Predicting Long-horizon Futures by Conditioning on Geometry and Time

Tarasha Khurana Deva Ramanan

Carnegie Mellon University

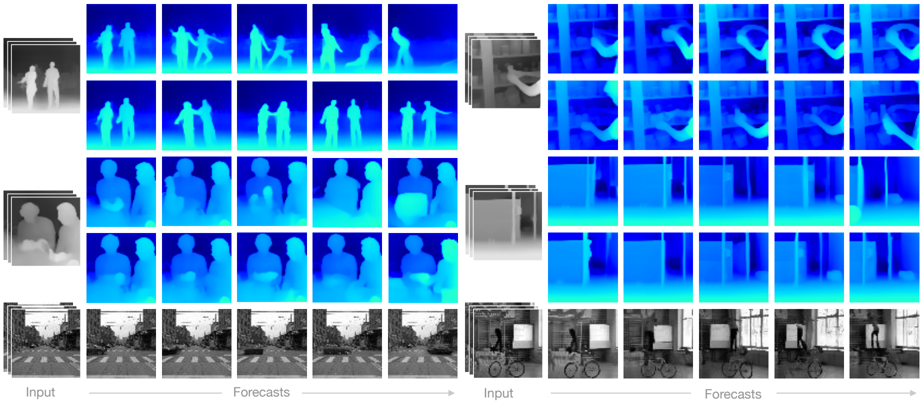

Fig. 1. Predicting long-horizon futures by conditioning on geometry and time. In this work, we focus on the task of forecasting sensor observations given the past. Since the unobserved future can unfold in multiple ways, we capitalize on the recent explosion in large-scale pretraining of 2D diffusion networks, which are able to model the multi-modal distribution of natural images. By introducing invariances in data and additionally learning to condition on frame timestamps, we are able to equip 2D diffusion models with the ability to perform predictive video modeling using moderately-sized training data. Since we are able to query arbitrary timestamps, we find new sampling schedules that perform better than traditional autoregressive / hierarchical sampling strategies. Here, we show two pseudo-depth **futures** each, given the **past** pseudo-depth for four scenes, along with forecasts from training with luminance.

Abstract. Our work explores the task of generating future sensor observations conditioned on the past. We are motivated by 'predictive coding' concepts from neuroscience as well as robotic applications such as self-driving vehicles. Predictive video modeling is challenging because the future may be multi-modal and learning at scale remains computationally expensive for video processing. To address both challenges, our key insight is to leverage the large-scale pretraining of image diffusion models which can handle multi-modality. We repurpose image models for video prediction by conditioning on new frame timestamps. Such models can be trained with videos of both static and dynamic scenes. To allow them to be trained with modestly-sized datasets, we introduce invariances by factoring out illumination and texture by forcing the model to predict (pseudo) depth, readily obtained for in-the-wild videos via off-the-shelf monocular depth networks. In fact, we show that simply modifying networks to predict grayscale pixels already improves the accuracy of video prediction. Given the extra controllability with timestamp conditioning, we propose sampling schedules that work better than the traditional autoregressive and hierarchical sampling strategies. Motivated by probabilistic metrics from the object forecasting literature, we create a benchmark for video prediction on a diverse set of videos spanning indoor and outdoor scenes and a large vocabulary of objects. Our experiments illustrate the effectiveness of learning to condition on timestamps, and show the importance of predicting the future with invariant modalities.

1 Introduction

Recent innovations in generative visual modeling have paved the way for a variety of applications. In this work, we focus on the task of conditionally generating (or forecasting) the future from past observations. Our motivation is from an embodied perspective. Evidence from neuroscience suggests *predictive coding* to be a fundamental phenomena for biological processing of visual streams [45]; specifically, biological agents process the future by first predicting what may occur and then updating predictions based on actual observations (similar to classic dynamic models such as kalman filters [26, 28]). Predictive modeling is the backbone of autonomous systems such as self-driving vehicles that forecast environment motion for downstream applications like motion planning [9, 29].

Why is this hard? One of the challenges in operationalizing such a predictive task is that the future is inherently *multi-modal*; consider an outdoor scene of a busy intersection where cars may continue straight or turn. Encoding such uncertainty has been a notorious challenge, but recent generative modeling techniques such as diffusion networks provide an attractive formalism for generating multiple *samples* from the multi-modal future. As such, our work follows a growing body of work on video-based diffusion models [5,6,22,66]. But crucially, rather than generating video samples unconditionally or conditioned on textual prompts, we generate future frames conditioned on past observations. However, this introduces a significant practical challenge of satisfying compute demands that are required for learning from massive-scale video datasets.

Our approach relies on two key insights. First, we take the view that accurate video prediction can be achieved by using recent 2D image diffusion models [47] alone. This is because such models are trained on a massive scale of image data that (inevitably) contains multiple stages or instances of *temporal* events (c.f. Fig. 2). We add a control mechanism to image diffusion models in the form of *timestamps* that help build a temporal understanding, and are fairly easy to obtain. Moreover, by training on videos with differing framerates, our timestamp-conditioned model can support a variety of video prediction tasks including short-horizon forecasting, autoregressive long-horizon forecasting, and even frame interpolation (by conditioning on fractional timestamps). This flexibility to sample an arbitrary timestamp in the future lets us probe newer (and stronger) sampling schedules, other than just autoregressive and heirarchical sampling that is most commonly used by prior work [17, 22, 58].

Modalites Our second key insight is motivated by embodied applications such as robotics / self-driving vehicles. Oftentimes, we are not concerned with the photometric properties of the future (e.g., "what will be the color of this car?") but rather geometric properties (e.g.,"*where* will this car be?") [30]. Geometric processing of depth sensors is commmon in point cloud processing [39, 62, 63] & occupancy forecasting [2, 27, 38] from 3D LiDAR sweeps, and legged locomotion using only egocentric depth [1, 11]. However, such depth data is not as widely as available as passive camera imagery. To leverage the latter, we show that one can use (pseudo) depth, which can readily be obtained at-scale for videos by running recent monocular depth estimators [4]. We show that simply choosing to forecast in grayscale rather than color already simplifies the forecasting problem to a great degree. More importantly, introducing invariances in data allows us to finetune image diffusion models with only 1000 videos in about 7 hours (11 hours for training them from scratch with same data)!

Contributions In summary, we present a video prediction diffusion network that can be efficiently fine-tuned from foundational image networks by additionally conditioning on frame timestamps. The flexibility in sampling an arbitrary future, allows us to propose stronger sampling schedules than prior work. We also demonstrate that our design choices allow our model to be trained on a modest but diverse set of ∼1000 videos from the TAO dataset [12], that encompasses a variety of indoor and outdoor scenes, spanning a large vocabulary of objects. We use a variety of baselines [13, 17, 58] (including nonlinear regression, constant and linear prediction) to illustrate the effectiveness of different modalities. To illustrate the effectiveness of multi-modal forecasting, we make use of probabilistic (top-K) metrics developed in the forecasting community [10].

2 Related work

Extracting priors from image diffusion models Denoising diffusion models [20, 53] have emerged as an expressive and powerful class of text-to-image generative models. Because of the massive scale of data used to train models like Stable Diffusion [47], Imagen [48] and DALL·E [44], numerous follow-up works have investigated and built upon their rich representations. Specifically for novel-view synthesis, a few works [37, 49] aimed at extracting geometric, pose priors from Stable Diffusion [47] for object or scene-level novel-view-synthesis. Other sparse-view 3D reconstruction works [55, 73, 74] also draw motivation from the same concept for distilling the information from image diffusion into 3D models. A new paradigm of text/image-to-3D assets emerged, where many works [42, 51, 54] iteratively enforced 3D consistency from the outputs of image diffusion models, whereas others repurposed the image models for directly predicting tri-plane representations [23]. In fact, a dedicated study was conducted for understanding the 3D priors learnt by image diffusion models [72].

Similarly for the task of video or motion diffusion, some works [52, 58, 65] have attempted to "inflate" image diffusion models to suit video generation, with normalization tricks, a general phenomenon that has apppeared before for designing convolutional video understanding architectures [8]. This also extends to the task of 3D motion generation, be it for humans [14] or object trajectories [3, 16]. In a similar spirit, we address the task of video forecasting, emphasizing the fact that in order to repurpose 2D diffusion models to suit the video-based task of forecasting given the past, it is important to extract and control the axis of time, by explicit conditioning on fractional timestamps.

Video diffusion models For video diffusion, algorithms have been built on top of recurrent or 3D architectures, including 3D convolutions [22], and RNNs [68], usually coupled with large-scale training datasets. Apart from these, there has been a meteoric rise in recent developments in dedicated text-to-video diffusion models, ranging from industrial-scale pretraining [6,15,19,34], to multi-modality conditioning and generation networks [66]. Some of these methods are even designed for extremely-long autoregressive video generation [17,58,69]. We instead explore the setting where in addition to a moderately-sized data, only limited training resources are available for building a model that conditions on an input timestamp, instead of text (therefore, find the open-sourced Stable Video Diffusion [5] to be out of resource bounds). We also find better sampling schedules than autoregressive and hierarchical sampling.

Training with masked-autoencoder objectives The ground-breaking findings from learning self-supervisable representations with masked autoencoders [18], have recently been adopted by image and video transformer architectures [25, 56, 60, 70, 71], and diffusion models designed for a variety of tasks [61, 67]. Although we do not explicitly train in the fashion of masked autoencoders, we touch upon a similar finding when designing the timestamp conditioning mechanism for optimizing the forecasting performance at inference.

Forecasting for autonomous systems In robotics, an important precursor to motion planning is forecasting what the scene and its agents will look like in the future [10, 24, 64]. In self-driving, this spans the field of point cloud [62, 63], and recently, occupancy forecasting [2, 27, 38]. Forecasting videos of depth has a direct analogue to works that forecast range images of point clouds from LiDAR sensors [39]. For the task of legged locomotion in quadrupeds, egocentric-depth is increasingly becoming the sole modality that robots rely on [1,11]. This is largely for the reason that depth acts a low-level actionable cue that helps generalization across a vast set of diverse environments for robot navigation. We are motivated by this, and explore forecasting future geometries for use in autonomous systems.

3 Method

We lean on recent image-to-image diffusion architectures, specifically Zero-1-to-3 [37], trained for changing the camera viewpoint of an object given its RGB

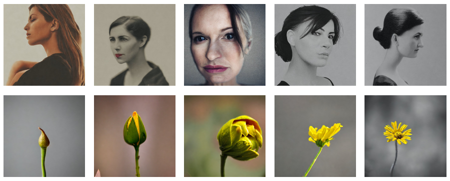

Fig. 2. Using 2D diffusion models for video prediction As part of designing the *video* prediction architecture, we make the important design choice of using *image* diffusion models. Owing to the scale of data such models are trained on, we can expect them to understand indepedent stages of *temporal* events such as 'turning head from left to right', and 'flower bud opening up'. We show individual frames prompted from Stable Diffusion v2. We propose to add a control knob to image models in the form of timestamps that helps in temporal understanding.

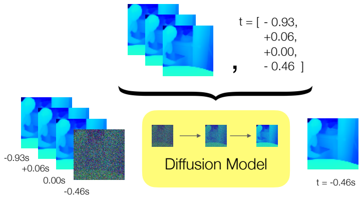

Fig. 3. High-level architecture We use a diffusion model that conditions on three video frames, their corresponding timestamps and a query timestamp. It generates a single video frame for the query. We adopt the two-stream conditioning from image-to-image models [37], and (1) channel-concatenate the context frames with the noisy input to diffusion model, and (2) CLIP-encode the context frames for cross-attention across the UNet layers. Context and query timestamps are positionally encoded and concatenated with CLIP embeddings.

image. We repurpose its image and camera pose conditioning for the task of timestamp-conditioned video forecasting given multiple past contexts.

3.1 Problem formulation

Given a set of context frames $\mathbf{c} \in \mathbb{R}^{K \times H \times W \times C}$ from a video of a (static or dynamic) scene, our goal is to generate a frame $\mathbf{x} \in \mathbb{R}^{1 \times H \times W \times C}$ for the same scene but from a different point in time, t. Let all timestamps in consideration be $\mathbf{t} \in \mathbb{R}^{K+1}$. Then, we want to learn a function g that generates an estimate of the unobserved frame $\mathbf{x}$ given context frames $\mathbf{c}$ and timesteps $\mathbf{t}$,

$$\hat{\mathbf{x}} = g(\mathbf{c}, \mathbf{t}) \tag{1}$$

Since, $\hat{\mathbf{x}}$ is unobserved, it inherently follows a multi-modal distribution, making its prediction underconstrained. To this end, we exploit pretrained large diffusion models like Stable Diffusion [33,47] that can model and sample from such multi-modal distributions of natural images. We can use single-frame 2D diffusion models for the task of video prediction, as their large-scale pretraining likely covers the space of temporal events and the different stages of their unfolding. In Fig. 2, we show different stages of two temporal events, prompted from Stable Diffusion v2. However, such architectures are not straight-forward to use, as time-conditioned video prediction demands for two new capabilities: first, the ability to generate a new frame that is consistent with the historical context frames $\mathbf{c}$, and second, the ability to listen to the continuous valued timestamps, $\mathbf{t}$.

Given the above, we formalize the task of video prediction in the context of diffusion models as follows. Given a dataset of videos with a known FPS, we extract snippets of length $K + 1$ and construct a training sample as $\{\mathbf{x}, \mathbf{c}, \mathbf{t}\}$. Using this training data, we start from the natural scene-level data distribution learnt by Stable Diffusion Image Variations [31,33] and finetune it for controlling both the conditioning with the context frames, and timestamp scalars. Architecturally (ref. Fig. 3), we use a denoising UNet ϵ_θ [37], that looks at 64×64 images. For any timestep $i \sim [0, 1000]$, we train ϵ_θ with the well-adopted noise prediction objective for diffusion training,

$$\min_\theta \mathbb{E}_{z \sim x, i, \epsilon \sim \mathcal{N}(0,1)} ||\epsilon - \epsilon_\theta(z_i, i, f(\mathbf{c}, \mathbf{t}))||_2^2. \tag{2}$$

where $f(\mathbf{c}, \mathbf{t})$ is the conditioning embedding discussed in the following subsections. At inference, we start from pure gaussian noise, and iteratively denoise it, steering the denoised image in the direction of the conditioning embedding.

3.2 Conditioning on context views

We use a two-stream image conditioning protocol from prior work [37] but modify it to suit our multi-frame setting. For conditioning on low-level features of the input context frames (such as depth, texture, and motion patterns of scene actors), we concatenate the K frames with the noisy input image to the UNet. For conditioning on higher-level features of the input context frames (such as the scene elements, contextual background, and observed camera trajectory), we pass the context frames through the CLIP image encoder [43] to get their image embeddings. We additionally construct a "residual" CLIP embedding for the target frame, by learning the weights on K embeddings and taking their weighted average. Intuitively, this "guides" the target image with a residual embedding that can be hooked onto, in order to generate the prediction.

3.3 Conditioning on timestamp scalars

In addition to building a conditioning mechanism for the context views, we also need to let the denoising UNet know, which timestamps the context frames belonged to, and which timestamp we are probing for. To accomplish this, we positionally encode the timestamp scalar with sinusoidal embeddings,

$$\gamma(t) = (\sin 2^0 \pi t, \cos 2^0 \pi t, \ldots, \sin 2^{L-1} \pi t, \cos 2^{L-1} \pi t) \tag{3}$$

This ensures that even if every timestamp value is not seen during training, any high-frequency variation of it can be approximated in the frequency domain at inference. We concatenate this with CLIP embeddings, and cross-attend them at every residual block in the UNet architecture.

Even though at *inference* this method addresses forecasting, we *train* it for a 'random timestamp prediction' objective (*i.e.*, the order of K frames and their timestamps can be arbitrary), instead of the task of forecasting itself. We detail more results from this finding in Sec. 4.4.

3.4 Stitching together a video from individual frames

At inference, we generate long-horizon forecasts by predicting one frame at-a-time, which means that our model has to be queried more than once. Consider the case where we want to predict depth maps for T timesteps in the future. Prior work for long-horizon generation tends to make use of T sequential autoregressive next-frame predictions [17,58], or $\log(T)$ hierarchical [17,22,52] predictions that first predict a low framerate future that is iteratively refined into $2\times$ higher framerate predictions (until T frames are generated). However, both sampling strategies have their drawbacks; autoregressive prediction may suffer from "drift" as the historical window of frames (to be conditioned on) will eventually contain only predicted frames rather than actual ground-truth histories. On the other hand, hierarchical sampling may not exhibit enough temporal coherence.

Interestingly, because our approach explicitly conditions on both input and output timestamps when making predictions, our trained model can support both such sampling strategies in addition to other more flexible approaches. We describe two such flexible approaches, which Sec. 4.2 shows perform better than the conventional sampling. First, given pairs of past frames and their timestamps, $\{\mathbf{c}_{-k:-1}, \mathbf{t}_{-k:-1}\}$, one can directly jump to all futures $t \in [1, T]$ independently. We term this ***Direct*** sampling. While this predicts more plausible futures because 'real' historical frames are used for conditioning, generated frames aren't temporally coherent (every frame might be sampled from a different future).

To improve temporal consistency, we propose *mixing* forecasts from direct sampling (which are accurate but temporally inconsistent) with forecasts from autoregressive sampling (which are temporally consistent but not as accurate as they are conditioned on the previously-predicted past, $\{\mathbf{c}_{t-k:t-1}, \mathbf{t}_{t-k:t-1}\}$). This means that for outputs x_D^T and x_A^T generated from direct and autoregressive sampling respectively, we can linearly combine these two inference pathways during the reverse diffusion process similar to classifier-free guidance [21],

$$x_D^t = g(\mathbf{c}_{-k:-1}, \mathbf{t}_{-k:-1}) \qquad x_A^t = g(\mathbf{c}_{t-k:t-1}, \mathbf{t}_{t-k:t-1})$$

$$x_M^t = x_A^t + w_m \cdot (x_D^t - x_A^t) \tag{4}$$

where w_m is the mixing guidance and g is a generative model. We term this sampling schedule, ***Mixed*** sampling. Intuitively, this makes samples from direct inference more coherent, and samples from autoregressive inference more plausible, as they now condition on a 'real' past. This also curbs the tendency of autoregressive inference to blow up at longer horizons as the output sample can now always fall back on predictions with direct inference.

Training details For all experiments in this work, $K = 3, L = 160, w_m = 2.0$. We train our architecture with classifier free guidance, *i.e.* we randomly remove the conditioning to generate unconditional frames (which can be used as a guidance signal during inference [21]). During training, the diffusion model predicts noise, and we set the probability of dropping the conditioning for classifier free guidance to 10%. During inference, we use a guidance of 2.0 for all experiments, with DDPM sampling for 40 iterative denoising steps. We do not perform diffusion in the latent space, but train and evaluate on images of size 64 × 64 using the Stable Diffusion Image Variations [31] UNet. To circumvent the use of VAE, we learn two new convolutional layers at the start and end of the UNet that help the input image to adjust to the weights of the latent space diffusion model, similar in spirit to prior works [40,57] that also do not depend on the VAE. We learn all new layers 10× faster than other layers, for training from scratch. We train the network with a batch size of 12 for 10k iterations (which takes ∼7 hours on 8 NVIDIA RTX A6000s), using AdamW with $\beta_1 = 0.95, \beta_2 = 0.999, \epsilon = 1e^{-8}$ and weight decay of $1e^{-6}$, with a learning rate of $1e^{-4}$.

4 Experiments

4.1 Benchmarking Setup

Datasets To cover a wide range of dynamic environments from a number of domains like activity recognition and self-driving, we use the large-vocabulary diverse tracking dataset, TAO [12]. TAO is a collection of seven different datasets that is originally used for multi-object tracking. For its unconstrained *dynamic* nature of videos, we repurpose it for predictive modeling. For rigid scenes, we also include video sequences from Common Objects in 3D (CO3Dv2) [46]. CO3Dv2 is a collection of 19k video sequences spanning objects from 51 MS-COCO [35] categories, designed for use in object-level 3D reconstruction and new-view synthesis of *static* scenes. We experiment with three different modalities: RGB videos, their luminance channels and most importantly, sequences of *pseudo-depth*, where the pseudo-depth is obtained from a single-frame monocular depth estimator, ZoeDepth [4], that predicts metric depth for scenes. We randomly sample the input and output frames in a window of 8s across the entire length of a video and shuffle the frame ordering for training. For dataset splits of TAO and using metric depth from CO3Dv2, please see supplement.

Evaluation settings For benchmarking, we consider two settings. First, we evaluate single-frame forecasting. Because this is a scalable evaluation, we benchmark all baselines discussed below and do all ablations for the setting where methods are asked to generate a single prediction for either the future +1s or +10s with input frames given at {-1.0, -0.5, 0}s. Note the forecasting windows are motivated by and reminiscent of motion planning benchmarking [7,30].

Second, we evaluate multi-frame forecasting for up to +10s long horizon. This setting allows us to empirically evaluate the proposed direct and mixed sampling schedules. The input is still provided at {-1.0, -0.5, 0}s and samplers generate predictions for future {+1, +2, +3, ..., +10s}.

Metrics For evaluating depth prediction across both TAO and CO3Dv2 datasets, we adopt the scale and shift invariant L1 error on relative depth maps from monocular depth estimation literature [32], where scale and shift are computed as a minimization of the following least squares objective:

$$(s,t) = \arg\min_{(s,t)} \sum_{i=1}^{M} (s\mathbf{d}_i + t - \mathbf{d}_i^*)^2 \tag{5}$$

Here, $\mathbf{d}_i$ is the set of per-pixel predicted depths, and $\mathbf{d}_i^*$ are the corresponding groundtruth values. Using Eq. 4, the L1 error is computed as $e = \frac{1}{M}\sum_{i=1}^{M} |s\mathbf{d}_i + t - \mathbf{d}_i^*|$. For evaluating both grayscale and RGB modalities, we follow prior work in novel-view synthesis [41,73] and compute the peak-signal-to-noise ratio (PSNR), which measures mean color difference. We take motivation from the forecasting literature in the autonomous driving domain [10] and use Top-k versions of both L1 and PSNR metrics: we take k samples from the model and report the best L1 / PSNR of k. When benchmarking multi-frame depth forecasting, we compute an average trajectory error (ATE, *i.e.* L1 error across the entire predicted sequence), and compute the Top-k errors across a set of k trajectories.

Baselines We compare to state-of-the-art video prediction architectures MCVD [58], FDM [17] and RIVER [13] and construct three simple baselines for video prediction: (1) constant past which predicts the current frame as the future, (2) linear extrapolation from the two temporally closest context frames, and (3) non-linear regression, which is trained for the task of forecasting the next +1.0s using our architecture but without cross-attention layers (therefore, no conditioning) with an L2 loss on the predicted depth from diffusion model. We retrain MCVD [58], FDM [17] and RIVER [13] on our TAO pseudo-depth dataset and use them at inference for single-frame forecasting given three past frames. For MCVD, we use the 'concat' variant as it has lower memory requirements.

Finally, in the setting where the scene is rigid but camera has a non-zero motion, like in CO3Dv2, we compare to a state-of-the-art method for sparse (3-) view reconstruction, SparseFusion [73], on the task of novel-view depth synthesis. Here, we evaluate on a randomly sampled set of test sequences from the core subset proposed in a prior work [46]. This subset consists of 10 object categories from CO3Dv2. All experiments, including qualitative analysis, on CO3Dv2 against SparseFusion can be found in the supplement.

4.2 Comparison to state-of-the-art

We begin the quantitative analysis by comparing our method to MCVD [58], FDM [17] and RIVER [13] for future timestamp prediction in dynamic videos.

Short horizon forecasting We evaluate our method and all baselines for single-frame +1s forecasting in Tab. 1. We find that our method outperforms state-of-the-art video prediction methods, MCVD [58], FDM [17] and RIVER [13]. We posit that against MCVD, our randomized frame prediction objective during training and additional conditioning on timestamps, helps in learning better temporal coherence across frames. FDM, specifically, is not designed for scenes that have dynamic actors, so may perform suboptimally when learning to handle dynamics. RIVER's bottleneck is video prediction in a significantly low dimensional latent space which results in imprecise reconstructions at inference.

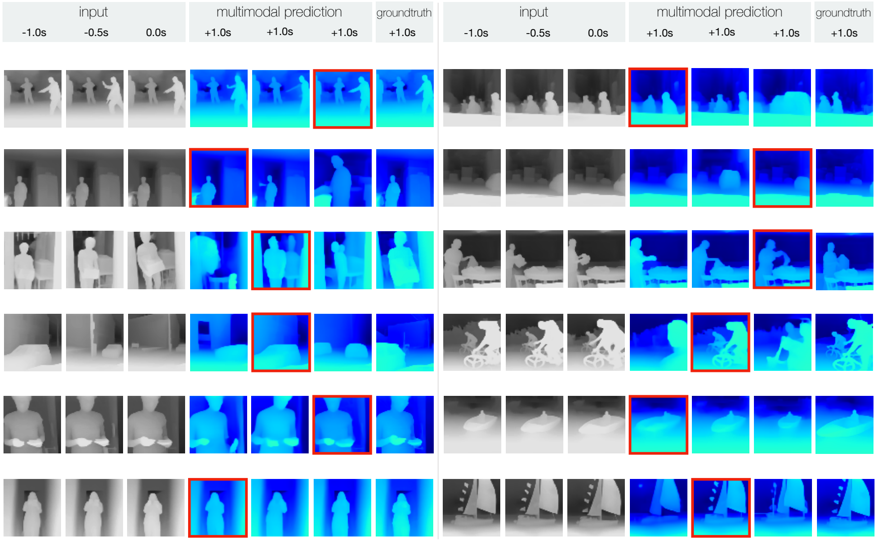

Fig. 4. Qualitative analysis of single-frame short horizon forecasting We show examples of **input**-**output**-groundtruth triplets. Given 3 past frames as input, we show 3 different samples of the future from our diffusion network, and the corresponding groundtruth. Prediction highlighted in red is the closest to groundtruth. Despite learning from only 1000 videos and training for only 7 hours, our method learns to generate multiple realistic futures and listens to low-level details in the historical context frames (e.g., scene structure, actors performing events, and overall camera motion). For reference, the events across examples in row major form could be described as, 'playing in field', 'crossing road', 'doing laundry', 'driving (front view)', 'exiting room while holding a box', 'picking up from table', 'driving (side view)', 'biking', 'fidgeting', 'boating with camera zooming in', 'standing in hallway', 'sailing'.

When comparing our method to simple baselines such as the (non-learned) constant past and linear extrapolation, and the unimodal non-linear regression, it becomes readily apparent that, (1) both constant past and linear extrapolation are *strong* baselines for scenes that are static and have been captured by a stationary camera, and (2) regression, expectedly, stands out as an even stronger baseline (often used by pioneering work in occupancy forecasting [30]) but regresses to the mean of multi-modal distribution of possible futures. This mean-seeking behaviour still suffices for most scenes and metrics (such as our *mean* Top-1 L1 error), but our method provides the increased capability of sampling multiple futures which reduces the probabilistic Top-5 L1 further. An in-depth qualitative analysis of all baselines along with our method, and a training / inference runtime analysis, can be found in the supplement.

Long horizon forecasting First, we evaluate the single-frame forecasting for +10s using three different sampling schedules as discussed in Sec. 3.4. Note

Table 1. Comparison to state-of-the-art We evaluate future depth prediction for +1s against state-of-the-art video prediction methods by retraining them for pseudo-depth prediction, and against other simple or non-learned baselines. We find that our method beats prior work with a substantial margin.

Method	Top-1 L1	Top-3 L1	Top-5 L1
Linear extrapolation	21.25	21.25	21.25
Non-linear regression	7.96	7.96	7.96
Constant past	**7.15**	7.15	7.15
RIVER [13]	10.82	10.32	10.17
MCVD [58]	10.54	7.83	7.12
FDM [17]	9.99	7.78	7.24
Ours	8.40	**6.93**	**6.59**

Table 2. Single-frame long horizon forecasting We evaluate future depth prediction for +10s against the discussed baselines. Given our timestamp conditioning, we are able to explore more flexible sampling schedules like direct and mixed, which perform better than the widely used autoregressive sampling.

Method	Top-1 L1	Top-3 L1	Top-5 L1
Linear extrapolation	21.80	21.80	21.80
Non-linear regression	14.76	14.76	14.76
Constant past	**11.61**	11.61	11.61
Ours (autoreg.)	12.93	11.24	10.77
Ours (direct)	12.65	11.13	10.65
Ours (mixed)	12.39	**10.97**	**10.51**

that in the single-frame case, direct and hierarchical sampling are equivalent as the first lowest framerate layer of hierarchical sampling generates the +10s frame *directly* from the given inputs. Compared to the baselines in Tab. 2, we find that our proposed mixed sampling strategy performs the best at the probabilistic L1, while surprisingly constant prediction suffices for the Top-1 metric.

Second, in Tab. 3, we benchmark different sampling schedules discussed for the multi-frame forecasting case with Top-k ATE, where samplers predict a 1fps sequence up to 10s in the future. First, we find that directly jumping to a future frame, performs better than the conventional autoregressive and hierarchical sampling schedules. Specifically, for autoregressive sampling, the error in prediction starts adding up as the diffusion models starts conditioning on predicted frames rather than the groundtruth past. For hierarchical sampling, the future is coarsely decided by the first set of predictions. After this, intermediate frames can only be interpolated and the future cannot be refined. Finally, for mixed sampling, we find that it produces more accurate and coherent futures as it benefits from the advantages of both direct & autoregressive sampling (Fig. 5).

4.3 Comparison between different modalities

We also explore luminance and RGB modalities for single-frame +1s video prediction. Specifically, instead of pseudo-depth, we train our model for luminance and RGB prediction under the short-horizon forecasting setting. When evaluating RGB, we factor out the luminance channel from the prediction and use that for benchmarking against our luminance prediction model. In Tab. 4, we see that introducing invariances in the input data (such as learning from luminance rather than a combination of color and texture), helps in making forecasting easier. Quantitatively, the Top-5 PSNR increases by a large margin of ∼2.1 points.

We also compare our depth and RGB prediction models by running Stable Diffusion Depth2Img on the predicted depth. We find that, (1) our depth is

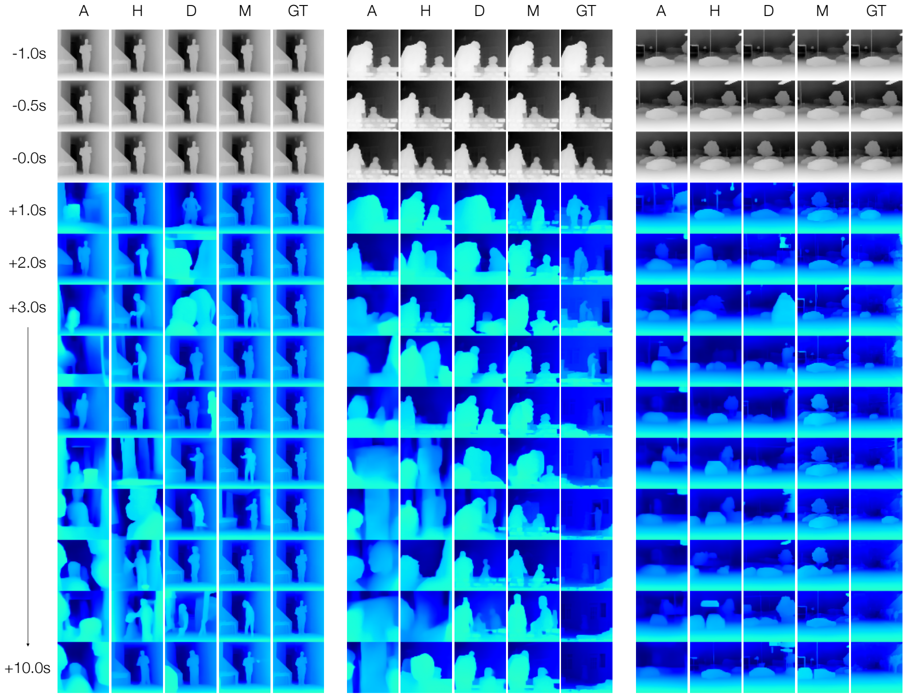

Fig. 5. Comparison between sampling strategies We qualitatively analyse the **predictions** from the four discussed sampling strategies given same **past** alongside the **G**round**T**ruth: **A**utoregressive and **H**ierarchical [17, 22, 58], and **D**irect and **M**ixed, which are enabled by our timestamp conditioning. As detailed in Sec. 3.4, we find that autoregressive sampling suffers from "drift", and the performance of hierarchical sampling is governed by its first layer of forecasts (*i.e.* lacks flexibility). While direct sampling does better, it cannot produce coherent futures. Concretely, we propose mixed sampling, which mixes both the coherence of autoregressive and the accuracy of direct samples. For reference, the samples from left to right could be described as, 'standing in hallway', 'interaction between two people', 'side-view from a driving car'.

readily usable for downstream tasks, and (2) it is infact easier to do RGB prediction by learning to forecast scene depth first! For details on the depth2img parameters and text prompts used, please see supplement.

Finally, we compare all three modalities for the task of pseudo-depth forecasting. This requires running ZoeDepth [4] on our predictions from the luminance and RGB models. We once again find that it is easier to directly learn to forecast depth, without depending on color or scene texture.

4.4 Architecture ablations

We ablate our design decisions in Tab. 5 for +1s forecasting. For a fair comparison with MCVD, FDM and RIVER and to see how much performance boost we

Table 3. Multi-frame long horizon forecasting We evaluate multiple sampling strategies for generating a sequence of future depths upto +10s. We evaluate with Top-k ATE and find that our proposed mixed sampling, which is able to generate accurate and coherent futures, performs the best of all.

Method	Top-1 ATE	Top-3 ATE	Top-5 ATE
Ours (autoreg.)	15.20	13.56	13.06
Ours (hierar.)	15.15	13.77	13.32
Ours (direct)	13.54	12.73	12.43
Ours (mixed)	**12.16**	**11.73**	**11.58**

Table 4. Comparison between different modalities. We quantitatively enable a fair comparison between modalities by evaluating them for either pseudo-depth, luminance, or RGB forecasting. We consistently find that invariant modalities like depth and luminance perform drastically better than RGB at video prediction. **L**uminance and **C**olor models are evaluated with PSNR and **D**epth with L1.

Method	Top-1	Top-3	Top-5	Evaluation
Ours-L	**16.32**	**17.07**	**17.33**	Luminance
Ours-RGB	12.16	14.47	15.24	
Ours-D	**16.28**	**16.44**	**16.50**	Color
Ours-RGB	14.10	15.40	15.80	
Ours-D	**8.40**	**6.93**	**6.59**	Pseudo-depth
Ours-L	22.68	19.17	17.61	
Ours-RGB	27.05	20.88	19.33	

get from the Stable Diffusion Image Variations weights, we attempt to train from scratch. Surprisingly, this training does not take much longer than finetuning (11 hours for training from scratch vs. 7 hours for finetuning), and performs remarkably well (still better than the state-of-the-art). We further attempt to reduce the number of parameters in our network by removing 1 convolution block each from the UNet encoder and decoder. This brings number of parameters closer to the state-of-the-art video prediction models, and training the smaller model from scratch still beats all baselines. For exact parameter counts, see supplement.

Next, we find that the CLIP embedding is essential to conditioning on the past context frames and results in a drop of ∼1.4 points if ablated. Finally, we ablate the design decisions for the timestamp conditioning.

Anchoring timestamps When designing the timestamp conditioning, we find that it helps to condition on relative rather than absolute timestamps. This includes, "anchoring" timestamps to a constant frame in the input such that that frame always occurs at t=0s. For our experiments, we choose the third context frame as anchor, and this frame at timestamp +0s becomes the 'current' frame for the diffusion network. This practice has recently been adopted by methods [59] that use diffusion models for conditioning on 3D cues such as camera pose.

Timestamp randomization One of our key insights is that training directly for the task of forecasting is sub-optimal to training for a random frame prediction objective. Specifically, the drop in performance is rather significant (∼1.8 Top-1 L1 points). This aligns with the insights from masked autoencoder literature [18, 56, 60] where randomization in masking results in better representations. Analogously, destroying structure in the data and making the final task harder for the diffusion models, helps in building robust temporal understanding.

Table 5. Architecture ablations. We ablate our method under the single-frame +1s forecasting setting with L1 error. We assess the benefits from using pretrained weights [31], a large model, and CLIP embeddings for context frames. We additionally investigate the design choices in creating the timestamp conditioning, by using relative timestamps and randomizing their order. Ablations indicate that all design choices play a crucial role.

Method	Top-1	Top-3	Top-5
Ours	8.40	**6.93**	**6.59**
- pretrained weights	**7.89**	7.04	6.78
- 2 × conv blocks	8.50	7.27	6.89
- CLIP embedding	9.19	8.25	7.95
- timestamp anchoring	9.00	7.08	6.62
- random timestamps	10.24	7.89	7.31

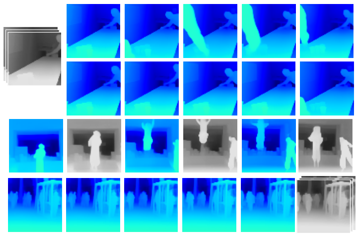

Fig. 6. Video applications. We show examples of how the formulation of our method unlocks multiple video applications: variable framerate forecasting (top row at 1FPS, second row at 5FPS), (third row) frame interpolation given the frames in gray, and (last row) backcasting at 5FPS given the future. For reference, events from top to bottom could be described as, 'playing pool', 'jumping', 'walking on a busy street'.

4.5 Applications

In Fig. 6, we show qualitative examples of different applications our approach can be used for: (1) generating videos at varying framerates for different horizons given the same context frames, (2) frame interpolation at fractional timestamps between the given context frames, and (3) looking back in the past with negative timestamps given the future frames as context.

5 Discussion

We focus on the problem of predicting the future from past sensor observations, and take motivation from the neuroscience literature on predictive coding. Since the future is multi-modal and can therefore unfold in multiple ways, we lean on the explosive advancements in large-scale pretraining of diffusion models, that can internally represent such multi-modal distributions. With two key modifications to *image* diffusion networks, we come up with a method for predictive *video* modeling. We find that for training with moderately-sized datasets, it helps to introduce invariances in the data – such as forecasting only pseudo-depth or luminance of real-world images. Physical quantities like pseudo-depth are readily usable by downstream tasks in robot autonomy (locomotion and planning) as they represent the time-to-contact. We introduce a mechanism for diffusion models to condition on a frame's timestamp. This allows models to perform better at the task of forecasting (especially when they are *not trained* for forecasting). Timestamp conditioning also lets us come up with flexible sampling schedules for long-horizon forecasting. We find that these new sampling schemes perform better than conventional autoregressive or hierarchical sampling strategies.

Acknowledgments We thank Nupur Kumari and Jason Zhang for insightful comments. Tarasha Khurana is supported by a funding from Bosch Research.

References

1. Agarwal, A., Kumar, A., Malik, J., Pathak, D.: Legged locomotion in challenging terrains using egocentric vision. In: Conference on Robot Learning. pp. 403–415. PMLR (2023) 3, 4
2. Agro, B., Sykora, Q., Casas, S., Urtasun, R.: Implicit occupancy flow fields for perception and prediction in self-driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1379–1388 (2023) 3, 4
3. Ahn, H., Mascaro, E.V., Lee, D.: Can we use diffusion probabilistic models for 3d motion prediction? arXiv preprint arXiv:2302.14503 (2023) 4
4. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 3, 8, 12
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 2, 4
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators 2, 4
7. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021) 8
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 4
9. Casas, S., Sadat, A., Urtasun, R.: Mp3: A unified model to map, perceive, predict and plan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14403–14412 (2021) 2
10. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8748–8757 (2019) 3, 4, 9
11. Cheng, X., Shi, K., Agarwal, A., Pathak, D.: Extreme parkour with legged robots. arXiv preprint arXiv:2309.14341 (2023) 3, 4
12. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 436–454. Springer (2020) 3, 8, 20, 21
13. Davtyan, A., Sameni, S., Favaro, P.: Efficient video prediction via sparsely conditioned flow matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23263–23274 (2023) 3, 9, 11, 21, 22, 23
14. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion

model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023) 4

15. Girdhar, R., et al.: Emu video: Factorizing text-to-video generation by explicit image conditioning (2023) 4
16. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17113–17122 (2022) 4
17. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible diffusion modeling of long videos. Advances in Neural Information Processing Systems **35**, 27953–27965 (2022) 3, 4, 7, 9, 11, 12, 21, 22, 23
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 4, 13
19. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 4
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 7
22. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 2, 3, 4, 7, 12
23. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023) 3
24. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023) 4, 22
25. Huang, B., Zhao, Z., Zhang, G., Qiao, Y., Wang, L.: Mgmae: Motion guided masking for video masked autoencoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13493–13504 (2023) 4
26. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering **82**(Series D), 35–45 (1960) 2
27. Khurana, T.: Argoverse 2 challenge on 4d occupancy forecasting at the workshop on autonomous driving, cvpr. In: https://eval.ai/web/challenges/challenge-page/1977/overview (2023) 3, 4
28. Khurana, T., Dave, A., Ramanan, D.: Detecting invisible people. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3174–3184 (2021) 2
29. Khurana, T., Hu, P., Dave, A., Ziglar, J., Held, D., Ramanan, D.: Differentiable raycasting for self-supervised occupancy forecasting. In: European Conference on Computer Vision. pp. 353–369. Springer (2022) 2
30. Khurana, T., Hu, P., Held, D., Ramanan, D.: Point cloud forecasting as a proxy for 4d occupancy forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1116–1124 (2023) 3, 8, 10
31. Labs, L.: Stable diffusion image variations. In: https://huggingface.co/lambdalabs/sd-image-variations-diffusers (2022) 6, 8, 14

32. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019) 8
33. Lee, T.S., Mumford, D.: Hierarchical bayesian inference in the visual cortex. JOSA A **20**(7), 1434–1448 (2003) 5, 6
34. Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., Wang, J.: Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. arXiv preprint arXiv:2309.00398 (2023) 4, 22
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 8
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 24
37. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023) 3, 4, 5, 6
38. Mahjourian, R., Kim, J., Chai, Y., Tan, M., Sapp, B., Anguelov, D.: Occupancy flow fields for motion forecasting in autonomous driving. IEEE Robotics and Automation Letters **7**(2), 5639–5646 (2022) 3, 4
39. Mersch, B., Chen, X., Behley, J., Stachniss, C.: Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In: Conference on Robot Learning. pp. 1444–1454. PMLR (2022) 3, 4
40. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023) 8
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020) 9, 20
42. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 3
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 6
44. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 3
45. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience **2**(1), 79–87 (1999) 2
46. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10901–10911 (2021) 8, 9, 20, 21, 24
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3, 5, 23

48. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 3
49. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single real image. arXiv preprint arXiv:2310.17994 (2023) 3
50. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) 20
51. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 3
52. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 4, 7
53. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ICLR (October 2021), https://arxiv.org/abs/2010.02502 3
54. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023) 3
55. Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J.B., Durand, F., Freeman, W.T., Sitzmann, V.: Diffusion with forward models: Solving stochastic inverse problems without direct supervision. arXiv preprint arXiv:2306.11719 (2023) 3
56. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022) 4, 13
57. Turner, K.: Decoding latents to rgb without upscaling. In: Huggingface 8
58. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022) 3, 4, 7, 9, 11, 12, 21, 22, 23
59. Wang, J., Rupprecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9773–9783 (2023) 13
60. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023) 4, 13
61. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., Feichtenhofer, C.: Diffusion models as masked autoencoders. arXiv preprint arXiv:2304.03283 (2023) 4
62. Weng, X., Nan, J., Lee, K.H., McAllister, R., Gaidon, A., Rhinehart, N., Kitani, K.M.: S2net: Stochastic sequential pointcloud forecasting. In: European Conference on Computer Vision. pp. 549–564. Springer (2022) 3, 4
63. Weng, X., Wang, J., Levine, S., Kitani, K., Rhinehart, N.: 4d forecasting: Sequantial forecasting of 100,000 points. In: Euro. Conf. Comput. Vis. Worksh. vol. 3 (2020) 3, 4
64. Wilson, B., Qi, W., et al.: Argoverse 2.0: Next generation datasets for self-driving perception and forecasting. In: NeuRIPS Datasets and Benchmarks Track (Round 2) (2021) 4

65. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023) 4
66. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. arXiv preprint arXiv:2306.00943 (2023) 2, 4
67. Yan, H., Liu, Y., Wei, Y., Li, Z., Li, G., Lin, L.: Skeletonmae: graph-based masked autoencoder for skeleton sequence pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5606–5618 (2023) 4
68. Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. Entropy **25**(10), 1469 (2023) 4
69. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv preprint arXiv:2303.12346 (2023) 4
70. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023) 4
71. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023) 4
72. Zhan, G., Zheng, C., Xie, W., Zisserman, A.: What does stable diffusion know about the 3d scene? arXiv preprint arXiv:2310.06836 (2023) 3
73. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12588–12597 (June 2023) 3, 9, 20, 21
74. Zou, Z.X., Cheng, W., Cao, Y.P., Huang, S.S., Shan, Y., Zhang, S.H.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. arXiv preprint arXiv:2308.14078 (2023) 3

Appendix

In this appendix, we extend our discussion of the performance of our method on predicting diverse future geometries, both qualitatively and quantitatively.

A Dataset splits and evaluation

We use the diverse TAO [12] dataset for learning dynamism in unconstrained scenes. Since TAO is a tracking *benchmark*, its training set is smaller than the validation or test sets. For this reason, we train on the validation set of TAO (∼1000 videos) and report all results on one randomly sampled subsequence each, from the train set (containing about 500 videos). The randomly sampled set is fixed across all experiments for fair comparison.

In the case of evaluation on rigid scenes, we use CO3Dv2 [46]. Although CO3Dv2 has groundtruth depth that is obtained from COLMAP [50], it is not dense. For this reason, we still run ZoeDepth on CO3Dv2 and use those pseudo-depth maps for training our method, but use the valid depths from groundtruth for computing the probabilistic L1 metric on CO3Dv2 for both our method and the baseline, Sparsefusion [73]. In the following section, we analyse depth forecasting on CO3Dv2 qualitatively and quantitatively.

B Novel-view synthesis

We consider the case where a moving camera captures a static scene. In literature, this has been studied under the umbrella of novel-view synthesis from dense [41] or sparse views [46, 73]. The setting we evaluate (context from {-1s, -0.5s, 0s} and prediction at +1s) falls under sparse view reconstruction/synthesis. We use a variant of our model trained on CO3Dv2 alongside TAO. Note that the state-of-the-art method, SparseFusion [73], which we use as a baseline has access to future camera pose for rendering the novel-view from its reconstruction, whereas for our method, the camera pose is *unknown*. Along with the scene, it is sampled from the timestamp-conditioned diffusion model during inference. Despite this disadvantage, we find that our method predicts plausible depths for the objects, *in addition* to the depth predictions for the object backgrounds, which is ignored by SparseFusion. We cover some qualitative analysis in Fig. 7.

In Tab. 6, we formally evaluate the task of novel-view synthesis. Since CO3Dv2 has multi-view object data captured in the form of videos, we structure this problem as, given frames at -1.0s, -0.5s, 0.0s, we want to predict the frame at +1.0s. For our method, only the future timestamp is available. For SparseFusion, instead of future timestamp, future camera pose information is available. Quantitatively, we find that our method performs better than SparseFusion on a few categories (`donut`, `apple`, `ball`, `suitcase`, etc.) because of more smooth depth forecasts (ref. Fig. 7). Other than that, for categories where camera viewpoint matters more (`hydrant`, `bench`, `plant` etc.) for rendering the geometry, SparseFusion does better.

Table 6. Novel-view synthesis results on Co3Dv2 core subset. We evaluate our method for the task of novel-view depth synthesis with Top-1 L1 error on normalized depth, against a recent approach for 3-view reconstruction. Over the set of categories in the core subset of CO3Dv2, we see that SparseFusion performs better overall. Unlike SparseFusion, our method does not have the access to future camera pose or object mask. Despite this, it is able to generate plausible depth maps for object turn-table sequences in Co3Dv2. We only compute the metric on valid groundtruth depths inside the given object mask in CO3Dv2, without penalizing the background forecasts.

Method	Donut	Apple	Hydrant	Vase	Cake	Ball	Bench	Suitcase	Teddybear	Plant	Overall
SparseFusion	11.54	28.94	**19.04**	**14.29**	26.28	27.64	**75.89**	34.32	40.04	**71.53**	**34.95**
Ours	**7.22**	**19.23**	30.39	21.82	**19.57**	**20.65**	91.83	**33.56**	**38.44**	75.23	35.79

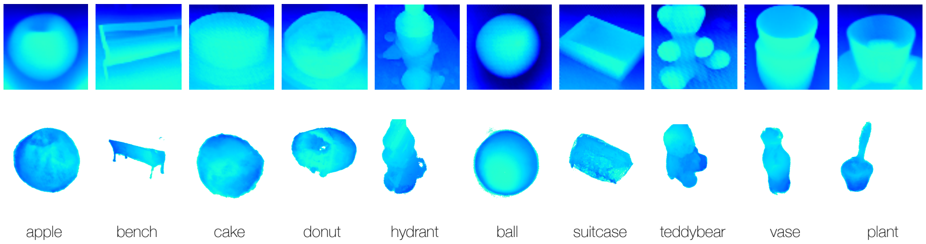

Fig. 7. Qualitative comparison to novel-view synthesis We train and qualitatively evaluate our method on CO3Dv2. From the core subset [46] of 10 categories in CO3Dv2, we show the visualization of novel-view synthesis from both our method (**top**) and SparseFusion [73] (**bottom**). While SparseFusion has access to the parameters of both the input and new (or future) view, these are implicitly estimated by our method from the camera trajectory encoded in the past frames. Therefore, our method does not rely on known camera poses! Qualitatively, our method performs favourably on the task of new-view synthesis from 3 input views, while handling dynamics and backgrounds in general for a wide variety of scenes.

More importantly, the extension of our method for the task of novel-view synthesis coupled with its performance on forecasting for dynamic scenes, we show that we can handle object backgrounds, and dynamic video settings such as in TAO [12], unlike methods for sparse-view static object/scene reconstruction like SparseFusion.

C Qualitative comparison with baselines

In Fig. 9-12, we qualitatively compare our method to all baselines discussed in Tab. 1 in main paper. It can be seen that predicting the most recent past frame as the future serves as a strong baseline. Non-linear regression, regresses to the mean of the future distribution. FDM [17], RIVER [13], MCVD [58] and our method instead, sample *modes* of the future distribution. Linear extrapolation

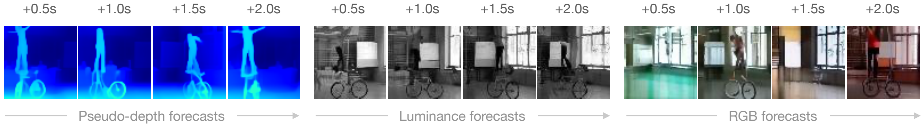

Fig. 8. With same training data/architecture/duration, a depth or luminance model learns better temporal coherence than RGB.

is not shown but it serves as a strong baseline when the scene is static. Overall, we see that our method produces more realistic and diverse outputs, as compared to MCVD [58] which usually does not diverge much from the input views. RIVER [13] struggles to learn temporal coherence because of its processing in the low-dimensional latent space, and FDM [17] is not able to learn precise object boundaries likely because it is not designed to handle dynamic scenes.

Mean vs. mode-seeking behavior Fig. 9, row 1 shows how the non-linear regression baseline hallucinates multiple possible futures, thereby introducing artifacts because of this phenomenon (e.g., multiple people are visible in the output). In contrast to this, our method and other state-of-the-art approaches are able to sample multiple futures separately, commonly referred to as the mode-sampling or mode-seeking behavior.

Depth vs. luminance vs. RGB Fig. 8 shows a qualitative comparison between forecasts from different modalities; we see that RGB forecasting tends to be noisy. Temporal coherence is better learned with invariant modalities such as pseudo-depth or luminance. While many recent works do show successful RGB video generation [24, 34], they typically train on far more data than us (days of compute on 10 million videos vs 7 hours of compute on 1000 videos).

D Comparison to state-of-the-art on long-horizon forecasting

In Tab. 7, we compare the performance of our method shown in the main paper, by retraining FDM [17] and RIVER [13] for +10s forecasting. Note that this is not an apples-to-apples comparison to our method, as even for the case where we want to predict just the +10s frame with the baselines, they are forced to predict every intermediate (0s to 10s) frame because this is the only way to reach the future +10s. On the other hand, when we evaluate our method for single-frame +10s forecasting, we directly jump to that timestamp.

Quantitatively we see that, (1) errors are higher when methods are used for predicting sequences of future frames, rather than when evaluated for a single timestamp in future, and (2) across the discussed settings, our method performs the best of all with the proposed mixed sampling.

Table 7. Long horizon forecasting We evaluate future depth prediction for +10s against FDM and RIVER, two state-of-the-art methods for video generation in the single-frame (with L1) and multi-frame (with ATE) settings. Given our timestamp conditioning, we are able to explore more flexible sampling schedules like mixed sampling, which performs better than the widely used autoregressive sampling strategies for FDM and RIVER.

Method		L1			ATE	
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
FDM [17]	16.05	13.15	12.29	16.57	14.38	13.79
RIVER [13]	13.21	11.71	11.26	13.34	12.28	11.93
Ours	**12.39**	**10.97**	**10.51**	**12.16**	**11.73**	**11.58**

E Memory requirements and speed

In Tab. 8, we detail the memory and speed requirements of our method and its variants along with the state-of-the-art for the task of +1s single-frame forecasting. First, we find that at inference, our method samples the fastest from the diffusion model. Second, FDM [17] uses the least amount of memory as it has the smallest model. RIVER [13] also uses lesser memory for a lighter architecture since it learns video generation in significantly low dimensional latent space. While these methods allow for a smaller memory footprint, as seen qualitatively and quantitatively, none of them is able to learn persistence and temporal coherence of objects and scenes. For a fair comparison to baselines, we see that a variant of our model that is not initialized with the Stable Diffusion Image Variations weights finishes training in 11 hours, still better than all baselines. Another variant of our model that has lesser parameters and is more comparable to baselines, is much faster to both train and sample from.

Table 8. Resource requirements of baselines for single-frame +1s forecasting.

Method	Params. (M)	Mem. (GB)	Train (hrs.)	Test (s)
RIVER [13]	236	12	32	6.90
MCVD [58]	565	19	66	12.50
FDM [17]	**80**	**8**	72	24.41
Ours	860	21	**7**	4.09
Ours (scratch)	860	21	11	4.09
Ours (small, scratch)	399	16	8	**3.78**

All numbers are provided for batch size = 1. For RIVER, a VQ-GAN needs to be trained whose number of parameters (68M) are added to the RIVER parameters (168M). Note that all our variants quantitatively perform better than the state-of-the-art as shown in the main paper, and these differences in the training and inference resources are even more pronounced when the state-of-the-art methods are used for multi-frame long-horizon future generation.

F Details on Stable Diffusion Depth2Img

In the main paper, we show that given the same amount of training resources, it is better to train a depth video prediction diffusion model and use this 'temporally-aware' depth in conjunction with a single-frame depth-to-image model (such as Stable Diffusion Depth2Img [47]) than an RGB video prediction model. To get this RGB image for every predicted *future* depth frame, we input the RGB image at timestamp `t = 0s` (which is the last input timestamp), alongside every predicted depth frame from the future, into the Stable Diffusion Depth2Img model one-at-a-time. We use the LLaVA [36] model to caption the RGB image at `t = 0s` which is input as the text prompt for the depth-to-image generation. We use `‘ugly looking, bad quality, cartoonish’` as the negative text prompt. The guidance scale is set to 5.0 and the conditioning strength is set to 0.3.

F.1 LLaVA prompting

To get the text prompt from LLaVA, we use the HuggingFace `llava-hf/llava-1.5-7b-hf` weights, and use the input prompt for LLaVA as, `"<image>\nUSER: Caption the image in one long sentence.\nASSISTANT:"`. All text returned after this prompt is used as input to Stable Diffusion Depth2Img. The RGB images are resized at a 512 × 512 resolution, and a max output length of 500 characters is used.

G Limitations

Our method suffers from two important limitations. First, our method is biased towards hallucinating people and cars. For the other categories, the future is rather difficult to forecast. This limitation arises from the fact that TAO, which is used for training, has $\sim$52% of the objects as people, with the second most common category being cars. However, when even a small finetuning set of varied objects from CO3Dv2 [46] are used, our method does perform well on forecasting the future for those categories.

Second, we find that the pseudo-depth produced by our method (and others) is low-fidelity, lacking details that otherwise appeared in the inputs. We posit that this because although neural networks are universal function approximators, they struggle with modeling high-frequency functions.

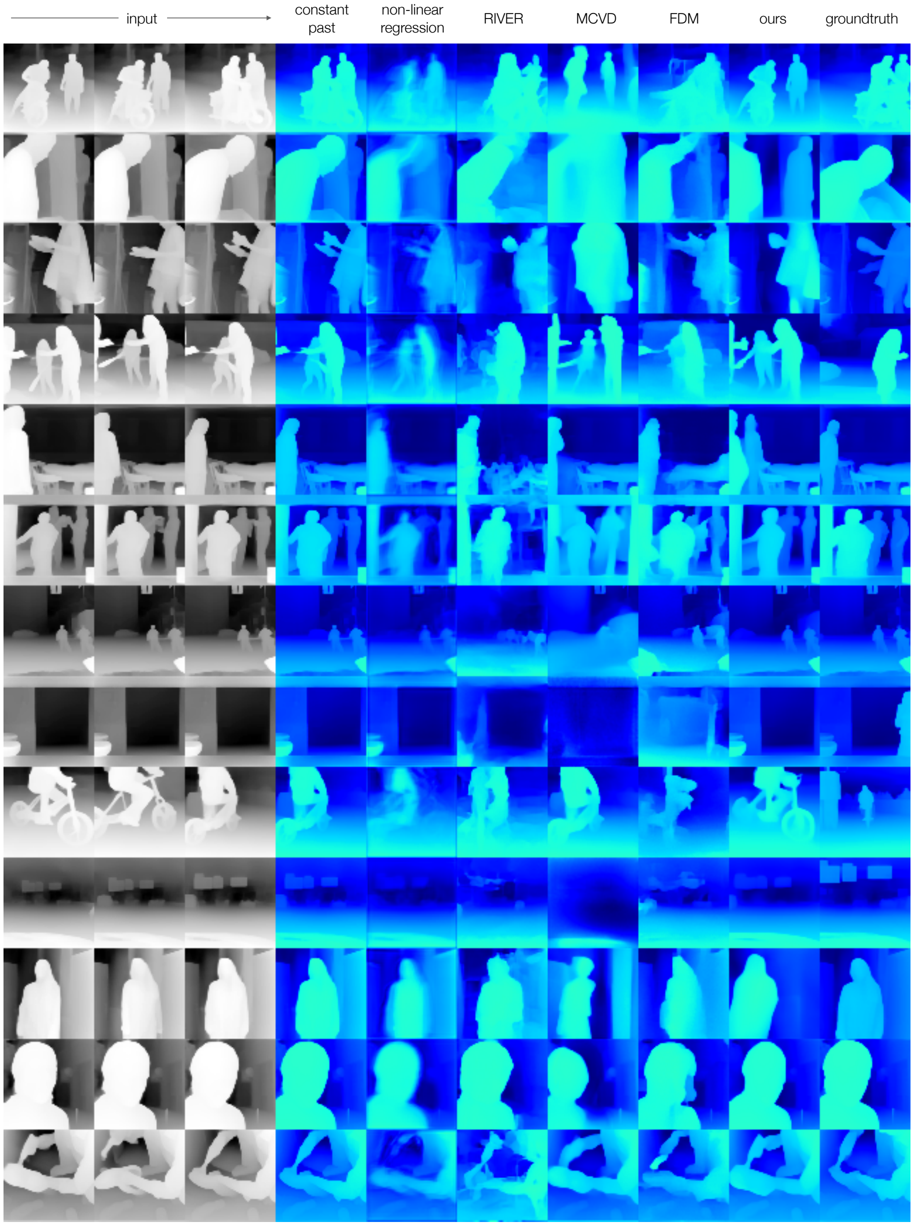

Fig. 9. Qualitative comparison to baselines (1 of 4). We compare to all baselines for the task of short-horizon forecasting on TAO-val set. Given inputs at -1.0, -0.5, 0.0s in gray, methods predict future pseudo-depth at +1.0s. Lighter color is closer depth.

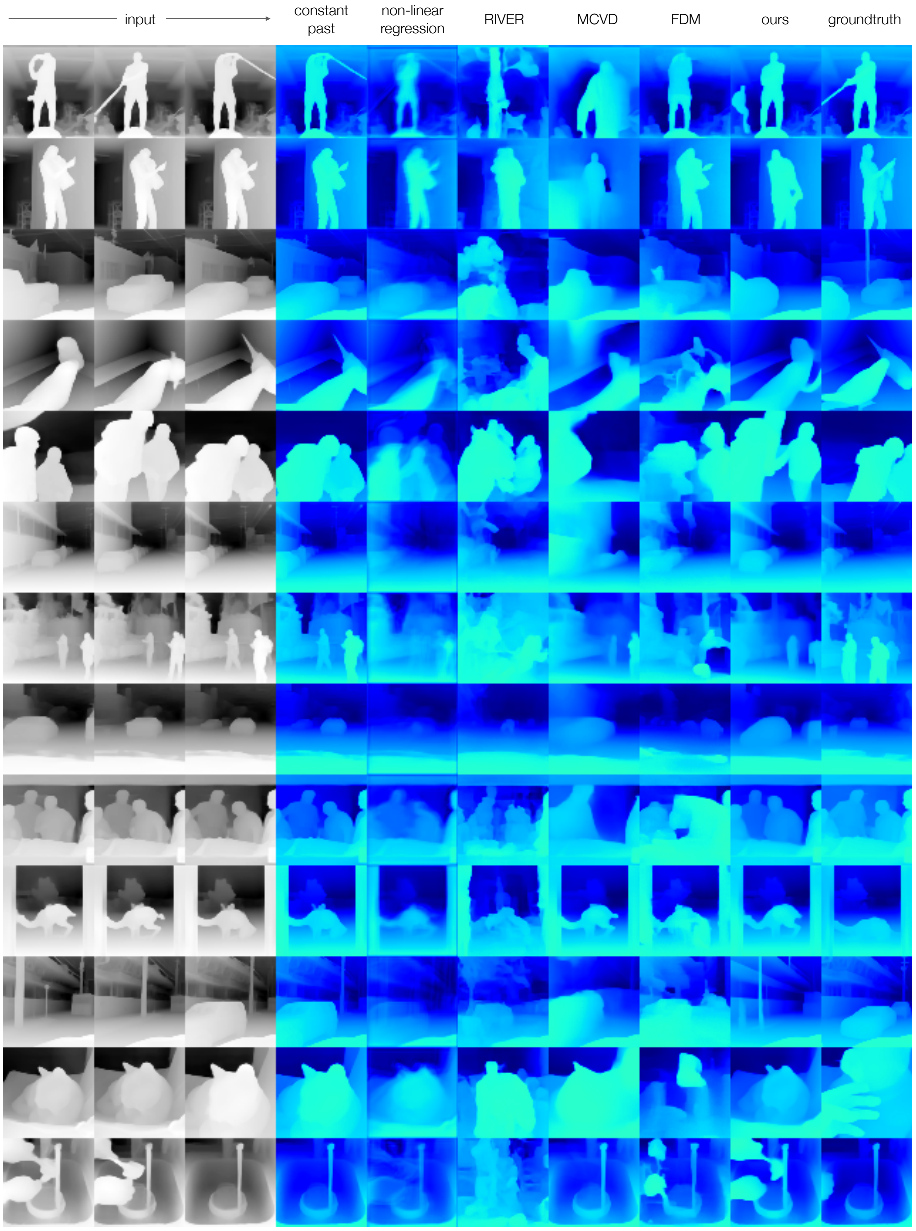

Fig. 10. Qualitative comparison to baselines (2 of 4). We compare to all baselines for the task of short-horizon forecasting on TAO-val set. Given inputs at -1.0, -0.5, 0.0s in gray, methods predict future pseudo-depth at +1.0s. Lighter color is closer depth.

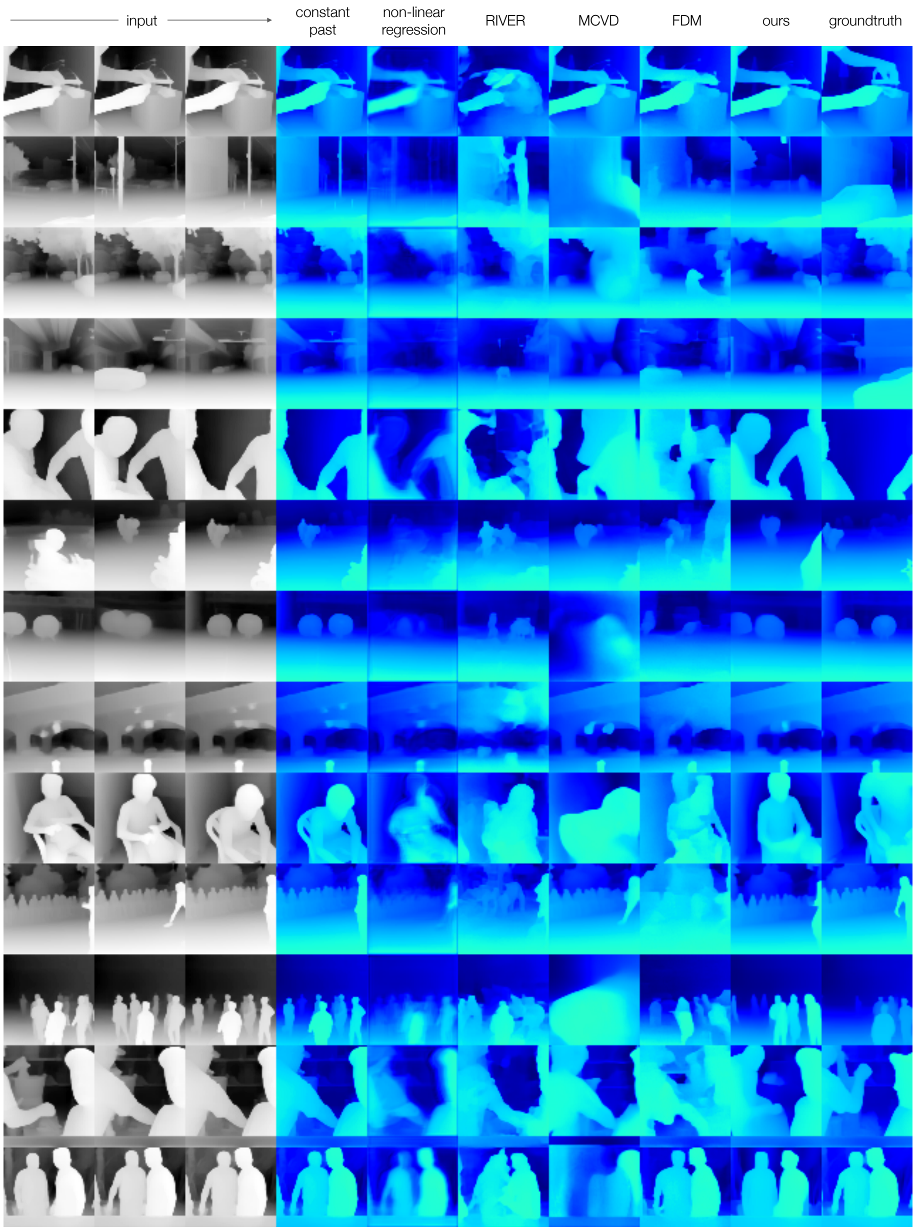

Fig. 11. Qualitative comparison to baselines (3 of 4). We compare to all baselines for the task of short-horizon forecasting on TAO-val set. Given inputs at -1.0, -0.5, 0.0s in gray, methods predict future pseudo-depth at +1.0s. Lighter color is closer depth.

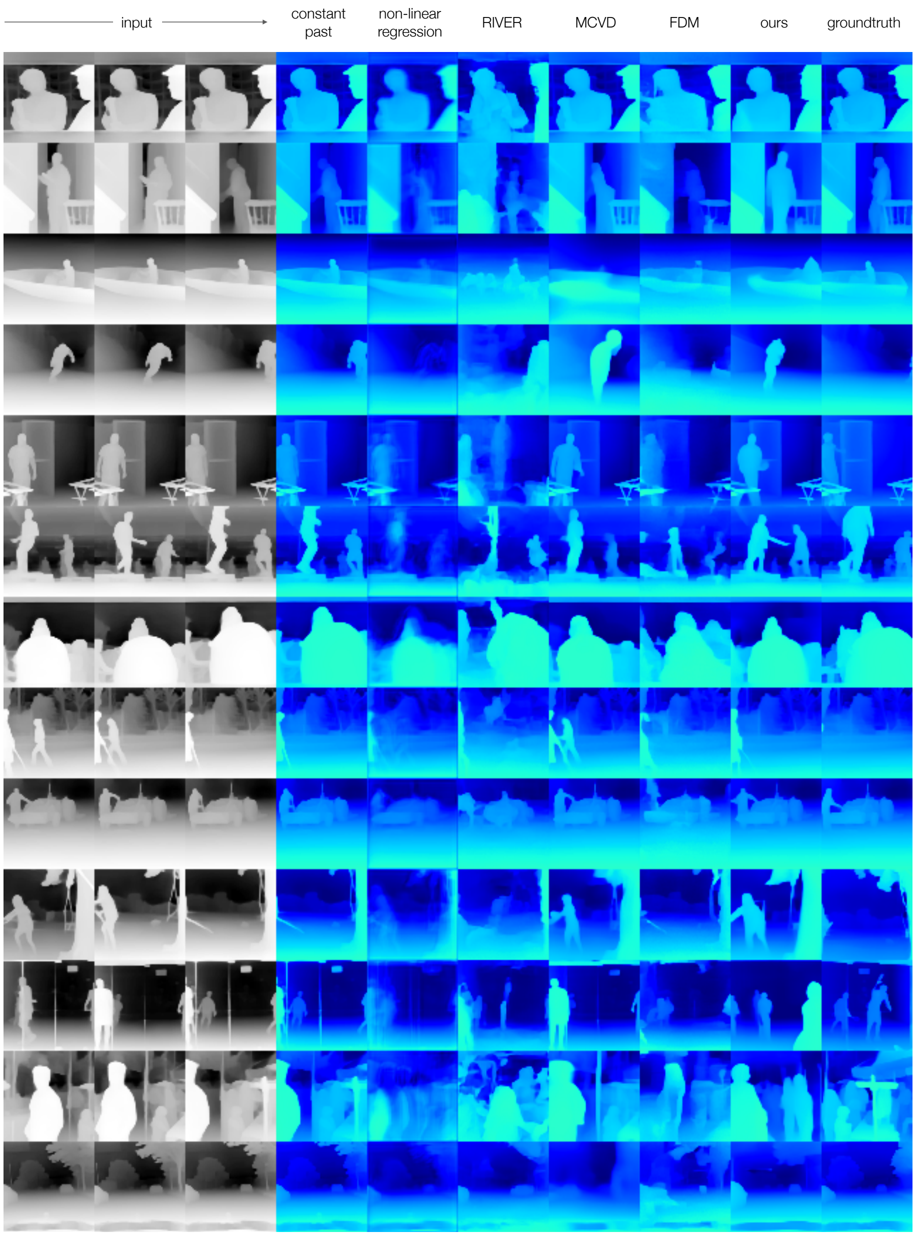

Fig. 12. Qualitative comparison to baselines (4 of 4). We compare to all baselines for the task of short-horizon forecasting on TAO-val set. Given inputs at -1.0, -0.5, 0.0s in gray, methods predict future pseudo-depth at +1.0s. Lighter color is closer depth.