmarks (Section 7).* We discuss how an LfV benchmark should be designed (Section 7.1), and present details of relevant existing benchmarks whilst suggesting LfV-specific improvements (Section 7.2).
- *Challenges & Opportunities (Section 8).* We discuss the key challenges and opportunities for future LfV research, as identified during our analysis of the literature. First, we give high-level recommendations for future LfV research (Section 8.1). Second, we discuss in more detail how video foundation model techniques can be benefit LfV (Section 8.2). Third, we identify paths towards overcoming key LfV challenges (Section 8.3). Finally, we direct attention towards challenges that may not be resolved via naive scaling to internet video data (Section 8.4).
- *Conclusion (Section 9).* We conclude a summary of the key takeaways of the survey.

Contributions. We now summarize the key contributions of this survey.

- *Advocacy for Learning from Videos (LfV):* This survey highlights the promise of LfV methodologies for developing general-purpose robots, and should serve to encourage further research in the area. Specifically, we advocate for approaches that can scale to large, diverse internet video datasets.
- *Formalization of Fundamental Concepts:* We explicitly discuss and formalize fundamental LfV concepts and notions in a single contained document.
- *Enumeration and Taxonomization:* We synthesise the relevant literature into useful and comprehensive taxonomies. This can facilitate a holistic understanding of the LfV research landscape, and provides a structured framework in which future LfV research can be placed and assessed.
- *Critical Analysis:* Throughout the survey, we conduct critical analyses and discussions of existing approaches; identifying their advantages, disadvantages, and corresponding promising future directions.
- *Identification of Key Challenges and Opportunities:* After a thorough analysis of the LfV setting and literature, we outline key challenges and opportunities for future LfV research. This should serve to encourage further progress in LfV along these directions.

2 Background

In this section we introduce relevant formalisms and concepts related to reinforcement learning (RL) (Section 2.1), and discuss relevant prior work in ML (Section 2.2) and LfV-related surveys (Section 2.3).

2.1 Reinforcement Learning

In this section, we first formalise the reinforcement learning (RL) setting. We then summarise some RL concepts introduced by Wulfmeier et al. [2023] that we make use of throughout the rest of the survey. Specifically, we will introduce the notion of a reinforcement learning *Knowledge Modality (KM)*, and will refer to several *mechanisms of transfer* for transferring pretrained KMs to a downstream domain.

Formalism. In reinforcement learning (RL), an agent observes its environment, takes an action, and receives a reward after the state of the environment changes. This can be formalised as a Markov Decision Process (MDP) consisting of the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the transition probability $p(s_{t+1}|s_t, a_t)$ of reaching state $s_t + 1$ from state s_t when executing action a_t. The agent's behaviour is given in terms of a policy $\pi(a_t|s_t)$. When aspects of the state cannot be observed, the environment is termed a Partially Observable MDP (POMDP), and the agent only has access to observations $o_t \in \mathcal{O}$ that are partial mappings of the state $o_t = f(s_t)$. Unless stated otherwise, we will always assume a generalist robot must operate in a POMDP. Throughout this paper we will simplify the notation and denote any agent observations as s_t.

The agent aims to maximise its sum of discounted future rewards, commonly referred to as its 'returns'. This is captured by the objective in Equation 1:

$$J(\pi) = \mathbb{E}_{[p(s_0), p(s_{t+1}|s_t, a_t), \pi(a_t|s_t)]_{t=0...T}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{1}$$

where γ is the discount factor, $r_t = r(s_t, a_t)$ is the reward, and $\rho(s_0)$ is the initial state distribution. The optimal policy $\pi^*(a_t|s_t)$ maximises Equation 1.

Reinforcement Learning Knowledge Modalities. An RL Knowledge Modality (KM) is some function learned from data that represents specific types of RL-related knowledge. We now give a brief summary of the KMs we will refer to throughout this survey. Notably, we use the notion of a KM to help define our main taxonomy of the LfV-for-robotics literature in Section 5.2.

Policy, $\pi(a_t|s_t)$. This is a mapping from states to actions directly describing the agent's behaviour. $\pi(a_t|s_t)$ can be learned via imitation learning from (s_t, a_t) tuples or via offline RL [Levine et al., 2020] from trajectories of (s_t, a_t, r_t, s_{t+1}) tuples.

State or State-Action Value Function, $V_\pi(s_t)$ *or* $Q_\pi(s_t, a_t)$. These functions map from states or state-action pairs to the expected future return when acting under a particular policy π. The optimal value functions V^* and Q^* are the value functions for the optimal policy π^*. Given transition tuples (s_t, a_t, r_t, s_{t+1}) for Q, or (s_t, r_t, s_{t+1}) for V, the value functions can learned using Monte Carlo estimates (determined over complete trajectories), or Temporal Difference estimates (determined over individual transitions) [Sutton and Barto, 2018]. The value function estimates the quality of states or actions, and thus can be used to aid planning [Schrittwieser et al., 2019] or learning of the policy [Lillicrap et al., 2015, Schulman et al., 2017].

Dynamics Model, $p(s_{t+1}|s_t, a_t)$. This function predicts the next state given the current state and action. It captures the underlying transition dynamics of the agent's embodiment and its environment, and can be learned using transition tuples of the form (s_t, a_t, s_{t+1}). A dynamics model can be used: (i) for planning – for example, via model predictive control (MPC) [Garcia et al., 1989, Hafner et al., 2018]; (ii) to generate synthetic data for improved sample efficiency [Sutton, 1990, Janner et al., 2019]; or (iii) to help generate value gradients for policy training by backpropagating through a learned dynamics model [Hafner et al., 2023].

Reward Model, $r(s_t, a_t)$. This function predicts the reward provided by the environment for a specific state-action pair. Note, other formulations, such as $r(s_t, s_{t+1})$ or $r(s_{t+1})$ are also acceptable. Depending on the formulation, models of the reward may be learned via supervised learning on tuples containing some, or all, of the following information: (s_t, a_t, s_{t+1}). In some cases, reward models can be inferred from expert reward-free trajectories, e.g., via inverse reinforcement learning [Arora and Doshi, 2021].

Representations. More generally, useful knowledge can be stored in the form of learned representations. Such representations can be learned via any learning objective, including whilst learning any of the above KMs. Representation transfer (see more details in the following paragraph), is a common method for transferring knowledge from one dataset to another setting [Nair et al., 2022, Majumdar et al., 2023].

Transfer Mechanisms. RL KMs can be learned from a source MDP or dataset and subsequently transferred to a target MDP. In the case of LfV, this involves learning a KM from D_{video} and adapting it to the robot MDP using D_{robot}. There are several possible transfer mechanisms here. Wulfmeier et al. [2023] distinguish between direct and indirect mechanisms.

Direct transfer mechanisms. (1) *Generalisation and zero-shot transfer* involves using the pre-trained KM directly in the target setting, without any further fine-tuning or adaptation [Escontrela et al., 2023]. (2) *Fine-tuning* involves continuing the training of the KM on data from the target domain. (3) *Representation transfer* involves a new model being defined that is composed of parts of the pretrained KM and elsewhere is randomly initialised [Nair et al., 2022, Majumdar et al., 2023] (see Figure 7). This new model is then trained using data from the source domain. We note that, unlike [Wulfmeier et al., 2023], we may use the terms fine-tuning and representation transfer interchangeably in this survey. (4) *Hierarchy: conditioning* involves using the pretrained KM to condition a new KM being trained in the target MDP [Schmidt and Jiang, 2023, Wang et al., 2023a, Wen et al., 2023]. (5) *Hierarchy: composition* involves composing solutions from multiple pretrained KMs [Wulfmeier et al., 2019]. (6) *Meta-learning* methods pretrain the KM such that it can adapt quickly to unseen domains. General categories of meta-learning include gradient-based approaches [Finn et al., 2017] and in-context learning [Akkaya et al., 2019, Laskin et al., 2022].

Indirect transfer mechanisms. (1) *Transfer via (auxiliary) objectives:* Here, the pretrained KM is used to help define a learning objective for the downstream KM. For example, one can distill knowledge by training with regularisation towards the previous KM [Tirumala et al., 2022, Ye et al., 2023]. (2) *Transfer via data (e.g., by shaping exploration):* Here, for example, we can use previous policies for data collection in the target domain, and learn a new KM off-policy from this experience [Campos et al., 2021].

2.2 General Machine Learning

Here, we give a general overview of machine learning literature relevant to the goal of utilising internet video data to help obtain generalist robots. This provides additional context to the video foundation model research (Section 4) and LfV-for-robotics research (Section 5.2) we review later in this survey.

Foundation Models. Progress in deep learning has continually been driven by scaling up datasets and model sizes [Krizhevsky et al., 2012, He et al., 2016, Radford et al., 2019, Brown et al., 2020]. 'Foundation models' are large models pretrained on large diverse datasets. They demonstrate relatively general capabilities (within their respective domains). Whereas previously ML research would train task-specific models, it is now common to employ general foundation models (either zero-shot or finetuned) to solve wide varieties of downstream tasks. It is important to note here that, in addition to scaling, improved model architectures (i.e., transformers [Vaswani et al., 2017] and diffusion models [Ho et al., 2020]) have been crucial for advances in foundation models. We now give an overview of foundation model research across several domains.

Language. In the domain of natural-language-processing (NLP), foundational large language models (LLMs) have led to remarkable progress [OpenAI, 2023]. Huge, diverse text datasets scraped from the internet have been used to pretrain transformer models with billions of parameters via simple self-supervised next-token prediction objectives [Brown et al., 2020, Narang and Chowdhery, 2022]. This pretraining scheme allows the model to learn vast amounts of declarative and procedural knowledge from internet data. These models can be further finetuned via supervised learning on smaller high-quality datasets, and via reinforcement learning from human or AI feedback (RLHF) [Ouyang et al., 2022, Bai et al., 2022]. This finetuning results in commercially viable models [OpenAI, 2023, Team et al., 2023a] that can be used, for example, as a chatbots or coding assistants.

Images. Image understanding and generation has similarly benefited from foundational training schemes. Deep learning exploded in popularity after successful image recognition results were obtained via the use of the relatively large ImageNet dataset [Deng et al., 2009, Krizhevsky et al., 2012]. Image recognition was further improved via training dual-encoder image-language models on large image-text datasets scraped from the internet [Radford et al., 2021]. More recently, text-conditioned image generation has excelled due to the use of large diffusion models and internet-scale data [Betker et al., 2023]. Recent efforts have sought to train foundational vision-language models (VLMs) that can take images and text as input, and output text [Alayrac et al., 2022, OpenAI, 2023, Team et al., 2023a]. Results here are promising, but these models still struggle with fine-grained and spatial visual understandings [Cui et al., 2023]. Note, we discuss *video* foundation models efforts in Section 4.

Agents. Preliminary efforts have been made to develop agentic foundation models that can output low-level actions. When such models handle visual and textual inputs, they are sometimes referred to as vision-language-action models (VLAs). Three distinct approaches have been seen here. (1) An internet-pretrained VLM can be finetuned on action-labelled robot data [Brohan et al., 2023].

(2) A sequence model can be jointly pretrained on internet data and action-labelled agentic data [Reed et al., 2022, Sohn et al., 2024]. (3) A large model can be trained solely on agentic (robot) data [Brohan et al., 2022, Team et al., 2023b]. We note that the agentic datasets used in these approaches have been small relative to internet-scale datasets. In general, these approaches have used supervised learning (i.e., imitation learning) on the agentic data. In related work, LLMs have been prompted to act as agents or planners [Ahn et al., 2022, Yao et al., 2022, Park et al., 2023].

Deep Reinforcement Learning. Progress in deep reinforcement learning, including the development of improved on-policy [Schulman et al., 2017] and off-policy [Lillicrap et al., 2015, Haarnoja et al., 2018] continuous control algorithms, has been promising for robotics. However, online RL is impractical for real-world robotics; it is time-consuming, costly, and potentially dangerous. Nevertheless, solutions have been proposed in the RL literature to address these issues.

Offline RL algorithms have been proposed for learning from previously collected data and avoiding issues with online learning [Levine et al., 2020, Kumar et al., 2020, Zhou et al., 2021]. Model-based RL approaches – which learn an explicit model of their environment – have been proposed to improve data-efficiency (and overall performance) versus model-free counterparts [Janner et al., 2019, Hafner et al., 2023]. When learning from pixel observations, efficiency can be improved via auxiliary learning objectives [Yarats et al., 2019], pretraining on prior image data [Wang et al., 2022a], and data-augmentation [Yarats et al., 2021]. Recently, the use of modern diffusion architectures has improved learning from diverse offline data [Wang et al., 2022c], whilst the use of transformer architectures has allowed for in-context learning [Lee et al., 2023, Laskin et al., 2022]. Elsewhere, curiosity-based exploration methods have previously been proposed to accelerate online RL [Burda et al., 2018, Pathak et al., 2017].

Scaling Robot Learning. A lack of available data, combined with the impracticalities of real-world online RL, has limited progress in robot learning. We now discuss a number of proposed approaches that seek to scale up robot learning in light of these issues.

Simulation. The use of simulation has often been proposed as a solution to the difficulties of real-world RL. This has led to impressive results in narrow settings, including legged locomotion [Zhuang et al., 2023] and drone racing [Kaufmann et al., 2023a]. However, simulated learning presents a number of issues. (1) Innaccuracies in low-level simulation physics creates a 'sim-to-real' gap [Zhao et al., 2020] that must be overcome. A common solution here is to employ domain randomization in simulation [Tobin et al., 2017]. (2) Manually creating a suitable diversity of simulated environments and tasks for generalist robotics is a challenge. Recent works seek to tackle this using procedurally generated environments [Deitke et al., 2022], or LLM-assisted environment design [Xian et al., 2023]. (3) Even if the previous issues were resolved, we would still lack a policy capable of collecting high-quality data in the simulated environment. Solutions here have included using humans to collect the data [Mees et al., 2022], or using LLMs with access to privileged simulation information as policies [Ha et al., 2023].

Scaling real world data collection. Recent efforts have sought to collect larger real-world robot datasets. Brohan et al. [2022] use human teleoperation to collect a dataset of 130k short episodes covering 700 table-top manipulation tasks. Khazatsky et al. [2024] collect a similar sized dataset, but with improved diversity. Padalkar et al. [2023] pool data from 33 different academic labs to create a dataset consisting of 22 different embodiments, 500 skills, and 150,000 tasks across more than 1 million episodes. Other work has investigated methods for automating data collection to improve scalablity versus human teleoperation [Bousmalis et al., 2023, Ahn et al., 2024, Yang et al., 2023a]. For example, Ahn et al. [2024] use VLMs and LLMs to orchestrate a fleet of data collecting robots. Finally, we note that several commercial robotics companies have demonstrated evidence of infrastructure for large-scale robot data-collection [Sohn et al., 2024, Jang, 2024].

Internet data. Robot learning can be aided via the use of internet data. Note, this may be done indirectly via the use of pretrained foundation models. Image and video data has been used to pretrain visual representations for robotics [Wang et al., 2022a, Nair et al., 2022]. Foundational VLMs and LLMs have been used to help define reward functions for the robot learner [Tam et al., 2022, Du et al., 2023c, Yu et al., 2023c, Klissarov et al., 2023]. LLMs have been employed as planners in long-horizons tasks, leaving execution of low-level skills to specialized robot controllers [Ahn et al., 2022, Huang et al., 2022]. Finally, as touched on above, internet data has been used to help train agentic robot foundation models [Brohan et al., 2023, Sohn et al., 2024]. We elaborate on how internet *video* data has been used to aid robot learning in Section 5.2.

2.3 Related Surveys

Here, we note prior surveys and works with content directly related to ours. This is done to guide the reader to other relevant works, and to provide further context to the contributions of our work.

Reinforcement and robot learning. In the realm of RL, Wulfmeier et al. [2023] highlight the promise of transferring pretrained knowledge from a source to a target domain. The notions of RL KMs and mechanisms of transfer are introduced here, as we detail in Section 2.1. Other RL-centric surveys relevant to the LfV setting include a review of offline RL methods [Prudencio et al., 2023], an analysis of zero-shot generalization in RL [Kirk et al., 2023], and surveys regarding learning from demonstration data [Argall et al., 2009, chaandar Ravichandar et al., 2020]. Robot learning and RL in robotics have previously been surveyed by Kober et al. [2013], Peters et al. [2016], whilst Kroemer et al. [2021] focus on robot learning for manipulation.

Foundation models for robotics. Other surveys have emphasised the promise of foundation models for robotics. Hu et al. [2023b] assess how foundational VLMs and LLMs can be used for robotics, and additionally detail methods for developing robot-specific foundation models. Yang et al. [2023d] assess how foundation models can be used for general decision making applications, including robotics. They briefly touch on LfV techniques, including the use of video generation models as policies (see Section 5.2.2).

Machine learning for video. There exist relevant surveys of the video ML literature. Schiappa et al. [2023] present an analysis of methods for self-supervised learning from video data. Ming et al. [2024] survey existing video prediction methods. Tang et al. [2023] survey methods for utilising LLMs in video understanding, whilst Zhang et al. [2024] provide more details on how LLMs can be adapted to additional modalities.

Learning from video for robotics. There has also been work directly focused on LfV. First, we note Torabi et al. [2019], who review imitation learning over observational data. However, this focuses solely on methods that assume access to expert demonstrations (and thus are not scalable to internet video). Recently, Yang et al. [2024] advocate for the use of video (and video generation methods, in particular) as a unified interface to absorb internet knowledge and represent diverse tasks. Relevant to our work, they discuss the robotics-relevant information in video, several use-cases of video generation for robotics, and challenges in video generation that are applicable to LfV. Unlike this work, we: (i) focus solely on generalist robotics applications (versus the other potential applications of video); (ii) more broadly assess different methods for using video data in robotics (i.e., we go beyond video generation methods); and (iii) present a thorough survey of LfV-related literature.

Finally, the existing work most relevant to ours is a recent survey on video-based learning approaches for robot manipulation [Eze and Crick, 2024]. In contrast to this work, we place a stronger emphasis on the goal of obtaining generalist robots, and on the promise of approaches that can scale to large-scale internet video data. This includes presenting a more holistic analysis of the literature (including dedicating Section 4 to video foundation models), and providing different and richer taxonomies that are better suited to this emphasis.

3 LfV-for-Robotics: Preliminaries

Whilst our main review of the LfV-for-robotics literature is presented in Section 5, here we provide useful preliminary information. This includes discussions of crucial LfV concepts, which lay foundations for LfV research. We first formalise and add clarifications regarding the LfV and generalist robot settings we are interested in (Section 3.1). We then discuss the exciting potential benefits video data may provide to robotics (Section 3.2), before discussing key fundamental and practical challenges in LfV (Section 3.3). Finally, we provide recommendations regarding how LfV methods should be evaluated in light of these benefits and challenges (Section 3.4).

3.1 The LfV Setting

Formalising the LfV setting. In the learning from videos for robotics setting, we will assume access to a video dataset D_{video}. Here, we denote a video clip as $\tau = (s_0, s_1, ..., s_T)$, where τ is the full clip and each s is an observation in the form of an image. We use s to represent images for compatibility with the RL observation notation established in Section 2.1. Optionally, D_{video} may come paired with language annotations or annotations from other modalities. In general, we will also assume access to a robot dataset D_{robot}. This dataset usually contains trajectories of transition tuples (s_t, a_t, r_t, s_{t+1}) – though r_t may be missing. The goal of LfV is to learn from D_{video} in order to improve performances versus when learning only from D_{robot}.

LfV for generalist robotics. Although D_{video} may come from any source, in this survey we are primarily interested in methods that can leverage large-scale video data gathered from the internet. We will generally assume that such an internet dataset consists primarily of videos of humans and has good coverage over the full range of physical tasks that humans commonly perform.

A 'generalist robot' is a general-purpose robot that can perform a diverse range of physical tasks in unstructured real-world settings. Such settings are POMDPs where the robot must rely heavily on visual observations. Throughout this survey, unless stated otherwise, we assume the generalist robot has an embodiment and affordances similar to those of a human. We also assume we want the robot to perform tasks similar to those humans commonly perform, and to do so in a comparable manner. Under these assumptions, internet video data can be particularly useful for training a generalist robot; it provides a large and diverse set of videos featuring embodiments similar to the robot, which are performing tasks relevant to the robot.

We note some limitations to these assumptions. First, for certain robotics tasks, non-human-like embodiments may prove more effective. Second, the robot may need to perform specific tasks that are not commonly performed by humans. In these cases, internet video will be less useful for the specific embodiment and task. Nevertheless, it can still be generally informative about the world and physical behaviour. Moreover, we believe aiming for humanoid robots that can perform human-like tasks is a good starting point for generalist robot efforts.

3.2 Potential Benefits

Robotic datasets are expensive to acquire and thus are currently task-specific or relatively narrow [Padalkar et al., 2023]. In contrast, diverse video data is freely available in vast quantities on the internet. To achieve our goal of obtaining generalist robots, we advocate for the use of methods that can leverage this data in a scalable manner. In this section, we briefly outline the specific benefits we hope to obtain from methods that do so.

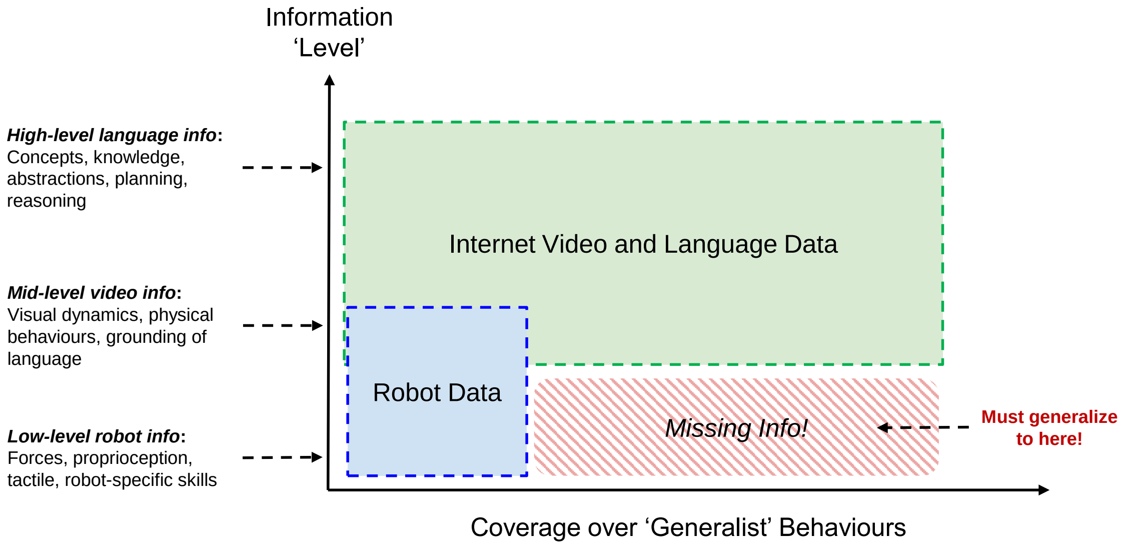

Information 'Level'
High-level language info: Concepts, knowledge, abstractions, planning, reasoning
Mid-level video info: Visual dynamics, physical behaviours, grounding of language
Low-level robot info: Forces, proprioception, tactile, robot-specific skills
Internet Video and Language Data
Robot Data
Missing Info!
Must generalize to here!
Coverage over 'Generalist' Behaviours

Figure 2: Generalization in the Learning from Videos (LfV) setting. The x-axis indicates the full set of behaviours expected from a generalist robot. The y-axis indicates the 'levels' of information contained in data. The figure demonstrates that internet data has better coverage over desired behaviours than small robot datasets, but lacks crucial low-level information essential to robotics. Generalising beyond the robot data despite this missing low-level information is a key LfV challenge. See Sections 3.2 and 3.3 for further discussion.

Generalization beyond $\mathbf{D}_{\text{robot}}$. LfV offers the exciting possibility of generalization beyond narrow robot datasets D_{robot} to the full space of tasks covered in the video data D_{video}. We now explain the rationale behind this expectation. First, consider a D_{robot} that is sufficiently diverse such that it contains most of the low-level skills or ‘atomic’ actions that will ever be required from the robot (e.g., specific grasping motions or locomotion skills). Now consider a task unseen in D_{robot} but seen in D_{video}. In this case, our combined dataset ($D_{\text{combined}} = D_{\text{robot}} + D_{\text{video}}$) contains all information required to know how to complete the the task. D_{robot} provides information regarding how to execute the low-level skills, whilst D_{video} provides higher-level information regarding how to complete the task (e.g., visually, what movements are required and what steps are involved). Thus, a suitable LfV method may be capable of generalising beyond D_{robot} to solve the task. There is preliminary evidence of this in LfV-related literature [Brohan et al., 2023, Du et al., 2023a, Wu et al., 2023a, Wang et al., 2023a]. Figure 2 explores this generalization setting in more detail.

Emergent capabilities. Learning from internet-scale video may yield capabilities qualitatively beyond what can be obtained when learning only from a narrow D_{robot}. We expect this for two reasons. First, in other domains, large quantities of internet data have allowed for unexpected ‘emergent’ capabilities [Radford et al., 2019, Brown et al., 2020]. Second, diverse internet video paired with language annotations offers a path towards stitching together the lower-level knowledge obtained from robotic and video data with the rich abstractions and comprehensive world knowledge obtained from textual data [OpenAI, 2023].

Improvements in-distribution of $\mathbf{D_{robot}}$. Finally, we also expect video data to yield improvements in tasks that (to some extent) are in-distribution of the robot dataset. First, utilising a large video dataset can allow for improved data-efficiency with respect to D_{robot} [Nair et al., 2022]. Second, LfV approaches may obtain higher absolute task performance (e.g., higher success rates) in settings in-distribution of D_{robot}, versus non-LfV approaches [Wu et al., 2023a].

3.3 Challenges

There are several challenges that may be encountered when attempting to learn from internet video for robotics. Here we discuss the key fundamental and practical LfV challenges (see Figure 3 for visualizations).

Missing action labels. A major challenge is that raw video data lacks the action labels required by existing methods for learning from demonstrations (e.g., imitation learning [Brohan et al., 2022] and offline RL [Chebotar et al., 2023]). Moreover, adding robotic action labels retrospectively to internet video data is generally not an option. This is because the low-level actions of a robot are incompatible with the unrestricted action space of heterogeneous internet video data. One solution to this problem is to use *alternative action representations* (see Section 5.1.1). Pure video data also lacks other RL-relevant metadata, such as reward labels (which can inform on the quality of the data), goal labels (which can be useful for goal-conditioning), or end-of-episode labels.

Distribution-shift. There may be a distributional shift between an internet video dataset and the downstream robot domain. This can include differences in physical embodiments, camera viewpoints, or environments. Additionally, humans in videos may be performing behaviours which are sub-optimal or irrelevant to the downstream robot task. These shifts present a challenge to deep learning methods. Mitigations to this problem include: (i) scaling to ever larger and diverse video data to aid generalization, (ii) leveraging robot data in addition to internet videos, and (iii) using explicit methods to address the distribution shifts (see Section 5.1.2).

Missing low-level information. Videos are missing crucial low-level information for robotics. For example, for certain skillful or dexterous behaviours, robots require low-level percepts such as tactile sensing, proprioception, or depth sensing. This information is not available in internet video. Thus, whilst internet video and text data may have excellent coverage over behaviours and concepts relevant to a generalist robot, it does not do so for the lowest levels of information. A key challenge in LfV is to obtain generalization beyond D_{robot}, despite the missing low-level information in D_{video}. This challenge is visualized in Figure 2.

Controllability, stochasticity, and partial observability. A challenging aspect of modelling behaviour purely from observational data is that the boundaries between the agent and the environment are unknown. In particular, it becomes impossible to distinguish which parts of transitions are affected (i.e., controlled) by the agent's actions and which effects are simply to due to the external environment or noise. This can be problematic for methods that attempt to extract action information from video (Section 5.1.1). Furthermore, stochasticity and partial-observability of the underlying environment can make it challenging to perform accurate video prediction.

High-dimensionality, noise, and redundancy. Methods that learn from or generate video data are typically computationally demanding due to the high-dimensional nature of video data. Additionally, video can contain significant noise and redundant information. This is in stark comparison to language data, which is highly compressed and structured. These characteristics make it more difficult to extract meaningful information from video data. To overcome this, many works attempt to learn and operate in a latent representation-space of video [Yan et al., 2021, Bardes et al., 2023], rather than in raw pixel-space.

Dataset limitations. Finally, the limitations of existing video datasets present a practical challenge to LfV methods. First, we note that open-source curated video datasets are still small relative to internet scales (see Figure 9). Second, we note that language annotations of video are often sparse, coarse, noisy, or non-existent. We give details on desiderata for video datasets, and methods for curating improved video data, in Section 6.

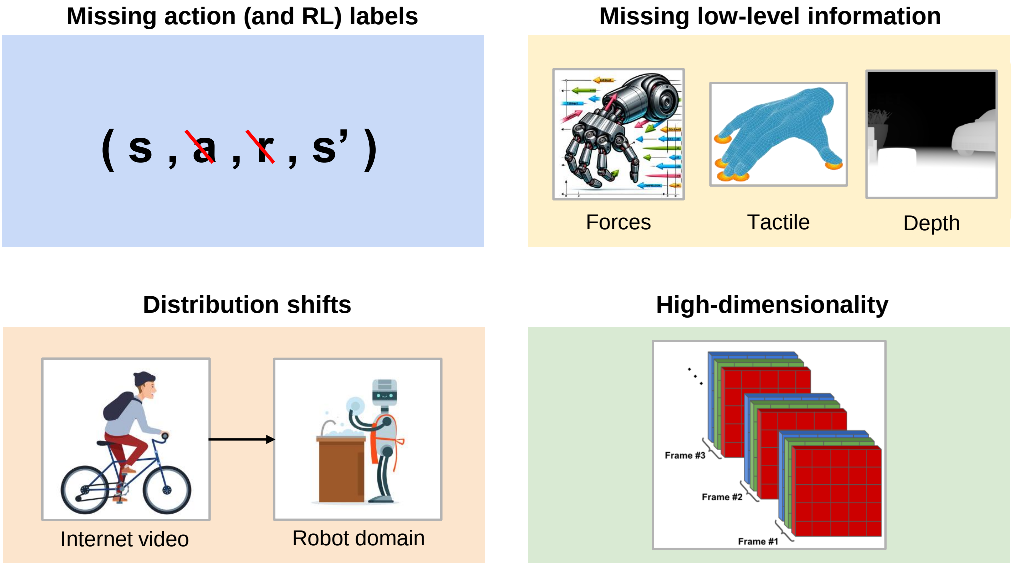


Figure 3: Key challenges in LfV (see Section 3.3) are visualised, including: missing information in video, LfV distribution shifts, and the high-dimensional nature of video data.

3.4 Evaluating LfV Methods

Here, we give guidelines on how LfV methods should be evaluated. These recommendations are predicated on our goal of scaling to diverse internet video data to help us obtain generalist robots. They are also informed by the LfV potential benefits and challenges discussed in the previous sections. We encourage the reader to bear the following criteria in mind when assessing the LfV literature reviewed in Section 5.

Scalability. First, we should assess the *scalability* of the LfV method. Below we outline characteristics of an LfV method that influence its scalability. These characteristics cannot usually be measured quantitatively, but, to an extent, can be assessed qualitatively.

1. *Can the method scale to diverse internet data, and to the generalist robot setting?* First, we note we are interested in LfV methods that can leverage the full scale of internet video, and extract as much information as possible from this heterogeneous data. We thus consider methods that make limiting assumptions on the nature of the video data to be less promising (e.g., some methods assume minimal domain shift between the video data and the robot domain [Stadie et al., 2017, Xiong et al., 2021, Baker et al., 2022]). Second, we note we

are interested in LfV methods that can be applied to unstructured generalist robot settings. Some past LfV methods are only applicable in narrow, contained robot settings [Torabi et al., 2018b, Stadie et al., 2017, Sermanet et al., 2018].

2. *Can the method benefit from advances in video foundation modelling?* There are many commercial forces that will encourage improvements in video foundational modelling in the coming years. LfV methods that can benefit from these advances [Du et al., 2023a, Yang et al., 2023c, Sohn et al., 2024] (by leveraging either the improved models or the improved techniques and datasets) are particularly promising.

Downstream performance gains. Second, we should evaluate the extent to which the LfV method can provide the potential benefits (see Section 3.2) of LfV. Here, we can obtain more concrete quantitative metrics. We now detail these metrics (which are inspired partially by those outlined by Wulfmeier et al. [2023] in the RL transfer setting):

1. *Performance in-distribution of $\mathcal{D}_{robot}$.* We can measure the performance (e.g., task success rate) of the robot in settings in-distribution of $\mathcal{D}_{robot}$, after training on $\mathcal{D}_{video}$ and a fixed-size $\mathcal{D}_{robot}$.

2. *Data-efficiency in-distribution of $\mathcal{D}_{robot}$.* We can measure the amount of data in $\mathcal{D}_{robot}$ required to reach a certain performance in settings in-distribution of $\mathcal{D}_{robot}$, when also training on $\mathcal{D}_{video}$.

3. *Generalization beyond $\mathcal{D}_{robot}$.* We can measure the performance of the robot in settings out-of-distribution of $\mathcal{D}_{robot}$, after training on $\mathcal{D}_{video}$ and a fixed size $\mathcal{D}_{robot}$.

These metrics cover the benefits of *generalization beyond* $\mathcal{D}_{robot}$ and *improvements in-distribution of* $\mathcal{D}_{robot}$, as outlined in Section 3.2. However, the benefit of *emergent capabilities* is not accounted for. Indeed, this benefit likely must be evaluated qualitatively. We finally note that it will be beneficial to use these metrics to compare LfV methods to non-LfV baselines (i.e., to equivalent methods that do not use video data). This will provide a concrete measure of the benefits $\mathcal{D}_{video}$ is providing.

4 Towards Video Foundation Models

We established some preliminaries for the LfV setting in the previous Section 3. We now turn our attention to video foundation modelling as a general means for extracting robotics-relevant knowledge (e.g., knowledge regarding physics and behaviours) from large-scale internet video data. In this section, we will review the general video foundation model literature, in advance of reviewing literature that applies video data specifically to robotics in Section 5.

Now, foundation models are large deep learning models trained on large, diverse datasets. They maintain general knowledge and capabilities useful in a wide range of downstream settings. This has most notably been demonstrated in natural language processing with large language models (LLMs) [OpenAI, 2023, Team et al., 2023a]. In this section, we focus on *video foundation models* (see Figure 4). We are interested in these models due to the following LfV use-cases:

- *Using pretrained video foundation models:* A pretrained video foundation model can be adapted for downstream robot applications. For example, a video prediction model can be used as a robot dynamics model [Yang et al., 2023c].

- *Using video foundation model techniques and datasets:* Techniques and datasets used originally for video foundation models can be customized for robotics purposes. For example, a robot foundation model can be trained on both video and robot data, using techniques inspired by video foundation modelling [Sohn et al., 2024].

We categorise the literature here according to three specific functionalities a video foundation model can provide: (i) video encoding (Section 4.1), (ii) video prediction (Section 4.2), and (iii) video-to-text generation (Section 4.3). We explain in each subsection why these functionalities are relevant to LfV. We note that, in practice, some models may perform multiple of these functions (e.g., any-to-any sequence models can perform video prediction and video-to-text generation [Liu et al., 2024b]).

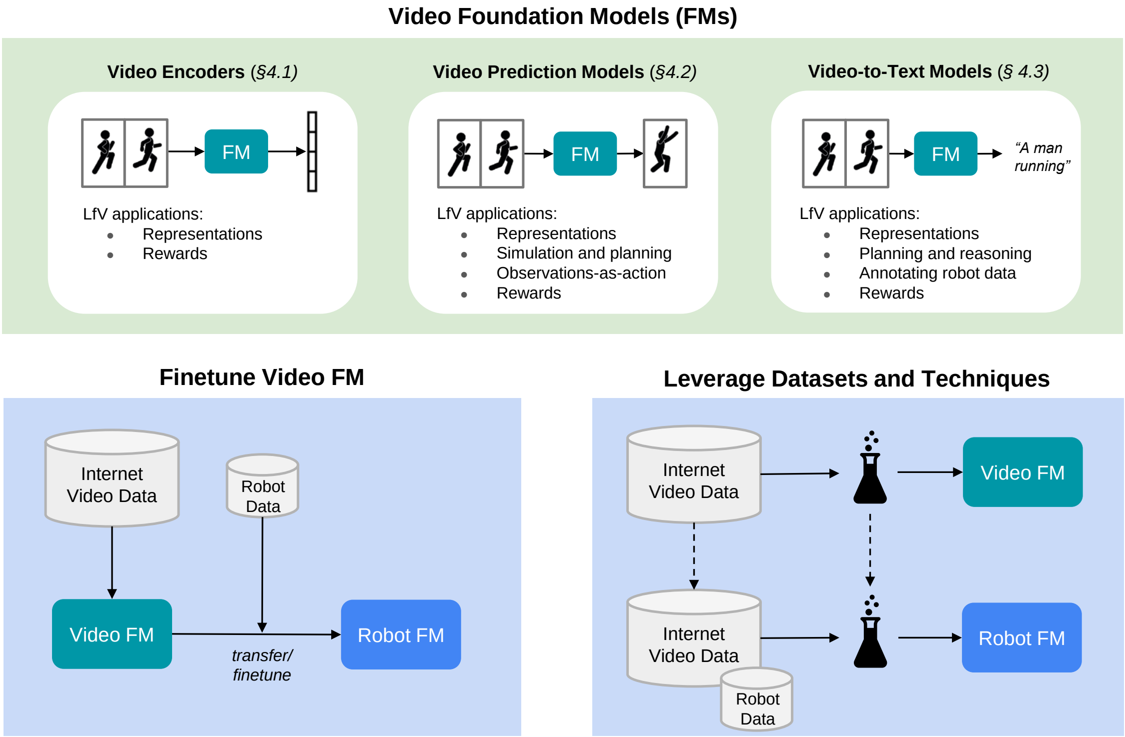


Figure 4: Video Foundation Models (see Section 4) for LfV. The top green box presents different categories of video foundation models and their potential utility to robotics. The bottom blue boxes illustrate two approaches by which video foundation modelling can contribute to LfV: (left) pretrained video foundation models can be finetuned into robot foundation models; (right) video foundation model techniques and datasets can be used to train robot foundation models.

4.1 Video Encoders

Foundational video encoders can provide robust video representations useful for downstream robotic applications. Representation transfer approaches (see Section 5.2.1) can take a pretrained video encoder and finetune it into an RL KM (e.g., into a policy) that can be used in the robot domain. Several works have used frozen pretrained video representations to help define a robot reward function [Fan et al., 2022, Sontakke et al., 2023, Nair et al., 2022] (see Section 5.2.4). In this section, we first give a broad overview techniques that can be used to train video encoders, before giving more details on recent promising techniques for training foundational video encoders.

Video Representation Learning Overview: Derive Y from $\bar{X}$, apply a supervised loss. We now give a broad overview of techniques that can be used to learn video encoders. Let $\bar{X}$ be a video clip X paired with some corresponding labels and data modalities. Most methods here can be framed as first deriving a 'label' Y from $\bar{X}$ and then using Y to define a learning objective. We now give an overview of the various ways in which Y can be defined and used for video representation learning.

- **How to Define Y?** (1) *Transforms of X:* Commonly, we can derive a pseudo-label from the video X itself. This transform could be some augmentation of the video (e.g., spatial or temporal shifts), a masked version of the video, or a slice of the video. Schiappa et al. [2023] identify a number of categories of augmentation that can be used for video data, including view augmentations and spatio-temporal-augmentations. (2) *Paired information:* Y can be derived from additional information paired with the video. This can include labels or data of another modality. Labels relevant to robotics could include the class of action performed in the video [Kay et al., 2017], object masks and bounding boxes [Ding et al., 2023a], or human hand poses and affordances [Shan et al., 2020]. Different modalities may include corresponding audio, speech transcriptions, language annotations, or 3D depth information [Grauman et al., 2021]. Finally, we note that occasionally X may be paired with another video, for example an equivalent video from a different viewpoint [Grauman et al., 2023]. (3) *Meta-data:* Meta-data associated with the video can also be used for Y. Some examples here include appearance statistics and playback speed [Schiappa et al., 2023].

- **How to Use Y?** (1) *Prediction:* Here, the learning objective is to predict Y. One option is to predict Y from X. For example, predicting a language description of a video [Karamcheti et al., 2023], or predicting the next frame in a video (i.e., video prediction [Gao et al., 2022]). Another variation is to predict Y from Y, i.e., auto-encoding. Here Y is encoded into a bottleneck (a compressed representation of the video) and then decoded [Tong et al., 2022, Villegas et al., 2022]. (2) *Joint-embeddings:* Instead of learning to predict the raw version of Y, we can define objectives where the loss is calculated via embeddings of X and Y. Contrastive learning approaches have been heavily explored here [Schiappa et al., 2023]. Noise-constrastive estimation objectives (NCE) [Oord et al., 2018] have been popular, most commonly used to contrast the video with positive and negative pairs of language descriptions [Xu et al., 2021, Zhao et al., 2024, Papalampidi et al., 2023]. Other contrastive objectives include those that use binary cross-entropy losses, or score-based approaches [Schiappa et al., 2023]. Non-contrastive joint-embedding approaches also exist. BYOL [Grill et al., 2020] define Y_1, Y_2 (two augmentations of X) and train an 'online' network to predict 'target' network embedding of Y_1 from Y_2. V-JEPA [Bardes et al., 2023] similarly define a learning objective by predicting the embedding of X from an embedding of Y.

The taxonomy presented above covers many schemes that can be used to train video encoders. However, it may neglect some learning objectives. We identify temporal difference learning from video [Bhateja et al., 2023] as one neglected objective, however there may be others.

Foundational Video Encoders: Learning Objectives. We now detail promising learning objectives and techniques for training foundational video encoders from internet-scale video data. We focus primarily on *video-first* architectures that jointly represent the spatio-temporal information in video [Arnab et al., 2021, Yu et al., 2023c, Zhao et al., 2024]. In contrast to models that rely on strong hand-crafted inductive biases to process video [Simonyan and Zisserman, 2014, Girdhar et al., 2017], we consider video-first methods to be more promising: when scaled appropriately to internet video data, these methods may learn richer, more informative representations (since they are not limited by strong inductive biases).

- **Video-language objectives.** Text annotations are useful for learning semantic representations of video. We note several objectives in the literature that leverage text annotations: (i) video-text contrastive losses [Xu et al., 2021], (ii) video-text matching [Li et al., 2023c], (iii) masked language modelling [Li et al., 2023c], and (iv) video-to-text losses [Papalampidi et al., 2023]. Video-text contrastive NCE losses have been very popular [Xu et al., 2021, Zhao et al., 2024, Papalampidi et al., 2023, Li et al., 2023c, Wang et al., 2023f]. Wang et al. [2023f] accelerate contrastive training by randomly masking input video in the early stages of training. Papalampidi et al. [2023] use a learning schedule, moving from shorter to longer videos over the course of the contrastive training. Note, the strategy for sampling negative examples and mini-batches is crucial for contrastive approaches. Zhao et al. [2024] alternate between mini-batches from different datasets during training, whilst Xu et al. [2021] use retrieval augmented sampling to chose mini-batches. Meanwhile, Bagad et al. [2023] artificially create negative samples to improve temporal representations. Elsewhere, other video-text losses have generally been used in conjunction with an NCE loss. Li et al. [2023c] combine video-text contrastive, video-text matching, and masked language modelling objectives. Yan et al. [2022], Papalampidi et al. [2023] combine a video-to-text objective with the video-text contrastive objective.

- **Masked auto-encoding (MAE).** Video MAE involves encoding and reconstructing a masked video, with the reconstruction loss acting as the learning objective [Tong et al., 2022, Wang et al., 2023c, Girdhar et al., 2022]. Tong et al. [2022] tokenise the video and mask in token space. Two important design decisions are made in this masking scheme: (i) due to temporal redundancy in video, an extremely high masking ratio (90-95%) is employed; (ii) a tube masking scheme – which extends masks along the temporal dimension – mitigates issues related to easy reconstruction in areas with minimal motion. Wang et al. [2023c] extend this approach, performing masking in the decoder to further improve computational efficiency. Li et al. [2023c] perform semantic masking, where semantically relevant parts of images (as determined by an image-language model) are prioritized for masking. Related methods use masked distillation schemes [Wang et al., 2022b], or vector-quantized auto-encoding [Yu et al., 2023a] (both of which we discuss in more detail below).

- **Vector-quantized (VQ) auto-encoding.** VQ-AEs [van den Oord et al., 2017] use codebooks to encode video to discrete representations in their AE bottleneck. The reconstruction objective can either be a pixel-error loss (i.e., a VQ-VAE [van den Oord et al., 2017]), or an adversarial loss (i.e., a VQ-GAN [Esser et al., 2020]). When deciding the encoder-decoder architecture, crucial decisions must be made regarding how to fuse the spatio-temporal information in video. A simple choice is to encode each frame individually using a 2D VQ-AE [Seo et al., 2022b]. However, this can neglect the importance of fusing information along the temporal dimension. To fuse spatio-temporal information in video, 3D convolutions have been explored [Yan et al., 2021, Yu et al., 2022], as have various spatio-temporal attention schemes [Arnab et al., 2021, Bertasius et al., 2021]. When training a video VQ-AE, Yu et al. [2023c] improves upon Yu et al. [2022] by introducing lookup-free quantization and using temporally causal 3D convolutions. Villegas et al. [2022] use a C-ViViT (a causal variation of the ViVit architecture [Arnab et al., 2021]) encoder-decoder, compressing video in space and time, whilst staying auto-regressive in time. Bruce et al. [2024] improve the computational efficiency of their VQ-AE via the use of a ST-transformer [Xu et al., 2020] encoder-decoder. These VQ-AEs models can be framed as video tokenizers and are often used to provide a compressed latent space for video prediction models [Yan et al., 2021] (see Section 4.2 for more details).

- **Distillation losses.** A number of works have explored the use of student-teacher distillation losses [Wang et al., 2022b, Zhao et al., 2024, Li et al., 2023c]. Wang et al. [2022b] pretrain separate image and video teachers (which provide better guiding spatial and temporal features respectively) and train a student to reconstruct the teacher's features. Zhao et al. [2024] freeze a video encoder from an initial video-text contrastive learning stage, and use it to provide a distillation loss during a second MAE stage. Li et al. [2023c] uses a frozen image-language model to provide a distillation loss throughout training, improving the learning of semantic features.

- **Joint-embedding Prediction Architectures (JEPA).** JEPA approaches have recently been applied to video [Bardes et al., 2023] (a brief description of JEPA approaches is found in the above paragraph). Versus pixel-based approaches, JEPA methods may better mitigate issues related to the high-dimensionality of video and noise in video. Bardes et al. [2023] show their video JEPA method to be computationally efficient and competitive under frozen downstream evaluation settings and in tasks requiring fine-grained understandings.

We note that the different objectives above can be combined to complement each other. For example, video-text objectives that capture semantic features (but suffer from noisy labels), can be combined with masked modelling objectives that better capture low-level features (and do not require language labels) [Li et al., 2023c, Zhao et al., 2024]. This is often done in a multi-stage training frame-work [Zhao et al., 2024]. Other multi-stage pipelines have been employed, including: pretraining on diverse lower-quality data before finetuning on higher-quality data [Wang et al., 2023c], or starting with shorter videos before moving to longer videos [Liu et al., 2024b].

Finally, we note that these *video-first* techniques still often bootstrap from image data; image data is generally more available and comes with improved language annotations. Some techniques here include: using image data jointly with video data during training [Papalampidi et al., 2023]; using pretrained image models to provide a distillation loss [Li et al., 2023c, Wang et al., 2022b]; or adapting pretrained image models into video models [Yan et al., 2022, Yang et al., 2023e].

SOTA models. Here, we give brief details on some representative state-of-the-art (SOTA) video encoder models.

- *VideoPrism [Zhao et al., 2024]:* A 1B parameter model is trained initially via video-text contrastive learning and then via masked video modelling. The training dataset consists of 36M well-captioned videos (closed-source) and 582M video clips with noisy parallel text (composed from open-sourced datasets [Stroud et al., 2020, Zellers et al., 2021, Wang et al., 2023f] and additional closed-source data).

- *VideoMAE-V2 [Wang et al., 2023c]:* A 1B parameter model is first trained using a masked auto-encoding objective. This is done on an unlabelled video dataset of 1.35M clips obtained from public datasets [Kay et al., 2017, Goyal et al., 2017, Gu et al., 2018, Bain et al., 2021], and from uncurated videos crawled from Instagram. After this intial stage, the model is supervised finetuned on labelled data from Kay et al. [2017].

- *Short/Long-ViViT [Papalampidi et al., 2023]:* A 1B parameter model is pretrained using a video-language contrastive loss. The pretraining dataset consists of 27M video-text pairs including closed-source data and data from the open-sourced Miech et al. [2019]. The model is then combined with a pretrained LLM head and finetuned on a video-to-text objective.

- *InternVid [Wang et al., 2023f]:* 50M video-text pairs from the InternVid dataset [Wang et al., 2023f] are used to train a model via a video-language contrastive loss. The model is open-sourced.

- *UMT [Li et al., 2023c]:* A 300M parameter model is trained in two stages. First, unmasked token alignment is performed using a pretrained image-language model as a teacher and high quality video data based on Kay et al. [2017]. Next, video-text contrastive, video-text matching, and masked language modelling objectives are added. The second stage leverages 25M image-text and video-text [Bain et al., 2021, Sharma et al., 2018] pairs. The model is open-sourced.

- *V-JEPA [Bardes et al., 2023]:* A 630M parameter model is trained using a JEPA objective. The dataset consists of 2M videos from [Miech et al., 2019, Kay et al., 2017, Goyal et al., 2017]. The model is open-sourced.

Discussion. Promising initial steps have been taken towards obtaining foundational video encoders [Zhao et al., 2024]. However, a major bottleneck to further advances is the computational difficulties of processing large-scale video data. Indeed, many of the methods above employ specific mechanisms (e.g., masking [Wang et al., 2023c]) to mitigate computational demands. Papalampidi et al. [2023] perform a relevant analysis of the trade-offs between model accuracy and reducing computational demands. Another bottleneck is the quality of the video datasets themselves. Video-text datasets are used to train several leading models, but language annotations are often of low quality. Finally, we note that representing long video sequences remains a major challenge [Papalampidi et al., 2023]. Improved architectures (e.g., recent state-space models [Li et al., 2024]) may help with both computational demands and long-term video modelling.

4.2 Video Prediction Models

We dedicate this section to discussing models that can perform next-frame video prediction $p(s_{t+1}|s_{t-k:t})$. From internet video, video prediction models can learn information regarding world dynamics and human behaviours. They can thus be useful for downstream robotics in the following ways:

- *Dynamics:* A video prediction model can be adapted into a robotics dynamics model to serve as a planner [Du et al., 2023b] or simulator [Yang et al., 2023c] (see Section 5.2.3).

- *Policies:* The video prediction objective implicitly allows the model to learn the distribution of behaviours in the video dataset [Escontrela et al., 2023]. As such, video prediction models can act as policies by generating videos of proposed action-sequences [Du et al., 2023a] (see Section 5.2.2).

- *Representations:* Due to the relevant information they represent, video predictors can be used for LfV representation transfer approaches [Wu et al., 2023a].

- *Rewards:* Finally, a reward signal can be defined that encourages the robot to match the behaviour expected by the video predictor [Escontrela et al., 2023] (see Section 5.2.4).

We elaborate more on how video prediction models can be applied to robotics in Section 5.2. We note that LfV methods that can utilise video prediction models are particularly promising as video generation has a number of commercial applications which will likely drive progress in capabilities in the near-future [Brooks et al., 2024].

In this section, we give an overview of literature relevant to foundational video prediction models. We focus on the most promising recent techniques: diffusion, autoregressive transformers, and masking transformers. We note that we are most concerned with models that can perform *next-frame video prediction* $p(s_{t+1}|s_{t-k:t})$. However, also relevant are models that can more generally perform some form of *conditional video generation* $p(\tau|c)$ (where τ is a video clip and c is some conditioning information) [Yang et al., 2024].

Technique: Diffusion. Diffusion models [Ho et al., 2020] have become popular in video generation due to their expressiveness and controllability. Indeed, many SOTA video generation models leverage diffusion [Brooks et al., 2024, Bar-Tal et al., 2024]. We now give an overview of diffusion-based video prediction methods.

Pixel vs latent-space prediction. Some works perform the diffusion process directly in pixel-space [Ho et al., 2022b, Singer et al., 2022, Ho et al., 2022a], whilst others do so in a learned latent space [Brooks et al., 2024, Bar-Tal et al., 2024, Blattmann et al., 2023a]. Blattmann et al. [2023b], Zhou et al. [2022] use a pretrained image VQ-AE to define the latent space, but add temporal layers into the pretrained decoder to reduce flickering artifacts. More information on learning VQ-AEs for video can be found in Section 4.1. Predicting in latent space aids computational efficiency, but can result in worse alignment of the generated video with the text prompt. Zhang et al. [2023a] combine both pixel-level and latent diffusion into a hybrid model in an attempt to obtain the benefits of both.

Leveraging image data. Video data is more limited in quantity and quality than image data. In particular, video often comes with inferior text annotations. As such, most diffusion video prediction models attempt to utilise image data in some way. Ho et al. [2022b,a] jointly train on image and video data. Bar-Tal et al. [2024], Ge et al. [2023], Blattmann et al. [2023b] take a pretrained image diffusion model and "inflate" it into a video model. This involves modifying the architecture to include temporal connections and finetuning it on video data. Meanwhile, Dai et al. [2023] factorise the video generation process; first generating an image with a pretrained image diffusion model, then generating a video conditioned on the image.

Video generation pipelines. The pipeline for generating video with diffusion models often involves some combination of the following steps [Ho et al., 2022a, Ge et al., 2023, Zhou et al., 2022]: key-frame generation, interpolation between key-frames, and spatial super-resolution upsampling. Usually, a separate diffusion model is used for each step. The hierarchical approach of generating key-frames before interpolating can serve to simplify generation of longer videos. However, Bar-Tal et al. [2024] recently generate the entire temporal duration in a single forward pass, obtaining improved global temporal consistency. Meanwhile, the use of 'cascaded' spatial super resolution can help keep the base model simple, reducing computational requirements and aiding generalization abilities [Ho et al., 2022a].

Other techniques and findings. Ge et al. [2023] find that naive extension of the image noise prior to video leads to sub-optimal performance. They instead use a scheme that better preserves natural correlations in video. Blattmann et al. [2023a] perform an in-depth analysis of the effects of data-quality on video diffusion models, finding pretraining on large diverse data and finetuning on smaller high-quality data to be an effective scheme. Beyond video prediction, video diffusion models can be fused for video editing, video in-filling and in-painting [Brooks et al., 2024, Bar-Tal et al., 2024], and can be flexibly conditioned on different input modalities [Chen et al., 2023c, Xing et al., 2024, Wang et al., 2024b].

Technique: Autoregressive and masking transformers. These transformer-based methods have recently achieved promising results in video prediction. We now give an overview of different methods seen in the literature.

Pixel vs latent-space predictions. Whilst early work made predictions directly in pixel-space [Weissenborn et al., 2019], subsequent methods do so in a learned latent space [Yan et al., 2021, Yu et al., 2023a, Ge et al., 2022, Kondratyuk et al., 2023]. This latent space is generally pretrained via vector-quantized auto-encoding (VQ-AE), providing a discrete token-space suitable for the transformer architecture. Predicting in a learned latent space provides a number of benefits, including improving computational efficiency and removing pixel-level noise and redundancy from the video prediction objective [Yan et al., 2021]. Deciding the architecture of the VQ-AE tokenizer is an important design decision [Yu et al., 2023a]. State-of-the-art tokenizers jointly represent spatio-temporal video using 3D convolutions [Yan et al., 2021, Yu et al., 2022, Ge et al., 2022] or causal spatio-temporal attention mechanisms [Villegas et al., 2022, Bruce et al., 2024]. We provide more details on various VQ-AE architectures in Section 4.1.

Autoregressive training. Autoregressive next-token prediction is a standard objective for training transformer sequence models. However, in video the embedding of a single image frame may consist of multiple tokens. Thus, decisions must be made regarding the order in which the tokens are predicted. One option is to simply flatten the discrete tokens in 'raster scan' order, and predict these tokens sequentially [Yan et al., 2021, Seo et al., 2022b, Hu et al., 2023a]. Rather than sequentially decoding, other approaches [Bruce et al., 2024] predict the next-frame autoregressively using MaskGit [Chang et al., 2022a] scheduled parallel decoding. MaskGit inference involves decoding all masked tokens of an image simultaneously, then refining the image iteratively conditioned on the previous generation (note, this draws strong parallels to the diffusion process). This parallel decoding process is more computationally efficient than standard autoregressive prediction. However, the sampling process must be carefully designed to ensure the generated tokens are consistent with each other.

Masked decoding training. Rather than training via autoregressive next-frame prediction, prior works mask part of the video and train the model to decode the masked tokens in parallel [Yu et al., 2022, 2023c, Gupta et al., 2022]. Usually, decoding is performed following a MaskGit process. This mask decoding training allows for multiple tasks to be performed at inference time, including: video prediction, video in-painting, frame-interpolation, and video editing [Yu et al., 2022]. Masked models are also more computationally efficient and do not suffer from the 'drifting' effect that can occur in auto-regressive models. However, they may suffer from sampling bias introduced by the independence assumptions within individual sampling steps [Yang et al., 2024].

Other notable techniques and findings. Several techniques have been proposed to improve long-horizon video predictions. Yan et al. [2023] use aggressive spatio-temporal compression to aid consistency in long-horizon predictions. Ge et al. [2022] improve long video prediction by first predicting key-frames, then interpolating. Meanwhile, different transformer architectures have been proposed that may reduce computational requirements on longer sequence lengths [Dao et al., 2022, Hawthorne et al., 2022, Liu et al., 2024b]. Recent advances in state-space models may be relevant here [Gu and Dao, 2023, Tong et al., 2022]. Elsewhere, Kondratyuk et al. [2023], Liu et al. [2024b] use separate tokenizers for different modalities, allowing their transformer prediction model to input and output other modalities (e.g., text), in addition to video.

Technique: Other. Here, we briefly outline less promising video prediction techniques from the literature. Initial video prediction techniques were based on recurrent or convolutional architectures [Srivastava et al., 2015, Lotter et al., 2016, Chiappa et al., 2017]. However, these often make a limiting assumption that the environment is deterministic. Nevertheless, Gao et al. [2022] demonstrate that a CNN trained with an MSE loss can act as a strong baseline. Latent-variable (VAE) video models [Babaeizadeh et al., 2017, Denton and Fergus, 2018, Villegas et al., 2019] attempt to account for stochasticity in videos, but tend to generate blurry outputs due to limited representational power and underfitting. An interesting extension of these approaches is the use of hierarchical VAE models [Saxena et al., 2021, Wu et al., 2021]. GAN-based methods [Vondrick et al., 2016, Clark et al., 2019, Tulyakov et al., 2018] can produce more realistic videos, but are known to suffer from training instability and limited generation diversity. Elsewhere, Wu et al. [2022], Jiang et al. [2023] propose object-centric video prediction models, Lin et al. [2022] use neural radiance field (NeRF) representations, and Whitney et al. [2023] learn a 3D particle-based simulator from RGB-D videos.

Conditioning video predictions. The use of conditioning information can simplify the video prediction problem and allow for more control over generated videos. Such conditioning is valuable for downstream robotics usage; it can allow us to simulate the effects of different action strategies. Yang et al. [2024] outline the various popular conditioning schemes for video generation models $p(\tau|c)$. Here, we briefly outline the different modalities of information that can be used as conditioning information.

Language is a popular choice for conditioning video generations. Indeed, text-to-video models $p(\tau|c = text)$ represent much of the state-of-the-art in video generation [Brooks et al., 2024, Bar-Tal et al., 2024, Kondratyuk et al., 2023]. Language can allow for flexible and intuitive control over generated video, at varying degrees of detail.

Images and Video can be used to condition video predictions. The next-frame prediction problem requires conditioning on the previous frame(s) $p(s_{t+1}|s_{t-k:t})$ [Gao et al., 2022]. Video in-filling predicts a video that joins initial and final conditioning frames $p(s_{t+1:t+H-1}|s_t, s_{t+H})$ [Höppe et al., 2022]. Video editing, in-painting, and stylization applications $p(\tau|\bar{\tau})$ are all conditioned on an initial input video $\bar{\tau}$ [Brooks et al., 2024, Bar-Tal et al., 2024].

Action information can be used to condition video predictions $p(s_{t+1}|c = \{s_{t-k:t}, action\})$. Bruce et al. [2024] use a learned single-step latent action-space to condition next-frame predictions. Yang et al. [2023c] use robot actions (when labels are available) to condition next-frame predictions. The next-frame predictions of stochastic video prediction models can be conditioned via their latent variable [Rybkin et al., 2018]. We outline various action representations that may be suitable for conditioning video predictions in Section 5.1.1.

Other. Many other modalities have been used to control video generations. Briefly, recent methods have used hand-drawn sketches [Wang et al., 2024b], depth information [Xing et al., 2024], (hand-drawn) motion guidance [Wang et al., 2024b], motion priors [Chen et al., 2023c], or structured scene information [Wang et al., 2023e].

SOTA models. Here, we give details of representative state-of-the-art video prediction models. Note, we focus on models trained on diverse internet data that are capable of generating realistic videos.

- *Sora [Yang et al., 2023c]:* This recent model represented a major breakthrough in video generation. Whilst experimental results (at the time of writing) are limited, released videos qualitatively demonstrate a large jump in capabilities. Improvements include higher visual quality, physical realism, and longer generations (up to 1 minute) versus previous models. The model employs a transformer-based latent diffusion architecture [Peebles and Xie, 2023]. However, a number of limitations are noted. The model: (i) can hallucinate, generating video with inaccurate physics and incorrect cause-and-effect relationships; and (ii) struggles to recreate precise descriptions or spatial details in its prompt.

- *Lumiere [Bar-Tal et al., 2024]:* This is a space-time UNet [Ronneberger et al., 2015] diffusion model. It is trained to generate the entire temporal duration of the video through a single pass of the model. It can generate video up to 80 frames at 16 fps (5 seconds long). No details are provided regarding the datasets used or model parameter counts.

- *VideoPoet [Kondratyuk et al., 2023]:* An 8B parameter decoder-only transformer is trained to input and output several data modalities. This includes images, videos, text, and audio. Separate auto-encoders are used to tokenise each modality. The model is trained on 1B image-text pairs and 270M videos from public internet and other sources (the curated dataset is closed source). Like many diffusion models, this transformer model can be used for various applications, such as text-to-video, image-to-video, video stylization, and video outpainting.

- *Emu Video [Dai et al., 2023]:* Here, the video generation process is factorized into two steps: first an image is generated conditioned on text, then a video is generated conditioned on the text and image. This cascade of diffusion models totals to 6B parameters. The training dataset consists of 34M licenced video-text pairs.

- *Runway [Esser et al., 2023]:* Several closed-source video generation models are available as commercial products [Wang et al., 2024a, pik, moo]. One representative example with published details is the Runway Gen-1 model, which is a text-guided diffusion model trained on (closed source) large-scale image and video data [Esser et al., 2023].

- *Stable Video Diffusion (SVD) [Blattmann et al., 2023a]:* This is a leading open-source video generation model. The 1.5B parameter model is trained in three stages: (i) image pretraining; (ii) video pretraining on curated datasets of up to 50M samples; and finally (iii) finetuning on 250k high-quality video samples. However, the training data is closed-source. Another notable open-source model is that of Hong et al. [2022].

Discussion. First, we note there are various pros and cons between diffusion, autoregressive transformer, and masked transformer video prediction methods. Diffusion can model continuous spaces and sample multiple frames in parallel. However, sampling speeds are slow and generating long sequences is a challenge. Autoregressive models are easier to train than diffusion, and scale well with context length. However, autoregressive decoding is computationally expensive and predictions can suffer from a drifting effect. As noted, versus autoregressive models, masked models are more computationally efficient but can suffer from sampling bias. Yang et al. [2024] advocate for improved future models that combine the advantages of these different schemes.

Now, whilst progress in foundational video prediction models has been promising [Brooks et al., 2024], a number of challenges remain. This includes issues with computational efficiency, long video generation, limited generalization, and hallucination. Nevertheless, there are promising avenues here. Improved architectures can deal with computational issues [Liu et al., 2024b] and long video generation [Yan et al., 2023]. Yang et al. [2024] note that RL finetuning [Black et al., 2023a] may improve hallucination issues. In general, gains in generated-video quality may be obtained via improved datasets (following the dataset desiderata outlined in Section 6.1). More speculative directions include leveraging 3D information to improve the physical realism of video generations [Zhen et al., 2024], or exploring the use of hierarchy and temporal abstractions for improved long video generation.

A future direction of particular interest to robotics is to pursue methods that allow for fine-grained, single-step conditioning of video predictions. Such conditioning simplifies adaptation of a video prediction model into a robot dynamics model. Bruce et al. [2024] use learned latent action representations, and we outline various other action representations suitable for conditioning video predictions in Section 5.1.1. Finally, we advocate for progress in open-sourced foundational video prediction models – open-sourced models will make LfV research more accessible to the wider community.

4.3 Video-to-Text Models

Here we refer to models with video-to-text capabilities. A capable video-to-text model can perform, for example, video question answering or video summarization. There are numerous commercial applications for such models and we are likely to see increasingly capable models in the near future. Certain models with video-to-text capabilities can benefit from internet text-only data, in addition to video data and video-text data (e.g., 'any-to-any' sequence models [Liu et al., 2024b]); from this text-only data, these models can learn powerful capabilities and knowledge similar to those seen in LLMs [OpenAI, 2023].

Now, (capable) video-to-text foundation models could be valuable to robotics in several ways:

- *High-quality representations:* A capable video-to-text model will have robust, high-quality video representations. Versus a video-only model, it will likely have improved high-level semantic representations. Versus an image-language model, it will have improved temporal-dynamic representations. Robotics can bootstrap from such models via representation transfer [Brohan et al., 2023], or by adding robot data directly into the model's pretraining corpus [Reed et al., 2022].

- *Grounded reasoning and planning:* LLMs have proven useful as planning modules in robotics [Ahn et al., 2022], but their lack of grounding in the physical world is limiting. In contrast, video-to-text models can perceive the environment through information-rich video, potentially allowing for improved and closed-loop reasoning and planning.

- *Annotating robot data:* High-quality language annotations can act as valuable conditioning information in many ML domains [Betker et al., 2023, Brooks et al., 2024]. Robotics is no different [Team et al., 2023b]. Capable video-to-text models could serve as useful language annotators for robotic datasets.

- *Rewards:* A sufficiently capable video-to-text model can provide reward or value estimates for a robot learner. For example, this could be done through a visual-question answering framework [Du et al., 2023c], or via an RL-from-AI-feedback framework [Klissarov et al., 2023] (see Section 5.2.4).

In this section, we give an overview of existing video-to-text methods and models. We do so with an emphasis on methods promising for obtaining foundational video-to-text models. We note that research here is preliminary: low-quality video captions and the difficulties of machine learning with video data (e.g., due to issues related to high-dimensionality and noise) mean progress lags behind foundation models in other domains.

Zero-shot combination of pretrained models. Due to the difficulties of training a monolithic video-to-text model, previous work has explicitly decomposed the problem, assigning the sub-tasks to frozen pretrained models. These pretrained models often communicate via language. A simple example is given by Chen et al. [2023b]. Here, an image-language model answers questions about individual video frames, and an LLM synthesized this information to produce a global summary of the video. Zeng et al. [2022] take a similar approach, making use of a wider range of pretrained models, including audio-language models and object detectors. Li et al. [2022c] solve multimodal problems, including video-to-text tasks, using pre-trained models as "generators" or "scorers" and composing them via a closed-loop iterative consensus optimization. Whilst these compositional approaches can be effective, their modular structure and lack of end-to-end video training mean they can lack nuanced video representations and understandings.

Leveraging pretrained LLMs via adaptors and finetuning. Alayrac et al. [2022], Li et al. [2022a] introduce schemes for adapting pretrained LLMs to be additionally conditioned on image inputs. Recent improvements in open-source LLMs [Touvron et al., 2023] have seen these approaches extended into the realm of video. Such approaches typically involve the following steps: (i) obtain a pretrained LLM and an (often pretrained) video encoder; (ii) define an adaptor module to channel information from the video encoder output into the LLM; and finally (iii) finetune the combined model on video-text data. We now give more details on these methods.

Pretrained LLMs. Works in this space [Zhang et al., 2023b, Li et al., 2023a] have often leveraged open-source LLMs from the LLaMa [Touvron et al., 2023] family. More recent SOTA open-source LLMs would also be suitable [Jiang et al., 2024, Team et al., 2024].

Video encoders. The encoder used can be frame-based (i.e., each frame is encoded separately) [Maaz et al., 2023] or can jointly encode spatio-temporal information [Papalampidi et al., 2023]. Image pretrained encoders based on CLIP [Radford et al., 2021] have commonly been used [Maaz et al., 2023, Li et al., 2023b]. We note that any of the state-of-the-art encoder from Section 4.1 may be suitable (as demonstrated by [Zhao et al., 2024, Papalampidi et al., 2023]), though encoders pretrained with language-based losses likely are more suitable.

Adaptors. Tang et al. [2023] identify two main categories of adaptors used in the literature. (1) Connective adaptors connect the outputs of the video encoder to the input space of the LLM. This has been done, for example, with linear projections [Chen et al., 2023a], MLP layers [Lin et al., 2023], and more complex adaptors such as Q-formers [Zhang et al., 2023b]. (2) Insertive adaptors insert the outputs of the video encoder into the internal layers of the LLM. This can be done, for example, via cross-attention [Alayrac et al., 2022, Papalampidi et al., 2023]. Note that, if the LLM remains frozen downstream, then the adaptors must map video information directly into the embedding spaces of the LLM. This may be limiting as a frozen LLM lacks rich representations of lower-level video information. We direct the reader to Zhang et al. [2024] for more information on different adaptor types.

Training pipelines. Once the new video-to-text architecture is defined, a common training scheme is: (i) convert large diverse video-text data into token sequences and perform next-token prediction pretraining, then (ii) perform supervised instruction-tuning on small high-quality instruction datasets [Lin et al., 2023, Zhang et al., 2023b]. Following trends in LLMs, future work may investigate a third RLHF finetuning stage [Kaufmann et al., 2023b]. Finally, we note that during training the LLM and/or video encoder may be finetuned [Lin et al., 2023], or kept frozen [Maaz et al., 2023].

Natively multi-modal models. The previously discussed methods have involved combining and finetuning pretrained models not initially intended for video-to-text purposes. Here, we highlight end-to-end training pipelines that (loosely) are more *natively* multi-modal. We note this is a particularly preliminary area of research. Nevertheless, recent trends towards any-to-any sequence models – where video-to-text is formulated as an interleaved video and text sequence modelling problem – are promising. Liu et al. [2024b], Team et al. [2023a], Jin et al. [2023] all train any-to-any (or any-to-text) autoregressive transformers via next token prediction. This requires the use of modality specific encoders and decoders. Liu et al. [2024b] use a pretrained VQ-GAN to tokenise individual video frames. Note, in practice, these methods may still initialise their model with a pretrained LLM [Jin et al., 2023], or perform an initial stage of training on text-only data [Liu et al., 2024b].

SOTA models. Here we give brief details on some representative state-of-the-art video-to-text models.

- *Video-LLaVa [Lin et al., 2023]:* A pretrained LLM [Chiang et al., 2023] is connected to pretrained video and image encoders [Zhu et al., 2023a] via a two layer connective adaptor. The new architecture is first trained on 558k image-text pairs [Liu et al., 2024c] and 702K video-text pairs [Luo et al., 2023], before instruction tuning on 665k image-text [Liu et al., 2023b] and 100k video-text [Maaz et al., 2023] instructional examples. The video encoder remains frozen, but the LLM is finetuned. The model only takes eight frames of video input. The authors note the final model has difficulty with temporal relationships and spatio-temporal localization. The model is open-source.

- *VideoChatGPT [Maaz et al., 2023]:* This model is intialized from LLaVa [Liu et al., 2023b], an LLM adapted to take images as input. Both the visual encoder and LLM from LLaVa are inherited. To adapt the visual encoder to video, frames are encoded individually before spatial and temporal pooling. The pooled features are projected into the LLM input space using a linear projection. Video-to-text training is performed on a curated dataset of 100,000 custom video-text instruction pairs. Both the image encoder and LLM are frozen during the video-text training. The authors note that the final model struggles to understand subtle temporal relationships and the visual details of small objects. The model is open-sourced.

- *LWM [Liu et al., 2024b]:* Here, ring-attention [Liu et al., 2023a] is used to scale a transformer model up to million-length sequences. The model is initialized using a Llama-2 7B [Touvron et al., 2023]. A first stage of autoregressive text-only training is performed with a curriculum of increasing context lengths. The second stage of training incorporates interleaved video/images with text. A pretrained VQ-GAN is used to tokenize the images and videos. Note, this is an any-to-any sequence model that can also perform image and video generation. The model is open-sourced.

- *VideoPrism+PALI [Zhao et al., 2024]:* The 1B parameter VideoPrism video encoder [Zhao et al., 2024] (see Section 4.1) is connected to the PaLM-2 LLM [Anil et al., 2023]. This is done following the scheme introduced by Alayrac et al. [2022], where a randomly initialized Perceiver-resampler projects video representations into the LLM embedding space, with randomly initialized cross-attention layers added at each layer of the LLM (i.e., this is an insertive adaptor). The backbones of the video encoder and LLM are kept frozen during the video-to-text training. Video-text data from 10 different datasets are employed, including some open-sourced datasets [Grauman et al., 2021, Damen et al., 2018, Monfort et al., 2021]. This totals to 4.4M video clips.

- *Short/LongViViT [Papalampidi et al., 2023]:* The Short/LongViViT video encoder is first pretrained via a video-language contrastive loss (see Section 4.1) [Papalampidi et al., 2023]. The frozen encoder is then plugged into a frozen pre-trained LLM for video-to-text adaptation. As per the previous model, the adaptation scheme follows Alayrac et al. [2022]. The video-to-text model is trained with an autoregressive video-captioning loss, using close to 27M (primarily closed source) video-text pairs.

Discussion. Whilst there has been progress in video-to-text models, state-of-the-art capabilities are still limited. Schemes that combine pretrained models in a zero-shot manner can be useful for some applications, but a lack of end-to-end training means they do not maintain rich video representations. Previous methods that leverage pretrained LLMs have generally only fine-tuned on relatively small amounts of video-text pairs and can suffer from hallucinations. More generally, video-to-text models currently struggle with spatial relationships, fine-grained understandings, and long-term understandings.

A key bottleneck to progress here is the quality (and quantity) of available video-text data. Versus existing image datasets, available video datasets are smaller and can have inferior text annotations. Annotations are often sparse, and can be coarse or noisy. Improved annotations will be essential for improving video-to-text capabilities. This will enhance fine-grained understandings, which is particularly relevant to robotic applications. We give details on methods for captioning video data in Section 6.2. Progress in long video understanding is not only bottlenecked by data quality, but also computational burdens. More efficient architectures for handling longer contexts are promising here [Liu et al., 2023a, Gu and Dao, 2023, Balažević et al., 2024]. Finally, we advocate for more research into 'natively' multi-modal models, where the full training pipeline is designed specifically for video-to-text purposes.

5 LfV-for-Robotics: Methods

We have previously outlined preliminary information for the LfV-for-robotics setting (Section 3), and reviewed video foundation models as a general-purpose approach for extracting knowledge from internet video (Section 4). In this section, we will review literature that has specifically utilised video data for robotics. We begin by reviewing some common methods used to mitigate LfV challenges (Section 5.1). We then move to our primary categorization and description of the LfV-for-robotics literature (Section 5.2), where we classify methods according to which component of the RL algorithm benefits from the use of video data.

5.1 Mitigating LfV Challenges

We outlined key LfV challenges in Section 3.3. In this section, we detail two categories of techniques that each address a key challenge: (1) The use of alternative action representations to mitigate missing action labels in video data (Section 5.1.1). (2) The use of certain representations that explicitly address LfV distribution-shift issues (Section 5.1.2). These techniques reoccur across various LfV methods, and are usually used as a single component within a larger LfV pipeline. As such, here we have separated out details of these techniques into contained sections ahead of our main analysis of the LfV literature in Section 5.2.

5.1.1 Recovering Action Information from Video

Transition tuples of robot data commonly come in the form of (s_t, a_t, r_t, s_{t+1}). With this data, some of the key RL knowledge modalities (KMs) (see Section 2.1) can be trained. For example, one can use behaviour cloning or RL to train an action-generating policy $\pi(a_t|s_t)$, use supervised learning to train an action-conditioned dynamics model $p(s_{t+1}|s_t, a_t)$, or TD-learning to train an action-conditioned value function $Q(s_t, a_t)$. However, video data transition tuples come in the form of (s_t, s_{t+1}) – i.e., video data is missing action labels, meaning one cannot naively use it to train the KMs listed above.

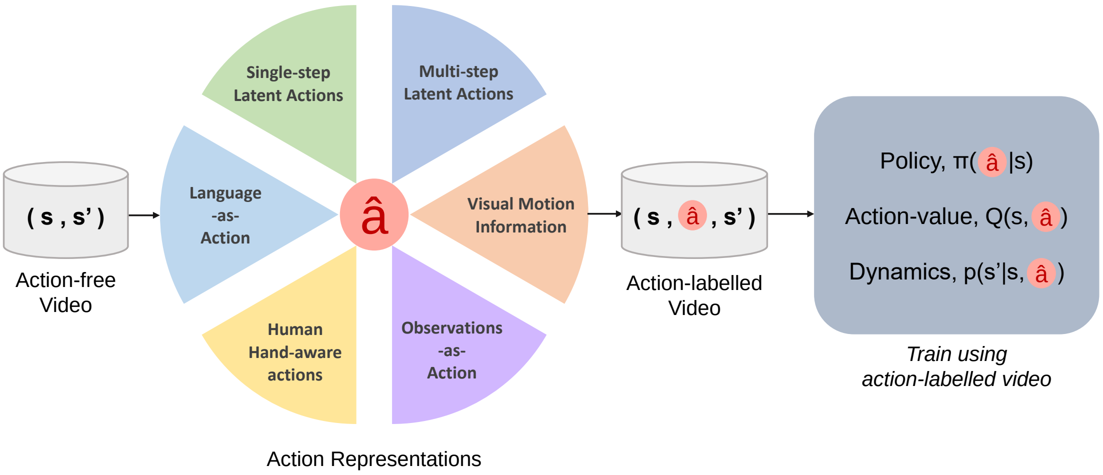
In this section, we review works that use *alternative representations of actions* to directly address this missing action label problem. These methods define or learn some representation space that is analogous to the notion of an action. We use $\hat{a} \in \hat{\mathcal{A}}$ to denote a single alternative action and its underlying alternative action space. An intuitive example here is the use of language to represent action information; e.g., $\hat{a} =$ "pick up the cube" . Once $\hat{\mathcal{A}}$ is learned or defined, video data can be relabelled from (s_t, s_{t+1}) to $(s_t, \hat{a}_t, s_{t+1})$. This relabelled data can be used to train an alternate action version of an RL KMs (see Figure 5a). In the literature, alternative action policies $\pi_{\text{alt}}(\hat{a}_t|s_t)$ [Schmidt and Jiang, 2023], dynamics model $p_{\text{alt}}(s_{t+1}|\hat{a}_t, s_t)$ [Bruce et al., 2024], and value functions $Q_{\text{alt}}(s_t, \hat{a}_t)$ [Bhateja et al., 2023] have all been trained. These alternative action KMs can be useful downstream for representation transfer [Bhateja et al., 2023, Schmidt and Jiang, 2023], or if a mapping from alternative to robot action space $f : \mathcal{A} \rightarrow \hat{\mathcal{A}}$ is obtained [Wang et al., 2023a, Wen et al., 2023, Du et al., 2023a].

Before detailing the different categories of alternative action representations $\hat{\mathcal{A}}$ in the literature, we first touch on some preliminary discussion points:

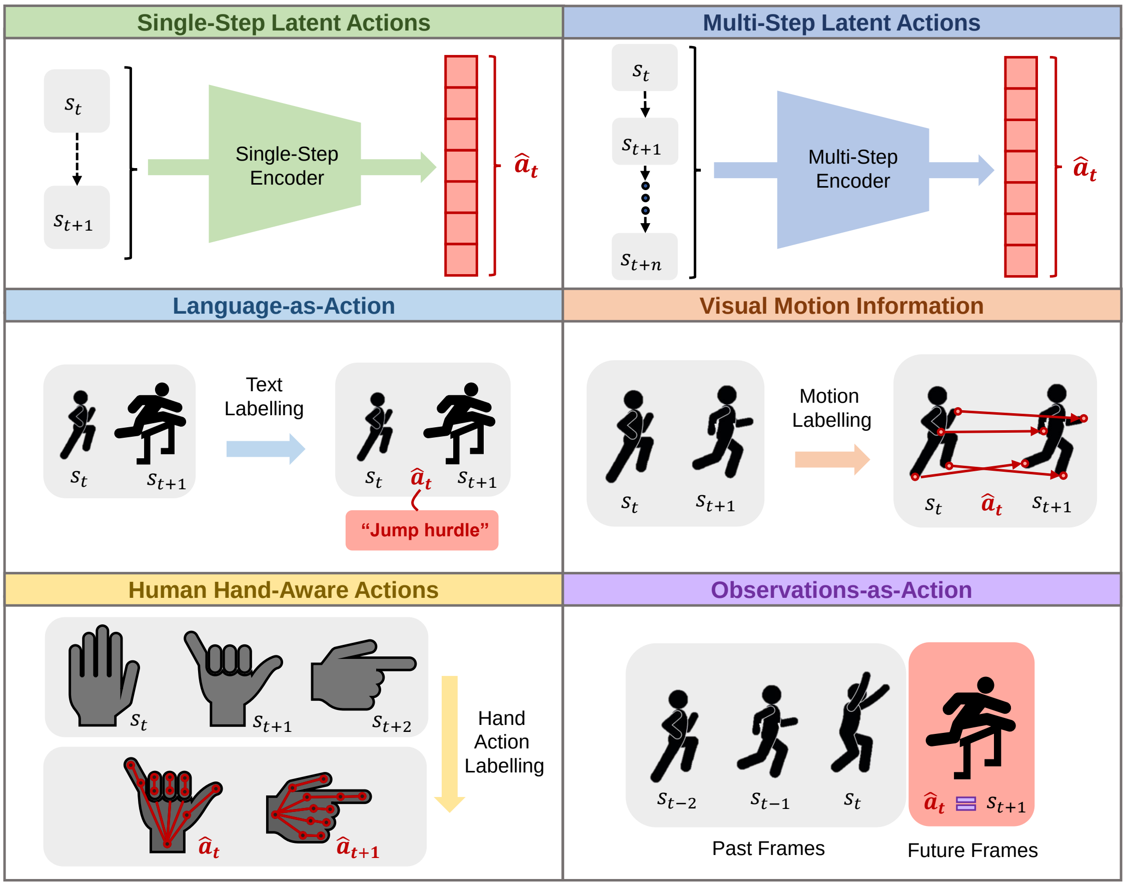
- *What are some desirable properties of alternative actions?* First, we note that $\hat{\mathcal{A}}$ must be inferable purely from video data (or video data paired with labels). Beyond this, we identify several additional desirable properties. (1) $\hat{\mathcal{A}}$ can be used to label transitions from (s_t, s_{t+1}) to $(s_t, \hat{a}_t, s_{t+1})$, enabling meaningful learning of alternative-action RL KMs. (2) $\hat{\mathcal{A}}$ should contain information regarding what will happen in future frames of a video – i.e., it contains information regarding future 'actions' taken in the video. As a consequence, learning $\pi(a_t|\hat{a}_t, s_t)$ should be easier than learning $\pi(a_t|s_t)$. (3) $\hat{\mathcal{A}}$ should capture a high-level, general notion of an action, making it transferable across the specific underlying action spaces of different human and robot embodiments. (4) $\hat{\mathcal{A}}$ should be interpretable and disentangled. In particular, $\hat{\mathcal{A}}$ should have a consistent meaning across the state space. This can allow for learning of $\pi(a_t|\hat{a}_t, s_t)$ in a more data-efficient manner [Rybkin et al., 2018].

- *What are other notable characteristics of alternative action representations?* We note two distinct characteristics that vary across the different categories of alternative action representations. (1) Learned vs defined: $\hat{\mathcal{A}}$ can be on a spectrum from being entirely learned from data (e.g., single-step latent actions), to simply being manually defined (e.g., observations-as-action). (2) Time-horizons: $\hat{\mathcal{A}}$ can contain only short horizon information (e.g., single-step latent actions) or longer horizon information (e.g., multi-step latent actions).

We now outline the main categories of $\hat{\mathcal{A}}$ we have identified in the literature (see Figure 5b). Here, we focus on detailing how $\hat{\mathcal{A}}$ can be defined and learned from video data. More details on how alternative action representations can be used downstream for robotics are found throughout Section 5.2.



(a) Utilising action representations to overcome the missing action-label problem in LfV. Action-free videos can be labelled with alternative action representations $\hat{a}$. This labelled video data can then be used to train an alternative action RL knowledge modality (e.g., a policy, value function, or dynamics model).



(b) Categories of action representations for LfV. This survey identifies and analyses six distinct categories of action representations that can be learned or obtained from video datasets. Details of each category are elaborated upon in Section 5.1.1.

Figure 5: Recovering action information from video (Section 5.1.1). (a) An overview of why action representations are useful in LfV. (b) Visualisations of the different categories of action representations that can be used.

Single-step latent actions. Here, we consider learned action representations where, like true robot actions, each 'latent' action corresponds to exactly one transition. Thus, the latent action $\hat{a}_t$, which occurs between observations s_t and s_{t+1}, contains information about the true action a_t. Methods that learn such a latent representation commonly do so using a next-state prediction objective $p(s_{t+1}|\hat{a}_t, s_t)$, such that $\hat{a}_t$ is informative for the predictions of the forward dynamics model (FDM) [Edwards et al., 2018, Rybkin et al., 2018, Schmidt and Jiang, 2023, Bruce et al., 2024].

Edwards et al. [2018] introduce the concept of latent policies, i.e., policies $\pi_{\text{alt}}(\hat{a}_t|s_t)$ that produce actions in a latent action space. However, their method can only learn discrete latent action spaces, and the proposed learning objective is susceptible to mode collapse [Struckmeier and Kyrki, 2022, Schmidt and Jiang, 2023]. Approaches to learn continuous latent action spaces [Rybkin et al., 2018, Schmeckpeper et al., 2019, Chang et al., 2022b, Schmidt and Jiang, 2023, Bruce et al., 2024] generally work as follows: (1) A latent inverse dynamics model (IDM) encodes past frames $\{s_{t-k}, \ldots, s_t, s_{t+1}\}$ to infer $\hat{a}_t$. (2) $\hat{a}_t$ is then passed, along with $\{s_{t-k}, \ldots, s_t\}$, to an FDM $p(s_{t+1}|\hat{a}_t, \{s_{t-k}, \ldots, s_t\})$ which predicts s_{t+1}. (3) This setup is trained end-to-end using the FDM prediction error. Additionally, a regularising objective or mechanism is usually employed to prevent the IDM from copying the entire observation s_{t+1} into the latent action. For this purpose, Rybkin et al. [2018] regularise latent actions towards a Gaussian prior and additionally enforce a latent action composability loss. Schmidt and Jiang [2023], Bruce et al. [2024] use a vector-quantization bottleneck to constrain the flow of information. Rybkin et al. [2018], Schmidt and Jiang [2023] both show their regularization techniques to improve disentanglement of the latent action representations and thus improve downstream performance.

We note some potential limitations of these approaches. First, learned latent action spaces jointly model changes caused by the agent and the external environment. Often, we would like to only represent changes due to actions taken by a single agent. Second, these approaches model environment transitions on a visual level, thus they may omit low-level information relevant to robotics (such as forces). We finally note that more research into learning single-step latent action-spaces from realistic internet video is required.

Multi-step latent actions. Here, we discuss learned action representations $\hat{a}$ that contain information about multiple time steps (rather than a single timestep). One line of work here uses variational auto-encoding of video trajectories to learn *latent plans*; i.e., compressed representations of video. This has been explored extensively with action-labelled demonstrations [Lynch et al., 2020, Cui et al., 2022, Rosete-Beas et al., 2022], but less so with action-free video. Wang et al. [2023a] train latent plan representations from action-free videos via auto-encoding, but use 3D human hand trajectories as the decoder target (rather than raw video).

There are a number of other approaches. When videos are labelled with language descriptions, representations of video trajectories have been learned using video-language contrastive losses [Fan et al., 2022, Lifshitz et al., 2023, Sontakke et al., 2023]. Chane-Sane et al. [2023], Chen et al. [2021a] obtain informative representations of video clips by performing supervised contrastive learning on a dataset with clip-level action labels. Xu et al. [2023] learn representations of video clips using a self-supervised learning clustering framework. Other approaches for learning multi-step latent actions are seen in Tomar et al. [2023], Pertsch et al. [2022], Cai et al. [2023].

We note that these methods are often used for *video-as-instruction* approaches that specify the task via video, encoding the video instruction into $\hat{\mathcal{A}}$ before feeding it to a low-level robot policy (see Section 5.2.2). Finally, we note that any video encoding technique from Section 4.1 could apply here, though certain representations may have more suitable properties than others.

Observations-as-action. A simple approach is to use future observations (i.e., future images in the video) as the action representation: future observations (or an encoding thereof) provide information on what actions will be taken next in the video. A benefit here is that this $\hat{\mathcal{A}}$ is directly available from raw video; unlike latent action representations, no additional learning steps are required. Observations-as-actions can be used on varying time-horizons, as we now describe.

Next-observation-as-action: These methods use the next observation s_{t+1} as $\hat{a}_t$; a proposed s_{t+1} provides clear information about the action that should be taken at s_t. A number of works train alternative action policies $\pi(s_{t+1}|s_t)$ from video data to propose observations the robot should reach in the next time-step. Such policies can, for example, be learned via supervised video prediction objectives [Du et al., 2023a, Thomas et al., 2023].

Observations-as-subgoals: A subgoal can be viewed as a high-level action. In 'Observations-as-subgoals' methods [Black et al., 2023b, Park et al., 2024, Bhateja et al., 2023], an observation (or embedding thereof) from H time-steps into the future is used as a sub-goal (or alternative-action) representation. In this framework, a sub-goal policy $\pi(s_{t+H}|s_t)$ can be trained to propose an image observation the robot should reach [Black et al., 2023b]. Some simple strategies for defining sub-goals in the video data include: choosing a fixed time-horizon H [Black et al., 2023b, Du et al., 2023a], or randomly sampling observations beyond the current timestep [Bhateja et al., 2023]. More complex strategies include using key-frame identification to identify bottleneck states in video [Pertsch et al., 2019]. Liu et al. [2023c] identify critical states in videos, but require access to reward labels.

Language-as-action. Natural language can be used as a flexible and meaningful high-level action-space (e.g., $\hat{a}_t =$ "pick up the cube"). Some video datasets come with language annotations that can directly provide such language action information [Grauman et al., 2021]. General-purpose language annotations can be further processed to convert them into a more suitable form. Mu et al. [2023] use LLMs to perform additional processing on Ego4D [Grauman et al., 2021] language descriptions to convert them into more usable language plans. If the video does not come with language descriptions, VLMs, LLMs, and other off-the-shelf models (such as object detectors), can be used to label videos with language-action information (see Section 6.2 for information regarding methods for captioning video). Once the video dataset is annotated with language actions, it can be used, for example, to train a high-level policy $\pi_{\text{alt}}(\hat{a}_t|s_t)$ that outputs language actions [Mu et al., 2023, Du et al., 2023b], or to train language-conditioned video predictors $\pi(s_{t+1}|s_t, \hat{a}_t)$ [Du et al., 2023a]. Language actions are also generally useful as they allow for easy interfacing with other language models [Du et al., 2023b]. However, language is coarse and language actions may not contain important lower-level action information.

Visual motion information. Other works have used visual motion information in video to define alternative action-spaces. Wen et al. [2023] do so by labelling video data with 2D point trajectories. Specifically, random points on objects are sampled and tracked throughout the video using an off-the shelf point tracker [Karaev et al., 2023]. Here an alternative action policy $\pi_{\text{alt}}(\hat{a}_t|s_t)$ is trained to predict future point trajectories $\hat{a}_t$, and a low-level policy $\pi(a_t|\hat{a}_t, s_t)$ is trained to decode these point trajectories to robot actions. Yuan et al. [2024] follow a similar approach, but use 3D point trajectories. The 3D annotations are obtained either from 3D annotated datasets or videos with depth information (but they note depth estimation [Bhat et al., 2023] techniques could also be used). Elsewhere, Ko et al. [2023] use an off-the-shelf model [Xu et al., 2022] to predict optical flow between two images, giving a pixel-level dense correspondence map between two frames. This map can be used to infer robot action without any action labels. In a related approach, Nasiriany et al. [2024] represents actions via visual arrows in images, allowing a VLM to select actions via an iterative refinement procedure. Finally, Wang et al. [2023b] use structure-from-motion [Schonberger and Frahm, 2016] to help recover action information, Yuan et al. [2021] use motions of object-centric representations, and Yang et al. [2023c] optionally use camera frame motion/angle information to condition video predictions.

Human-hand-aware actions. Off-the-shelf human-hand detection models [Rong et al., 2020, Shan et al., 2020] can extract hand poses or affordances representations from videos which can act as an alternative action representation [Bharadhwaj et al., 2023, Bahl et al., 2023, Shaw et al., 2022, Qin et al., 2022, 2021]. For example, $\hat{a}_t$ can be defined as the pose that should be reached at time $t+1$. Human-to-robot retargeting can be used to convert human poses to a robot action-space [Qin et al., 2021, Shaw et al., 2022, Sivakumar et al., 2022], whilst affordance representations can naturally transfer across embodiments [Bahl et al., 2023]. Bharadhwaj et al. [2023] use object masks in addition to human poses to define a $\hat{\mathcal{A}}$. We give more details on methods for detecting human hand poses and affordances in Section 5.1.2.

IDM pseudo-action labels. Finally, though not technically an *alternative* action representation, we mention these methods here as they also address the missing action label problem. Several works train an inverse dynamics model $p^{-1}(a_t|s_t, s_{t+1})$ on action labelled robot data, and use it to provide pseudo-action labels for action-free video data [Baker et al., 2022, Torabi et al., 2018a, Schmeckpeper et al., 2020]. However, these simple approaches are unlikely to scale to diverse internet video. This is because they require either: (i) minimal domain-shift between the video data and the robot domain (e.g., Baker et al. [2022] assume identical embodiments), or (ii) an explicit mechanism to deal with domain-shift that may not scale to diverse internet video [Schmeckpeper et al., 2020] (see Section 5.1.2).

Discussion. Alternative action representations are promising for tackling the problem of missing action labels in video data. We now discuss which categories of alternative action representations can be most useful for LfV.

How scalable is $\hat{\mathcal{A}}$ to diverse internet video? The degree to which $\hat{\mathcal{A}}$ is readily available in video, or can scale to more unstructured heterogeneous video data, is important. *Observations-as-action* are immediately available in videos. *Language-as-action*, whilst very useful, requires the video data to be paired with suitable language descriptions that may not always be available. *Single-step* and (some) *multi-step latent actions* can be learned purely from raw unlabelled video, but this learning step adds extra complexity and these methods are largely untested on diverse internet video. *Visual motion information* and *human-hand-aware* action spaces can provide useful inductive biases, but it also unclear how well they will scale to more unstructured video datasets.

How easy is it to decode the robot action from $\hat{\mathcal{A}}$? Alternative action policies $\pi_{\text{alt}}(\hat{a}_t|s_t)$ can be used help to obtain a downstream robot policy. This is commonly done via a hierarchical conditioning approach $\pi(a_t|s_t, \pi_{\text{alt}}(\hat{a}_t|s_t))$ or representation transfer $\pi_{\text{alt}}(\hat{a}_t|s_t) \rightarrow \pi(a_t|s_t)$ (see Section 5.2 for more details). In these cases, the ease of decoding a_t from $\hat{a}_t$ is likely a good proxy for downstream performance gains. We now discuss properties of $\hat{a}_t$ that may simplify this decoding: (1) Decoding may be simplified if $\hat{a}_t$ contains more mutual information with a_t. Considering the nature of the information in a_t, this perhaps advocates for the use of shorter-horizon, or less abstracted, versions of $\hat{\mathcal{A}}$. However, we note that more abstracted action-spaces can be beneficial in long-horizon tasks [Du et al., 2023b], and may suffer less from missing low-level information in video. (2) Minimising the information in $\hat{a}_t$ (as seen in Rybkin et al. [2018], Schmidt and Jiang [2023]) could simplify decoding and improve decoding generalization by reducing spurious correlations in the decoding. However, both (1) and (2) are relatively untested in the literature.

Levels of action information in video. Video is missing important low-level information for robotics (e.g., it lacks explicit force information). Thus, alternative action representations obtained from video may not be fully informative for decoding robot actions (and so, in many cases, the decoder should be defined as $\pi(a_t|\hat{a}_t, s_t)$ rather than $\pi(a_t|\hat{a}_t)$). This perhaps suggests that video is best suited for extracting slightly abstracted or longer-horizon action representations. Indeed, research has explored the use of video as a mid-level action representation within a hierarchical framework that also leverages a high-level language-as-action model and a low-level robot action model [Du et al., 2023b].

5.1.2 Representations to Address Distribution-Shift

Distribution shift between internet videos and the target robot domain poses a challenge to LfV approaches (see Section 3.3). For example, a common shift is the embodiment gap between humans and robots. Such distribution shift can hinder our ability to transfer knowledge from the video data to the robot.

We are aware of three strategies that may help overcome these distribution shift issues.

1. *Scaling to large diverse datasets:* Scaling a deep learning method to ever larger and diverse internet video data may help. Higher coverage in the pretraining video dataset will minimise unseen shifts encountered in the downstream robot domain. Additionally, the increased scale of pretraining may improve the overall generalization abilities of the model.

2. *Leveraging in-domain robot data:* In-domain robot data can be incorporated into pretraining, or the model can be finetuned on robot data. Robot data contains information about the robot domain that is not covered in the video data.

3. *Explicitly addressing distribution-shifts:* Techniques can be used to explicitly address distribution-shifts between the video data and the robot domain. For example, embodiment invariant representations can be learned [Schmeckpeper et al., 2020].

Ideally, we hope to overcome distribution shifts using strategies (1) or (2), as methods from strategy (3) may be less scalable to diverse internet video. Nevertheless, a number of LfV works have explored strategy (3), and we review these methods in this section. We identify four main categories of methods, which we now elaborate on.

Method: Human-(hand)-aware approaches. In LfV, we are most often interested in learning from human behaviour; we wish for our robots to replicate this behaviour. In particular, for robot manipulation, we are interested in replicating the behaviour and effects of the human hand. As such, there is a line of LfV research that explicitly detects human embodiment-centric information in video, and subsequently transfers this information to the robot.

What types of human embodiment-centric information can be detected?

- *Poses*: Several works explicitly estimate the pose of the human body or hand in videos. This involves estimating the positions and orientations of various joints or key points on the body. The human hand is often represented using the MANO hand model [Romero et al., 2022]. The human-body can be represented by, for example, by the SMPL/SMPL-X models [Loper et al., 2023, Pavlakos et al., 2019]. In the LfV literature, often an off-the shelf model (e.g., OpenPose [Cao et al., 2017]) is used to crop around the hand, and a second off-the-shelf model (e.g., FrankMocap [Rong et al., 2020]) is used to detect the hand pose from the cropped image [Sivakumar et al., 2022, Shaw et al., 2022]. Once detected, the human-hand pose may need to be retargeted to the robot embodiment. This can be achieved by directly optimising a loss function [Shaw et al., 2022], or by training a retargeting network to minimise a retargeting loss function [Sivakumar et al., 2022].
- *Affordances*: Affordances are a common notion in robotics [Ardón et al., 2020]. A common affordance detected in human videos in the LfV literature is the combined use of: (i) grasp/contact points (i.e., where on the object does the hand make contact?); and (ii) post-grasp/contact trajectories (i.e., once contact is made, where will the hand move?) [Bahl et al., 2023, Mendonca et al., 2023]. This affordance information can be extracted from videos using an off-the-shelf model [Shan et al., 2020], and using further tricks to obtain accurate, usable affordance labels [Bahl et al., 2023]. Note, 2D affordances can be converted to 3D using depth estimation [Mendonca et al., 2023].
- *Masks and Bounding boxes*: Masks or bounding boxes of human hands and objects can be used [Bharadhwaj et al., 2023]. These can be obtained from videos using off-the shelf models [Kirillov et al., 2023, Zhang et al., 2022] or via labelled datasets [Darkhalil et al., 2022].

How to use this human embodiment-centric information? With a video dataset labelled with human-centric information, one of the following can be performed: (1) With the labelled dataset, the process for extracting the human-centric information can be distilled into a single model [Mendonca et al., 2023] (i.e., a model is trained to predict the label given an image). This can be useful when the original extraction method is convoluted or unlikely to generalize. It may also be a good auxiliary representation learning objective for robot manipulation [Bahl et al., 2023]. (2) The extracted information can be treated as an 'alternative action representation' $\hat{\mathcal{A}}$ [Bharadhwaj et al., 2023, Bahl et al., 2023, Shaw et al., 2022, Qin et al., 2022, 2021] (see Section 5.1.1 for more details on alternative actions).

Limitations. Human hand poses can be directly retargeted to certain robot hands, and the affordances detailed above are (in theory) embodiment agnostic. However, we note there still can be difficulties transferring this information to the robot. For example, a model trained to propose future poses solely on human images is unlikely to generalise zero-shot to robot images. This has led to tricks being used in the LfV literature, such as predicting affordances only when the embodiment is out of frame [Bahl et al., 2023], or in-painting human embodiments over robot embodiments [Bharadhwaj et al., 2023].

Method: Learned invariant representations. Here we refer to methods that *learn* a representation invariant to a specific distribution shift between the video dataset and robot domain. These methods generally rely on access to data from both distributions. We now briefly list some methods seen in the LfV literature, before commenting on limitations.

Methods. (1) Domain confusion techniques have been used to learn representations that are invariant across viewpoints [Stadie et al., 2017] and embodiments [Schmeckpeper et al., 2020]. (2) Contrastive learning techniques can also be used to encourage invariance across a particular axis; Sermanet et al. [2018] learn viewpoint invariant representations given time-aligned videos from different viewpoints. In related work, Aytar et al. [2018] use temporal distance classification to learn representations that generalise across visual changes. (3) Temporal cycle-consistency objectives have been used to learn embodiment invariant representations; Zakka et al. [2021] do so using videos of different embodiments performing the same task. (4) Factorized representations have also been proposed. Schmeckpeper et al. [2019], Shang and Ryoo [2021] factorise a learned representation into two components: one that is common across distributions, and one that is unique to each distribution. Chang et al. [2023] use embodiment segmentation and in-painting removal to obtain explicit factorized representations of the agent and environment. (5) Image translation can be used to convert an image from the source distribution (e.g., human embodiment) to the target (e.g., robot embodiment). This can be done, for example, using Cycle-GANs [Smith et al., 2019] or diffusion in-painting [Bahl et al., 2022, Bharadhwaj et al., 2023].

Limitations. However, there are some inherent limitations to the scalability of these approaches. (1) Invariant representations can forego useful information. For example, there may be meaningful differences between human and robot embodiments that the agent should be aware of (note, factorized representations [Schmeckpeper et al., 2019, Shang and Ryoo, 2021] attempt to tackle this issue). (2) These methods can have overly strict requirements on the video data. This limits their ability to scale to diverse internet data. For example, Sermanet et al. [2018] assumes access to multiple videos of the same scene but from different viewpoints. (3) Many of these methods only provide invariance to a single type of distribution shift (e.g., embodiment differences only). In reality, there may be many distinct shifts between a video and the robot setting.

Method: Transferable abstractions. Some methods exploit abstractions that naturally transfer well from human videos to the robot. Sieb et al. [2020], Kumar et al. [2022] use object-centric graphical representations to imitate human videos. Nagarajan and Grauman [2021] learn object-centric activity-context priors from human videos. Karnan et al. [2021], Xiong et al. [2021] detect key-points to compare human and robot videos. As touched on above, Bahl et al. [2023], Mendonca et al. [2023] leverage embodiment-agnostic affordances. These methods all respectively benefit from the use of off-the-shelf object, key-point, and human-hand detectors. Meanwhile, language has also been used as an abstraction that naturally transfers well from human videos to the robot setting [Chen et al., 2021a, Mu et al., 2023, Pertsch et al., 2022].

Method: Other approaches. Kim et al. [2023] use eye-in-hand demonstrations to bypass embodiment differences. Young et al. [2020] have humans use a manipulator similar to the robot whilst collecting video demonstrations. These approaches help bypass LfV distribution shifts, but do not present a method that can scale to internet video data.

Discussion. These methods highlight the trade-off that must be made between: (i) imposing structure and inductive biases to improve performance in narrow settings, and (ii) opting for end-to-end learning methods that are more scalable, but are less effective in the small data regime or in narrow settings. As the central focus of this survey is on scaling to diverse internet data, in order to tackle unstructured environments with generalist robots, we generally advocate for methods following the spirit of (ii). Through this lens, we regard some of the methods reviewed in this section to be less promising. For example, we have commented on the limitations of 'learned invariant representations' above.

Nevertheless, human and object-centric information is certainly relevant to robot manipulation. It remains to be seen how far these approaches can take us. Rather than explicitly using these representations downstream, a more flexible approach may be to instead bootstrap from these representations during video pretraining. For example, human and object-centric labels could be used: (i) to provide auxiliary learning objectives, or (ii) to inform automated language annotation efforts. Improving the quality of off-the-shelf models may be a worthwhile direction, given the importance of these models for obtaining human and object-centric labels from video.

5.2 Applications to RL Knowledge Modalities

This section presents our main analysis of the LfV-for-robotics literature. We taxonomise this literature according to the downstream RL knowledge modality (KM) that most directly benefits from the use of video data (see Figure 2.1). The KMs we consider here are: representations (Section 5.2.1), policies (Section 5.2.2), dynamics models (Section 5.2.3), reward functions (Section 5.2.4), and value functions (Section 5.2.5). For more general information on these KMs, we refer the reader back to Section 2.1. Within each category, we aim to summarise the methods and findings in the literature, before concluding with a brief discussion of advantages, disadvantages, and future directions.

Before beginning our review, we briefly comment on minor inconsistencies in our taxonomy. First, *representations* are not an individual RL KM per se [Wulfmeier et al., 2023], and are in fact

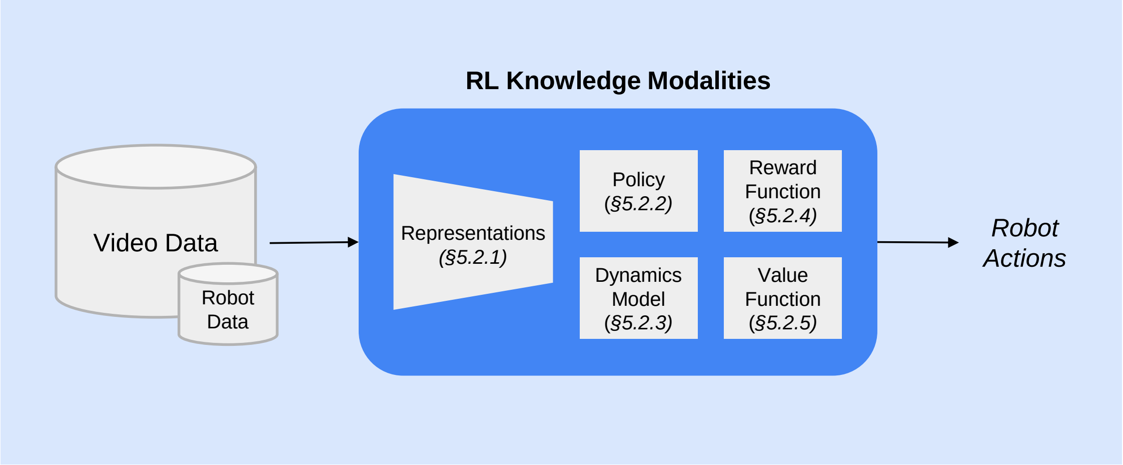


Figure 6: LfV applications to RL Knowledge Modalities (Section 5.2). Video data can be leveraged to help learn a Reinforcement Learning (RL) Knowledge Modality (KM) that can be applied in the downstream robot domain. We classify the LfV-for-robotics literature according to which KM benefits from the use of video data, resulting in the taxonomy illustrated in the above figure.

a sub-component of each of the other KMs in our taxonomy. However, the methods described in Section 5.2.1 “Representation Transfer” form a distinct cluster of the literature and we consider it worthwhile to dedicate to them a separate, self-contained section. Second, the video-pretrained *value functions* seen in the methods in Section 5.2.5 “Value Functions” are often not used as value functions downstream. Nevertheless, we consider it worthwhile to detail methods for pretraining value functions from video in a self-contained section. Overall, despite these inconsistencies, we believe our presented categorisation offers a useful and distinct organisation of existing work for the reader.

5.2.1 Representation Transfer

Representation transfer is a simple and effective method for utilising video data for downstream robot learning. Here, visual representations are first pretrained on video data using some learning objective. Then, the representations are transferred – either frozen or to be finetuned – to aid downstream learning of an RL KM. Figure 7 visualizes this process. Most commonly in the LfV literature, the pretrained representations have been used to help learn a policy during downstream imitation learning or RL (in the RL case, representation transfer may also be applied to the value function). For the methods below, the reader should assume representations are being transferred to a policy, unless stated otherwise.

In this section, we give an overview of different pretraining schemes used for representation transfer, along with corresponding downstream results and findings, as seen in the LfV literature. We note, that many LfV methods can framed as performing some form of representation transfer. Thus, to simplify this section, and avoid overlap with other sections, here we focus on methods where the video pretrained model does *not* closely resemble an RL KM.

Pretraining: Datasets. Intuitively, videos of humans interacting with objects have been a popular choice for pretraining representations for robot manipulation. Ego-centric video data has been popular [Grauman et al., 2021, Damen et al., 2018, 2022, Goyal et al., 2017]. Other human video datasets relevant to robot manipulation that have been used include: those focusing on human hands [Shan et al., 2020], and general video of humans completing tasks [Miech et al., 2019]. Whilst static image datasets (e.g., ImageNet [Deng et al., 2009]) seem less relevant, they have generally been shown to aid learning of robust visual features and have been explored in the LfV literature [Dasari et al., 2023, Zhao et al., 2022]. Meanwhile, other works seek to combine datasets to improve data diversity [Majumdar et al., 2023, Dasari et al., 2023, Radosavovic et al., 2022]. Majumdar et al. [2023] include ego-centric navigation videos in their pretraining data to aid downstream performance in navigation tasks. Others include robot-centric videos in their pretraining dataset; Dasari et al. [2023] use robot manipulation data, and Shafiullah et al. [2023] use video of humans operating a robot-like gripper.

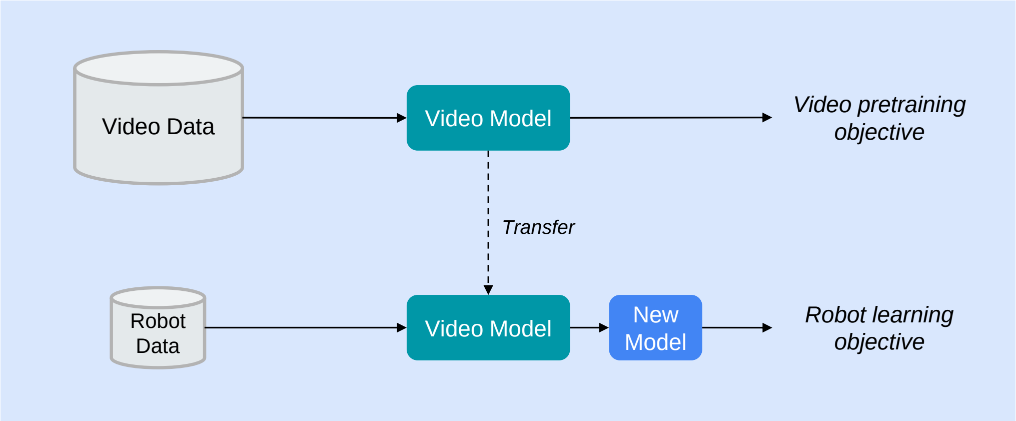


Figure 7: Representation transfer (Section 5.2.1). This figure presents a visualization of the LfV representation transfer pipeline. First, visual representations are pretrained on video data using some learning objective. Subsequently, the representations are transferred – either kept frozen or to be finetuned – to facilitate downstream learning of an RL KM.

Pretraining: Learning objectives. A number of learning objectives have been explored in the literature. Simple frame-level self-supervised learning (SSL) objectives such as masked-autoencoding (MAE) [Radosavovic et al., 2022] and contrastive learning [Chen et al., 2021b, Jing et al., 2023] are popular baseline. Zhao et al. [2022] compare a wider range of frame-level SSL objectives. To leverage the time-dimension in video, time-contrastive learning has been proposed to encode temporal dynamics information [Ma et al., 2022, 2023, Nair et al., 2022]. Image-language contrastive learning [Ma et al., 2023], video-language alignment [Nair et al., 2022] and video captioning [Karamcheti et al., 2023] objectives can exploit language labels, aiding the learning of useful semantic features. Since different objectives can learn different types of features, previous works have sought to combine multiple objectives [Nair et al., 2022, Ma et al., 2023, Karamcheti et al., 2023]. This often involves pairing an objective that learns low-level features (e.g., MAE to learn fine-grained spatial features) with one that learns higher-level semantic features (e.g., language captioning) [Karamcheti et al., 2023].

The objectives above can be considered as 'standard' auxiliary representation learning objectives, in that they only serve to learn representations for downstream representation transfer. However, any pretraining objective could be used, including those that result in a model that resembles an RL KM. In representation transfer frameworks, Bhateja et al. [2023] use a TD learning pretraining objective, Schmidt and Jiang [2023] pretrain a model to predict latent actions, Bahl et al. [2023] predict affordances from videos, whilst Seo et al. [2022a] use a video prediction pretraining objective to initialise the representations of a dynamics model.

Pretraining: Open-source models. Below, we outline some representative open-source video-pretrained models intended for LfV representation transfer. We note that, for all models, the image encoders used do not encode videos in a meaningful sense along the temporal dimension, but simply encode single images.

- *MVP [Radosavovic et al., 2022]:* The model is trained with an MAE objective. The pretraining dataset consists of 4.5M images, obtained from several sources including image data [Deng et al., 2009] and manipulation-centric human video [Damen et al., 2018, Goyal et al., 2017, Shan et al., 2020, Grauman et al., 2021]. The model is 307M parameter Vision Transformer (ViT).

- *R3M [Nair et al., 2022]:* The model is trained with a combination of time-contrastive, video-language alignment, and L1-sparsity objectives. The dataset includes 3,500 hours of video from Ego4D [Grauman et al., 2021]. The model is a convolutional ResNet-50 [He et al., 2016].

- *LIV [Ma et al., 2023]:* The model is trained with a combination of time-contrastive and image-language contrastive objectives. The video dataset used is EpicKitchens (which includes 20M frames) [Damen et al., 2022]. The model is a convolutional ResNet-50 [He et al., 2016].

- *Voltron [Karamcheti et al., 2023]:* The model is trained with an MAE and language generation objective. The video dataset used is Something-Something-v2 [Goyal et al., 2017]. The model employs a ViT architecture.

- *VC-1 [Majumdar et al., 2023]:* The model is trained with an MAE objective. The dataset includes 5.6M images, obtained from image data [Deng et al., 2009] and various video datasets, including manipulation-centric human video [Grauman et al., 2021, Shan et al., 2020, Goyal et al., 2017, Damen et al., 2018] and ego-centric navigation videos [Zhou et al., 2018]. The model employs a ViT architecture.

Downstream: Results and findings. There is a substantial body of prior work studying representation transfer from video for robotics. Several works perform large-scale analyses. This includes assessing: (i) the effects of different pretraining design choices – such as datasets, architectures, and training objectives [Jing et al., 2023, Silwal et al., 2023]; and (ii) how downstream performance is effected by distribution-shifts [Zhao et al., 2022, Burns et al., 2023], the nature of the downstream task [Majumdar et al., 2023], and the choice of downstream policy learning algorithm [Hu et al., 2023c].

Below, we attempt to give a concise overview of key findings from the literature. Note that findings can be conflicting due to differences in experimental setups. As such, we encourage the reader to find more details via the corresponding references where appropriate. These key findings are as follows:

1. Large-scale visual pretraining has consistently improved performance versus learning from scratch [Radosavovic et al., 2022, Nair et al., 2022, Parisi et al., 2022]. This includes aiding downstream data-efficiency [Nair et al., 2022] and downstream generalization [Zhao et al., 2022].

2. There is not yet a universally superior pretrained representation. Pretrained representations tend to work best in the domains they were designed for [Majumdar et al., 2023].

3. Besides size, pretraining data diversity is also crucial [Majumdar et al., 2023, Dasari et al., 2023].

4. There is evidence that control-centric video datasets (e.g., ego-centric human video) can be particularly beneficial for downstream robot manipulation [Jing et al., 2023]. However, standard, non-control-centric image datasets (e.g., ImageNet) have been shown to be surprisingly competitive in some cases [Dasari et al., 2023, Burns et al., 2023].

5. The segmentation ability of a ViT model is strong predictor of O.O.D performance [Burns et al., 2023].

6. Although like-for-like comparisons are not provided, there is evidence that control-centric pretraining objectives, such as TD-learning [Bhateja et al., 2023] or predicting affordances [Bahl et al., 2023], can outperform standard SSL objectives.

7. Versus freezing the representations, downstream finetuning can be beneficial when there is a large amount of finetuning data, but can give negative results in the few-shot setting [Majumdar et al., 2023]. In the few-shot setting, the use of an SSL objective may improve finetuning performance [Ma et al., 2023].

8. We briefly mention some other notable findings here. Sometimes performance in simulation has been a good proxy for real robot performance [Silwal et al., 2023], whilst on other occasions it has not [Dasari et al., 2023]. Majumdar et al. [2023], Jing et al. [2023] find increasing model size to improve downstream performance. Jing et al. [2023] find contrastive learning to outperform MAE. Parisi et al. [2022] find that features from early convolutional layers are better for fine-grained control task, whilst features from later layers are better for semantic tasks. Hu et al. [2023c] find that linear probing of representations can help predict downstream performances.

Discussion. Representation transfer has been demonstrated to be a simple and effective method for leveraging video data to aid downstream robotic performances. We now note some gaps in the literature that point towards promising research directions.

Jointly representing spatio-temporal information. Representations that jointly embed spatio-temporal video information have been neglected in the LfV literature. Almost all of the works discussed above learn encoders that separately embed each image. As touched on in Section 4, the video ML literature has demonstrated the benefits of jointly encoding spatio-temporal information [Yu et al., 2023a, Villegas et al., 2022].

Neglected control-centric learning objectives. Pretraining objectives specifically designed to represent control-centric information seem neglected in the literature. Initial promising works here includes using TD learning objectives [Bhateja et al., 2023], predicting alternative action representations (see Section 5.1.1) in video [Schmidt and Jiang, 2023, Bahl et al., 2023], or video prediction objectives [Wu et al., 2023a]. Further research down these avenues will likely yield promising results.

Neglected data. First, we note that larger-scale internet-scraped video-text datasets (see Section 6.3) used to train video foundation models have yet to be employed in the LfV representation transfer literature. Second, pretraining objectives that leverage other modalities or labels paired with video, beyond language annotations, could be useful. The use of widely available audio data has been neglected in the LfV literature. Leveraging object-centric labels for auxiliary objectives – for example, by predicting object bounding boxes or segmentation masks, or predicting human hand poses [Bahl et al., 2023] – may yield improved features for robot manipulation. The commonly used Ego4D dataset [Grauman et al., 2021] includes several additional modalities (such as 3D meshes or eye-gazes) that have yet been exploited.

Representation transfer from video foundation models. The capabilities of video foundation models are ever improving [Zhao et al., 2024, Brooks et al., 2024] (see our analysis of video foundation models in Section 4). Leveraging them for LfV representation transfer is a highly promising direction. In related work, promising results have been obtained using foundational image-language models for representation transfer [Brohan et al., 2023].

5.2.2 Policies

Ultimately, the goal of LfV is to use video data to help obtain a policy $\pi(a_t|s_t)$. In this section, we detail literature where $\pi(a_t|s_t)$ is the RL KM that most directly benefits from the use of video data. We first identify and detail several distinct categories of methods (see Figure 8), before moving into a brief discussion.

Method: Joint training on video and robot data. Here we refer to methods that train a monolithic model jointly on video and robot data. Recent trends towards self-supervised any-to-any sequence modelling [Liu et al., 2024b], and towards the use of cross-robot embodiment data [Team et al., 2023b], suggest these methods may become increasingly popular and effective. However, these approaches are still in their nascent stages. Reed et al. [2022] train a sequence model on multi-modal data, including robot data, but do not use video data. Sohn et al. [2024] recently train an 8 billion parameter any-to-any transformer on text, images, videos, robot actions, and a range of numerical sensor readings, but experimental details are not published. These methods are promising due to their simplicity and scalability. They can scale well with both increasing robot and video dataset sizes, and can hope to obtain positive transfer via the use of multiple modalities of data.

Method: Representation transfer. Here, we touch on methods that pretrain a model on video, and finetune it on robot data to output robot actions (i.e., to act as a policy). Wu et al. [2023a] pretrain a transformer video prediction model, and supervised finetune it on robot data to additionally output actions. Schmidt and Jiang [2023] finetune a pretrained latent action policy to output robot actions via online RL. Note, most methods detailed in Section 5.2.1 can be framed as falling under this category [Nair et al., 2022, Karamcheti et al., 2023, Radosavovic et al., 2022]. In related work, Brohan et al. [2023] take a VLM pretrained on image-text data and finetune it to output robot actions.

Method: Alternate-action policies $\pi_{\text{alt}}(\hat{a}_t|s_t)$. In Section 5.1.1, we outline different *alternative action* representations $\hat{\mathcal{A}}$ seen in the LfV literature. Such representations can be used to train an alternate-action policy $\pi_{\text{alt}}(\hat{a}_t|s_t)$ from video data. We give more details on how $\pi_{\text{alt}}(\hat{a}_t|s_t)$ can be used downstream in the following paragraph. Here, we detail methods for learning $\pi_{\text{alt}}(\hat{a}_t|s_t)$, categorized by the type of $\hat{\mathcal{A}}$ used.

Single-step latent actions. Edwards et al. [2018], Schmidt and Jiang [2023], Bruce et al. [2024] all label expert video data with latent actions using a learned latent inverse dynamics model. This labelled video data is used to train a latent-action policy $\pi_{\text{alt}}(\hat{a}|s)$ using behaviour cloning.

Multi-step latent actions. Wang et al. [2023a] learn a latent planner that takes in the current image and goal image and outputs a 'latent plan'. Pertsch et al. [2022] train two separate high-level policies from video data to give representations of the actions that should be taken in future time-steps. We note that, in the literature, multi-step latent action representations are mostly used downstream for 'video-as-instructions' methods (see below), rather than to train a policy $\pi_{\text{alt}}(\hat{a}|s)$. Considering video is well suited to provide abstracted multi-step action representations, more research into training a $\pi_{\text{alt}}(\hat{a}|s)$ here could be fruitful.

Language-actions. Ajay et al. [2023] use a pretrained LLM as an alternative action policy to generate language actions that condition video predictions for a 'video-as-policy' method (see below). Elsewhere, to obtain language-policies, Yang et al. [2023c], Du et al. [2023b] supervised finetune internet-pretrained VLMs on language-labelled video data. Mu et al. [2023] use LLMs to convert language descriptions of video into language plans, and the new language plans are used to finetune an LLM language planner.

Observations-as-action. (1) 'Next-observation-as-action' policies: These methods train policies on video data to propose the next observation (or sequence of observations) the robot should observe. The observation may be in the form of an image, or an image representation. Later we detail related 'video-as-policy' methods that propose video trajectories that can be mapped to actions [Du et al., 2023a]. Thomas et al. [2023] learn a planner via a video prediction objective that proposes video trajectories in image embedding space. (2) 'Sub-goal' policies: These methods train a policy that proposes a sub-goal image (or image representation). Black et al. [2023b] supervise finetune an image-editing diffusion model on video data to edit the current image observation into a subgoal image. From action-free data, Park et al. [2024] learn a high-level policy that outputs image representations as subgoal. For more details on how sub-goals can be identified in the video data, see Section 5.1.1.

Visual motion information. From videos labelled with 2D point trajectories, Wen et al. [2023] train a policy to output point trajectories which can act as sub-goals for a low-level robot policy. Yuan et al. [2024] follow a similar scheme, but use RGB-D data, enabling the use of 3D point trajectories.

Human-hand-aware actions. Several works have labelled videos with affordances or hand pose information (see Section 5.1.2) and trained a "policy" to propose an affordance/pose given an image observation [Bahl et al., 2023, Shaw et al., 2022, Qin et al., 2022, 2021, Peng et al., 2018]. Bharadhwaj et al. [2023] learn from video a plan predictor that predicts future hand and object masks.

Method: Alternative action decoders $\pi(a_t|\hat{a}_t, s_t)$. Here, we refer to methods that train a low-level robot policy $\pi(a_t|\hat{a}_t, s)$ to be conditioned on an alternative action representation – i.e., it must decode $\hat{a}_t$ to a_t. There are two ways a policy $\pi(a_t|\hat{a}_t, s_t)$ can be used downstream.

(1) Hierarchical conditioning via $\pi_{alt}(\hat{a}_t|s_t)$. If an alternative-action policy $\pi_{\text{alt}}(\hat{a}_t|s_t)$ has been trained from video, then a decoding policy $\pi(a_t|\hat{a}_t, s_t)$ can be used to decode robot actions from the alternative-action policy outputs: $\pi(a_t|\pi_{\text{alt}}(\hat{a}_t|s_t), s_t)$. This decoding policy can either be learned through data, or can be manually crafted. When learning the decoder, a common approach is to label a robot dataset with alternative actions – obtaining tuples of the form $(s_t, a_t, \hat{a}_t, s_{t+1})$ – and learn the mapping via supervised learning or reinforcement learning. For example, this has been done with single-step latent actions [Schmidt and Jiang, 2023, Bruce et al., 2024], multi-step latent actions [Wang et al., 2023a], and point-trajectory latent actions [Wen et al., 2023]. This decoding can also be framed as a goal-conditioned behaviour cloning problem [Black et al., 2023b]. In 'video-as-policy' methods [Du et al., 2023a], the decoding can be framed as an inverse dynamics model $p^{-1}(a_t|s_t, s_{t+1})$. These approaches are often compositional; $\pi_{\text{alt}}(\hat{a}_t|s_t)$ (trained primarily with video data) and $\pi(a_t|\hat{a}_t, s_t)$ (trained with labelled robot data) are trained separately then combined. To improve composition, Ajay et al. [2023] enforce consistency between neighbouring levels of the hierarchy via iterative refinement. In contrast to compositional approaches, Liu et al. [2022] learn the mapping $\pi(a_t|\hat{a}_t, s_t)$ via a decoupled generative adversarial training scheme. Other works explore obtaining the mapping $\pi(a_t|\hat{a}_t, s_t)$ without the use of robot data. Here, Ko et al. [2023] infer low-level robot actions from optical flow, Yuan et al. [2024] infer actions from 3D flow

predictions, whilst Nasiriany et al. [2024] map from robot action to visual arrows and back. Other works have retargeted human-hand poses to robot poses [Qin et al., 2021, Sivakumar et al., 2022].

(2) Video-as-instructions. Here, a video instruction can first be embedded into a compressed representation $\hat{a}_t$, allowing the low-level policy to more easily learn to follow the instructions [Chane-Sane et al., 2023, Cai et al., 2023, Lifshitz et al., 2023, Xu et al., 2023]. The video usually comes in the form of a human (or robot) demonstration.

Method: Policy-as-video. These methods use video prediction models (pretrained on video data) as robot policies. Du et al. [2023a] introduce this paradigm. They (i) use a language-conditioned video predictor to generate plausible video trajectories that complete a language-specified task, then (ii) use an action-decoding IDM (trained on robot data) to extract actions from the video. Note, these video-as-policy methods fit into the *hierarchical conditioning via* $\pi_{alt}(\hat{a}_t|s_t)$ scheme outlined above (here the generated video is the alternate action). Ajay et al. [2023] extend the approach in Du et al. [2023a] by firstly using LLMs as language planners to condition video predictions, and secondly by enforcing consistency between the LLM plans, the video generations, and the IDM actions. Du et al. [2023b] improve the action decoding model by switching the IDM objective to a more flexible goal-conditioned behaviour cloning objective. Ko et al. [2023] show that actions can be decoded without requiring any robot data via the use of optical flow. Due the computational costs of these policy-as-video methods, Ye et al. [2023] distill behaviours into a simpler policy. We note that the video prediction models in video-as-policy methods can also be framed as dynamics models (indeed, Section 5.2.3 details how such video predictors can be used as simulators [Yang et al., 2023c], or for planning [Du et al., 2023b]).

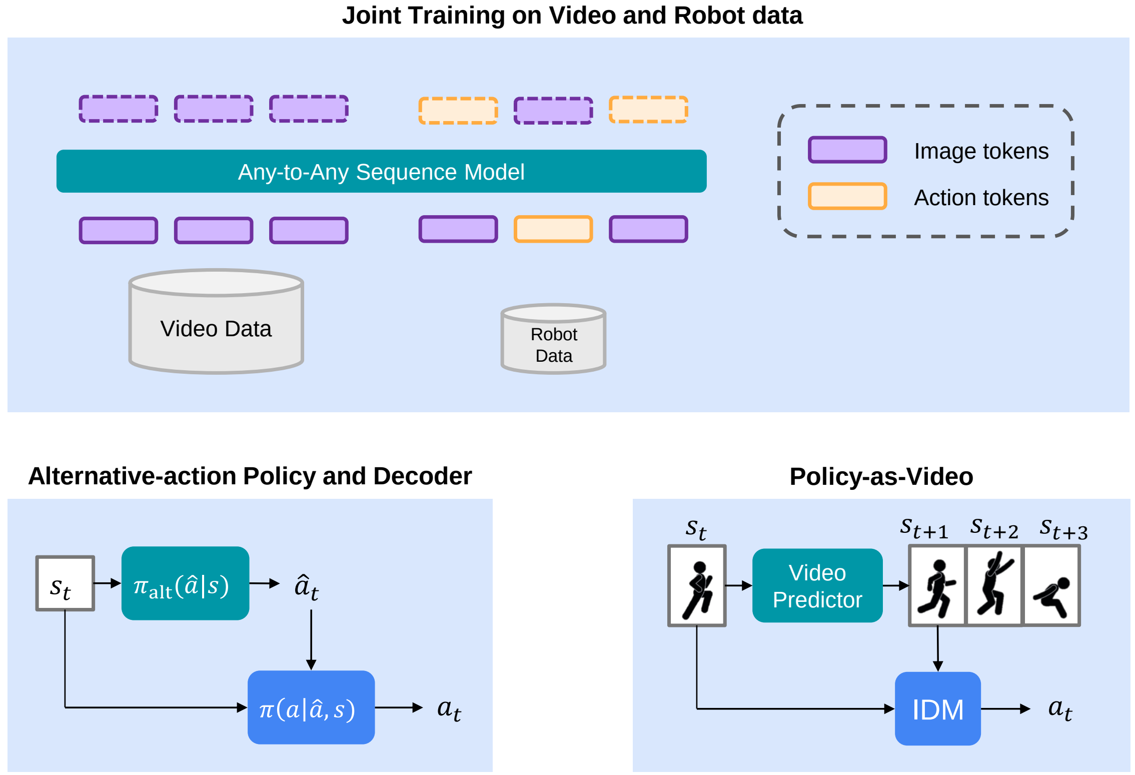
Method: IDM pseudo-actions. These methods train $p^{-1}(a_t|s_t, s_{t+1})$, an inverse-dynamics model (IDM) on an action-labelled robot dataset, then use the IDM to label transitions in a video dataset with pseudo-actions. The pseudo-labelled video data can then be used to help train the policy $\pi(a|s)$. Baker et al. [2022] relabel Minecraft videos from Youtube to increase the size of their dataset, using the relabelled data for offline behaviour cloning. However, as mentioned in Section 5.1.1, naive IDM pseudo-action approaches are unlikely to scale to diverse internet videos.

Results: Performance gains. The methods in this section show promising signs of achieving the potential benefits of LfV (see Section 3.2). Results have demonstrated improved generalization beyond the robot dataset [Du et al., 2023a, Wang et al., 2023a, Black et al., 2023b, Bruce et al., 2024, Thomas et al., 2023], improved long-horizon performances [Ajay et al., 2023], and improved efficiency during online learning [Schmidt and Jiang, 2023, Ye et al., 2023]. However, many only do using toy video datasets; for example, using robot data stripped of action labels [Schmidt and Jiang, 2023, Wen et al., 2023, Ko et al., 2023, Thomas et al., 2023], or human videos in the same environment as the downstream robot [Wang et al., 2023a]. This is in contrast to the setting we are interested in: scaling to diverse internet video. When adding diverse human videos to pretraining, performance gains have generally been modest [Wu et al., 2023a, Du et al., 2023a, Ajay et al., 2023, Shaw et al., 2022, Black et al., 2023b]. This may be because these methods have thus far only used smaller-scale video data (relative to internet-scale).

Discussion. We discuss the advantages and disadvantages of targeting the policy KM, before highlighting promising directions.

Advantages and disadvantages. As our end-goal is to obtain a policy, this is one of the most promising RL KMs to target. However, it is difficult to obtain a policy from a video dataset D_{video} alone; video lacks lower-level information (see Section 3.3) a robot policy may require, in addition to lacking action labels. Thus, some robot data D_{robot} is often needed to help train the policy. These pros and cons contrast those of reward functions. LfV reward functions (see Section 5.2.4) can perform well after being trained only on video data [Sontakke et al., 2023, Escontrela et al., 2023]. However, once the LfV reward function is obtained, one must perform significant additional online training to obtain a robot policy.

Promising directions. In general, we consider methods that are most scalable to internet video, and most generally applicable to the generalist robot setting, to be most promising. 'IDM pseudo actions' are not scalable to diverse internet video. Vanilla 'video-as-instructions' are not scalable to the generalist robot settings as they rely on human-provided video demonstrations. However, methods that can naturally improve with advances in video foundation modelling are particularly


Joint Training on Video and Robot data

Alternative-action Policy and Decoder

Policy-as-Video

Figure 8: Learning policies from video (Section 5.2.2). Promising approaches for leveraging video data to help learn a robot policy are visualised. Any-to-any sequence models can be jointly trained on video and robot data, and can predict robot actions downstream [Sohn et al., 2024]. An alternative-action policy $\pi_{\text{alt}}(\hat{a}|s)$ can be trained from video and used to condition an alternate-action decoder $\pi(a|\hat{a}, s)$ trained on robot data [Schmidt and Jiang, 2023, Wang et al., 2023a, Wen et al., 2023]. 'Policy-as-video' methods are a distinct instance of the alternate-action setup where $\hat{a}$ is a generated video.

promising (e.g., 'joint training on video and robot data', 'representation transfer', and 'video-as-policy' methods). It remains to be seen whether methods that utilise alternative action representations will scale well and ultimately prove to be beneficial.

An interesting point to discuss is the contrast between monolithic approaches (e.g., 'joint training on video and robot data') and compositional approaches (e.g., hierarchical approaches that combine a $\pi_{\text{alt}}(\hat{a}|s)$ with a $\pi(a|\hat{a}, s)$). It is unclear yet which of these approaches are more promising and may best attain the potential benefits of LfV. Future research should seek to answer this question. More detailed discussions on monolithic versus compositional approaches are found in Section 8.2.

There are a number of interesting directions for monolithic approaches here. First, any-to-any foundation models trained on both video and robot data have been underexplored. Second, as noted in Section 5.2.1, the use of video foundation models for LfV representation transfer is promising but underexplored.

There are also directions for improving compositional approaches. One direction is to develop improved alternative action representations (see more discussion in Section 5.1.1). In general, training latent action policies [Schmidt and Jiang, 2023, Bruce et al., 2024] from large-scale video seems promising but underexplored. Elsewhere, 'Policy-as-video' methods can be improved by addressing foundational video prediction model issues (see Section 4.2). Finally, an interesting direction is to explore the extent to which either π_{alt} or π is the performance bottleneck in hierarchical conditioning $\pi(a_t|\pi_{\text{alt}}(\hat{a}_t|s_t), s_t)$ approaches.

5.2.3 Dynamics Models

Video prediction models, denoted as $p(s_{t+1}|s_t)$, capture temporal dynamics information – information about the dynamics and physics of the world. They capture information relevant to, and perform a functionality similar to, a robot dynamics model $p(s_{t+1}|s_t, a_t)$. As such, a line of LfV research has sought to use video prediction objectives on video datasets to improve the learning of robot dynamics models. This can involve pretraining $p(s_{t+1}|s_t)$ on D_{video} and adapting it into $p(s_{t+1}|s_t, a_t)$ using D_{robot}. Alternatively, it could involve jointly pretraining on video and robot data.

It is worth noting that a standard video prediction model may have several deficiencies when it comes to employing it as a dynamics model. These include:

1. *Action-conditioning:* A standard dynamics model $p(s_{t+1}|s_t, a_t)$ should be conditioned on low-level robot actions a_t. A video predictor $p(s_{t+1}|s_t)$ trained on internet video cannot be robot action-conditioned (though it could be conditioned on an alternative-action representation to give $p_{\text{alt}}(s_{t+1}|s_t, \hat{a}_t)$; see more on alternative actions in Section 5.1.1).

2. *Missing low-level information:* Video only contains visual information. However, to make accurate predictions, a robotic dynamics model may need to account for low-level information not contained in images – such as forces and tactile information.

3. *Distribution shifts*: There may be distribution shifts between the video data and the target robot domain. Thus, a pretrained video predictor may not transfer well zero-shot to the robot domain.

Much of the research in this section investigates mechanisms to address these issues. We now present details of such methods. We begin by detailing the various video prediction pretraining schemes seen in the LfV literature, before outlining how the pretrained models can be adapted and utilized for downstream robotic settings.

Pretraining: Architectures and Datasets. Here, we outline the model architecture and dataset combinations used to train video prediction models in the LfV literature. Some works simply use toy simulated video datasets [Seo et al., 2022a], or small-scale custom human videos [Schmeckpeper et al., 2019]. However, others have scaled to larger internet video datasets [Yang et al., 2023c, Bruce et al., 2024].

Due to their SOTA video generation abilities and flexible conditioning (i.e., language conditioning), video diffusion architectures have been a popular architectural choice. These diffusion models have been scaled to increasingly large human video datasets [Yang et al., 2023c, Du et al., 2023a,b]. UniSim [Yang et al., 2023c] train a 5.7B parameter diffusion model on a combination of human and robot video datasets. This includes the use of 3.5M ego-centric human videos from Ego4D [Grauman et al., 2021], 13M internet videos, and 800M internet text-image pairs.

Autoregressive transformer architectures have also been employed [Hu et al., 2023a, Wu et al., 2023a, Bruce et al., 2024]. Inputs (i.e., image sequences and text) are first embedded into tokens and the next frame is predicted autorgressively. Bruce et al. [2024] introduce a spatial-temporal transformer [Xu et al., 2020] tokenizer to incorporate temporal dynamics into embeddings, and predict next frames autoregressively using MaskGit [Chang et al., 2022a]. They train an 11B parameter foundation model on 200,000 hours of publicly available internet gaming video, showing their architecture to scale well with additional computational resources. Elsewhere, in the domain of self-driving, Hu et al. [2023a] train an auto-regressive transformer on 4,700 hours of proprietary driving video, also demonstrating predictable improvements as compute and model size is scaled.

Other LfV works [Seo et al., 2022a, Mendonca et al., 2023] base their video predictor on the Dreamer world-model architecture [Hafner et al., 2023]. This architecture – a recurrent state-space model that operates in a learned latent space – has achieved SOTA performances in model-based RL [Hafner et al., 2023]. Wu et al. [2023b] adopt a specialized version of the Dreamer architecture that separates modelling of context information and dynamics information, allowing them to avoid underfitting when training on human videos.

Pretraining: Action-conditioning. In pretraining, the action-conditioning problem mentioned above can mitigated by either: (1) pretraining jointly on both video and robot data, allowing the video predictor to be natively conditioned on robot actions [Yang et al., 2023c, Sohn et al., 2024], or (2) pretraining an alternative-action-conditioned video predictor $p_{\text{alt}}(s_{t+1}|s_t, \hat{a}_t)$ [Yang et al., 2023c, Bruce et al., 2024]. Alternative-action-conditioning allows for control over video generations, and alternative actions can often be mapped to robot actions [Rybkin et al., 2018, Schmidt and Jiang, 2023]. Thus, a $p_{\text{alt}}(s_{t+1}|s_t, \hat{a}_t)$ can be more immediately usable as a dynamics model than $p(s_{t+1}|s_t)$. We now outline the different action conditioning schemes seen in the LfV literature.

Language-as-action. As videos are often paired with text descriptions, a popular choice is to train text-conditioned video predictors [Yang et al., 2023c, Du et al., 2023b]. Language can allow for intuitive and meaningful control over the generated video. Note, these language-actions are generally temporally extended – i.e., a single language 'action' may result in multiple timesteps of generated video.

Single-step latent actions. Video predictors have been trained to be conditioned on learned latent actions [Rybkin et al., 2018, Bruce et al., 2024], allowing predictions to be conditioned at each time-step.

Observations-as-action. The video predictor could be conditioned on a goal image, encouraging generation of a video that connects the current observation to the goal image [Du et al., 2023a].

Conditioning on multiple action-types. UniSim [Yang et al., 2023c] train a large-scale video diffusion model to be conditioned on several different action-spaces. This is achieved by jointly pretraining on video and robot data. The model is trained to be conditioned on robot actions (obtained from robot data), language actions (obtained from language-labelled videos and images), and camera motion actions. We note that the robot action-conditioning is only meaningful when applied to robot video.

Other. Mendonca et al. [2023] pretrain a video predictor to be conditioned on future grasp-location and post-grasp waypoint affordance information. Yuan et al. [2021], Wang et al. [2023b] use motion information for conditioning. As touched on above, Yang et al. [2023c] can condition generations on camera motion and camera pose information.

Downstream: Adapting the video predictor. A pretrained video predictor may require adaptation before it can be used as a dynamics model. An obvious way to adapt a video predictor is to finetune it on robot data [Du et al., 2023a, Ajay et al., 2023]. This could involve adapting the model's input-space to allow conditioning on robot actions – though we note there are few works that finetune an action-free video predictor $p(s_{t+1}|s_t)$ into an action-conditioned dynamics model $p(s_{t+1}|s_t, a_t)$. Mendonca et al. [2023] pretrain on video with affordance-based conditioning, and add optional robot action-conditioning during finetuning to create a hybrid action space. Seo et al. [2022a] find naive finetuning on robot data to result in the erasure of pretraining knowledge. Instead, they leverage the pretrained video prediction model by 'stacking' a new action-conditioned model on top of it. In related work, Yang et al. [2023b] demonstrate that the score function of a large pretrained video diffusion model can act as a probabilistic prior to guide the generations of a task-specific small video model.

Note, where applicable, adaptation can involve obtaining a mapping from the pretraining alternative actions $\hat{a}$ to robot actions a. For example, Rybkin et al. [2018] perform MPC using latent action-conditioning and learn a mapping from $\hat{a}$ to a to execute these plans. We direct the reader to Section 5.2.2 for more information on methods that map from $\hat{a}$ to a.

Downstream: Using the dynamics model. Once the dynamics model has been obtained (e.g., by adapting a video predictor), it has been used in several ways in the LfV literature:

- *As a simulator*: $p(s_{t+1}|s_t, a_t)$ can be used as a simulator to generate synthetic data. UniSim [Yang et al., 2023c] do so to: (i) train a policy with RL using synthetic rollouts; and (ii) train a policy with behaviour cloning via hindsight relabelling of synthetic trajectories.
- *As a differentiable simulator*: Similarly, the dynamics model can be used as a differentiable simulator. Seo et al. [2022a], Wu et al. [2023b] backpropagate through model-generated rollouts as part the Dreamer model-based RL algorithm [Hafner et al., 2023].
- *For planning*: A common choice is to use the dynamics model for planning. Several LfV works use the model for standard MPC [Rybkin et al., 2018, Mendonca et al., 2023]. Du et al. [2023b] use language actions to condition video generations and perform a tree-search to choose suitable video-based plans.

We finally note that these use-cases often require evaluation of the model-generated trajectories; i.e., reward or return estimates are required. Some works learn these estimates via a downstream finetuning stage, making use of reward-labelled robot data [Seo et al., 2022a]. Mendonca et al. [2023], Rybkin et al. [2018] use representational similarity between the current observation and a goal image to obtain a reward signal. Yang et al. [2023c] pretrain a model from scratch, whilst Du et al. [2023b] similarly finetune a VLM to predict the number of steps till task completion in a video. Note, other methods for learning reward functions (see Section 5.2.4) or value functions (see Section 5.2.5) from video may be applicable here.

Downstream: Performance Gains. Promisingly, there is evidence in the literature that the use of video data can boost the performance of model-based RL approaches. Seo et al. [2022a] demonstrate improved robot data-efficiency, but use a toy video dataset. When pretraining with large human video datasets, several works have demonstrated moderate performance gains [Mendonca et al., 2023, Wu et al., 2023b]. Yang et al. [2023c], Du et al. [2023b] demonstrate that large-scale pretrained video diffusion models can beat strong baselines, notably in long-horizon tasks. We note however that there has yet to be concrete evidence of video pretraining allowing for significant generalization beyond the robot dataset (one of the key potential benefits of LfV, as outlined in Section 3.2).

Discussion. Recent breakthroughs in video generation [Brooks et al., 2024], combined with recent progress in the LfV literature [Yang et al., 2023c, Bruce et al., 2024], make using video data to help learn dynamics models is an attractive direction. We now briefly touch on related discussion points.

Advantages and Disadvantages. Video prediction is highly analogous to visual dynamics modelling, and dynamics models are very useful for obtaining a robot policy. As such, targeting the dynamics model KM with video data is a promising LfV direction. However, there are a number of challenges here. First, like policies, dynamics models often require low-level information unavailable in video. Second, foundational video prediction models are prone to hallucinations. For example, generated video may have unrealistic physics [Brooks et al., 2024, Yang et al., 2024]. Similarly, the video predictor may not fully understand what the robot can or cannot control in the environment, leading to it generating imagined video plans where uncontrollable elements behave more favourably than they realistically would [Yang et al., 2022].

Improving and leveraging foundational video predictors. Improvements in video prediction will be key to advancing LfV dynamics model approaches. This should include addressing the issues of hallucination mentioned above. We note, however, that state-of-the-art foundational video prediction models (see Section 4.2) are either closed source or have yet to be used for LfV robotics applications. Besides leveraging state-of-the-art video prediction models, another interesting direction is to train foundational video prediction models customized for LfV purposes [Yang et al., 2023c, Sohn et al., 2024]. This could involve experimenting with scalable alternative action conditioning techniques (Bruce et al. [2024] present a promising option here). Finally, we note that incorporating 3D information could improve the applicability of video prediction models to low-level robotics [Zhen et al., 2024].

In this section, we have seen that a video prediction model can be used as a simulator [Yang et al., 2023c] or for planning purposes [Du et al., 2023b]. In the previous section, we additionally saw that video predictors can be employed for 'video-as-policy' approaches [Du et al., 2023a], or finetuned to directly output actions [Wu et al., 2023a]. It is not yet clear which of these is the most effective method for utilising foundational video prediction models for robotics.

Other promising directions. (1) Improved mechanisms for efficiently adapting a video predictor into a dynamics model should be investigated. (2) Hierarchy in world models has often been advocated for [LeCun, 2022]. This may be particularly relevant to the LfV setting. Videos lack low-level information, and maintaining a hierarchy where higher levels are learned from video data and the lower-levels are learned from robot data may be suitable. (3) It is worth noting that approaches that learn dynamics models from video are particularly promising for autonomous driving, due to the existence of large driving video datasets [Hu et al., 2023a, Caesar et al., 2019]. (4) An interesting but underexplored direction is to combine standard analytic simulation [Todorov et al., 2012, Makoviychuk et al., 2021] with video prediction simulation [Yang et al., 2023c] to obtain the benefits of both. (5) Finally, using foundational video predictors to generate synthetic data is an interesting direction. This can include generating synthetic video rollouts, or augmenting existing data using video editing capabilities (similar to the image editing performed in [Yu et al., 2023b]).

5.2.4 Reward Functions

The reward function is an essential component of an RL algorithm. However, manual design and implementation of a reward function can be difficult for a number of reasons. First, in the real world, complex sensing systems may be required to track reward-relevant information. Second, reward shaping can be tricky, even for seemingly simple behaviours [Popov et al., 2017]. Thus, manual reward design is not scalable to real-world generalist robot settings. LfV research has sought to tackle this issue by extracting visual reward functions from video data. In this section, we detail any method that uses video data to help relabel transition tuples (s_t, a_t, s_{t+1}) to (s_t, a_t, r_t, s_{t+1}), thus allowing the tuples to be used for online or offline RL.

Extracting reward functions from video. We now describe the main clusters of methods for extracting and constructing reward functions from video data.

- **Video-language model rewards.** These methods specify the task via language and use a video language model (VidLM) to provide a reward signal. These methods are promising for two reasons. First, language is a simple and intuitive way to specify a task. Second, foundational VidLMs are likely to continue to improve into the future. More details on foundational VidLMs can be found in Section 4. We identify two categories of methods here.

  (1) *Video-text similarity*: A dual encoder video-language model can be trained to embed videos and language into the same representation-space [Xu et al., 2021, Zhao et al., 2024]. From such a model, a reward can be defined as the similarity (in embedding space) between a language task-description and a video of the robots behaviour. Fan et al. [2022], Ding et al. [2023b] do so in MineCraft, leveraging internet videos to train the dual-encoder. Sontakke et al. [2023] pretrain the dual-encoder on a human video dataset to provide rewards for simulated robot manipulation tasks. We note that dual-encoder image-text rewards have been studied in detail [Baumli et al., 2023].

  (2) *Visual question answering*: A video-to-text model (see Section 4.3) can provide rewards via visual question answering (VQA). A simple technique here is to use the model as a success checker. Here the model is passed a video of the robots attempt and asked to classify whether the task has been completed; the answer can then be converted into a sparse reward. A denser reward could be achieved by asking the model to score the robots progress, or by using the model to provide feedback within an 'RL-AI-F' framework [Klissarov et al., 2023]. In the literature, such VQA rewards have been used with images as input [Du et al., 2023c], whilst Yang et al. [2023a] use video data to finetune an image-based VQA success checker. However, the limited capabilities of current video-to-text models has thus far prevented successful use of video-based VQA rewards.

- **Video-predictors as reward functions.** A video prediction model $p(s_{t+1}|s_t)$ can be pretrained on video data and converted into a reward function. These methods define a reward based on the likelihood of the robot video under the video predictor – this implicitly encourages the robot to match the behaviour distribution of the video data. Zhu et al. [2023b] define the reward as the difference between the achieved s_{t+1} of the agent and the predicted s_{t+1} from the video predictor. Escontrela et al. [2023] define the reward as the probability of s_{t+1} according to an autoregressive transformer video predictor. Huang et al. [2023b] pretrain a diffusion video predictor, and define the reward as the negative of its conditional entropy, arguing that this can capture complex behaviour distributions. Thus far, works in this space have used toy and expert video data. However, language-conditioned video prediction may allow these approaches to scale to diverse, non-expert internet video [Escontrela et al., 2023]. Overall, these methods are promising as they will improve with advances in foundational video prediction models [Brooks et al., 2024].

- **Representational similarity to a reference.** Here, we refer to methods that define the reward as the similarity between the robot's observation and a reference observation (i.e., a goal image or demonstration video). These methods require the use of a visual representation that can provide meaningful similarity comparisons. We outline two distinct lines of work here.

  (1) *Standard deep representations:* When a goal image is provided, representational similarity between the current observation and the goal image can define the reward. This has been done using representations learned from video data (see Sections 4.1 and 5.2.1 for details on learning visual representations from video). We note that some representation spaces can be more effective than others. A principled approach is to use representations obtained via time-contrastive objectives [Ma et al., 2022]; these representations can give implicit measurements of the temporal distance between two images. Related to this, Quasimetric functions can be learned from video to provide a meaningful distance metric [Wang et al., 2023d]. Hu et al. [2023c] provide a detailed study of the efficacy of many different types of representations learned from video. They find that many standard representations can be effective, though masked autoencoding-based representations perform poorly.

(2) *Representations that address LfV distribution shifts:* When comparing a robot video to a human reference video, distribution-shifts (such as embodiment differences) can prevent meaningful comparison. Here, we detail methods that use representations designed to explicitly ignore such distributions shifts when defining their reward. A number of methods have been proposed to overcome the LfV embodiment gap. Zakka et al. [2021] use embodiment agnostic representations obtained from a temporal-cycle consistency loss. Kumar et al. [2022], Sieb et al. [2020] use object-centric graphical representations. Mandikal and Grauman [2022], Qin et al. [2021] compare robot and human hand poses via pose retargeting. Kim et al. [2023] use task labels on human and robot videos to learn representations that ignore nuisance details, such as embodiment differences. Elsewhere, Sermanet et al. [2018], Aytar et al. [2018] compare the similarity between the robot and a reference video using viewpoint invariant representations. Other related research encourages the robot to match human behavioural priors using embodiment agnostic affordances [Bahl et al., 2023], or factorized representations [Chang et al., 2023, Shang and Ryoo, 2021]. We give more details on these representations in Section 5.1.2, where we note that representations that explicitly address LfV distribution shifts often rely on assumptions that limit their scalability to internet video and unstructured robotic environments.

- **Potential-based shaping with value functions.** A value function $V(s_t)$ pretrained from video can be used to provide a dense reward via 'potential-based shaping'. Such rewards are based on the difference in estimated value between the new state and the previous state: $r_t = V(s_{t+1}) - V(s_t)$. In the LfV literature, these rewards have been defined using value functions pretrained on video data via TD learning [Chang et al., 2022b], or time-contrastive learning [Ma et al., 2022]. We refer the reader to Section 5.2.5 for more details on how to train value functions from video data.

- **Generative adversarial imitation.** Generative adversarial imitation learning [Ho and Ermon, 2016] approaches have been used to encourage the robot to match the behaviour distribution of a video dataset during online learning [Torabi et al., 2018b]. This has required the use of representations that ignore LfV distribution shifts (see Section 5.1.2), such as viewpoint invariant representations [Stadie et al., 2017], or human-to-robot hand pose retargetting [Qin et al., 2021].

- **Other methods.** There are several other distinct methods worth noting. A task classifier can be trained on task-labelled video data to provide downstream rewards [Chen et al., 2021a, Shao et al., 2020]. Several methods use the number of steps-to-completion of the video as a proxy label to train a video reward model [Yang et al., 2023c, Edwards and Isbell, 2019]. Other methods encourage similarity to a video-obtained behavioural prior: such as a video-pretrained policy [Ye et al., 2023], or a human affordance distribution [Bahl et al., 2023].

Downstream Usage: Mechanisms of transfer. The simplest way to use an LfV-obtained reward function is zero-shot, as the task reward [Escontrela et al., 2023]. Some works further finetune the reward function on in-domain robot data to improve its accuracy [Sontakke et al., 2023]. Other works use the LfV reward as an exploration or shaping bonus, in addition to a sparse task reward [Ye et al., 2023, Chang et al., 2022b]. Adeniji et al. [2023] pretrain the policy using the LfV reward, before finetuning the policy on the true task reward. Although the literature mainly explores using LfV rewards for online RL, it is worth noting they could also be used to provide reward labels for an offline RL dataset.

Discussion. We first discuss the advantages and disadvantages of learning reward functions from videos (versus learning other RL KMs), before discussing promising directions.

Advantages. Learning capable reward functions purely from video data may be more feasible than for other RL KMs. This is for two reasons. (1) Reward functions can often perform well whilst taking only visual information as input. This is unlike policies or dynamics models which ultimately may need access to non-visual information unavailable in video. LfV reward functions can thus more reasonably be used zero-shot after video-only pretraining [Escontrela et al., 2023]. (2) We note that evaluation is often easier than generation. Thus, the task of an 'evaluating' reward function may be easier to learn from passive video data than that of a 'generating' policy or dynamics model.

Disadvantages. Our ultimate goal is to obtain a generalist robot policy. If only a reward function is extracted from the video data, we may still require prohibitive quantities of robot data to learn the generalist policy. This is in contrast to other LfV methods (e.g., policies or dynamics models) which may better reduce demands on the robot data and aid generalization beyond the robot data.

Promising directions. (1) The most promising approaches for learning reward functions from video are likely those that can leverage internet-pretrained video foundation models. 'Video-language model rewards' and 'Video-predictors as reward functions' will continue to improve as progress in video foundation modelling accelerates. (2) One promising avenue is to combine LfV reward functions with other LfV KMs. For example, finetuning an LfV policy with online RL and LfV rewards. (3) An underexplored direction is to investigate how LfV rewards can be used to augment reward labels in offline RL settings. (4) Video-language reward functions may prove useful for detecting safety-related metrics when deploying robots in the real world [Guan et al., 2024]. (5) Many LfV reward functions are differentiable, which makes them suitable for use in certain model-based RL algorithms [Hafner et al., 2023, JyothirS et al., 2023]. (6) Video-to-text models may be suitable for providing shaped reward functions via RL from AI feedback (RL-AI-F) frameworks [Klissarov et al., 2023].

5.2.5 Value Functions

Value functions (see Section 2.1) are an essential component of most deep RL algorithms [Schulman et al., 2017, Haarnoja et al., 2018]. There is a small but distinct line of LfV research which pretrains models that closely resemble value functions from video data. We note that commonly these pretrained models are not used as the value function downstream, but rather have been used, for example, to provide representations or rewards. As such, many of these methods have been touched on in previous sections. Nevertheless, it is useful to detail methods that pretrain value functions from video in a self-contained subsection. We do so here.

Pretraining: TD learning. The temporal difference (TD) learning objective is commonly used to learn value functions in RL [Sutton and Barto, 2018]. Here, we outline work that has used the TD objective on video data. Now, versus TD learning on robot data, video data poses additional challenges due to its missing labels. Namely, video is missing important action labels, reward labels, and goal labels. Below, we elaborate on these issues and on the corresponding solutions seen in the LfV literature.

Missing action labels. State-action value functions $V(s_t, a_t)$ (i.e., Q-functions [Sutton and Barto, 2018]) are often more useful than state value functions $V(s_t)$. The lack of action labels in video is an issue if we wish to obtain $V(s_t, a_t)$. Solutions have been proposed in the literature, most of which leverage some form of alternative action representation (see Section 5.1.1). Bhateja et al. [2023], Ghosh et al. [2023] pretrain a value function from video that is conditioned on sub-goal alternative actions. Here the sub-goals are observations sampled from a future timestep in the video. Edwards et al. [2020] similarly use a 'next-observation-as-action' approach, learning a Q-function conditioned on s_{t+1}, and employing a cycle-consistency objective to ensure the corresponding policy proposes physically plausible next-observation actions. Chang et al. [2022b] use a single-step latent action to condition the Q-function. Chang et al. [2020] use an IDM pseudo-action labelling approach to label navigation video data with actions before TD-learning.

Missing reward and goal labels. Missing reward labels in video is an issue as TD learning requires reward information. Meanwhile, missing goal labels is problematic as a generalist robot should be able to perform many tasks, and so its corresponding value function should be task/goal-conditioned. We now outline solutions to these issues seen in the LfV literature. To obtain reward labels, Bhateja et al. [2023], Ghosh et al. [2023], Park et al. [2024] use hindsight goal relabelling, where the goal observation g is sampled either from a future frame or another video. From this goal label, a sparse reward is defined as $r = (s == g)$ (where s is the current observation). In a semantic visual navigation setting, Chang et al. [2020] leverage object labels and off-the-shelf object detectors to provide goal and reward labels in video data. Elsewhere, other works assume a single task setting [Edwards et al., 2020], or assume access to reward labels in the video data [Edwards et al., 2020, Chang et al., 2022b]. Note, these are not scalable LfV assumptions. Although not seen in the above works, we note that the LfV reward functions from Section 5.2.4 could be applicable here. Additionally, video-to-text models (see Section 4.3) could be used to provide goal labels in textual form.

Pretraining: Time-contrastive learning. Time contrastive objectives can be used to learn implicit value functions from video [Ma et al., 2022]. Importantly, these objectives do not require action labels for video pretraining. The objective induces a temporally smooth representation, and a value function can thus be defined by measuring the distance between the current observation and a goal image in the embedding space. Similarly, Quasimetric functions [Wang et al., 2023d] and temporal cycle-consistency objectives [Zakka et al., 2021] can be learned from video to provide meaningful distance metrics. Note, the requirement for goal images is a limitation of these approaches.

Pretraining: Other methods. Edwards and Isbell [2019] use the number of timesteps remaining in video as heuristic value labels, and regress a value function to these labels. However, this approach assumes expert behaviour in the video. Du et al. [2023b] take a similar approach, finetuning a VLM to give heuristic value estimates. Liu et al. [2023c] use critical state identification to aid value prediction, but this assumes access to reward labels in the video data.

Downstream Usage. We now briefly outline how video pretrained value functions have been used downstream in the literature.

As a value function. The pretrained value function can be used directly downstream if: (i) it was pretrained to be conditioned on robot actions (i.e., $V(s_t, a_t)$) [Chang et al., 2020]; or (ii) it is alternative action-conditioned (i.e., $V_{\text{alt}}(s_t, \hat{a}_t)$) and a mapping from alternative to robot action can be obtained; or (iii) a dynamics model is available, allowing for planning with a unconditioned $V(s_t)$ [Du et al., 2023b, Chang et al., 2022b].

Representation transfer. After video pretraining, the value function may require finetuning on robot data, either to improve in-domain performance, or to allow for robot action-conditioning. Along these lines, Bhateja et al. [2023] initialise their downstream value function and policy representations from a video pretrained value function.

Potential-based reward shaping. A reward function can be defined as: $r = V(s_{t+1}) - V(s_t)$, and be used for downstream online RL [Edwards and Isbell, 2019, Ye et al., 2023, Chang et al., 2022b]. It may be desirable to do this if: (i) the pretrained value function is not fully reliable, but can provide useful auxiliary rewards to guide exploration; or (ii) the value function is not action-conditioned, so is not immediately usable for Q-learning.

TD bootstrapping. If the pretrained value function is not action-conditioned it can still be used to accelerate downstream TD learning by using its estimates for the bootstrap term in the bellman backup [Edwards and Isbell, 2019].

Discussion. Relative to the RL KMs in the previous sections, there has been less research into learning value functions from video. Moreover, many of the methods in this section do not actually use the pretrained model as a value function downstream. This is perhaps due to notable challenges that may come with learning value functions from videos. First, as noted above, there are issues related to missing action, reward, and goal information in video. Second, value functions estimate returns under a particular policy, but the behaviour in video data is highly multi-modal. Third, TD learning from videos may run into common issues seen in the offline RL literature [Levine et al., 2020] – such as overestimating the returns of out-of-distribution actions. An interesting future direction here is to use corresponding solutions from the offline RL literature [Kumar et al., 2020, Zhou et al., 2021]. We also note that many of the works in this section employ toy video datasets or toy downstream settings [Edwards et al., 2020, Chang et al., 2022b]. Nevertheless, Bhateja et al. [2023] have shown that TD learning from large-scale human video is a promising direction for real-world robotics – one that warrants further research.

6 Datasets

Previously, we've discussed methods for learning from video data, including methods for training video foundation models (see Section 4) and methods for utilising video data for robot learning (see Section 5.2). We now turn our attention to the video datasets themselves. These datasets are the foundation of LfV methods. In this section, we will discuss desired properties of video datasets, summarise methods for curating video data, and review existing datasets and their limitations (see Figure 9).

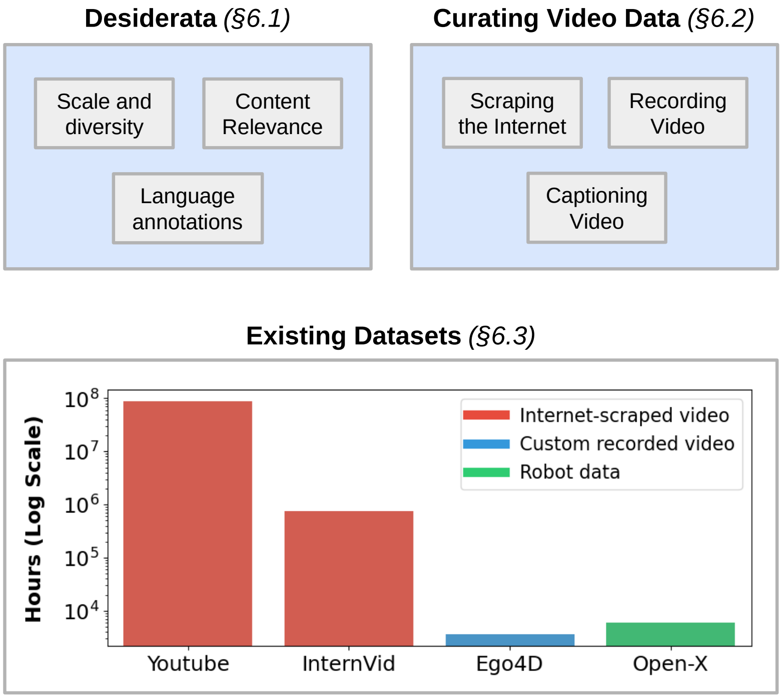


Figure 9: LfV Datasets (Section 6). In this survey, we review dataset desiderata (Section 6.1), methods for curating video data (Section 6.2), and details of existing datasets (Section 6.3). The plot at the bottom compares the sizes of the largest datasets in three different categories. InternVid [Wang et al., 2023f], Ego4D [Grauman et al., 2021], and Open X-Embodiment [Padalkar et al., 2023] are the largest curated internet-scraped video, custom-recorded video, and robot datasets, respectively (to the best of our knowledge). The plot demonstrates that internet video is more abundantly available than robot data by several orders of magnitude. We estimated the hours in the Open X-Embodiment dataset by assuming the average trajectory length is 10 seconds (at the time of writing, the dataset contained ∼2 million trajectories). The number of hours of video on YouTube is based on rough calculation [Sjöberg, 2023].

6.1 Desiderata

We are interested in methods that can use available internet video data to help us obtain generalist robots. We now discuss the properties and characteristics desired from our LfV video dataset.

Scale and Diversity. Crucially, the dataset should be large in *scale* and high in *diversity*. Scale can be measured in terms of total video duration. Diversity refers to variety in the content of the data. As shown in a number of other domains, increased scale and diversity of training data can reliably improve deep learning performance and generalization [Brown et al., 2020]. LfV is promising precisely because of the massive scale and diversity of the video data available on the internet.

Content Relevance. It is also crucial for the content of the video data to be relevant to the downstream robotic setting, i.e., the video data should encapsulate information useful to the robot. This includes information regarding the dynamics of the world and how embodied agents can complete physical tasks. Crucially, the video data should have good coverage over the tasks and environments that a generalist robot is likely to encounter upon deployment.

Language Annotations. Language annotations are textual descriptions (‘captions’) associated with segments of videos (‘clips’). Such annotations can be very useful: First, language conditioning is a powerful method for eliciting specific behaviours from generative models [Betker et al., 2023, Brooks et al., 2024]. Second, including language in the training process can aid the learning of abstract, semantic representations of video. This is important as video data is high-dimensional and noisy, making it difficult to learn such representations from scratch.

Our desiderata for such language annotations are as follows. First, annotations should be *accurate* and well-*aligned* with the events in the video. Second, we desire descriptions of *varied granularity.* For robotics, we are particularly interested in lower-level descriptions, such as detailed descriptions of the actions being performed in the video, or fine-grained descriptions of spatial information and object relations. Meanwhile, higher-level descriptions can also be useful for learning more abstracted representations. Third, additional forms of diversity in the language annotations may also be useful, such as syntactic diversity [Bansal et al., 2023].

Other desiderata. Finally, we briefly note some less important desiderata not covered above. (1) *Continuity* within a clip – i.e., an absence of sharp scene transitions – can be important; most LfV methods assume this property in their video data [Bruce et al., 2024, Yang et al., 2023c]. (2) Higher-*resolution* video can be beneficial for learning finer-grained understandings and features. (3) Video *length* can be an important characteristic; training on longer clips may benefit long-horizon representations useful for complex, multi-stage tasks. (4) Finally, video can be paired with other information – besides language annotations – that may be useful. Object bounding boxes, segmentations, or human-poses estimates may aid representation learning efforts [Bahl et al., 2023], or inform automated language captioning efforts. Other modalities such as audio, or depth/3D information [Grauman et al., 2021], can provide useful information beyond what is contained in the RGB pixels.

6.2 Curating Video Data

Here, we give concrete details on techniques from the literature for curating video datasets. We discuss techniques for scraping video from the internet, techniques for recording custom video, and techniques for (manually and automatically) annotating video.

Scraping Video from the Internet. Diverse, large-scale video datasets can be obtained from internet repositories of pre-existing video data. Techniques here focus on ensuring only relevant and high-quality videos are scraped from these repositories. A typical pipeline for scraping internet video involves: (i) formulating a pool of query prompts used to search the repository (e.g., YouTube) for candidate videos; (ii) post-processing the pool of candidate videos; and (iii) optionally re-annotating the videos to improve the quality of labels. In the bullets below, we describe methods for constructing the query pool and filtering the initial pool of candidate videos. We leave details regarding re-annotation till the following paragraph.

- *Select internet repository.* To obtain internet videos, one must first select a repository to search. YouTube is a popular choice due to the scale and diversity of its videos [Wang et al., 2023f, Xue et al., 2022, Zellers et al., 2021, Baker et al., 2022]. These repositories generally provide an API that can be used to search for and download video.

- *Obtain raw videos: Query Formulation.* Query strings (e.g., "A cat walking on a piano") must be constructed to search the repository for video. When the goal is to construct a large-scale diverse dataset, a diverse pool of relevant queries should be formulated. To help construct the query pool, past works have: obtained lists of relevant behaviours from surveys of human time-use [Grauman et al., 2021, Wang et al., 2023f]; generated prompts based on desired nouns and verbs [Goyal et al., 2017, Wang et al., 2023f]; crowd-sourced prompts from human workers [Sermanet et al., 2023, Goyal et al., 2017]; leveraged listed activities from WikiHow [Miech et al., 2019]; and used LLMs to parse action prompts from text corpora [Wang et al., 2023f]. Once the pool is constructed, a sanity check can be performed by visualising the resulting candidate videos across broad 'categories' [Nagrani et al., 2022, Miech et al., 2019, Goyal et al., 2017]. With the query pool finalized, an initial raw video dataset can be obtained by collating the top k search results for each query [Miech et al., 2019]. Once the initial raw video pool is obtained, various post-processing steps can be performed (as described in the following bullets).

- *Post-process: Split into clips.* Raw videos from the Internet may contain multiple sharp, discrete 'cuts' between scenes. It may be beneficial to post-process videos into clips, which contain no cuts. Blattmann et al. [2023a] do so using an off-the-shelf cut-detection model.

- *Post-process: Filter via metadata.* Internet videos naturally come paired with rich metadata. Approaches to identify and remove low-quality video based on this metadata include; filtering for popular and reputable publishers [Xue et al., 2022], filtering out videos with low view count [Miech et al., 2019, Nagrani et al., 2022], filtering out videos with low resolution [Xue et al., 2022], filtering out videos with too few captions [Miech et al., 2019, Xue et al., 2022], and filtering out older videos [Nagrani et al., 2022].

- *Post-process: Filter via automated metrics.* Other automated metrics can be useful for filtering video data. Optical flow can be used to filter out static scenes [Blattmann et al., 2023a]. Image-language model embeddings [Radford et al., 2021] can be used to measure the alignment between a video and its caption and provide an 'aesthetic score' [Schuhmann et al., 2022] used to filter out 'uninteresting' or lower-quality videos [Blattmann et al., 2023a]. Shan et al. [2020] obtain a metric for video 'realism' via a pretrained neural network operating over sub-samples of the video.

Recording New Video. Manually recording custom videos can be an effective (but expensive) means of collecting video relevant to specific tasks [Sermanet et al., 2023, Grauman et al., 2023] or embodiments [Sermanet et al., 2023, Shafiullah et al., 2023]. Here, we briefly describe procedures for recording new video data, as seen in the existing literature.

- *Participant recruitment.* To record large-scale custom video data, past works have relied on crowd-sourced workers. In the literature, participants have been recruited through online crowd-sourcing platforms [Goyal et al., 2017], or locally via the institutions collecting the data [Grauman et al., 2021]. Diversity in the recruited participants and their geographical locations is important to ensure sufficient diversity in the recorded video content. Existing work has surveyed potential participants to ensure a diverse mix of professions, hobbies, ages, genders, and nationalities [Damen et al., 2018, Grauman et al., 2021].

- *Unscripted data collection.* A simple approach to recording new videos is to ask participants to record their daily lives [Damen et al., 2018, Grauman et al., 2021]. This process will naturally cover many relevant human activities. Here, it is particularly important to carefully select participants to ensure sufficient diversity in the recorded video.

- *Scripted data collection.* Video collection can otherwise be scripted, by providing text instructions to the participant. These instructions could be short-horizon [Goyal et al., 2017], or long-horizon and involving multiple subgoals [Sermanet et al., 2023]. Scripted data collection allows for more control over what behaviours are represented in the dataset, and can be used to collect rarer behaviours, such as skillful behaviours [Grauman et al., 2023]. Scripted methods rely on constructing a relevant and diverse pool of instructions to guide the video collection; techniques for formulating internet search queries, as outlined in the previous paragraph, are relevant here.

- *Other considerations.* (1) Video can be recorded from a first-person or third-person viewpoint. While both can be useful, first-person video is less common on the internet, and can be particularly relevant to robotics. Thus, recording new first-person videos may be particularly useful. (2) Diversity in the dataset can be improved via the use of a range of different commercial-grade cameras [Grauman et al., 2021]. (3) When the video data is being collected for downstream robotics, it can be beneficial to use different embodiments whilst recording first-person videos. Sermanet et al. [2023] mix a diversity of embodiments, including human arms, human-controlled manual grippers, and robot arms. Shafiullah et al. [2023] use a low-cost human-controlled gripper that closely resembles the embodiment of the downstream robot.

Manual Captioning. We outlined the benefits language captions (i.e., textual descriptions of videos) can provide in Section 6.1. We now describe manual techniques for captioning video data. These may be performed by third-party workers or (where applicable) the person recording the video.

- *Instructing human annotators.* In the simplest scheme, human annotators are instructed to describe the contents of the video. To improve upon this, more specific instructions (potentially in the form of examples or via a tutorial [Blattmann et al., 2023a, Damen et al., 2018]) can enhance annotation quality and avoid issues with variability in style and quality between individual annotators. Elsewhere, annotators can be asked to provide captions beyond just describing the video. For example, Grauman et al. [2023] have third-party annotators provide expert commentaries, giving details on 'how' the activity is executed rather than 'what' it entails. They additionally prompt video-recording participants to explain their thought processes while performing actions. After initial annotation is performed, Maaz et al. [2023] prompt workers to augmenting the descriptiveness of video captions with additional details.

- *Spoken vs. typed annotations.* Compared to typed annotations, Damen et al. [2018] find spoken annotations to result in higher quantities and quality of captions. Damen et al. [2022] improve the quality of retrospective spoken annotations by allowing participants to pause the video whilst annotating, resulting in better temporal alignment of annotations and fewer missed events.
- *Annotating manually recorded video.* In the case of manually-recorded videos, annotations can be provided online (e.g., a narration by the person recording the video) [Grauman et al., 2023] or offline (e.g., retrospectively by a third-party annotator after the recording) [Grauman et al., 2021, Damen et al., 2018]. Online annotations may be more cost effective, as they can be performed by the same person whilst recording the video. However, offline annotations may be more accurate and detailed as the third-party worker can fully focus on the annotation task. It is also possible to combine both approaches [Grauman et al., 2023].

Automated Captioning. Manual human annotations are expensive (e.g., the Ego4D dataset required 250,000 hours of human annotator effort [Grauman et al., 2021]). For larger-scale datasets, automated annotation pipelines are promising. Below, we describe various methods for obtaining information about video content in an automated fashion.

- *Automatic Speech Recognition (ASR), metadata, and others.* A common strategy is to use ASR to convert speech in the video's audio into text, yielding information about the contents of the video [Miech et al., 2019, Xue et al., 2022, Zellers et al., 2021]. However, raw ASR captions can be noisy and often unrelated to the contents of the video. Another source of information is corresponding metadata, such as descriptions, tags, and titles [Wang et al., 2023f]. Elsewhere, Nagrani et al. [2022] obtain video captions by starting with a dataset of high-quality image-caption pairs, then mining video clips with similar frames and transferring the captions to those clips ('Transfer' in Table 1). Finally, captions can be obtained from the Alt-text HTML attribute associated with web images and videos [Bain et al., 2021] ('Alt-text' in Table 1).
- *Off-the-shelf vision models.* Information about the video can be extracted using off-the-shelf vision models. An obvious option is to use video-to-text captioning models (see Section 4.3) to obtain video-captions [Blattmann et al., 2023a]. However, these models are currently unreliable, and thus have not often been used in the literature. An alternative is to use more reliable image-captioning models. One heuristic for short clips is to caption the video based on the image caption of its middle frame [Blattmann et al., 2023a]. For longer videos, multiple frames in the video can be captioned, and these can be synthesised into a video caption (e.g., by an LLM) [Maaz et al., 2023]. Beyond captioning models, off-the-shelf object detectors can be employed to gather information about the video contents [Zeng et al., 2022, Maaz et al., 2023]. The methods in this bullet are referred to as 'Generated' in Table 1.
- *LLM processing.* If information about the video has been gathered from various sources, an LLM can be used to synthesise this information in a single coherent caption. Blattmann et al. [2023a], Maaz et al. [2023] summarise keyframe-based and full-clip captions into a single video caption. Wang et al. [2023f] summarise several keyframe captions into an overall caption. LLM post-processing can also be used to filter out inconsistent captions across sources or frames [Maaz et al., 2023]. Overall, combining multiple sources of information followed by LLM processing may result in more detailed and accurate captions, versus relying on a single captioning method.

6.3 Existing Datasets

In this section, we provide an overview of existing video datasets relevant to LfV. We aim to highlight datasets that satisfy key desiderata from Section 6.1, and thus are promising for training foundational video and/or robotics models.

Overview of Existing Datasets. Table 1 presents details of our chosen representative set of existing video datasets. We now briefly discuss these datasets, and their characteristics, in more detail.

Large-scale internet-scraped datasets. This category covers datasets constructed by searching web repositories with a carefully-designed pool of diverse text prompts. These can be very large, spanning up to several decades worth of video data [Wang et al., 2023f, Xue et al., 2022]. They are also diverse, capturing human behaviours in tasks and environments sampled from the lives and activities of a global population. Several of these datasets are constructed using prompt-based searches of YouTube [Wang et al., 2023f, Miech et al., 2019, Stroud et al., 2020], using the 'Query Formulation' techniques outlined in Section 6.2. To find videos, other datasets leverage YouTube categories [Xue et al., 2022], aggregate video from previously existing datasets [Zellers et al., 2021], or scrape video from various webpages [Bain et al., 2021]. Due to the large size of the datasets, their language captions are usually obtained via automated methods (see Section 6.2). This includes leveraging ASR [Xue et al., 2022], video meta-data [Stroud et al., 2020], or pretrained captioning and language models [Wang et al., 2023f]. We note that these automated annotations may be noisy, innacurate, or irrelevant compared to manual annotations. In the literature, these large internet video datasets have commonly been used in initial attempts at training large-scale video foundation models [Zhao et al., 2024, Wang et al., 2023c] (see Section 4).

Manually collected datasets. The second category includes datasets manually recorded by human participants. These are generally not as large or diverse as the largest web-obtained video datasets. However, many contain content highly relevant to robotics. Ego4D [Grauman et al., 2021] contains egocentric video of diverse participants going about their daily lives. Ego-Exo4D [Grauman et al., 2023] contains paired egocentric and third-person video of skilled activities. RoboVQA [Sermanet et al., 2023] contains video of teleoperated robots and humans performing long-horizon household tasks. Something-something V2 [Goyal et al., 2017] contains short clips of object-centric actions. Epic-Kitchens 100 [Damen et al., 2022] contains long-form video of cooking activities in real households. Here, it is common for captions to be provided manually by crowd-sourced workers, which can be more accurate than automated captioning methods. Due to their robotics-relevant contents, these datasets have often been used in past LfV research [Nair et al., 2022, Yang et al., 2023c, Wu et al., 2023a].

Other annotations. All of the datasets in Table 1 come with text annotations. However, other annotations are also possible. Whilst manually collecting video, Grauman et al. [2021, 2023] also occasionally collect corresponding audio, 3D meshes, eye-gazes, stereo, and synchronized video from different viewpoints. Meanwhile, Shan et al. [2020] provide human-hand-centric labels for 100k images in their video dataset. However, large-scale internet video datasets generally lack these other annotations as they are either costly and time-consuming to add (e.g., manually adding bounding boxes), or the information is not available with the internet video (e.g., 3D meshes do not come with standard internet videos).

Discussion. As seen in Table 1, the largest internet-curated video datasets are two orders of magnitude larger than the largest manually-recorded video datasets. Yet, these internet video datasets still barely scratch the surface of the full range of video content available online (see Figure 9). As such, we advocate for continued efforts into curating ever-larger internet video datasets for LfV.

Crucially, new curation efforts should optimise for other key dataset desiderata (Section 6.1), in addition increasing size. Firstly, the dataset should have suitable diversity, content relevance, and quality. This can be ensured using improved query formulations and filtering mechanisms (Section 6.2). Second, the video data should come with sufficiently high-quality language annotations. We note that low-quality language annotations are one of the key limitations of current internet-curated video datasets. Improved automated annotation methods (Section 6.1) that scale with increased dataset size will be important here.

We also note some other less urgent limitations of current video datasets that should be addressed. Many existing datasets focus only on short clips which do not capture long-horizon behaviour [Wang et al., 2023f, Xue et al., 2022, Stroud et al., 2020]. Many datasets primarily use Youtube as their video repository [Wang et al., 2023f, Miech et al., 2019, Stroud et al., 2020]; since over-centralization could bake in systematic bias we note it could be beneficial to also leverage other video repositories (e.g., Instagram, Tiktok, or Chinese video servers such as Weibo).

[1] These datasets are not publicly available at time of writing.

Dataset Name	Content	Total Duration (h)	# Clips	Caption Type	Collection Method
InternVid [Wang et al., 2023f]	YouTube	760,000	230M	Generated	Internet
HD-VILA-100M [Xue et al., 2022]	YouTube	370,000	103M	ASR	Internet
YT-Temporal-180M [Zellers et al., 2021]	YouTube	-	180M	ASR	Internet
WTS-70M [Stroud et al., 2020]	YouTube	190,000	70M	Metadata	Internet
HowTo100M [Miech et al., 2019]	Instruction	134,000	136M	ASR	Internet
WebVid-10M [Bain et al., 2021][1]	YouTube	52,000	10M	Alt-text	Internet
VideoCC3M [Nagrani et al., 2022][1]	YouTube	18,000	6M	Transfer	Internet
100 Days of Hands [Shan et al., 2020]	Actions	3,100	27k	Metadata	Internet
Ego-4D [Grauman et al., 2021]	Everyday	3,600	28k	Manual	Manual
Ego-Exo-4D [Grauman et al., 2023]	Skilled	1,400	6k	Manual	Manual
SS-v2 [Goyal et al., 2017]	Actions	245	221k	Manual	Manual
RoboVQA [Sermanet et al., 2023]	Everyday	230	98k	Manual	Manual
Epic-Kitchens-10 [Damen et al., 2022]	Cooking	100	700	Manual	Manual

Table 1: Existing video datasets. Listed are: (top) large-scale internet-scraped video datasets, and (bottom) robotics-relevant, manually-recorded video datasets. The datasets are ordered by decreasing total video duration. Details regarding ‘Caption Type’ can be found in Section 6.2.

Finally, we note that whilst we primarily advocate for curating larger datasets, efforts to curate higher-quality (but smaller-scale) video dataset will still be useful. Such high-quality dataset can be used for finetuning after pretraining on larger, lower-quality data [Blattmann et al., 2023a]. Here, manual video recording and manual captioning techniques may still be a valid option.

7 Benchmarks

We previously reviewed LfV methods that can be used for robotics in Section 5. We now turn our attention to benchmarks that can be used to develop, evaluate, and compare such LfV methods. In general, such benchmarks can be crucial for catalysing rapid research progress in a given area [Deng et al., 2009]. Here, we first outline how an LfV benchmark should be designed (Section 7.1). We then review relevant existing benchmarks from the literature, commenting on limitations and proposing improvements (Section 7.2).

7.1 Designing Benchmarks for LfV

An LfV benchmark can serve to evaluate: (i) the capabilities of a policy (i.e., some action-generating model) obtained via an LfV approach; or (ii) the effectiveness of an LfV algorithm at producing a policy, under certain constraints (e.g., when constrained to use a fixed dataset). A suitable benchmark should provide metrics that allow us to compare LfV models and algorithms. Specifically, we are interested in evaluating how well an LfV method can provide the potential benefits (Section 3.2) and handle the challenges (Section 3.3) of LfV. These metrics can also serve as a target for future LfV research to optimise for.

In this section, we will describe different categories of LfV benchmarks and their corresponding desiderata. Here, we will assume a setting where a pre-existing video dataset $\mathcal{D}_{\text{video}}$ and a robot dataset $\mathcal{D}_{\text{robot}}$ are used to train some policy π_{lfv} via an offline learning method. We assume this offline setting as online learning is impractical in robotics, and because this simplifies the explanations presented below.

Categories of LfV benchmarks. All categories of benchmark we describe include a fixed set of evaluation environments $\mathcal{M}_{\text{eval}}$ in which an LfV policy π_{lfv} is to be evaluated. However, the categories differ based on whether they specify the datasets $\mathcal{D}_{\text{video}}$ and $\mathcal{D}_{\text{robot}}$ used during training. Concretely, the categories of benchmark we identify are as follows.

- $\boldsymbol{B}_{\text{e}} = \{\boldsymbol{\mathcal{M}}_{\text{eval}}\}$**.** Here, we can take a π_{lfv} trained on any $\mathcal{D}_{\text{video}}$ and $\mathcal{D}_{\text{robot}}$, and evaluate it on a fixed $\mathcal{M}_{\text{eval}}$.
- $\boldsymbol{B}_{\text{e-r}} = \{\boldsymbol{\mathcal{M}}_{\text{eval}}, \boldsymbol{\mathcal{D}}_{\text{robot}}\}$**.** Here, a fixed $\mathcal{D}_{\text{robot}}$ is paired with $\mathcal{M}_{\text{eval}}$ and should be the only robot data used by the LfV method. Any $\mathcal{D}_{\text{video}}$ can be used in combination with the $\mathcal{D}_{\text{robot}}$ to train π_{lfv}.

- $\boldsymbol{B_{\text{e-r-v}} = \{\mathcal{M}_{\text{eval}}, \mathcal{D}_{\text{robot}}, \mathcal{D}_{\text{video}}\}}$**.** Here, a fixed $\mathcal{D}_{\text{video}}$ and a fixed $\mathcal{D}_{\text{robot}}$ are paired with $\mathcal{M}_{\text{eval}}$. $\mathcal{D}_{\text{video}}$ and $\mathcal{D}_{\text{robot}}$ should be the only data used to train π_{lfv}. We note this setup precludes using models pre-trained on other datasets.

There are trade-offs between these categories of benchmark. Benchmarks that fix the datasets (e.g., $B_{\text{e-r-v}}$) can provide a fairer comparison of LfV algorithms. However, they may end up being 'toyish' and not perfectly analogous to the scaled-up LfV setting. On the other hand, whilst benchmarks that do not fix the data (i.e., B_{e}) cannot fairly compare LfV algorithms, they can be used to compare LfV models. This is useful as, ultimately, we care about the quality of the final policy model.

Desiderata and setups. We now give more details on the desiderata for each potential component ($\mathcal{M}_{\text{eval}}$, $\mathcal{D}_{\text{robot}}$, $\mathcal{D}_{\text{video}}$) of an LfV benchmark. This includes details regarding how each component should be setup in relation to each other, and how they should be setup to ensure the benchmark presents the key LfV challenges (see Section 3.3).

- $\boldsymbol{\mathcal{M}_{\text{eval}}}$**.** Given the scope of this survey, we would like $\mathcal{M}_{eval}$ to be analogous to the generalist robot settings we are interested in. We note a number of desiderata and considerations here. (1) *Relevance:* The environments and tasks should meaningfully resemble those we expect to face in the generalist robot setting. (2) *Diversity:* There should be sufficient diversity in $\mathcal{M}_{\text{eval}}$. This allows us to better measure how π_{lfv} will handle diverse and unseen real-world scenarios. (3) *Realism:* The benchmark physics should be sufficiently realistic and it should include challenges that may be faced in the real-world, such as noisy observations and stochastic environments. To evaluate the ability of the LfV method to tackle the challenge of missing low-level information in video, $\mathcal{M}_{eval}$ should require perception of information not contained in video.

- $\boldsymbol{\mathcal{D}_{\text{robot}}}$**.** Here, we focus on how $\mathcal{D}_{\text{robot}}$ should be defined in relation to $\mathcal{M}_{\text{eval}}$. First, we note that $\mathcal{D}_{\text{robot}}$ should be drawn from $\mathcal{M}_{\text{eval}}$. Specifically, it should be drawn from a *subset* of the tasks and environments in $\mathcal{M}_{\text{eval}}$, allowing us to directly measure performance both in and out-of-distribution of $\mathcal{D}_{\text{robot}}$. Second, to ensure LfV generalization (see Figure 2) is possible, it may be desirable to ensure the robot dataset contains all low-level, "atomic" actions the robot can perform. Third, there are decisions to be made regarding the nature of the behaviours in $\mathcal{D}_{\text{robot}}$. Behaviours could be expert or suboptimal, uni-modal or multi-modal. Ideally, an LfV method should be able to learn from multi-modal suboptimal behaviour in $\mathcal{D}_{\text{robot}}$ – thus, the benchmark should test for this.

- $\boldsymbol{\mathcal{D}_{\text{video}}}$**.** The primary criteria for $\mathcal{D}_{\text{video}}$ is that it does not come with any action labels. Other desiderata for depend on the extent to which we wish $B_{\text{e-r-v}}$ to be a toy 'sandbox' for testing LfV algorithms.

  (1) *Sandbox setups:* Here, we use assume a self-contained setting that allows us to control which LfV challenges are faced. There are a number of considerations here. First, following our assumption that internet video has good coverage over generalist behaviours (see Section 3.1), $\mathcal{D}_{\text{video}}$ should have good coverage over the environments and tasks in $\mathcal{M}_{\text{eval}}$. This allows for direct assessment of how well an LfV algorithm can utilise a suitable $\mathcal{D}_{\text{video}}$ to generalise beyond $\mathcal{D}_{\text{robot}}$. Second, we can control certain aspects of $\mathcal{D}_{\text{video}}$ to see how well the LfV method can overcome distribution shifts between $\mathcal{D}_{\text{video}}$ and $\mathcal{M}_{\text{eval}}$ (e.g., embodiment gaps, viewpoint differences, environment differences). Third, we can toggle the challenge of 'controllability' by managing the extent to which changes in the videos are due to effects beyond a single agent's actions.

  (2) *Approximating the scaled-up LfV setting:* We otherwise may wish for $B_{\text{e-r-v}}$ to be more analogous to the scaled-up LfV setting. In this case, we likely will have less control over the specific characteristics of $\mathcal{D}_{\text{video}}$. Here we should ensure $\mathcal{D}_{\text{video}}$ resembles internet video in terms of its scale, diversity, and content. Thus, $\mathcal{D}_{\text{video}}$ should ideally consist of real-world human videos scraped from the internet. Doing so will inherently present several LfV challenges, and allow evaluation of the scalability of the LfV method.

7.2 Existing Benchmarks

Here, we briefly highlight existing benchmarks suitable for each category outlined in the previous section. We close the section by commenting on limitations and gaps in the current selection of LfV benchmarks.

$\boldsymbol{B}_{\mathbf{e}} = \{\boldsymbol{\mathcal{M}}_{\mathbf{eval}}\}$**.** There are a number of benchmarks that provide an $\mathcal{M}_{\text{eval}}$ relevant to LfV research. (1) *Toy settings:* Though not realistic or directly relevant to the robot setting, simplified toy settings can be useful for prototyping and for controlled evaluation of specific LfV challenges. Relevant toy environments include video game settings and simple simulations [Chevalier-Boisvert et al., 2024, Bellemare et al., 2013, Tassa et al., 2018]. The diversity of certain complex or open-ended video game environments [Fan et al., 2022, Küttler et al., 2020] can provide an excellent setting in which to test LfV policy generalization. (2) *Robotics simulators:* More realistic and robotics-relevant simulated environments are also available. These include benchmarks focused on low-level motor control, commonly-used skills, and object interaction [Mees et al., 2022, Yu et al., 2020, Gu et al., 2023, Liu et al., 2024a, Kumar et al., 2024, Makoviychuk et al., 2021]. We note that whilst some benchmarks intrinsically encode task diversity [Yu et al., 2020, Gu et al., 2023] and environment diversity [Liu et al., 2024a, Xie et al., 2023], the scope of the diversity tends to be limited. (3) *Embodied AI simulators:* There exist simulated benchmarks for embodied AI which abstract away low-level control details to focus on higher-level planning [Puig et al., 2023, Kolve et al., 2017]. These benchmarks contain relevant settings, re-creating human households and tasks resembling everyday activities. (4) *Real-robot setups:* All the benchmarks discussed thus far are simulated, and thus may be limited in their realism. Real-world robot evaluations with low-cost hardware have been proposed to tackle this issue (e.g., Ahn et al. [2020]). However, real-world evaluation is more costly and time-consuming than simulation.

$\boldsymbol{B}_{\mathbf{e\text{-}r}} = \{\boldsymbol{\mathcal{M}}_{\mathbf{eval}}, \boldsymbol{\mathcal{D}}_{\mathbf{robot}}\}$**.** There are several benchmarks that pair a $\mathcal{D}_{\text{robot}}$ with an $\mathcal{M}_{\text{eval}}$. Most relevant are benchmarks that provide a $\mathcal{D}_{\text{robot}}$ that can be split such that there are settings in the corresponding $\mathcal{M}_{\text{eval}}$ that are 'held-out' from the dataset. This allows us to directly evaluate LfV generalization. There are some robotics simulator benchmarks that fit this description. CALVIN [Mees et al., 2022], LanguageTable [Lynch et al., 2023], and LIBERO [Liu et al., 2024a] all provide demonstration or play data for tabletop manipulation tasks, whilst Maniskill [Gu et al., 2023] includes mobile manipulation tasks. For other robotics simulators [Yu et al., 2020] and toy [Chevalier-Boisvert et al., 2024, Küttler et al., 2020] benchmarks, a $\mathcal{D}_{robot}$ does not exist but could be collected by available scripted policies. Other robot [Fu et al., 2020, Yarats et al., 2022] and video-game [Fan et al., 2022] benchmarks do include a paired dataset $\mathcal{D}_{robot}$, but it may be difficult to meaningfully split the data for generalization testing. Lastly, there exist relatively large real-world robot datasets which are not paired with any specific benchmark [Padalkar et al., 2023, Khazatsky et al., 2024]. These can be relatively diverse, containing a variety of embodiments, tasks, and environments. They are more analagous to the $\mathcal{D}_{robot}$ we would use in a realistic, scaled-up LfV setting. Thus, these datasets could be suitable for use in a real-world LfV benchmark.

$\boldsymbol{B}_{\mathbf{e\text{-}r\text{-}v}} = \{\boldsymbol{\mathcal{M}}_{\mathbf{eval}}, \boldsymbol{\mathcal{D}}_{\mathbf{robot}}, \boldsymbol{\mathcal{D}}_{\mathbf{video}}\}$**.** Currently, there are very few benchmarks that specify a fixed $\mathcal{D}_{\text{video}}$ along with a fixed $\mathcal{D}_{\text{robot}}$. In a Minecraft video game setting, Fan et al. [2022] provide a $\mathcal{D}_{\text{video}}$ of 730k YouTube videos. The scale of the data and diversity of the Minecraft environment provides a useful LfV setting, though there are obvious differences to real-world robotics. The only other example we are aware of is Xiong et al. [2022], who provide a $\mathcal{D}_{\text{video}}$ collected from human demonstrations, coupled with a tightly paired simulation environment $\mathcal{M}_{\text{eval}}$ and an associated $\mathcal{D}_{\text{robot}}$. However, their code is yet to be released. We note that many previous LfV works have contructed a $\mathcal{D}_{video}$ by stripping action labels from a $\mathcal{D}_{robot}$ [Seo et al., 2022a, Schmidt and Jiang, 2023, Wen et al., 2023]. This can allow for easy setup of a scenario where $\mathcal{D}_{\text{video}}$ has good coverage over $\mathcal{M}_{\text{eval}}$, but can neglect distribution shifts seen between the video data and the robot domain in realistic, scaled-up LfV settings.

Discussion. Here, we give recommendations for improvements in LfV benchmarking, based on limitations in the existing literature.

Improving the diversity in $\mathcal{M}_{eval}$. The diversity in the most suitable robotic simulators [Mees et al., 2022, Lynch et al., 2023, Liu et al., 2024a, Gu et al., 2023] is limited in terms of the tasks, environments, and objects presented. Improving the diversity of $\mathcal{M}_{\text{eval}}$ will better allow us to assess the applicability of the LfV method to the generalist robotic setting. One possibility here is to use procedural generation [Deitke et al., 2022] or LLM-assisted environment design [Xian et al., 2023] to improve diversity. Another is to aggregate multiple $\mathcal{M}_{\text{eval}}$'s within a common framework [Kumar et al., 2024] (though this can introduce a need for downstream cross-embodiment generalization, which is not strictly necessary in the LfV setting).

Establishing a fixed $B_{e\text{-}r}$ and $B_{e\text{-}r\text{-}v}$. We note that there are currently no well-established LfV benchmarks in these categories. An established LfV benchmark would bring an improved ability to compare LfV algorithms; past works have often chosen different $\mathcal{D}_{\text{video}}$'s [Nair et al., 2022, Yang et al., 2023c, Schmidt and Jiang, 2023] and thus are difficult to compare. When designing such an LfV benchmark, we recommend following the desiderata and setup details from Section 7.1. This encourages the design of benchmarks that can evaluate how well a method can provide the potential benefits of LfV (Section 3.2) and how well it can handle the challenges of LfV (Section 3.3). We may be particularly interested in evaluating generalization beyond $\mathcal{D}_{\text{robot}}$, and evaluating performances in the face of distribution-shift challenges and the challenge of missing low-level information in video. These benchmarks should provide the LfV evaluation metrics outlined in Section 3.4.

8 Challenges & Opportunities

We now provide a comprehensive discussion of challenges and opportunities for future LfV research, based on our analysis of the existing literature. First, we give high-level recommendations for future LfV research (Section 8.1). Second, we detail promising directions for utilising video foundation models and techniques for LfV (Section 8.2). Third, we highlight approaches for overcoming previously identified key LfV challenges (Section 8.3). We conclude by discussing other challenges in generalist robotics that are unlikely to be solved via scaling to larger datasets (Section 8.4).

8.1 High-level Recommendations

After our analysis of the LfV research in Section 5, we now provide some high-level recommendations for future LfV research.

Scalable approaches. We should focus on methods that can scale well to large, diverse internet video. Many previous LfV works make strong assumptions on the nature of the video data or the downstream robot setting [Baker et al., 2022, Stadie et al., 2017]. Others use strong inductive biases (e.g., methods that explicitly address distribution shifts from Section 5.1.2). These assumptions and inductive biases limit the scalability of the method. We advocate for methods that can use simple but general learning objectives to extract knowledge from diverse video. For example, video prediction objectives or other objectives and techniques used to the train video foundation models (see Section 4) are promising. We further discuss opportunities for utilising video foundation models in Section 8.2.

Targeting the key benefits of LfV. Future LfV research should be more focused on obtaining the most promising LfV benefits (see Section 3.2). This should include developing LfV methods that can better mitigate robot data bottlenecks and allow for generalization beyond the limited available robot data. This raises question of: *which RL KM should be targeted?* Much past LfV research has sought to extract reward functions from video (see Section 5.2.4). Whilst these approaches are useful, we advocate for targeting KMs that can better minimise reliance on robot data. Specifically, we believe the policy (Section 5.2.2) and dynamics model (Section 5.2.3) KMs are most promising. Dynamics models are highly useful for RL and can be partially learned via standard video prediction objectives. Meanwhile, obtaining a policy is ultimately what we care about, and thus deserves significant focus. We note that it may be practical to target multiple KMs. Related to these points, we finally note that online RL is impractical in robotics and thus we should focus more on LfV approaches that can operate in the offline learning setting.

Improved evaluation of LfV methods. We should more explicitly measure the extent to which an LfV method provides the potential benefits of LfV. In Section 3.4, we outline specific metrics that can be used here. It is often difficult to identify these metrics in existing LfV research. Tracking such metrics is informative for the community, and can serve as an optimization target to drive further advances. In Section 7.2 we advocate for designing improved LfV benchmarks that can offer these metrics off-the-shelf.

8.2 Video Foundation Models for LfV

A highly promising LfV direction is to utilise large-scale internet video data to help train foundation models for robotics. Here, we outline some related research avenues and discussion points.

Improving and utilising off-the-shelf video foundation models. Video foundation models are relevant to LfV for two reasons. First, LfV methods can adapt a pretrained video foundation model into an RL KM [Du et al., 2023a, Yang et al., 2023c]. Second, video foundation model datasets and techniques can be repurposed for LfV [Sohn et al., 2024, Bruce et al., 2024]. Thus, improving video foundation model techniques is a key LfV research avenue. Some crucial directions here include addressing issues related to dataset limitations and computational requirements (see Section 8.3). We more generally highlight promising directions for improving video foundation models throughout the discussions of Section 4.

Now we ask: how to best utilise an off-the-shelf video foundation model for LfV? As discussed in the previous section, we advocate for targeting the policy and dynamics model KMs. Whilst the video model could be used zero-shot in some cases, it will generally be beneficial to finetune it on robot data before deployment. A strong baseline could be to finetune the video foundation model into an action-outputting policy using offline robot data [Brohan et al., 2023, Wu et al., 2023a]. If the foundation model has video generation capabilities, another reasonable baseline is to finetune it into a robot dynamics model. Here, data-efficient [Hu et al., 2021] and computationally efficient [Yang et al., 2023b] adaptation mechanisms will be useful.

Customising video foundation model pipelines for LfV. Ultimately, the best LfV foundation models will be developed using a pipeline fully optimized for LfV purposes (versus training generic video foundation models before adapting them for LfV). We now touch on directions for developing video foundation models more customized for LfV purposes:

- Generic video foundation model capabilities could be improved in areas pertinent to robotics. For example, we could improve the physics realism of video prediction models (e.g., via RL training [Black et al., 2023a]), or the fine-grained understandings of video-to-text models (e.g., via improved fine-grained language captioning of training data).

- Domain-specific robot videos can be included in generic video foundation model pretraining. This may improve the performance of the pretrained model in the downstream robot setting.

- Custom, control-centric foundation models could be trained *solely* on video data. One promising avenue is to use alternate action representations (see Section 5.1.1) to integrate explicit action information into the model. For example, Bruce et al. [2024] use alternative action representations to allow for fine-grained action conditioning of video predictions, whilst Schmidt and Jiang [2023] learn a latent-action policy from video. The use of TD learning objectives on video data is also relevant [Bhateja et al., 2023]. Elsewhere, training on video paired with robotics-relevant low-level information could improve the models suitability to robotics. The use of 3D depth information [Zhen et al., 2024, Yuan et al., 2024] could improve 3D understandings. Here, auxiliary objectives that capture the robotics-relevant information may be useful. For example, predicting object masks or human-hand-centric information (see Section 5.1.2), or using masked modelling [Wang et al., 2023c] to aid learning of fine-grained features.

- Custom, control-centric foundation models could be trained on video *and* robot data. This could involve having the video foundation model predict robot actions [Sohn et al., 2024] or conditioning video predictions on robot actions [Yang et al., 2023c]. An interesting avenue here may be to leverage both available robot action labels and alternative action representations during pretraining. Elsewhere, the recent trend towards monolithic any-to-any sequence models [Liu et al., 2024b, Kondratyuk et al., 2023] is relevant here. Such any-to-any models can be trained jointly on text, video, and robot data, and can act as policies, dynamics models, and high-level planners [Sohn et al., 2024].

Monolithic vs compositional models. The exciting prospect of any-to-any sequence models brings us to the question of whether monolithic or compositional models are best suited for LfV. Here, an example monolithic model could be the any-to-any models discussed above. An example hierarchical compositional approach seen in LfV is that of Ajay et al. [2023], where a language model conditions predictions of a video generator, and an action model predicts actions from the generated video. Each approach has its pros and cons.

Monolithic models can be optimized end-to-end; simple end-to-end approaches have thus far proven very effective in deep learning. Monolithic models additionally may benefit from positive transfer between the different data modalities and tasks they handle. On the other-hand, Du and Kaelbling [2024] argue that compositional approaches are less computationally expensive, more data-efficient, and can obtain improved generalization. These advantages are very relevant to the LfV setting. Future LfV research should seek to better compare monolithic and compositional approaches. In particular, directly comparing generalization abilities (as per the LfV generalization setting visualized in Figure 2) would yield informative results.

Open-sourced video foundation models. Finally, we advocate for increased research into open-sourced video foundation models. Open-sourced models will make cutting-edge LfV research more accessible to academic researchers, accelerating progress in LfV.

8.3 Overcoming LfV Challenges

In Section 3.3, we outlined the key challenges in LfV. Here, we discuss promising directions for overcoming some of these challenges.

Improved datasets. Improving our video datasets is a reliable way to advance our video foundation models and LfV methods. Indeed, the recent advance in video generation demonstrated by Brooks et al. [2024] was likely largely driven by scaling and improved data, rather than major algorithmic improvements. Future data curation efforts should optimise for the dataset desiderata we outline in Section 6.1, and follow the recommendations given in the discussion of Section 6.3. In summary, we should allocate more of our resources into scraping larger-scale video data from the internet (whilst maintaining sufficient quality and diversity of content), and captioning this data with high-quality language annotations (following the techniques outlined in Section 6.2). Finally, we advocate for open-sourcing of any future curated video datasets.

Bridging the gap to low-level robot information. Internet video is not fully informative for robotics. Video only contains visual information, and lacks crucial low-level information (e.g., forces and tactile information). Additionally, internet video mainly consists of human embodiments; videos of specific robot embodiments are rare. Thus, we cannot rely entirely on internet video for robotics and will likely always require some additional robot data. A key LfV challenge is to leverage video data to help minimise the quantity of robot data required, despite the missing low-level information in the video data (see Figure 2).

Here, it is worth noting that certain settings may require more robot data than others. In general, we expect heavier requirements in settings where the robot is more dependent on low-level information not available in video. For example, dexterous manipulation [Akkaya et al., 2019] and skillful locomotion [Zhuang et al., 2023] require finegrained control informed by low-level precepts (e.g., touch and proprioception). These settings will require more robot data than settings with coarse action-spaces where visual observations alone are sufficient [Brohan et al., 2022].

The key question here is: what can be done to reduce demands on the robot data, in light of the missing low-level information problem? Whilst it is possible that scaling the robot and video datasets will implicitly solve this issue, here we discuss more explicit solutions.

(1) Efficiently incorporate low-level information into the video model. Here, we could try to (i) extract as much low-level, control-centric information as possible from the video data itself, or (ii) incorporate low-level information via the use of robot data. See relevant recommendations in Section 8.2. Alternatively, to reduce demands on expensive real-world robot data, we could make use of cheap simulated data to provide low-level robot information to the model.

(2) Bypass the need to incorporate low-level information into the video model. First, here one could use a coarse-action space (e.g., cartesian control for manipulation [Brohan et al., 2022]). As discussed above, this can reduce demands on robot data. Indeed, Ko et al. [2023] show that coarse actions can be inferred purely from video data. However, coarse action spaces can be limiting. One possibility is to begin with a coarse action-space, and finetune for finer-grained control with the additional collected robot data. Second, a factorized/hierarchical approach could be employed to bypass the need for the video model to account for low-level information. For example, a high-level policy trained on video can propose alternative actions, and a low-level policy trained on robot data can decode these to robot actions [Du et al., 2023a, Schmidt and Jiang, 2023, Wen et al., 2023] (see Section 5.2.2). As discussed in Section 8.2, such compositional approaches may provide benefits over monolithic approaches.

Recovering action information from video. The missing action label problem in LfV can be partially addressed using alternative action representations (see Section 5.1.1). Such representations can be used as a substitute for raw robot actions. This can be useful for training alternative actions versions of certain RL KMs, or for defining auxiliary learning objectives. These alternative actions approaches are promising and we encourage more research in the area. We recommend future research to: (1) Scale existing methods to realistic and diverse internet video. Many leading alternative action methods have yet to do so [Bruce et al., 2024, Wen et al., 2023]. We note that scaling to unstructured internet video will aggravate certain issues. This includes the issue of controllability (i.e., where it is impossible to distinguish which changes are controlled by the agent's actions and which are due to external environment factors). (2) Explicitly compare the pros and cons of different action representations. For example, compare robot action decoding accuracy from different alternative action representations, including measuring the data-efficiency and generalization abilities of the decoder. Here, it will also be interesting to evaluate how different action representations transfer across the varied action-spaces of different robots. (3) Develop improved alternative action representations, perhaps by combining the benefits of existing options.

Tackling distribution shift. The distributions shifts between internet video and the downstream robot domain represent a fundamental LfV challenge. Previous LfV works have attempted to address distribution shifts using specific algorithmic mechanisms (see Section 5.1.2). However, generally these have not proven scalable to diverse internet video or to generalist robot settings. Instead, we advocate for implicitly addressing these shifts by scaling to larger and more diverse internet video data, and, when possible, additionally training on robot data. Nevertheless, addressing LfV distribution shift in a robot data-efficient manner is very much an open problem.

Mitigating computational demands. The high computational demands of training on video data can be mitigated via improved architectures that process this data more efficiently [Wang et al., 2023c, Liu et al., 2023a, Balažević et al., 2024]. Of course, the ever decreasing cost of compute will also help [Moore, 1998]. We note that, in addition to improving computational efficiency, methods that operate in a learned latent space (rather than pixel space) are promising for mitigating issues related to noise and redundancy in video [Bardes et al., 2023, Yan et al., 2021].

8.4 Other Generalist Robotics Challenges: Is Scaling Enough?

Arguably, a lack of data is currently the main bottleneck to current generalist-robot efforts; observing results in other machine learning domains [Achiam et al., 2023, Betker et al., 2023, Yang et al., 2023c], we can be reasonably confident that scaling current robot learning approaches to larger, improved datasets would yield significant improvements in capabilities. In this work, we advocate for the use of internet video data to help us achieve this scaling. However, scaling alone may not take us all the way to generalist robots. Once the data bottleneck in robotics is adequately addressed, new bottlenecks will likely emerge.

Here, we note some fundamental and practical challenges that may not be solved via scaling to internet video data under the current paradigm. Instead, many of these challenges may also require algorithmic advances.

Safety and reliability. Safety is paramount when deploying robots in the real world. Unfortunately, deep learning models are known to generalise poorly to unseen scenarios. Whilst scaling to internet data may provide the model with improved 'common sense' – improving its reliability and generalization – this does not fundamentally solve the problem. Ensuring sufficient safety and reliability in high-risk applications will be a major future bottleneck to generalist robotics.

Explainability, interpretability, and uncertainty. We now note some techniques that may aid safety efforts. Explainability and interpretability methods can help us better understand the reliability of the model. Scaling to large language datasets (paired with video) can improve explainability [Lanham et al., 2023, Wayve, 2024] and interpretability [Zou et al., 2023]. However, current approaches are unconvincing and further algorithmic advancements are required. We note that the use of mechanistic and concept-based interpretability methods [Li et al., 2022b, Zou et al., 2023] are underexplored in robotics. Elsewhere, algorithmic advancements to improve the uncertainty-awareness and calibration of large-scale deep learning models could benefit safety efforts.

Long-horizon tasks and memory. Generalist robot may need to operate on significantly longer time horizons than current ML models. This presents two challenges. First, longer-horizon tasks may necessitate improved planning and reasoning abilities (we discuss 'reasoning' more below). Second, longer-horizon tasks require improved memory capabilities. Current solutions to longer-term memory involve scaling up the context-length of the model [Liu et al., 2024b], or storing and retrieving from simple memory databases [Park et al., 2023]. It is unclear whether these solutions will be sufficient to fundamentally address the problem.

Latency. Large foundation models currently have long inference times that limit their ability to perform real-time low-level control at high frequencies [Reed et al., 2022, Brohan et al., 2022]. Reducing inference latency in robot foundation models is an important practical challenge.

Continual learning and adaptation. In the unstructured real world, a generalist robot may regularly face novel scenarios. It must be able to adapt to these scenarios appropriately. Now, scaling to internet video data can improve generalization and in-context meta-learning abilities [Brown et al., 2020], to an extent. Nevertheless, obtaining true continual learning [Abel et al., 2024] abilities is still an open problem, likely requiring algorithmic advances.

Reasoning and causality. There exists much speculation and empirical evidence regarding the inability of deep learning methods to learn truly 'causal' models [Berrevoets et al., 2023] and perform true 'reasoning' [Huang et al., 2023a]. Such causal understandings and reasoning abilities may be crucial for a generalist robot. If deep learning is fundamentally limited in these regards, then scaling up to internet video data may not be sufficient to achieve general-purpose robots.

9 Conclusion

Developing general-purpose robots is a grand challenge in robotics. However, current learning approaches are bottlenecked by a lack of suitable robot data. Learning from Video (LfV) methods seek to alleviate this issue by leveraging existing video data. These methods are promising as internet video comes in vast quantities and contains information highly relevant to general-purpose robots.

In this survey, we conducted a comprehensive review of the LfV setting and existing LfV literature, providing the reader with useful taxonomies and discussions throughout. We focused on methods that have the potential to scale well to large, heterogeneous internet video datasets: following recent trends in machine learning, we consider these methods to be most promising for developing generalist robots. The key takeaways of this survey are summarised below:

- *LfV preliminaries.* In Section 3 we explored the exciting benefits that can be obtained from LfV (Section 3.2), key LfV challenges that stand in the way of these benefits (Section 3.3), and gave details on how LfV methods should be evaluated in light of these benefits and challenges (Section 3.4).

- *Foundation models from internet video.* Video foundation model techniques are highly promising for extracting knowledge from large, heterogeneous internet video datasets. We reviewed the video foundation model literature in Section 4, highlighting recurring issues related to low-quality video data and prohibitive computational requirements. In Section 8.2, we discussed directions for leveraging video foundation models for robotics. Developing models customized for robotics is a particularly promising direction. Here, recent monolithic any-to-any sequence models offer a clear path forward. It will also be beneficial to pursue compositional approaches in parallel.

- *Takeaways from the LfV-for-robotics literature.* Our analysis of the LfV-for-robotics literature (Section 5) yielded several important takeaways. First, future LfV research should prioritise scalability. In particular, we should avoid inductive biases that limit scalability (i.e., those seen in Section 5.1.2), and adopt simple, scalable learning objectives similar to those used for video foundation models (Section 4). Second, we should focus on methods that can best obtain the key LfV benefits, such as generalisation beyond the available robot data. This should involve targeting the most promising RL knowledge modalities (KMs): the policy and dynamics model. Improved benchmarks that quantitatively measure LfV benefits (see Section 7.2) will facilitate these efforts.

- *Alternative action representations.* We outline methods for extracting action representations from video – which serve to mitigate the LfV missing action label problem – in Section 5.1.1. These methods are promising. However, it is often unclear how well they will scale to heterogeneous internet video data. More research is needed in this area.

- *Datasets.* We advocate for allocating increased resources into curating improved video datasets (as per the dataset desiderata outlined in Section 6.1). Improving datasets is a reliable path towards boosting LfV performances. We outline suitable methods for curating internet video data in Section 6.2.

- *Is scaling enough?* Exploiting internet video will drive significant advancements in robotics. However, the generalist robot setting presents fundamental and practical challenges that may not be solved via naive scaling (as discussed in Section 8.4). As LfV research advances and the data bottleneck becomes less of an issue, efforts should refocused towards these challenges.

Overall, the analysis, taxonomies, and directions presented in this survey should serve as a valuable reference for future LfV research. This can catalyze further promising research in the area, and ultimately accelerate our progress towards developing general-purpose robots.

References

Moonvalley - animate your ideas. `https://moonvalley.ai/`. Accessed: 2024-04-04.

Pika - empowering creativity. `https://pika.art/home`. Accessed: 2024-04-04.

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ademi Adeniji, Amber Xie, Carmelo Sferrazza, Younggyo Seo, Stephen James, and P. Abbeel. Language reward modulation for pretraining reinforcement learning. *ArXiv*, abs/2308.12270, 2023. URL `https://api.semanticscholar.org/CorpusID:261075941`.

Michael Ahn, Henry Zhu, Kristian Hartikainen, Hugo Ponte, Abhishek Gupta, Sergey Levine, and Vikash Kumar. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on robot learning*, pages 1300–1313. PMLR, 2020.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Jayant Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego M Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:247939706`.

Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024.

Anurag Ajay, Seung-Jun Han, Yilun Du, Shaung Li, Abhishek Gupta, T. Jaakkola, Josh Tenenbaum, Leslie Pack Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *ArXiv*, abs/2309.08587, 2023. URL `https://api.semanticscholar.org/CorpusID:262012485`.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. URL https://api.semanticscholar.org/CorpusID:232417054.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyun Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:44061126.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2503–2516, 2023. URL https://api.semanticscholar.org/CorpusID:255440354.

Shikhar Bahl, Abhi Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *ArXiv*, abs/2207.09450, 2022. URL https://api.semanticscholar.org/CorpusID:248941578.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 01–13, 2023. URL https://api.semanticscholar.org/CorpusID:258180471.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *ArXiv*, abs/2206.11795, 2022. URL https://api.semanticscholar.org/CorpusID:249953673.

Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*, 2024.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. *arXiv preprint arXiv:2311.10111*, 2023.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Jeroen Berrevoets, Krzysztof Kacprzyk, Zhaozhi Qian, and Mihaela van der Schaar. Causal deep learning. *arXiv preprint arXiv:2303.02186*, 2023.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Homanga Bharadhwaj, Abhi Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. *ArXiv*, abs/2312.00775, 2023. URL `https://api.semanticscholar.org/CorpusID:265551754`.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Chethan Bhateja, Derek Guo, Dibya Ghosh, Anika Singh, Manan Tomar, Quan Ho Vuong, Yevgen Chebotar, Sergey Levine, and Aviral Kumar. Robotic offline rl from internet videos via value-function pre-training. *ArXiv*, abs/2309.13041, 2023. URL `https://api.semanticscholar.org/CorpusID:262217278`.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023a.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *ArXiv*, abs/2310.10639, 2023b. URL `https://api.semanticscholar.org/CorpusID:264172455`.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023b.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022. URL `https://api.semanticscholar.org/CorpusID:254591260`.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Krzysztof Choromanski, Tianli Ding, Danny Driess, Chelsea Finn, Peter R. Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Sergey Levine, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Pierre Sermanet, Jaspiar Singh, Anika Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Ho Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Ted Xiao, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818, 2023. URL `https://api.semanticscholar.org/CorpusID:260293142`.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL `https://api.semanticscholar.org/CorpusID:218971783`.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. 2024. URL `https://api.semanticscholar.org/CorpusID:267897982`.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. *ArXiv*, abs/1810.12894, 2018. URL `https://api.semanticscholar.org/CorpusID:53115163`.

Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? *ArXiv*, abs/2312.12444, 2023. URL `https://api.semanticscholar.org/CorpusID:266369884`.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2019. URL `https://api.semanticscholar.org/CorpusID:85517967`.

Shaofei Cai, Bowei Zhang, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. Groot: Learning to follow instructions by watching gameplay videos. *ArXiv*, abs/2310.08235, 2023. URL `https://api.semanticscholar.org/CorpusID:263908999`.

Víctor Campos, Pablo Sprechmann, Steven Hansen, Andre Barreto, Steven Kapturowski, Alex Vitvitskyi, Adria Puigdomenech Badia, and Charles Blundell. Beyond fine-tuning: Transferring behavior in reinforcement learning. *arXiv preprint arXiv:2102.13515*, 2021.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

Harish chaandar Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annu. Rev. Control. Robotics Auton. Syst.*, 3: 297–330, 2020. URL `https://api.semanticscholar.org/CorpusID:208958394`.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Learning video-conditioned policies for unseen manipulation tasks. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 909–916, 2023. URL `https://api.semanticscholar.org/CorpusID:258588267`.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315, 2022a. URL `https://api.semanticscholar.org/CorpusID:246680316`.

Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. *ArXiv*, abs/2006.10034, 2020. URL `https://api.semanticscholar.org/CorpusID:219721405`.

Matthew Chang, Arjun Gupta, and Saurabh Gupta. Learning value functions from undirected state-only experience. *ArXiv*, abs/2204.12458, 2022b. URL `https://api.semanticscholar.org/CorpusID:245064979`.

Matthew Chang, Aditya Prakash, and Saurabh Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *ArXiv*, abs/2305.16301, 2023. URL `https://api.semanticscholar.org/CorpusID:258888066`.

Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pages 3909–3928. PMLR, 2023.

Annie S. Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. *ArXiv*, abs/2103.16817, 2021a. URL `https://api.semanticscholar.org/CorpusID:232428118`.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023a.

Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023b.

Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023c.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021b. URL `https://api.semanticscholar.org/CorpusID:233024948`.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. *arXiv preprint arXiv:1704.02254*, 2017.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Zichen Jeff Cui, Yibin Wang, Nur Muhammad (Mahi) Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *ArXiv*, abs/2210.10047, 2022. URL `https://api.semanticscholar.org/CorpusID:252968170`.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

Sudeep Dasari, Mohan Kumar Srirama, Unnat Jain, and Abhinav Gupta. An unbiased look at datasets for visuo-motor pre-training. *ArXiv*, abs/2310.09289, 2023. URL `https://api.semanticscholar.org/CorpusID:263914653`.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20224–20234, 2023a.

Ziluo Ding, Hao Luo, Ke Li, Junpeng Yue, Tiejun Huang, and Zongqing Lu. Clip4mc: An rl-friendly vision-language model for minecraft. *ArXiv*, abs/2303.10571, 2023b. URL `https://api.semanticscholar.org/CorpusID:257632482`.

Yilun Du and Leslie Pack Kaelbling. Compositional generative modeling: A single model is not all you need. *ArXiv*, abs/2402.01103, 2024. URL `https://api.semanticscholar.org/CorpusID:267406745`.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *ArXiv*, abs/2302.00111, 2023a. URL `https://api.semanticscholar.org/CorpusID:256459809`.

Yilun Du, Mengjiao Yang, Peter R. Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Josh Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. *ArXiv*, abs/2310.10625, 2023b. URL `https://api.semanticscholar.org/CorpusID:264172935`.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023c.

Ashley D Edwards and Charles L Isbell. Perceptual values from observation. *arXiv preprint arXiv:1905.07861*, 2019.

Ashley D. Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Lee Isbell. Imitating latent policies from observation. *ArXiv*, abs/1805.07914, 2018. URL https://api.semanticscholar.org/CorpusID:29156793.

Ashley D. Edwards, Himanshu Sahni, Rosanne Liu, Jane Hung, Ankit Jain, Rui Wang, Adrien Ecoffet, Thomas Miconi, Charles Lee Isbell, and Jason Yosinski. Estimating q(s, s') with deep deterministic dynamics gradients. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:211258729.

Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and P. Abbeel. Video prediction models as rewards for reinforcement learning. *ArXiv*, abs/2305.14343, 2023. URL https://api.semanticscholar.org/CorpusID:258841355.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. URL https://api.semanticscholar.org/CorpusID:229297973.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

Chrisantus Eze and Christopher Crick. Learning by watching: A review of video-based learning approaches for robot manipulation. *arXiv preprint arXiv:2402.07127*, 2024.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180, 2022.

Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pages 11321–11339. PMLR, 2023.

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10406–10417, 2022. URL https://api.semanticscholar.org/CorpusID:249712367.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. URL `https://api.semanticscholar.org/CorpusID:834612`.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2021. URL `https://api.semanticscholar.org/CorpusID:238856888`.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

Lin Guan, Yifan Zhou, Denis Liu, Yantian Zha, Heni Ben Amor, and Subbarao Kambhampati. " task success" is not enough: Investigating the use of video-language models as behavior critics for catching undesirable agent behaviors. *arXiv preprint arXiv:2402.04210*, 2024.

Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR, 2023.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Danijar Hafner, Timothy P. Lillicrap, Ian S. Fischer, Ruben Villegas, David R Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *ArXiv*, abs/1811.04551, 2018. URL `https://api.semanticscholar.org/CorpusID:53280207`.

Danijar Hafner, J. Paukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse domains through world models. *ArXiv*, abs/2301.04104, 2023. URL `https://api.semanticscholar.org/CorpusID:255569874`.

Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. General-purpose, long-context autoregressive modeling with perceiver ar. In *International Conference on Machine Learning*, pages 8535–8558. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL `https://api.semanticscholar.org/CorpusID:219955663`.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *ArXiv*, abs/2309.17080, 2023a. URL `https://api.semanticscholar.org/CorpusID:263310665`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Zhibo Zhao, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023b.

Yingdong Hu, Renhao Wang, Li Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning*, 2023c. URL `https://api.semanticscholar.org/CorpusID:258048578`.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023a.

Tao Huang, Guangqi Jiang, Yanjie Ze, and Huazhe Xu. Diffusion reward: Learning rewards via conditional video diffusion. *arXiv preprint arXiv:2312.14134*, 2023b.

Wenlong Huang, F. Xia, Ted Xiao, Harris Chan, Jacky Liang, Peter R. Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:250451569`.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Eric Jang. All neural networks, all autonomous, all 1x speed. https://www.1x.tech/discover/all-neural-networks-all-autonomous-all-1x-speed, Feb 2024. Accessed: 2024-04-10.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *ArXiv*, abs/1906.08253, 2019. URL https://api.semanticscholar.org/CorpusID:195068981.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring visual pre-training for robot manipulation: Datasets, models and methods. *ArXiv*, abs/2308.03620, 2023. URL https://api.semanticscholar.org/CorpusID:254198890.

V JyothirS, Siddhartha Jalagam, Yann LeCun, and Vlad Sobal. Gradient-based planning with world models. *ArXiv*, abs/2312.17227, 2023. URL https://api.semanticscholar.org/CorpusID:266573170.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *ArXiv*, abs/2302.12766, 2023. URL https://api.semanticscholar.org/CorpusID:257205716.

Haresh Karnan, Garrett Warnell, Xuesu Xiao, and Peter Stone. Voila: Visual-observation-only imitation learning for autonomous navigation. *2022 International Conference on Robotics and Automation (ICRA)*, pages 2497–2503, 2021. URL https://api.semanticscholar.org/CorpusID:234790310.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023a.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023b.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

Moo Jin Kim, Jiajun Wu, and Chelsea Finn. Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations. *ArXiv*, abs/2307.05959, 2023. URL `https://api.semanticscholar.org/CorpusID:259836885`.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76: 201–264, 2023.

Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*, 2023.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Josh Tenenbaum. Learning to act from actionless videos through dense correspondences. *ArXiv*, abs/2310.08576, 2023. URL `https://api.semanticscholar.org/CorpusID:263908842`.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021.

Eric Krotkov, Douglas Hackett, Larry Jackel, Michael Perschbacher, James Pippine, Jesse Strauss, Gill Pratt, and Christopher Orlowski. The darpa robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pages 1–26, 2018.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Sateesh Kumar, Jonathan Zamora, Nicklas Hansen, Rishabh Jangir, and Xiaolong Wang. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:251223373`.

Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Abhishek Gupta, and Aravind Rajeswaran. Robohive: A unified framework for robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan A. Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. *ArXiv*, abs/2210.14215, 2022. URL `https://api.semanticscholar.org/CorpusID:253107613`.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *ArXiv*, abs/2306.14892, 2023. URL `https://api.semanticscholar.org/CorpusID:259262142`.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020. URL `https://api.semanticscholar.org/CorpusID:218486979`.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022a.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022b.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19891–19903, 2023c. URL `https://api.semanticscholar.org/CorpusID:257771777`.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022c.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila A. McIlraith. Steve-1: A generative model for text-to-behavior in minecraft. *ArXiv*, abs/2306.00937, 2023. URL `https://api.semanticscholar.org/CorpusID:258999563`.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. 2024b. URL `https://api.semanticscholar.org/CorpusID:268385142`.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024c.

Haozhe Liu, Mingchen Zhuge, Bing Li, Yuhui Wang, Francesco Faccio, Bernard Ghanem, and Jürgen Schmidhuber. Learning to identify critical states for reinforcement learning from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1955–1965, 2023c.

Minghuan Liu, Zhengbang Zhu, Yuzheng Zhuang, Weinan Zhang, Jianye Hao, Yong Yu, and J. Wang. Plan your target and learn your skills: Transferable state-only imitation learning via decoupled policy optimization. In *International Conference on Machine Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:247244605`.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR, 2020.

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *ArXiv*, abs/2210.00030, 2022. URL `https://api.semanticscholar.org/CorpusID:252683397`.

Yecheng Jason Ma, William Jiahua Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, 2023. URL `https://api.semanticscholar.org/CorpusID:258999195`.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Chris A Mack. Fifty years of moore's law. *IEEE Transactions on semiconductor manufacturing*, 24(2):202–207, 2011.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, P. Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? *ArXiv*, abs/2303.18240, 2023. URL `https://api.semanticscholar.org/CorpusID:257901087`.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. *ArXiv*, abs/2202.00164, 2022. URL `https://api.semanticscholar.org/CorpusID:237369373`.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *ArXiv*, abs/2308.10901, 2023. URL `https://api.semanticscholar.org/CorpusID:259336798`.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. URL `https://api.semanticscholar.org/CorpusID:182952863`.

RuiBo Ming, Zhewei Huang, Zhuoxuan Ju, Jianming Hu, Lihui Peng, and Shuchang Zhou. A survey on video prediction: From deterministic to generative approaches. *ArXiv*, abs/2401.14718, 2024. URL `https://api.semanticscholar.org/CorpusID:267301200`.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021.

Gordon E Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wen Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Y. Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *ArXiv*, abs/2305.15021, 2023. URL `https://api.semanticscholar.org/CorpusID:258865718`.

Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. *ArXiv*, abs/2110.07692, 2021. URL `https://api.semanticscholar.org/CorpusID:239009498`.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhi Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:247618840`.

Sharan Narang and Aakanksha Chowdhery. Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance. *Google AI Blog*, 2022.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Ho Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Manfred Otto Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *ArXiv*, abs/2402.07872, 2024. URL `https://api.semanticscholar.org/CorpusID:267627797`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL `https://api.semanticscholar.org/CorpusID:257532815`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joseph Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *ArXiv*, abs/2312.07395, 2023. URL `https://api.semanticscholar.org/CorpusID:266174654`.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Kumar Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:247292805`.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36, 2024.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 488–489, 2017. URL `https://api.semanticscholar.org/CorpusID:20045336`.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, P. Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37:178, 2018. URL `https://api.semanticscholar.org/CorpusID:52937281`.

Karl Pertsch, Oleh Rybkin, Jingyun Yang, Shenghao Zhou, Konstantinos G. Derpanis, Kostas Daniilidis, Joseph J. Lim, and Andrew Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In *Conference on Learning for Dynamics & Control*, 2019. URL `https://api.semanticscholar.org/CorpusID:218571383`.

Karl Pertsch, Ruta Desai, Vikash Kumar, Franziska Meier, Joseph J. Lim, Dhruv Batra, and Akshara Rai. Cross-domain transfer via semantic skill imitation. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:254636470`.

Jan Peters, Daniel D Lee, Jens Kober, Duy Nguyen-Tuong, J Andrew Bagnell, and Stefan Schaal. Robot learning. *Springer Handbook of Robotics*, pages 357–398, 2016.

Ivaylo Popov, Nicolas Heess, Timothy Lillicrap, Roland Hafner, Gabriel Barth-Maron, Matej Vecerik, Thomas Lampe, Yuval Tassa, Tom Erez, and Martin Riedmiller. Data-efficient deep reinforcement learning for dexterous manipulation. *arXiv preprint arXiv:1704.03073*, 2017.

Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, 2021. URL `https://api.semanticscholar.org/CorpusID:236986915`.

Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7:10873–10881, 2022. URL `https://api.semanticscholar.org/CorpusID:248392006`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Ilija Radosavovic, Tete Xiao, Stephen James, P. Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:252718704`.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley D. Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Trans. Mach. Learn. Res.*, 2022, 2022. URL `https://api.semanticscholar.org/CorpusID:248722148`.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:252367910`.

Oleh Rybkin, Karl Pertsch, Konstantinos G. Derpanis, Kostas Daniilidis, and Andrew Jaegle. Learning what you can do before doing anything. In *International Conference on Learning Representations*, 2018. URL `https://api.semanticscholar.org/CorpusID:60441438`.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *Advances in Neural Information Processing Systems*, 34:29246–29257, 2021.

Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023.

Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, 2019. URL `https://api.semanticscholar.org/CorpusID:209515451`.

Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. *ArXiv*, abs/2011.06507, 2020. URL `https://api.semanticscholar.org/CorpusID:226306712`.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. *ArXiv*, abs/2312.10812, 2023. URL `https://api.semanticscholar.org/CorpusID:266359570`.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609, 2019. URL `https://api.semanticscholar.org/CorpusID:208158225`.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Younggyo Seo, Kimin Lee, Stephen James, and P. Abbeel. Reinforcement learning with action-free pre-training from videos. *ArXiv*, abs/2203.13880, 2022a. URL `https://api.semanticscholar.org/CorpusID:247762941`.

Younggyo Seo, Kimin Lee, Fangchen Liu, Stephen James, and P. Abbeel. Harp: Autoregressive latent video prediction with high-fidelity image generator. *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3943–3947, 2022b. URL `https://api.semanticscholar.org/CorpusID:252280733`.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. *arXiv preprint arXiv:2311.00899*, 2023.

Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9866–9875, 2020. URL `https://api.semanticscholar.org/CorpusID:215413188`.

Jinghuan Shang and Michael S. Ryoo. Self-supervised disentangled representation learning for third-person imitation learning. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 214–221, 2021. URL `https://api.semanticscholar.org/CorpusID:236772575`.

Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40:1419 – 1434, 2020. URL `https://api.semanticscholar.org/CorpusID:220069237`.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:254408735`.

Maximilian Sieb, Zhou Xian, Audrey Huang, Oliver Kroemer, and Katerina Fragkiadaki. Graph-structured visual imitation. In *Conference on Robot Learning*, pages 979–989. PMLR, 2020.

Sneha Silwal, Karmesh Yadav, Tingfan Wu, Jay Vakil, Arjun Majumdar, Sergio Arnaud, Claire Chen, Vincent-Pierre Berges, Dhruv Batra, Aravind Rajeswaran, Mrinal Kalakrishnan, Franziska Meier, and Oleksandr Maksymets. What do we learn from a large-scale study of pre-trained visual representations in sim and real environments? *ArXiv*, abs/2310.02219, 2023. URL https://api.semanticscholar.org/CorpusID:263608779.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *ArXiv*, abs/2202.10448, 2022. URL https://api.semanticscholar.org/CorpusID:247011104.

Alice Sjöberg. How many videos are there on youtube? https://www.dexerto.com/entertainment/how-many-videos-are-there-on-youtube-2197264/, December 2023.

Laura Smith, Nikita Dhawan, Marvin Zhang, P. Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *ArXiv*, abs/1912.04443, 2019. URL https://api.semanticscholar.org/CorpusID:209140723.

Andrew Sohn, Anusha Nagabandi, Carlos Florensa, Daniel Adelberg, Di Wu, Hassan Farooq, Ignasi Clavera, Jeremy Welborn, Juyue Chen, Nikhil Mishra, Peter Chen, Peter Qian, Pieter Abbeel, Rocky Duan, Varun Vijay, and Yang Liu. Introducing rfm-1: Giving robots human-like reasoning capabilities. https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities/, March 2024. Accessed: 2024-03-29.

Sumedh Anand Sontakke, Jesse Zhang, S'ebastien M. R. Arnold, Karl Pertsch, Erdem Biyik, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *ArXiv*, abs/2310.07899, 2023. URL https://api.semanticscholar.org/CorpusID:263909538.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015.

Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017.

Jonathan C Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A Ross. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*, 2020.

Oliver Struckmeier and Ville Kyrki. Preventing mode collapse when imitating latent policies from observations. 2022.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Allison C. Tam, Neil C. Rabinowitz, Andrew Kyle Lampinen, Nicholas A. Roy, Stephanie C. Y. Chan, DJ Strouse, Jane X. Wang, Andrea Banino, and Felix Hill. Semantic exploration from language abstractions and pretrained representations. *ArXiv*, abs/2204.05080, 2022. URL https://api.semanticscholar.org/CorpusID:248085427.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023a.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy, 2023b.

Garrett Thomas, Ching-An Cheng, Ricky Loynd, Vibhav Vineet, Mihai Jalobeanu, and Andrey Kolobov. Plex: Making the most of the available data for robotic manipulation pretraining. *ArXiv*, abs/2303.08789, 2023. URL `https://api.semanticscholar.org/CorpusID:257532588`.

Dhruva Tirumala, Alexandre Galashov, Hyeonwoo Noh, Leonard Hasenclever, Razvan Pascanu, Jonathan Schwarz, Guillaume Desjardins, Wojciech Marian Czarnecki, Arun Ahuja, Yee Whye Teh, et al. Behavior priors for efficient reinforcement learning. *Journal of Machine Learning Research*, 23(221):1–68, 2022.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. URL `https://api.semanticscholar.org/CorpusID:2413610`.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Manan Tomar, Dibya Ghosh, Vivek Myers, Anca Dragan, Matthew E Taylor, Philip Bachman, and Sergey Levine. Video-guided skill discovery. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022. URL `https://api.semanticscholar.org/CorpusID:247619234`.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *International Joint Conference on Artificial Intelligence*, 2018a. URL `https://api.semanticscholar.org/CorpusID:23206414`.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *ArXiv*, abs/1807.06158, 2018b. URL `https://api.semanticscholar.org/CorpusID:49863329`.

Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. In *International Joint Conference on Artificial Intelligence*, 2019. URL `https://api.semanticscholar.org/CorpusID:173188327`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *ArXiv*, abs/1711.00937, 2017. URL `https://api.semanticscholar.org/CorpusID:20282961`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399, 2022. URL https://api.semanticscholar.org/CorpusID:252715594.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

Che Wang, Xufang Luo, Keith W. Ross, and Dongsheng Li. Vrl3: A data-driven framework for visual deep reinforcement learning. *ArXiv*, abs/2202.10324, 2022a. URL https://api.semanticscholar.org/CorpusID:247011862.

Chen Wang, Linxi (Jim) Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *ArXiv*, abs/2302.12422, 2023a. URL https://api.semanticscholar.org/CorpusID:257205825.

Jianren Wang, Sudeep Dasari, Mohan Kumar Srirama, Shubham Tulsiani, and Abhi Gupta. Manipulate by seeing: Creating manipulation controllers from pre-trained representations. *ArXiv*, abs/2303.08135, 2023b. URL https://api.semanticscholar.org/CorpusID:257505038.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023c. URL https://api.semanticscholar.org/CorpusID:257805127.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6312–6322, 2022b. URL https://api.semanticscholar.org/CorpusID:254408955.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*, pages 36411–36430. PMLR, 2023d.

Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024a.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023e.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Jian Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Y. Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv*, abs/2307.06942, 2023f. URL https://api.semanticscholar.org/CorpusID:259847783.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *ArXiv*, abs/2208.06193, 2022c. URL https://api.semanticscholar.org/CorpusID:251554821.

Wayve. Lingo-1: Exploring natural language for autonomous driving. https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/, 2024. Accessed: 2024-04-04.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *ArXiv*, abs/2401.00025, 2023. URL `https://api.semanticscholar.org/CorpusID:266693687`.

William F Whitney, Tatiana Lopez-Guevara, Tobias Pfaff, Yulia Rubanova, Thomas Kipf, Kimberly Stachenfeld, and Kelsey R Allen. Learning 3d particle-based simulators from rgb-d videos. *arXiv preprint arXiv:2312.05359*, 2023.

Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2318–2328, 2021.

Hongtao Wu, Ya Jing, Chi-Hou Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *ArXiv*, abs/2312.13139, 2023a. URL `https://api.semanticscholar.org/CorpusID:266374724`.

Jialong Wu, Haoyu Ma, Chao Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *ArXiv*, abs/2305.18499, 2023b. URL `https://api.semanticscholar.org/CorpusID:258967679`.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022.

Markus Wulfmeier, Abbas Abdolmaleki, Roland Hafner, Jost Tobias Springenberg, Michael Neunert, Tim Hertweck, Thomas Lampe, Noah Siegel, Nicolas Manfred Otto Heess, and Martin A. Riedmiller. Compositional transfer in hierarchical reinforcement learning. *arXiv: Learning*, 2019. URL `https://api.semanticscholar.org/CorpusID:213142736`.

Markus Wulfmeier, Arunkumar Byravan, Sarah Bechtle, Karol Hausman, and Nicolas Heess. Foundations for transfer in reinforcement learning: A taxonomy of knowledge modalities. *arXiv preprint arXiv:2312.01939*, 2023.

Zhou Xian, Théophile Gervet, Zhenjia Xu, Yi-Ling Qiao, Tsun-Hsuan Wang, and Yian Wang. Towards generalist robots: A promising paradigm via generative simulation. 2023. URL `https://api.semanticscholar.org/CorpusID:259202431`.

Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. *arXiv preprint arXiv:2307.03659*, 2023.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Y He, H Liu, H Chen, X Cun, X Wang, Y Shan, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834, 2021. URL `https://api.semanticscholar.org/CorpusID:231632575`.

Haoyu Xiong, Haoyuan Fu, Jieyi Zhang, Chen Bao, Qiang Zhang, Yongxi Huang, Wenqiang Xu, Animesh Garg, and Cewu Lu. Robotube: Learning household manipulation from human videos with simulated twin environments. In *Conference on Robot Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:250164181`.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

Hu Xu, Gargi Ghosh, Po-Yao (Bernie) Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL `https://api.semanticscholar.org/CorpusID:238215257`.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela M. Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. *ArXiv*, abs/2307.09955, 2023. URL https://api.semanticscholar.org/CorpusID:259982636.

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *ArXiv*, abs/2001.02908, 2020. URL https://api.semanticscholar.org/CorpusID:210116815.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. 2022. URL https://api.semanticscholar.org/CorpusID:254535696.

Wilson Yan, Yunzhi Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *ArXiv*, abs/2104.10157, 2021. URL https://api.semanticscholar.org/CorpusID:233307257.

Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *International Conference on Machine Learning*, pages 39062–39098. PMLR, 2023.

Jingyun Yang, Max Sobol Mark, Brandon Vu, Archit Sharma, Jeannette Bohg, and Chelsea Finn. Robot fine-tuning made easy: Pre-training rewards and policies for autonomous real-world reinforcement learning. *arXiv preprint arXiv:2310.15145*, 2023a.

Mengjiao Yang, Dale Schuurmans, P. Abbeel, and Ofir Nachum. Dichotomy of control: Separating what you can control from what you cannot. *ArXiv*, abs/2210.13435, 2022. URL https://api.semanticscholar.org/CorpusID:253098210.

Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B. Tenenbaum, and P. Abbeel. Probabilistic adaptation of text-to-video models. *ArXiv*, abs/2306.01872, 2023b. URL https://api.semanticscholar.org/CorpusID:259075709.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *ArXiv*, abs/2310.06114, 2023c. URL https://api.semanticscholar.org/CorpusID:263830899.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, P. Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *ArXiv*, abs/2303.04129, 2023d. URL https://api.semanticscholar.org/CorpusID:257378587.

Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *ArXiv*, abs/2302.03024, 2023e. URL https://api.semanticscholar.org/CorpusID:256615635.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:203737314.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *ArXiv*, abs/2107.09645, 2021. URL https://api.semanticscholar.org/CorpusID:236134152.

Denis Yarats, David Brandfonbrener, Hao Liu, Michael Laskin, Pieter Abbeel, Alessandro Lazaric, and Lerrel Pinto. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*, 2022.

Weirui Ye, Yunsheng Zhang, Mengchen Wang, Shengjie Wang, Xianfan Gu, Pieter Abbeel, and Yang Gao. Foundation reinforcement learning: towards embodied generalist agents with foundation prior assistance. *ArXiv*, abs/2310.02635, 2023. URL `https://api.semanticscholar.org/CorpusID:263620344`.

Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Kumar Gupta, P. Abbeel, and Lerrel Pinto. Visual imitation made easy. In *Conference on Robot Learning*, 2020. URL `https://api.semanticscholar.org/CorpusID:221095826`.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10459–10469, 2022. URL `https://api.semanticscholar.org/CorpusID:254563906`.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David C. Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. 2023a. URL `https://api.semanticscholar.org/CorpusID:263830733`.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023b.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023c.

Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *ArXiv*, abs/2401.11439, 2024. URL `https://api.semanticscholar.org/CorpusID:267069070`.

Haoqi Yuan, Ruihai Wu, Andrew Zhao, Hanwang Zhang, Zihan Ding, and Hao Dong. Dmotion: Robotic visuomotor control with unsupervised forward model learned from videos. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7135–7142, 2021. URL `https://api.semanticscholar.org/CorpusID:232146710`.

Kevin Zakka, Andy Zeng, Peter R. Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, 2021. URL `https://api.semanticscholar.org/CorpusID:235368061`.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023b.

Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022.

Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding. 2024. URL `https://api.semanticscholar.org/CorpusID:267760035`.

Tony Zhao, Siddharth Karamcheti, Thomas Kollar, Chelsea Finn, and Percy Liang. What makes representation learning from videos hard for control? 2022. URL `https://api.semanticscholar.org/CorpusID:252635608`.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification. *ACM Transactions on Graphics (TOG)*, 37:1 – 12, 2018. URL `https://api.semanticscholar.org/CorpusID:219893035`.

Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pages 1719–1735. PMLR, 2021.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023a.

Deyao Zhu, Yuhui Wang, Jürgen Schmidhuber, and Mohamed Elhoseiny. Guiding online reinforcement learning with action-free offline pretraining. *ArXiv*, abs/2301.12876, 2023b. URL `https://api.semanticscholar.org/CorpusID:256390557`.

Ziwen Zhuang, Zipeng Fu, Jianren Wang, Christopher Atkeson, Soeren Schwertfeger, Chelsea Finn, and Hang Zhao. Robot parkour learning. *arXiv preprint arXiv:2309.05665*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.