# DAM: DYNAMIC ADAPTER MERGING FOR CONTINUAL VIDEO QA LEARNING

**Feng Cheng**[†] **Ziyang Wang**[†] **Yi-Lin Sung** **Yan-Bo Lin** **Mohit Bansal** **Gedas Bertasius**
Department of Computer Science, UNC Chapel Hill
{fengchan,ziyangw,ylsung,yblin,mbansal,gedas}@cs.unc.edu
[†] Equal Contribution



## ABSTRACT

We present a parameter-efficient method for continual video question-answering (VidQA) learning. Our method, named DAM, uses the proposed **D**ynamic **A**dapter **M**erging to (i) mitigate catastrophic forgetting, (ii) enable efficient adaptation to continually arriving datasets, (iii) handle inputs from unknown datasets during inference, and (iv) enable knowledge sharing across similar dataset domains. Given a set of continually streaming VidQA datasets, we sequentially train dataset-specific adapters for each dataset while freezing the parameters of a large pretrained video-language backbone. During inference, given a video-question sample from an unknown domain, our method first uses the proposed non-parametric router function to compute a probability for each adapter, reflecting how relevant that adapter is to the current video-question input instance. Subsequently, the proposed dynamic adapter merging scheme aggregates all the adapter weights into a new adapter instance tailored for that particular test sample to compute the final VidQA prediction, mitigating the impact of inaccurate router predictions and facilitating knowledge sharing across domains. Our DAM model outperforms prior state-of-the-art continual learning approaches by 9.1% while exhibiting 1.9% less forgetting on 6 VidQA datasets spanning various domains. We further extend DAM to continual image classification and image QA and outperform prior methods by a large margin. The code is publicly available at: https://github.com/klauscc/DAM.

## 1 INTRODUCTION

The role of video in our lives has increased tremendously over the recent years, with millions of hours of video uploaded to the Internet daily. Due to such rapid video growth and the emergence of video-language models (Yu et al., 2021; Yang et al., 2022; Cheng et al., 2023; Wang et al., 2023d; Pramanick et al., 2023b;a), video question-answering (VidQA) has become one of the most important tasks in video understanding. However, modern VidQA models often assume static conditions with fixed training datasets. In contrast, many real-world applications increasingly demand adaptability to distribution shifts of continually arriving datasets. For instance, a VidQA model trained only on movie videos may struggle when questioned about instructional or social media videos due to stark domain disparities. Additionally, even within a single domain, a model trained on videos collected before 2020 may fail to answer questions about videos recorded in 2024 due to a substantial time difference between training and testing videos.

One could address these issues by fine-tuning a VidQA model each time new data is introduced. However, it would cause the model to forget previously acquired knowledge – a phenomenon commonly referred to as *catastrophic forgetting* (McClelland et al., 1995; McCloskey & Cohen, 1989). An alternative strategy is to retrain the model by incorporating both existing training data and the newly acquired data. However, training the model on the combined data is impractical due to the even larger computational cost (Zellers et al., 2021; Fu et al., 2021; Li et al., 2023c; Wang et al., 2022a). These challenges underscore the necessity for *continual VidQA learning*, where the VidQA

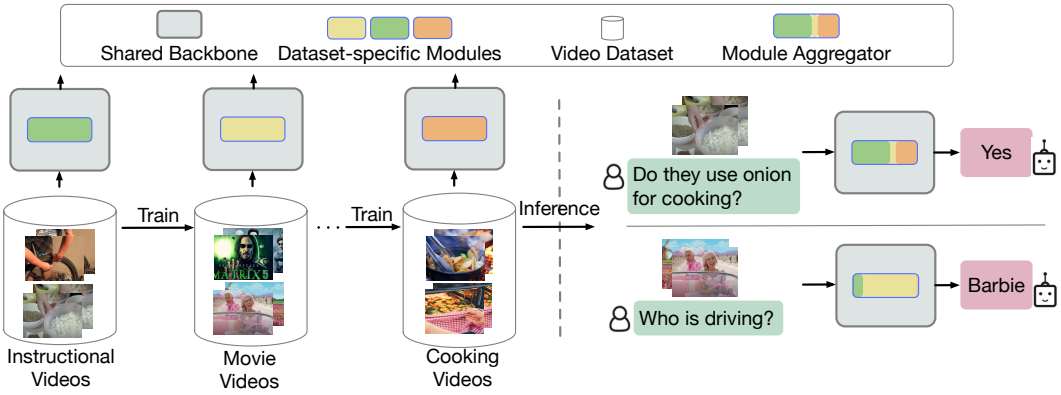


Figure 1: A high-level overview of our proposed Domain-Incremental Learning (DIL) framework for Video Questions-Answering (VidQA). Our model is continually trained on sequentially arriving datasets and evaluated on test samples with unknown dataset identities. Our framework **(i)** incorporates dataset-specific modules to allow specialization and mitigate forgetting, **(ii)** enables efficient adaptation to continually streaming datasets, **(iii)** ensures robustness to incorrect module selections, and **(iv)** facilitates knowledge-sharing across similar datasets.

model gradually learns to incorporate knowledge from continuously evolving video training data with minimal training cost.

In this work, we focus on the Domain-Incremental Learning (DIL) subproblem of continual learning (Kirkpatrick et al., 2017; Wang et al., 2023b), since it matches the above-discussed challenges of continuously adapting to datasets spanning different domains and time shifts. The key challenges in DIL arise from distribution shifts between sequentially-arriving training datasets. When the distribution shifts between datasets are large, the optimal representation for each distribution can be very different, thus leading to poor performance among regularization-based DIL methods (Kirkpatrick et al., 2017; Li & Hoiem, 2017), which use fully shared parameters across datasets. Recent prompt-based approaches (Wang et al., 2022b;e) alleviate this issue by using dataset-specific prompts independently trained on each dataset. During inference, these methods rely on a router function to predict the dataset identity and select the corresponding prompts. However, when distribution shifts between datasets are subtle, predicting dataset identity becomes challenging, adversely affecting the performance of such methods. Additionally, selecting individual dataset-specific prompts prevents knowledge-sharing between datasets, which may be suboptimal when the training datasets are similar. Thus, as shown in Fig. 1, an ideal DIL method should **(i)** incorporate dataset-specific modules to allow specialization and limit catastrophic forgetting, **(ii)** enable efficient adaptation to continually arriving datasets, **(iii)** be robust to incorrect dataset-specific module selections, and **(iv)** facilitate knowledge-sharing across similar domains.

Motivated by these observations, we propose **D**ynamic **A**dapter **M**erging (DAM), a highly-performant, generalizable, and parameter-efficient continual VidQA learning scheme. Our model consists of (i) an adapter for each continually arriving dataset, (ii) a non-parametric router, and (iii) a dynamic adapter merging module. Given a sequence of VidQA datasets spanning different data distributions, we begin by training a *dataset-specific adapter* for each dataset while freezing the pre-trained video-language backbone (e.g., CLIP (Radford et al., 2021) and DeBERTa (He et al., 2020)). Afterward, given a test sample from an unknown dataset during inference, we use a non-parametric video-language router to estimate probabilities for each dataset-specific adapter. These probabilities reflect the relevance of each adapter to that particular video-question input instance. Subsequently, the proposed dynamic adapter merging module merges all the adapter weights into a new adapter instance tailored for that particular test sample to compute the final VidQA prediction. As a result, even if the router produces partially inaccurate probabilities, DAM could still answer the VidQA problem as our dynamic merging scheme incorporates knowledge from multiple adapters, often including those associated with the correct domain. Therefore, the proposed dynamic merging scheme mitigates the impact of inaccurate router predictions and also facilitates knowledge sharing across distributions, thereby enhancing VidQA performance.

Our DAM method outperforms prior prompt-based DIL models (Wang et al., 2022b;e) by **9.1%** on 6 sequentially-introduced VidQA datasets from various domains while exhibiting **1.9%** less forgetting. DAM can also be easily extended to tasks such as image classification ( **+9.32%** on CORe50) and image question-answering (**+4.4%** on a benchmark with 4 datasets). Furthermore, we conduct extensive ablation studies to analyze the relationship between dynamic merging and router, elucidating the key success factors of our approach. We will release our code and pretrained models to enable the community to develop models for this emerging research area of domain-incremental VidQA learning.

## 2 Related Work

**Video Question Answering (VidQA)** represents a fundamental task in video-language understanding, aiming to answer natural language questions from video inputs. Most commonly used methods (Yang et al., 2022; Yu et al., 2023; Xiao et al., 2022; Cheng et al., 2023; Lei et al., 2021; Li et al., 2020; Miech et al., 2019; Sun et al., 2019) leverage video-language models (VLMs) with transformer architecture (Xiao et al., 2022; Lei et al., 2021; Cheng et al., 2023) and large pre-trained language models (Yang et al., 2022; Yu et al., 2023). FrozenBiLM (Yang et al., 2022) handles the multimodal input using a pretrained bidirectional language model and casts VidQA as a masked language modeling problem. SeViLA (Yu et al., 2023) builds upon a large image-language model, BLIP-2 (Li et al., 2023b), and extends it to accommodate video input for VidQA. However, none of these methods are designed to handle continual shifts in training data distribution, which is our focus in this work.

**Continual Learning (CL)** focuses on developing frameworks that can continually learn from streaming training datasets. This is a fundamental challenge for many deep learning methods due to *catastrophic forgetting* (McClelland et al., 1995). Continual learning methods can be categorized into regularization-based approaches (Kirkpatrick et al., 2017; Li & Hoiem, 2017), replay-based approaches (Cha et al., 2021a; Riemer et al., 2018), optimization-based approaches (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018) and representation-based approaches (Gao et al., 2023; Foret et al., 2020; Ermis et al., 2022; Douillard et al., 2022). Several recent CL approaches use pre-trained models for the vision-language domain, including CLiMB (Srinivasan et al., 2022) for task-incremental learning, VQACL (Zhang et al., 2023) and CL-CrossVQA (Zhang et al., 2022) for rehearsal-based Domain-Incremental Learning (DIL). Rehearsal-based methods require storing data of previously used training datasets, which may not be possible in real-world settings due to privacy or intellectual property concerns. In contrast, rehearsal-free CL approaches (Li & Hoiem, 2017; Smith et al., 2023b; 2021; Jung et al., 2023; Li et al., 2023d; Zuo et al., 2024; Wang et al., 2023c;a) do not require storing any previous training data. Among these, several recent prompt-based methods such as L2P (Wang et al., 2022e), DualPrompt (Wang et al., 2022d), S-Prompts (Wang et al., 2022b) and CODA-Prompt (Smith et al., 2023a) used visual prompts (Liu et al., 2023) prepended to a pre-trained transformer and extended prompt-based learning for continual learning scenarios. Unlike these prior prompt-based DIL methods, we propose dynamic adapter merging to alleviate the issues of inaccurate router predictions and enable cross-domain knowledge sharing.

**Model Merging** aims to merge multiple domain models into a single model that can be used for inference on these domains. The work in (Wortsman et al., 2022b; Ilharco et al., 2022b) computes the merged weights as an element-wise arithmetic mean of the weights of all domain models. Subsequently, several methods proposed to improve the performance of the model merging using techniques such as Fisher Merging (Matena & Raffel, 2022), RegMean (Jin et al., 2022), Git Re-Basin (Ainsworth et al., 2022), Task Arithmetic (Ilharco et al., 2022a) and TIES-Merging (Yadav et al., 2023). Model merging has been applied to many scenarios, including federated learning (McMahan et al., 2017), improving out-of-domain generalization (Cha et al., 2021b), and improving performance on a single target task (Gupta et al., 2020; Wortsman et al., 2022a). Recently, the method in (Guerrero-Peña et al., 2022) proposed a Sinkhorn re-basin network for replay-based class incremental continual learning but only experimented with small models (e.g., ResNet18 (He et al., 2016)) on small datasets (e.g., CIFAR-100 (Krizhevsky et al., 2009)). Unlike existing model merging methods that create a single merged model for all datasets, we dynamically generate a new model instance tailored for each test sample with minimal computational overhead.

# 3 TECHNICAL APPROACH

## 3.1 UNIFIED FORMULATION

We first consolidate recent DIL methods (Wang et al., 2022b;e) into a unified formulation. Specifically, most existing DIL methods share a common structure comprising a frozen pretrained backbone $f_\theta$ with parameters $\theta$, dataset-specific modules (i.e., prompts) $M = \{m_1, ..., m_T\}$, and a router. For the dataset arriving at time $t$, only the dataset-specific module $m_t$ is trained, while the backbone $f_\theta$ and all previously learned modules $m_1, ..., m_{t-1}$ are frozen to prevent forgetting. During inference, given a test sample $x$ with an unknown dataset identity, the inference process is formulated as Eqn. 1, where $f_{\theta,m_i}$ represents the pretrained backbone augmented with a dataset-specific module $m_i$, $p$ is the predicted probability depicting how relevant $x$ to each dataset-specific module, and $x$ and $y$ denote the input and output, respectively.

$$\begin{aligned} p &= \text{router}(f_\theta(x)) \\ i &= \text{argmax}(p) \\ y &= f_{(\theta,m_i)}(x) \end{aligned} \tag{1}$$

We identify suboptimal aspects in the formulation of Eqn. 1, notably (i) potential errors introduced by the router's incorrect probability predictions leading to the erroneous selection of a dataset-specific module $m_i$ and (ii) the lack of knowledge-sharing among modules $m_1, ..., m_T$.

Next, in Eqn. 2, we propose a more general formulation that replaces the argmax operation with a composer function. Rather than selecting a single module corresponding to the highest router probability, the composer aggregates multiple dataset-specific modules into one, offering robustness to imperfect router predictions and enabling knowledge sharing among dataset-specific modules.

$$\begin{aligned} p &= \text{router}(f_\theta(x)) \\ \boldsymbol{m'} &= \textbf{composer}(\boldsymbol{M}, \boldsymbol{p}) \\ y &= f_{(\theta,m')}(x) \end{aligned} \tag{2}$$

In the next section, we describe how we instantiate our above-described general formulation with specific modeling components.

## 3.2 DAM: DYNAMIC ADAPTER MERGING

Based on the formulation in Eqn. 2, we propose DAM , a framework that can learn to model sequentially streaming data $\mathbb{D}_t$ with little forgetting and minimal computational overhead. As shown in Fig. 2, our method consists of four main components: (i) a frozen pretrained video-language backbone $f_\theta$, (ii) dataset-specific modules $m_1, ..., m_T$ implemented as adapters (Houlsby et al., 2019) for each training dataset, (iii) a non-parametric video-language router that predicts probabilities for selecting the most relevant adapters for a given test-time VidQA input instance, and (iv) a dynamic adapter merging scheme as a composer to aggregate all the adapter modules. We now describe each component in more detail.

**Backbone.** Our backbone $f_\theta$ is a large-scale pretrained VidQA model (e.g. FrozenBiLM (Yang et al., 2022)), implemented with Transformers (e.g. CLIP ViT-L/14 (Radford et al., 2021) and DeBERTa-V2-XL (He et al., 2020)). In practice, our framework can be applied to arbitrary backbones as shown in Sec. 4.3.

**Dataset-Specific Modules.** We propose implementing our dataset-specific modules $m_1, ..., m_T$ using adapters (Houlsby et al., 2019). These adapter modules are then used to learn from sequentially streaming data $\mathbb{D}_t$, introduced to the model at time $t$. Compared to visual prompts, commonly used in DIL methods (Smith et al., 2023a; Wang et al., 2022b;e), adapters have several important advantages. First, the larger modeling capacity of adapters is beneficial as it enables us to effectively capture distribution-specific information from each dataset $\mathbb{D}_t$. Furthermore, the high parameter efficiency of adapters (e.g., $\sim 3\%$ of total parameters in a pretrained backbone) allows us to efficiently train our framework every time new data $\mathbb{D}_t$ is received.

Specifically, we use an adapter $A_t = \{A_t^{(\ell)}\}_{\ell=1}^{L}$ consisting of $L$ adapter layers for each sequentially streaming dataset $\mathbb{D}_t$. We use the standard adapter structure as in (Houlsby et al., 2019; Yang et al.,

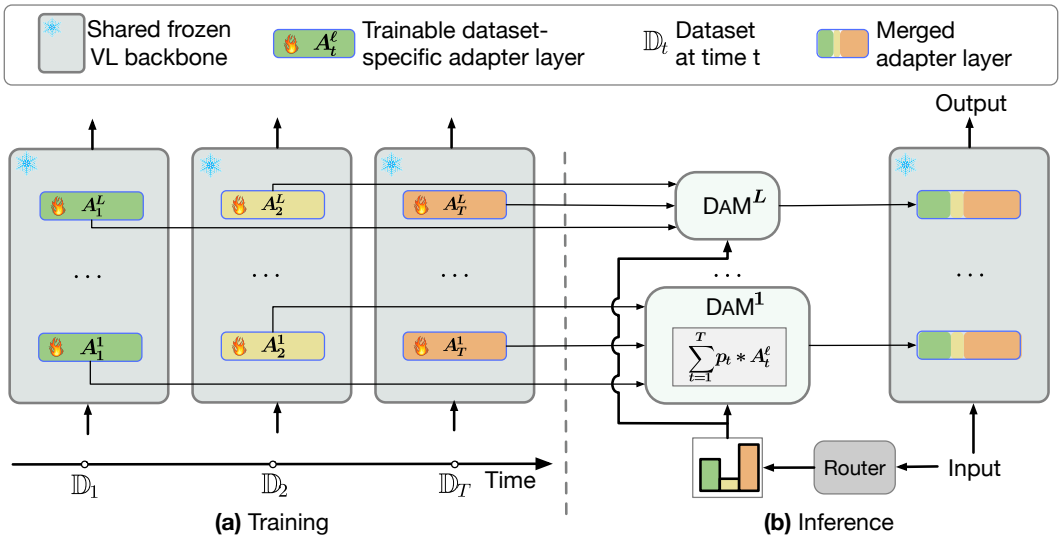


Figure 2: An overview of our Dynamic Adapter Merging (DAM) framework. **(a)** Our model is continually trained on sequentially arriving datasets $\{\mathbb{D}_1, ..., \mathbb{D}_T\}$. During training on dataset $\mathbb{D}_t$, we only train the adapter $A_t = \{A_t^{(\ell)}\}_{\ell=1}^{L}$ while keeping previously learned adapters fixed. **(b)** During inference, given a test sample (a video and a text question), we use the proposed router to predict the probability of each adapter being relevant to that particular input instance. Afterward, we dynamically merge multiple dataset-specific adapters in parameter space to reduce the impact of incorrect router predictions and leverage cross-domain VidQA cues. Finally, the pretrained backbone, together with the merged adapter, is used to make the final VidQA predictions.

2022), and insert an adapter layer $A_t^{(\ell)}$ after each self-attention and feed-forward network layer in our pretrained backbone. During training on dataset $\mathbb{D}_t$, we only train an adapter $A_t$ while keeping previously learned adapters fixed. This allows each adapter to focus on a single dataset, which is advantageous for maximizing dataset-specific performance while alleviating catastrophic forgetting. Additionally, to inherit previously acquired knowledge, we initialize $A_t$ with the weights of adapter $A_{t-1}$ trained in the preceding time step $t-1$, which we denote as *continual initialization*.

**Router.** To handle unknown dataset identity during inference, we employ a non-parametric router to predict the probability for each adapter, estimating their relevance to a given video-question input instance from an unknown distribution. Specifically, during training, we calculate the centroid $c_t = \frac{1}{N_t^s} \sum_{i=1}^{N_s} f_\theta(x_{t,i}^s) \in \mathbb{R}^d$ of each dataset $\mathbb{D}_t$ by averaging all multimodal video-language features extracted by the frozen pretrained backbone $f_\theta$ without adapters. Then, for a test sample $x$ during inference, we calculate adapter-specific probabilities $p_t = \frac{\exp(l_t/\tau)}{\sum_{i=1}^{T} \exp(l_i/\tau)}$. Here, $l_t = \cos(f_\theta(x), c_t)$ is the cosine similarity between a feature $f_\theta(x)$ and a centroid $c_t$, $T$ is the total number of datasets up to the current point, and $\tau$ is the temperature hyperparameter. We find our simple non-parametric router is more effective and computationally cheaper than other more complex design choices, including the ones used in prior DIL methods (Smith et al., 2023a; Wang et al., 2022e), as we will show in Sec. 5.2.

**Composer.** To improve our DIL framework's robustness to incorrect router predictions and enable knowledge-sharing across dataset-specific modules, we implement our composer function drawing the ideas from the model merging research community (Jin et al., 2022; Ainsworth et al., 2022; Yadav et al., 2023). In particular, recent model merging techniques (Wortsman et al., 2022a; Jin et al., 2022) have demonstrated the effectiveness of merging multiple domain models in the parameter space into a single model that generalizes to all the merged domains, thus, effectively eliminating the need for dataset identity predictions and naturally leveraging knowledge-sharing. However, a single fixed model may lack the expressiveness required to handle numerous domains, as observed in (Yadav et al., 2023), where the performance of the merged model drops significantly (e.g., $> 10\%$) when the number of domains is large (e.g., 8 domains).

Motivated by these considerations, we implement our composer using our proposed Dynamic Adapter Merging (DAM) scheme that dynamically merge multiple dataset-specific adapters for each test-time input instance (Figure 2b). Our composer is implemented through a simple instance-wise adapter weight merging using soft router-predicted probabilities. Note that all dataset-specific adapters share the same architecture, enabling element-wise merging of all adapters in their parameter space. Specifically, given adapter weights for all $T$ observed datasets $\mathcal{A} = \{A_1, \ldots, A_T\}$, and input-specific router probabilities $p \in \mathbb{R}^T$, the merged adapter weights $A_M = \sum_{t=1}^{T} p_t \cdot A_t$ are incorporated with the pretrained backbone for the final VidQA prediction.

Our dynamic adapter merging scheme is advantageous since it enhances robustness to incorrect dataset identity predictions. In particular, even when the router function produces partially incorrect dataset-identity probability predictions for the adapter selection, our dynamic merging scheme incorporates knowledge from multiple adapters, often including those associated with the correct domain. Additionally, such a dynamic adapter merging scheme facilitates knowledge sharing between dataset-specific adapters through parameter-space merging, proving beneficial when multiple datasets stem from similar domains. Unlike existing model merging techniques (Wortsman et al., 2022a; Jin et al., 2022), which produce a single fixed model for all test samples, our method produces a model that is uniquely tailored for each test sample, thus, offering greater modeling expressivity.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets and Metrics.** We perform experiments on 9 Video Question Answering (VidQA) datasets, which include iVQA (Yang et al., 2021), MSVD-QA (Xu et al., 2017), MSRVTT-QA (Xu et al., 2017), LSMDC (Maharaj et al., 2017), ActivityNet-QA (Yu et al., 2019), TGIF-QA (Jang et al., 2017), TrafficQA (Xu et al., 2021), EnvQA (Gao et al., 2021) and AGQA (Grunde-McLaughlin et al., 2021). MSVD-QA, MSRVTT-QA, and ActivityNet-QA involve social media videos, with ActivityNet-QA featuring notably longer videos (i.e., on average 2 minutes in length versus 30 second average duration of the videos in the first two datasets). iVQA, LSMDC, TGIF-QA, TrafficQA, EnvQA, and AGQA represent instructional, movie, short-GIF, traffic, virtual, and indoor human videos, respectively. We train the models on these sequentially arriving datasets. After training the model on the last dataset, we evaluate the resulting checkpoint on the test set of all trained datasets. During the evaluation, the dataset identity of each testing sample is assumed to be unknown. Following (Wang et al., 2022c;b), we use the **average accuracy** and **forgetting** as the evaluation metrics. See Appendix B for formal definitions.

**DIL Baselines.** For all of our continual learning baselines (including our approach), we use the recent FrozenBiLM (Yang et al., 2022) VidQA model, implemented using CLIP ViT-L/14 (Radford et al., 2021) and DeBERTa-V2-XL (He et al., 2020) video and language backbones and containing 1.2B parameters in total. In our comparisons, we include three recent Prompt-based methods *L2P* (Wang et al., 2022e), *CODA-Prompt* (Smith et al., 2023a), and *S-Prompts* (Wang et al., 2022b), and two regularization-based methods, *EwC* (Kirkpatrick et al., 2017) and *LwF* (Li & Hoiem, 2017). We also incorporate several naive baselines: (i) *Zero-Shot*, which directly evaluates the pretrained model on all datasets without any training, and (ii) *Seq-FT*, which sequentially finetunes the model on the sequentially arriving datasets. The upper bound performance for the adapter and prompt-based variants is obtained by multitask finetuning jointly on all training datasets.

**Model Merging Baselines.** We also compare with two model merging methods, Average Merging (Wortsman et al., 2022a; Ilharco et al., 2022a) and RegMean (Jin et al., 2022). In our implementation, both methods merge all the dataset-specific adapters into a single fixed adapter, which is then applied to all the test samples. This is in contrast to our framework, which produces a uniquely tailored adapter module for each test sample.

We refer readers to Appendix A for detailed implementations of our framework and all the baseline methods.

Table 1: Comparison with state-of-the-art on Domain-Incremental VidQA Learning. We obtain the upper-bound performance by multitask finetuning jointly on all the datasets. The zero-shot baseline directly evaluates the pretrained model on all datasets without any training, while the Seq-FT baseline sequentially finetunes a single model on all the datasets. We also reimplement prior methods (EwC, LwF, L2P, CODA-Prompt, S-Prompts) using our strong video-language backbone, as these methods were not initially designed for VidQA. All continual learning methods are trained sequentially from left to right in the table. Our proposed DAM outperforms the current state-of-the-art by 9.1% while exhibiting 1.9% less forgetting.

| Training Order: iVQA → MSVD → MSR-VTT → LSMDC → ANet → TGIF | | | | | | | |
|---|---|---|---|---|---|---|---|
| Method | Downstream VidQA Accuracy (Forgetting) (%) | | | | | | |
| | iVQA | MSVD | MSR-VTT | LSMDC | ANet | TGIF | Avg. |
| Zero-Shot | 26.8 | 33.0 | 15.0 | 51.5 | 25.5 | 41.9 | 32.3 |
| Seq-FT | 28.4 | 36.0 | 23.7 | 52.1 | 31.2 | 67.6 | 39.8 |
| *Multitask Finetuned (Upper-Bounds)* | | | | | | | |
| Adapters | 39.7 | 56.6 | 46.7 | 62.9 | 42.2 | 67.8 | 52.6 |
| Prompt Tuning | 35.0 | 49.0 | 37.1 | 57.4 | 33.9 | 59.2 | 45.3 |
| *Regularization-based methods* | | | | | | | |
| EwC | 29.9 | 39.3 | 25.5 | 54.9 | 32.4 | 67.5 | 41.6 (-10.9) |
| LwF | 28.3 | 38.2 | 25.8 | 56.4 | 33.6 | **68.5** | 41.8 (-10.7) |
| *Prompt-based methods* | | | | | | | |
| L2P | 32.8 | 43.3 | 32.1 | 54.8 | 27.2 | 54.4 | 40.8 (-4.6) |
| CODA-Prompt | 32.9 | 44.8 | 28.7 | 50.7 | 23.9 | 54.7 | 39.6 (-5.7) |
| S-Prompts | 31.8 | 45.5 | 30.2 | 54.9 | 27.9 | 56.1 | 41.1 (-4.2) |
| DAM (Ours) | **39.1** | **53.6** | **42.2** | **63.0** | **36.3** | 66.8 | **50.2** (**-2.3**) |

### 4.2 COMPARISON WITH STATE-OF-THE-ART

**Comparison with Domain-Incremental Learning (DIL) Methods.** Tab. 1 compares our method and state-of-the-art DIL approaches. Our findings demonstrate that our proposed DAM scheme outperforms the leading DIL method, S-Prompts, by a substantial margin of **9.1%** in average accuracy while also exhibiting **1.9%** less forgetting. Among prompt-based methods, L2P, CODA-Prompt, and S-Prompts show reduced forgetting compared to regularization-based methods EwC and LwF. However, these prompt-based methods achieve lower overall accuracy, which can be attributed to their smaller learning capacity. Overall, these results show the effectiveness of our proposed framework.

**Comparison with Model Merging Methods.** Next, we compare our method with two model merging methods, Average Merging (Wortsman et al., 2022a) and RegMean (Jin et al., 2022). For a fair comparison, all the methods merge the same set of adapters, each individually fine-tuned on each dataset without our continual initialization scheme. As shown in Tab. 2, DAM outperforms RegMean by **6.0%** and average merging by **7.5%** in average accuracy. We hypothesize that the significantly better performance of our model comes from the fact that DAM produces a custom model instance for each input instance. This makes our methods a lot more expressive than the existing model merging methods that use a single merged model instance for all data samples.

**Computational Complexity Analysis.** In addition to standard accuracy metrics, we also analyze the computational cost of our proposed approach. We note that each dataset-specific adapter in DAMcontributes merely **2.5%** of the pretrained model's parameters (CLIP-L/14 + DeBerTa-V2-XLarge), totaling 30M parameters. In terms of computational cost, merging adapter parameters incurs just **0.09** GFLOPs (30M *(2k-1), k=2 in our case), notably lower than the **162** GFLOPs required by CLIP-L/14 for a single image processing. Therefore, these results show that our proposed DAM can efficiently adapt to a reasonably large number of continually arriving datasets.

Table 2: Comparison with existing model merging techniques. For a fair comparison, all the methods merge the same set of adapters, each individually fine-tuned on each dataset without our continual initialization scheme (i.e., using random initialization). Our DAM outperforms existing model merging methods by a large margin on average.

| Method | iVQA | MSVD | MSR-VTT | LSMDC | ActivityNet | TGIF | Avg. |
|---|---|---|---|---|---|---|---|
| Multitask (upper-bound) | 39.7 | 56.6 | 46.7 | 62.9 | 42.2 | 67.8 | 52.6 |
| Avg. Merging | **38.0** | 45.7 | 27.7 | 54.5 | 27.0 | 56.6 | 41.6 |
| RegMean | 36.6 | 49.7 | 32.5 | 54.0 | 27.7 | 57.8 | 43.1 |
| DAM (Ours) | 36.5 | **51.6** | **39.5** | **63.0** | **36.5** | **67.7** | **49.1** |

Table 3: Domain-Incremental Learning on image classification. Our upper-bound is obtained by finetuning a shared adapter jointly on all domains. For a fair comparison, we de-emphasize S-Prompts with CLIP encoder since it is pretrained with much more data than the ImageNet-pretrained ViT-B/16 backbone used by our method.

| **Method** | **Backbone** | Buffer size | **Avg. Accuracy (%)** | |
|---|---|---|---|---|
| | | | CORe50 | DomainNet |
| Multitask (upper-bound) | ViT-B/16 | - | $94.59 \pm 0.21$ | 71.95 |
| DyTox | ViT-B/16 | 50/class | $79.21 \pm 0.10$ | 62.94 |
| LwF | ViT-B/16 | 0/class | $75.45 \pm 0.40$ | 49.19 |
| L2P | ViT-B/16 | | $78.33 \pm 0.06$ | 40.15 |
| S-Prompts | ViT-B/16 | | $83.13 \pm 0.51$ | 50.62 |
| S-Prompts | CLIP ViT-B/16 | | $89.06 \pm 0.86$ | 67.78 |
| DAM(Ours) | ViT-B/16 | 0/class | $\mathbf{92.45 \pm 0.25}$ | **69.23** |

### 4.3 GENERALIZATION TO IMAGES

To further showcase the generalizability of our approach, we extend DAM to two image tasks: 1) image classification and 2) image question-answering.

**Image classification.** We conduct experiments on two standard DIL benchmarks: CORe50 (Lomonaco & Maltoni, 2017) and DomainNet (Peng et al., 2019). CORe50 (Lomonaco & Maltoni, 2017) contains 50 categories across 11 domains. Following prior work, we continually train the model on 8 domains (120K images) and evaluate the trained model on the remaining 3 domains (40K images). DomainNet contains 345 categories across 6 domains. The DIL setup on DomainNet is the same as Wang et al. (2022b); Fini et al. (2022). Following standard evaluation protocol, we use ViT-B/16 (Dosovitskiy et al., 2020) pretrained on ImageNet as our backbone. As shown in Tab. 3, DAM surpasses the current state-of-the-art S-Prompts by **9.32%** and **18.61%** using the same ImageNet-pretrained ViT-B/16 backbone on CORe50 and DomainNet, respectively. These results suggest that our model can also be effectively applied to DIL image classification tasks.

**Image question-answering.** Next, we also extend our model to the visual question-answering (VQA) task on images. We integrate our proposed DAM and the best performing prompt-based baseline S-Prompts with the state-of-the-art VQA model, BLIP-2 (Li et al., 2023a), which uses CLIP ViT-G/14 (Radford et al., 2021) and FlanT5-XL (Chung et al., 2022) as its vision-language backbone and has 4.1B parameters in total. We then continually train both models on 4 mainstream VQA datasets: OK-VQA (Marino et al., 2019), aOK-VQA (Schwenk et al., 2022), GQA (Hudson & Manning, 2019) and VQAv2 (Goyal et al., 2017). The results are shown in Tab. 4. Our proposed DAM outperforms S-Prompts by **4.4%** with **1.2%** less forgetting, thus, demonstrating the generality of our approach.

Table 4: We extend our proposed DAM method to continual visual question-answering (VQA) task on image datasets. For these experiments, we use the recent BLIP-2 model (Li et al., 2023b) as our visual-language backbone. The proposed DAM outperforms the existing state-of-the-art method (S-Prompts) by 4.4% in average accuracy while exhibiting 1.2% less forgetting.

| Method | OK-VQA (test) | aOK-VQA (val) | GQA (val) | VQAv2 (val) | Avg. |
|---|---|---|---|---|---|
| Zero-Shot | 40.7 | 35.7 | 44.0 | 63.1 | 45.9 |
| Multitask (upper-bound) | 49.2 | 51.8 | 58.7 | 76.2 | 58.8 |
| S-Prompts | 42.9 (-5.3) | 46.1 (-2.2) | 47.3 (-7.1) | 65.3 (-6.0) | 50.4 (-5.2) |
| DAM | **45.1** (-4.1) | **50.4** (-1.4) | **54.1** (-4.6) | **69.8** (-6.4) | **54.8** (-4.0) |

Table 5: We investigate the number of dataset-specific adapters to merge for best performance. The Top-$K$ adapters are selected according to the highest router predicted probabilities. The first 4 rows depict the downstream VidQA accuracy, whereas the last row is the router accuracy. We highlight the largest accuracy gap between adapter merging and non-merging variants. Merging adapters is typically useful when the router makes many incorrect predictions.

| Top-$K$ | MSVD | MSR-VTT | ActivityNet | iVQA | TGIF | LSMDC |
|---|---|---|---|---|---|---|
| 1 (no-merging) | 49.0 | 40.4 | 37.4 | 37.5 | 66.3 | 62.9 |
| 2 | 53.6 | 42.2 | **36.3 (-1.1)** | 39.1 | 66.8 | 63.0 |
| 3 | 54.6 | **42.4(+2.0)** | 34.0 | 39.3 | **67.0(+0.7)** | 63.0 |
| 6 (merge all) | **54.9(+5.9)** | 41.9 | 33.0 | **39.6(+2.1)** | 66.9 | **63.1(+0.2)** |
| Router Acc | 51.0 | 69.6 | 76.4 | 81.6 | 96.1 | 100 |

# 5 ANALYSIS

## 5.1 ADAPTER MERGING ANALYSIS

In this section, we analyze the effectiveness of dynamic adapter merging. Specifically, in Tab. 5, we present a comprehensive breakdown of downstream VidQA accuracy and the router's accuracy on each dataset, considering various adapter merging variants. The table highlights an intriguing trend: as the router's accuracy decreases, the benefits derived from adapter merging become more pronounced. Specifically, when the router's accuracy is at **51.0%** and **69.6%**, adapter merging yields substantial downstream accuracy improvements of **5.9%** and **2.0%** on the MSVD and MSR-VTT datasets, respectively. In contrast, when the router approaches near-perfect accuracy, the gains from adapter merging become less significant (as seen with a marginal **0.2%** improvement on LSMDC).

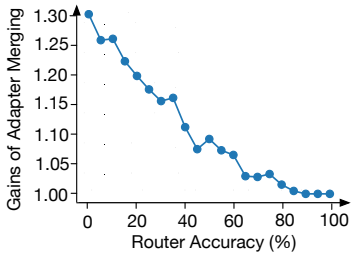


Figure 3: We study the normalized performance gain of dynamic adapter merging as a function of router accuracy. Our results show that dynamic adapter merging leads to a larger boost when the router is inaccurate.

To further validate this observation, Fig. 3 provides insights into the average performance gain of dynamic adapter merging over non-merging variants as a function of router accuracy. The data points are generated by creating a series of routers manually, each predicting dataset probabilities with a specified accuracy. The figure confirms the trend in Tab. 5, showcasing that adapter merging offers a **30%** relative improvement when the router's accuracy drops to **0%**.

Based on these results, we can conclude that our proposed DAM is particularly advantageous when dealing with many datasets. In such complex scenarios, dataset prediction becomes notably chal-

Table 6: We study the effectiveness of different router functions. Specifically, we incorporate router functions from several prior methods into our DAM method and measure our model's performance on the downstream VidQA task with each of these routers. Our results suggest that our non-parametric router function leads to the best downstream VidQA performance.

| Router | random | L2P's | CODA-Prompt's | S-Prompts' | GMM | Learnable MLP | Ours |
|---|---|---|---|---|---|---|---|
| router Acc. | 16.6 | 67.4 | - | 76.4 | 79.0 | 78.7 | **79.1** |
| VidQA Acc. | 40.2 | 48.6 | 45.3 | 49.7 | 49.4 | 48.9 | **50.2** |

Table 7: Domain-Incremental Learning (DIL) on 4 datasets from different domains. DAM has negligible forgetting rate on datasets with large domain gaps.

| Method | LSMDC | AGQA | Env-QA | TrafficQA($\frac{1}{2}$) | Avg. |
|---|---|---|---|---|---|
| Upper-Bound | 63.0 | 63.4 | 32.3 | 67.8 | 56.6 |
| DAM | 63.0 | 63.3 | 32.0 | 67.8 | 56.5 |
| Router Acc. of DAM | 100 | 99.9 | 99.2 | 99.7 | 99.7 |

lenging for the router. These collective findings underscore the practical significance and scalability of our proposed approach in real-world domain-incremental VidQA learning scenarios.

### 5.2 Router Analysis

In this section, we compare our router design with three router designs from prior DIL methods: L2P (Wang et al., 2022e), CODA-Prompt (Smith et al., 2023a), as well as Gaussian Mixture Model (GMM) and Learnable MLP. We incorporate these router functions into our DAM method and measure our model's performance on downstream VidQA task with each of these routers. We also measure the accuracy of each router function for correctly classifying the dataset/domain of a given VidQA input instance. Note that we cannot calculate CODA-Prompts' router's accuracy as it does not explicitly predict the domain identity. From Tab. 6, we observe that higher router accuracy typically leads to higher downstream VidQA accuracy, thus indicating the importance of an accurate router function. Second, we notice that jointly training router and domain-specific modules as was done in previous methods (L2P, CODA-Prompt) leads to worse downstream VidQA accuracy than disjoint training (S-Prompts, Ours). Lastly, our results suggest that despite the simplicity of our non-parametric router function, it produces the best performance.

### 5.3 Domain Analysis

In this section, we analyze the performance of our method through experiments on datasets with both large and small domain gaps.

**Large Domain Gap.** To validate the effectiveness of our framework on datasets with large domain/distribution gaps, we experiment with 4 datasets from 4 different domains: movie videos (LSMDC-QA), indoor human videos (AGQA), traffic videos (TrafficQA), and virtual videos (Env-QA). Tab. 7 presents DAM's vidQA accuracy and the router's domain identity prediction accuracy. We observe that DAM exhibits negligible forgetting on the 4 datasets. We attribute such good performance of our method to 1) dataset-specific adapters that can effectively specialize for modeling each dataset and 2) the high router's accuracy across most datasets in this setting. Consequently, these results indicate that our proposed DAM can be effectively applied to datasets with large domain gaps.

**Small Domain Gap.** Next, we evaluate our approach on datasets within the same domain but collected at different times. Such time-based distribution shifts are typically much smaller than for the previously considered datasets spanning entirely different domains (Tab. 7). Thus, such a setting necessitates knowledge sharing and the ability to handle incorrect router predictions. Specifically, we evaluate our model in this setting by dividing the iVQA dataset into 5 non-overlapping subsets based on the video upload date to YouTube. We continually train the model on these five subsets and

Table 8: We evaluate the ability of our framework to adapt to subtle time-distribution shifts. To do this, we divide the iVQA dataset into 5 subsets according to the video upload date to YouTube. We then train our model on these 5 sequentially arriving subsets. Our results indicate that our dynamic adapter merging scheme still works effectively, even when the dataset domains/characteristics are quite similar.

| Method | Multitask (upper-bound) | Router Acc. | DAM (no merging) | DAM (merging top-$k = 2$) | DAM (merging all) |
|---|---|---|---|---|---|
| VidQA Acc. | 39.8 | 43.8 | 37.1 | 38.4 | **39.3** |

Table 10: DAM benefits from the proposed continual initialization scheme.

| Method | iVQA | MSVD | MSR-VTT | LSMDC | ActivityNet | TGIF | Avg. |
|---|---|---|---|---|---|---|---|
| DAM | **39.1** | **53.6** | **42.2** | **63.0** | 36.3 | 66.8 | **50.2** |
| w/o continual initialization | 36.5 | 51.6 | 39.5 | **63.0** | **36.5** | **67.7** | 49.1 |

then evaluate on iVQA's original test set that spans 5 time distributions. Tab. 8 shows that unlike in the previous setting, in this case, the router attains an accuracy of only 43.8%. This can be explained by the fact that the dataset/domain-identity prediction problem becomes a lot more challenging due to minor distribution shifts between subsets. In this scenario, the model variant that merges all adapters surpasses the variant without merging by **2.2%** and experiences only 0.5% forgetting. This underscores the effectiveness of dynamic adapter merging and emphasizes the importance of knowledge sharing in settings where domains or datasets are similar.

### 5.4 Other Analyses

**Order of the Datasets.** In Tab. 9, we study how the order of the training datasets affects our model's performance. We randomly sample 5 different orders and train our framework on those orders. Based on the results, we observe that the performance of our approach is quite stable across all 5 orders (**50.56** $\pm$ **0.26%**). This indicates our method is robust to the order of training datasets.

**Continual Initialization Scheme.** In Section 3.2, we introduced a continual initialization scheme for initializing a current distribution-specific adapter using the weights of a previously learned adapter. In Tab. 10, we validate the effectiveness of this scheme and show that it leads to a notable **1.1%** average accuracy improvement. These improvements are particularly pronounced for the datasets that are used first, such as iVQA and MSVD. We posit that the benefits of continual initialization stem from the fact that the weights of continually learned adapters reside in a more similar parameter space. This phenomenon reduces interference disagreements when merging adapters (Yadav et al., 2023).

Table 9: Ablations on the order of datasets. We randomly sampled 5 orders and obtained average accuracies for each order. V: iVQA; D: MSVD; T: MSR-VTT; L: LSMDC; A: ActivityNet; G: TGIF.

| Domain Order | Avg. Acc (%) |
|---|---|
| V D T L A G | 50.2 |
| L T G D A V | 50.8 |
| V A D G T L | 50.4 |
| G T A V D L | 50.9 |
| V A G D T L | 50.5 |

## 6 Discussion and Conclusion

In this work, we investigate the challenging and relatively unexplored problem of rehearsal-free domain-incremental VidQA learning. Our proposed DAM framework outperforms existing state-of-the-art by **9.1%** with **1.9%** less forgetting on a benchmark with six distinct video domains. The proposed method DAM is simple and flexible, and we further extend it to image classification tasks and visual question-answering, demonstrating our method's generalization beyond video-level scenarios. Despite effective results, we also observe a few limitations of our proposed approach. Firstly, our approach employs a straightforward weighted averaging technique for merging adapter weights, leaving room for more advanced merging methods that could enhance knowledge sharing among

domains. Secondly, our validation encompasses a relatively small number of domains ($\leq$ 7 in our case), consistent with previous domain-incremental learning research. It would be valuable to assess the effectiveness of our method and existing domain-incremental learning methods across a more extensive domain spectrum, potentially involving a substantial number of domains (e.g., 100). We plan to explore these research directions in our future work.

**Acknowledgements.** We thank Md Mohaiminul Islam, Ce Zhang, Yue Yang, and Soumitri Chattopadhyay for helpful discussions. This work was supported by the Sony Faculty Innovation award, Laboratory for Analytic Sciences via NC State University, ONR Award N00014-23-1-2356, ARO Award W911NF2110220, DARPA ECOLE Program No. #HR00112390060, and NSF-AI Engage Institute DRL-2112635. The views contained in this article are those of the authors and not of the funding agency.

## APPENDIX

In this appendix, we present the following:

A. Implementation details.

B. Evaluation Metrics.

C. Extension to other types of continual learning.

D. Dataset descriptions.

## A IMPLEMENTATION DETAILS

**Details of our DAM approach.** Our choice for the VidQA model is FrozenBiLM (Yang et al., 2022), a state-of-the-art (SOTA) model in the VidQA domain. To align with this model, we utilize a vocabulary encompassing the 3635 most frequent answers. Adhering to the FrozenBiLM approach, we integrate adapters into each layer of the DeBERTa-XL (He et al., 2020) language model, employing a downsampling rate of 8. The loss function is the same as the original FrozenBiLM model, i.e., the cross-entropy loss between the predicted tokens and ground-truth answer tokens. For the initialization of dataset-specific adapters during the commencement of continual learning (first dataset), we use the weights from FrozenBiLM, which is pre-trained on a substantial dataset comprising 10 million video-text pairs (WebVid10M (Bain et al., 2021)). In the training of domain-specific adapters for each subsequent domain, we conduct 20 epochs of training with an initial learning rate of $5e-5$. The learning rate undergoes a linear warm-up for the first 2 epochs, followed by a linear decay to 0. Our proposed DAM introduces only two hyperparameters. Specifically, we set the temperature parameter ($\tau$) to 0.01 and merge top-$k$=2 adapters for the adapter merging process. We normalize the router's predicted probabilities by setting the sum of the top-$k$ probabilities to 1 and the remaining probabilities to 0.

**Network Structures:** Our frozen pretrained backbone is FrozenBiLM (Bain et al., 2021), comprising a language model DeBERTa-XL and a vision model CLIP-L/14. The input features to the router consist of the concatenation of the averaged hidden states from the 4th last layer of DeBERTa-XL and the averaged hidden states from the last layer of CLIP-L/14, without the incorporation of adapters. For each dataset, we employ an adapter comprising $L$ adapter layers, inserting an adapter layer after each self-attention layer and feed-forward network layer in DeBERTa-XL. Following (Yang et al., 2022; Houlsby et al., 2019; Yang et al., 2022), each adapter layer in our approach includes a downsampling and an upsampling linear layer, along with a residual connection. The linear layers are configured with an $8\times$ downsample scale to an intermediate hidden size, and the upsampler maps back to the original dimensionality.

**Continual Learning Baselines.** Since our work is the very first exploration of continual VidQA learning, we implement a number of continual learning baselines (focused on image classification) to VidQA task, including three recent Prompt-based methods L2P (Wang et al., 2022e), CODA-Prompt (Smith et al., 2023a), and S-Prompts (Wang et al., 2022b)and two regularization-based methods EwC (Kirkpatrick et al., 2017) and LwF (Li & Hoiem, 2017). For a fair comparison, we use the same pretrained model and preserve most hyper-parameter settings with our approach.

- **L2P** (Wang et al., 2022e). For the prompt settings, we set the prompt length to 10 and the size of the prompt pool to 6. The dimension of the prompt key is configured to be 3072, matching the dimension of the router input feature in our method. The prompt dimension is set to 1536, aligning with the input dimension of the frozen language model. We sweep the learning rate between $1e-2$ and $1e-5$ with an interval of $3.33\times$. The best performance is achieved with an initial learning rate 3e-3.
- **CODA-Prompt** (Smith et al., 2023a). For a fair comparison, we adopt the same prompt settings as our L2P baseline for CODA-Prompt. Following (Smith et al., 2023a), we apply orthogonality initialization to initialize the prompts, their keys, and their attention matrices. The dimension of prompt attention is set to 3072, consistent with the dimension of the prompt key. For optimal performance, we configure the learning rate to 1e-3.
- **S-Prompts** (Wang et al., 2022b). We use exactly the same prompt settings as in our implementation for L2P. For their K-Means router, we set $K = 3$ as the number of centroids for each domain and 1-first-nearest neighbor with the centroids to search for the best prompts.
- **EwC** (Kirkpatrick et al., 2017) and **LwF** (Li & Hoiem, 2017). We follow their original implementations, except that the regularization is only applied to adapters as all the other parameters are frozen. A single adapter is shared for all the domains.

## B Evaluation Metrics

Following (Wang et al., 2022b;d;e; Smith et al., 2023a), we employ standard evaluation metrics, including **average accuracy** and **forgetting**. The average accuracy metric evaluates both learning capacity and catastrophic forgetting, whereas the forgetting metric specifically addresses catastrophic forgetting. As an illustration, the pretrained zero-shot model attains 0% forgetting but may exhibit relatively low average accuracy.

Formally, let $S_{t,\tau}$ denote the accuracy on the $\tau$-th dataset after training on the $t$-th dataset (task). After the training on the $t$-th dataset, we compute the average accuracy $A_t$ and forgetting $F_t$ as follows:

$$A_t = \frac{1}{t}\sum_{\tau=1}^{t} S_{t,\tau} \tag{3}$$

$$F_t = \frac{1}{t}\sum_{\tau=1}^{t} \max_{\tau' \in \{1,\dots,t\}} (S_{\tau',\tau} - S_{t,\tau}) \tag{4}$$

Upon completion of training on all $T$ datasets, we report the final average accuracy $A_T$ and forgetting $F_T$.

## C Extension to Other Types of Continual Learning

To show the flexibility of our framework, we also extend DAM to two other types of continual learning: 1) Class-Incremental Learning (CIL) and 2) Task-Incremental Learning (TIL) on VidQA.

**CIL.** In standard CIL, there are no overlapping classes between tasks, and the training of each split or dataset is treated as a separate task. To simulate CIL, we treat each unique answer as a class, similar to the protocol commonly used in continual learning for image classification (Wang et al., 2022e). We conduct experiments in two distinct settings: 1) *MSRVTT-QA 10 subsets* , where the classes do not overlap between subsets, and 2) *4-Datasets* (iVQA, MSVD, LSMDC, ActivityNet), excluding samples with answers that overlap across datasets. In the first setting, the model is continually trained on these disjoint subsets of the data, while in the second setting, we train our model on the 4 continually arriving datasets. The results, presented in Tab. 11, show that DAM consistently outperforms S-Prompts (Wang et al., 2022b), achieving **18.2%** and **8.5%** improvement on average accuracy on MSRVTT-QA 10-tasks and 4-Datasets respectively.

Table 11: Class-Incremental Learning (CIL) experiments are conducted under two settings: continually training our model on 1) 10 data subsets of MSR-VTT without overlapping classes (answers), and 2) 4 sequentially arriving datasets (iVQA, MSVD, LSMDC, ActivityNet) that do not have any overlap between their classes (answers). The proposed DAM outperforms S-Prompts by a large margin in both settings.

| Method | MSRVTT-QA 10 subsets | | 4-Datasets | |
|---|---|---|---|---|
| | Average Acc. | Forgetting | Average Acc. | Forgetting |
| Multitask (upper-bound) | 47.3 | - | 51.6 | - |
| S-Prompts | 15.4 | -23.5 | 42.2 | -3.3 |
| DAM (Ours) | **33.6** | **-13.7** | **50.7** | **-0.9** |

Table 12: Application of our model to Task-Incremental Learning (TIL). Our proposed framework generalizes well to TIL achieving only $0.1\%$ lower accuracy than the upper-bound multitask learning baseline.

| Method | iVQA | MSVD | MSR-VTT | LSMDC | ActivityNet | TGIF | Avg. |
|---|---|---|---|---|---|---|---|
| Multitask (upper-bound) | 39.7 | **56.6** | 46.7 | 62.9 | 42.2 | 67.8 | **52.6** |
| DAM | **39.8** | 54.8 | **46.7** | **63.0** | **42.4** | **68.0** | 52.5 |

**TIL.** To extend our framework to TIL, we treat the training on each dataset as a task. Unlike DIL or CIL, in TIL, each test sample during inference is provided with a dataset identity. As shown in Tab. 12, DAM obtains only 0.1% lower accuracy than the upper-bound multitask learning baseline. This is because in this setting, DAM can always use the correct dataset-specific adapters, which are individually trained on their corresponding datasets and perform comparable to multitask training.

## D DATASET DESCRIPTIONS

**Video Question Answering(VidQA).** We perform experiments on 9 Video Question Answering (VidQA) datasets, which include iVQA (Yang et al., 2021), MSVD-QA (Xu et al., 2017), MSRVTT-QA (Xu et al., 2017), LSMDC (Maharaj et al., 2017), ActivityNet-QA (Yu et al., 2019), TGIF-QA (Jang et al., 2017), TrafficQA (Xu et al., 2021), EnvQA (Gao et al., 2021) and AGQA (Grunde-McLaughlin et al., 2021). MSVD-QA, MSRVTT-QA, and ActivityNet-QA involve social media videos, with ActivityNet-QA featuring notably longer videos (i.e., on average 2 minutes in length versus 30 second average duration of the videos in the first two datasets). iVQA, LSMDC, TGIF-QA, TrafficQA, EnvQA, and AGQA represent instructional, movie, short-GIF, traffic, virtual, and indoor human videos, respectively.

- **iVQA** (Yang et al., 2021) is an open-ended VidQA dataset with reduced language biases and high-quality redundant manual annotations. It contains 10K video clips and 10K questions, split into 6K/2K/2K for training/validation/testing.
- **MSVD-QA** (Xu et al., 2017) is an open-ended VidQA dataset based on Microsoft Research Video Description Corpus (Chen & Dolan, 2011). It contains 1.8K video clips and 51K question-answer pairs, split into 32K/6K/13K for training/validation/testing.
- **MSRVTT-QA** (Xu et al., 2017) is an open-ended VidQA dataset based on MSR-VTT dataset (Xu et al., 2016). It contains 10K video clips and 243K question-answer pairs, split into 158K/12K/73K for training/validation/testing.
- **ActivityNet-QA** (Yu et al., 2019) is an open-ended VidQA dataset based on long videos (Caba Heilbron et al., 2015) (averaging 180 seconds) and human annotation. It contains 5.8K video clips and 58K question-answer pairs, split into 32K/18K/8K for training/ validation/ testing.
- **TGIF-QA** (Jang et al., 2017) is an open-ended VidQA dataset based on the Tumblr GIF (TGIF) dataset (Li et al., 2016). It contains 46K GIFs and 53K question-answer pairs, split into 39K/13K for training/testing.

- **LSMDC-FiB** (Maharaj et al., 2017) is an open-ended video-conditioned fill-in-the-blank task that consists of predicting masked words in sentences that describe short movie clips (Rohrbach et al., 2015). It contains 119K video clips and 349K question-answer pairs, split into 297K/22K/30K for training/validation/testing.
- **Traffic-QA** (Xu et al., 2021) is a dataset designed for video QA, comprising 10,080 in-the-wild videos and annotated with 62,535 QA pairs. It serves as a benchmark for assessing the cognitive capability of causal inference and event understanding models in complex traffic scenarios. Our experiments focus on *setting-1/2*, where the model receives a question-answer pair as input and is tasked with predicting the correctness of the answer (yes or no).
- **Env-QA** (Gao et al., 2021) is a new video QA dataset to evaluate the ability of understanding the composition, layout, and state changes of the environment presented by the events in videos. It contains 23.3K videos collected in AI2-THOR simulator and 85.1K questions.
- **AGQA** (Grunde-McLaughlin et al., 2021) is a benchmark for compositional spatio-temporal reasoning. AGQA contains 192M unbalanced question answer pairs for 9.6K videos. We experiment on AGQA-v2 that contains a balanced subset of 2.27M question answer pairs to mitigate language bias.

**Visual Question Answering(VQA).** Follow (Li et al., 2023b), we evaluate our model on 4 mainstream VQA datasets: OK-VQA (Marino et al., 2019), aOK-VQA (Schwenk et al., 2022), GQA (Hudson & Manning, 2019) and VQAv2 (Goyal et al., 2017).

- **OK-VQA** (Marino et al., 2019) is a knowledge-based visual question-answering benchmark with 14k images and 14k questions.
- **aOK-VQA** (Schwenk et al., 2022) is an augmented successor of OK-VQA (Marino et al., 2019) and contains a diverse set of 25K questions requiring a broad base of commonsense and world knowledge to answer.
- **GQA** (Hudson & Manning, 2019) is a large-scale visual question-answering dataset with real images from the Visual Genome (Krishna et al., 2017) dataset and balanced question-answer pairs.
- **VQAv2** (Goyal et al., 2017) consists of 1.1M questions about COCO images (Chen et al., 2015) each with 10 answers. It is the balanced version of the original VQA (Antol et al., 2015) dataset.

## REFERENCES

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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*, pp. 1728–1738, 2021.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 9516–9525, 2021a.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021b.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10739–10750, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9285–9295, 2022.

Beyza Ermis, Giovanni Zappella, Martin Wistuba, Aditya Rawal, and Cedric Archambeau. Memory efficient continual learning with transformers. *Advances in Neural Information Processing Systems*, 35:10629–10642, 2022.

Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1675–1685, 2021.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. *arXiv preprint arXiv:2303.10070*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Fidel A. Guerrero-Peña, Heitor R. Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20237–20246, 2022. URL https://api.semanticscholar.org/CorpusID:255096607.

Vipul Gupta, Santiago Akle Serrano, and Dennis DeCoste. Stochastic weight averaging in parallel: Large-batch training that generalizes well. *arXiv preprint arXiv:2001.02312*, 2020.

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*, pp. 770–778, 2016.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022a.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022b.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

Dahuin Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11847–11857, October 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7331–7341, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023b.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23119–23129, 2023c.

Yujie Li, Xin Yang, Hao Wang, Xiangkun Wang, and Tianrui Li. Learning to prompt knowledge transfer for open-world continual learning, 2023d.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4641–4650, 2016.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, pp. 17–26. PMLR, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6884–6893, 2017.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023a.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023b.

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*, pp. 8748–8763. PMLR, 2021.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9374–9384, 2021.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023a.

James Seale Smith, Junjiao Tian, Shaunak Halbe, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2409–2419, 2023b.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35:29440–29453, 2022.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7464–7473, 2019.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022a.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu. Towards a general framework for continual learning with pre-training, 2023a.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023b.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning, 2023c.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022b.

Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, and Dacheng Tao. Continual learning with lifelong vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 171–181, 2022c.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022d.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022e.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2816–2827, October 2023d.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022b.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pp. 39–58. Springer, 2022.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9878–9888, 2021.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1686–1697, 2021.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023.

Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34:26462–26474, 2021.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

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.

Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19102–19112, 2023.

Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering. *arXiv preprint arXiv:2211.10567*, 2022.

Yukun Zuo, Hantao Yao, Lu Yu, Liansheng Zhuang, and Changsheng Xu. Hierarchical prompts for rehearsal-free continual learning, 2024.