

Learning Conditional Space-Time Prompt Distributions for Video Class-Incremental Learning

Xiaohan Zou*, Wenchao Ma, Shu Zhao
The Pennsylvania State University
{xfz5266, wmm5390, smz5505}@psu.edu

Abstract

Recent advancements in prompt-based learning have significantly advanced image and video class-incremental learning. However, the prompts learned by these methods often fail to capture the diverse and informative characteristics of videos, and struggle to generalize effectively to future tasks and classes. To address these challenges, this paper proposes modeling the distribution of space-time prompts conditioned on the input video using a diffusion model. This generative approach allows the proposed model to naturally handle the diverse characteristics of videos, leading to more robust prompt learning and enhanced generalization capabilities. Additionally, we develop a simple yet effective mechanism to transfer the token relationship modeling capabilities of pre-trained image transformers to spatio-temporal modeling in videos. Our approach has been thoroughly evaluated across four established benchmarks, showing remarkable improvements over existing state-of-the-art methods in video class-incremental learning. Code is available at <https://github.com/Renovamen/CoSTEP>.

1. Introduction

Recent studies [2, 33, 37, 45] have significantly advanced the performance of video understanding. Despite these advancements, integrating new data streams into existing models remains challenging, particularly when these streams introduce significant distribution shifts. This issue mainly stems from models overfitting to recent training data, leading to the forgetting of previously seen classes, which is named catastrophic forgetting [1, 27, 76]. While retraining models with a combined dataset of old and new data streams is a straightforward solution, it faces practical limitations such as memory constraints and privacy issues [48, 60, 71], which restrict access to comprehensive historical data. To address these limitations, Video Class-

Incremental Learning (VCIL) [44, 46, 47, 60, 61] aims to enable models to adapt to new tasks while retaining previous knowledge without using task identifiers during testing, thereby mitigating catastrophic forgetting.

In Class-Incremental Learning (CIL), a common approach is the use of a rehearsal buffer [3, 46, 49], which stores selected samples from past data for use with future tasks. However, its efficacy heavily relies on the buffer size, with smaller buffers causing significant performance drops, and explicitly storing past data raises privacy concerns. To overcome these limitations, recent studies in both image [65–67] and video domains [47, 61] have investigated encoding prior knowledge into pools of learnable prompts. These methods use a key-query mechanism that dynamically selects an appropriate prompt from the pool based on the input features, as shown in Figure 1 (a). While promising, the pool-based paradigm faces notable challenges. Firstly, as the number of tasks grows, it becomes necessary to expand the prompt pool to retain previous knowledge, which requires more memory [22]. Moreover, it impairs the model’s adaptability to new tasks, an issue that persists even with an enlarged prompt pool [54]. This issue arises from prompts being vastly outnumbered by data instances. As a result, each prompt must broadly represent a group of samples, restricting its ability to address individual instances [22].

In response, as depicted in Figure 1 (b), several studies in the image domain have shifted from a fixed-sized prompt pool to using a learnable prompt generator [22, 54, 57]. These approaches generate a prompt for each instance, enhancing the expressiveness of the prompts and improving plasticity. However, they remain ineffective for video tasks due to the spatio-temporal complexity of videos, necessitating a new perspective to enhance learning capacity. These approaches, aligned with *deterministic* prompt learning [73, 74], often lead to suboptimal diversity and generalizability of prompts [7, 38, 75], issues that are more severe in the video domain. They struggle to produce prompts that are diverse enough to handle the wide variability in video representations, as videos can differ significantly in

*Corresponding author.

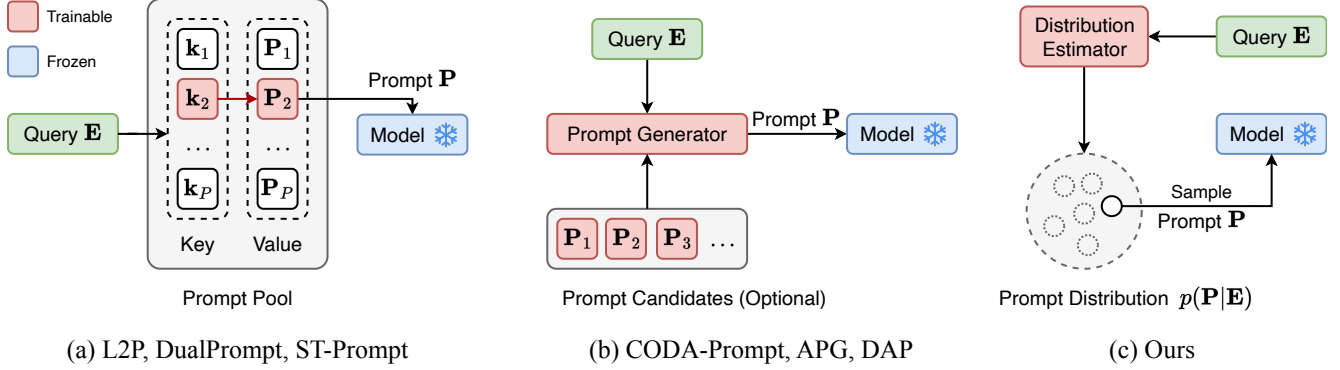


Figure 1. Previous prompt-based CIL approaches typically operate in a deterministic manner. They either (a) learn a pool of key-value pairs from which learnable prompts are selected or (b) generate prompts using a learnable prompt generator. Our approach (c) models the **distribution $p(P|E)$ of prompts conditioned on input features E** during training and samples a prompt P for each input from this distribution during inference.

both spatial and temporal aspects even within the same category. Additionally, the prompt learner tends to overfit to current tasks, which reduces its ability to transfer knowledge to new, unseen classes in future tasks. This is particularly problematic in VCIL, where learning video classification from scratch poses a greater challenge than with images, making knowledge transfer especially crucial.

Moreover, video tasks often require capturing critical spatio-temporal dynamics within specific patches across frames against static backgrounds. However, prevalent approaches [46, 47, 60, 61] often rely on operations applied directly to global feature vectors of sampled frames to generate video representations. These methods overlook the relationships between local patches across frames, potentially limiting their ability to fully capture detailed video features [77]. This prompts us to explore a more effective way to generate suitable video representations.

To this end, this paper proposes learning a **COnditional Space-TimE Prompt** distribution (CoSTEP) for VCIL, as depicted in Figure 1 (c). CoSTEP innovates by directly modeling the complex distribution of space-time prompts, conditioned on the input video during training, using a diffusion-based generative approach [11, 18, 63]. During inference, it generates a space-time prompt for each video from random Gaussian noise through a gradual diffusion process. This approach precisely captures the underlying prompt distribution and can sample a prompt tailored to the specific video input from the learned distribution, providing more nuanced guidance to the model. Additionally, it naturally injects noise during prompt learning, which encourages the model to learn diverse and informative prompts that are dispersed throughout the prompt space for each incremental task. As a result, it reduces overfitting to the current task and enhances generalization to future tasks, thereby

improving overall performance. Moreover, to capture rich spatio-temporal dynamics, CoSTEP employs a simple yet effective method to prompt the pre-trained Vision Transformer (ViT) [9] to infer relationships among local patches across frames. This incurs minimal computational cost and introduces no additional parameters, alleviating concerns about catastrophic forgetting in the temporal module.

Our main contributions are: (1) We introduce CoSTEP, a method that models the conditional distribution of prompts, enabling the generation of diverse and informative prompts for each video and improving generalization to new tasks. (2) We propose a simple yet effective prompting strategy that utilizes pre-trained image transformers to capture spatio-temporal relationships in videos. (3) We conduct a comprehensive evaluation of CoSTEP across four VCIL benchmarks, where our method achieves state-of-the-art (SoTA) performance on all.

2. Related Work

Image Continual Learning. Continual Learning (CL) has been extensively studied in the image domain. Previously, existing approaches can be grouped into regularization-based methods, exemplar-based methods, and architecture-based methods [40, 58]. For more detailed discussions on these methods, see Appendix 6. Recent years have seen an increasing trend in CL research focusing on prompting and pre-trained Vision Transformers (ViTs) [25, 54, 65–67, 72]. L2P [67] first introduces the concept of tuning a frozen ViT backbone for CL tasks using a prompt pool. DualPrompt [66] further develops this by learning task-invariant and task-specific prompts, respectively. However, the flexibility and learning capability of these approaches are limited by a fixed-sized prompt pool. To enhance this, approaches like CODA-Prompt [54],

DAP [22], and APG [57] use a learnable prompt generator that dynamically creates prompts from image features or patches, with some also incorporating a list of prompt candidates as additional input. In contrast, we capture the distribution in prompt space conditioned on the input, leading to improved generalization on new tasks and enhanced performance.

Video Continual Learning. CL in the video domain has recently gained attention [5, 6, 12, 24, 30, 31, 70]. TCD [44] computes time-channel importance for weighted distillation, vCLIMB [60] emphasizes temporal-consistency regularization in untrimmed videos. FrameMaker [46] learns to condense frames for representative videos of old classes to enhance memory efficiency. STSP [6] employs a spatial-temporal subspace projection method. Recently, PIVOT [61] and ST-Prompt [47] adapt prompt-based CL methods to the video domain, designing spatial and temporal prompts specifically for temporal modeling. However, their reliance on selecting prompts from a fixed-size pool reduces flexibility. Furthermore, these methods ignore the relationships between local patches across frames.

Prompt Learning. Prompt learning, widely used in natural language processing (NLP) [34], has inspired similar approaches in computer vision. VPT [21] explores prompting in pretrained vision models by learning a continuous prompt to minimize classification loss in downstream tasks. CoOp [74] and CoCoOp [73] optimizes prompts for vision-language models like CLIP [48]. Although effective, deterministic prompt learners often struggle with diversity and generalizability. To overcome these limitations, recent studies have begun exploring probabilistic prompt learning [7, 29, 35, 38]. However, these probabilistic methods often assume that the distribution of prompts or embeddings is simple, which is unsuitable for video tasks where the space-time prompt space is complex and not easily approximated by simple distributions. Consequently, we utilize a diffusion model, a more sophisticated generative model, to more accurately capture the prompt distribution.

3. Method

3.1. Problem Setting

In Video Class-Incremental Learning (VCIL), tasks are presented sequentially as $\{\mathcal{T}^1, \mathcal{T}^2, \dots, \mathcal{T}^K\}$, where each task \mathcal{T}^k is associated with a unique dataset $\mathcal{D}^k = \{(V_i^k, y_i^k), y_i^k \in \mathcal{L}^k\}$. This dataset comprises video instances V_i^k paired with their corresponding labels y_i^k . The labels in task k belong to a predefined label set \mathcal{L}^k and do not appear in previous tasks. Our primary goal is to develop a unified model that can accurately classify videos across

all encountered classes. This necessitating the model’s ability to prevent the loss of historical information, known as *catastrophic forgetting*.

3.2. Overview

We employ the prompt tuning paradigm [21] using a pre-trained Vision Transformer (ViT), which remains frozen and acts as a feature extractor. Traditional prompt learning optimizes a deterministic prompt $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{N_p}]$, where each \mathbf{p}_j is a vector in $\mathbb{R}^{1 \times D}$, with N_p representing the sequence length and D the embedding dimension (refer to Appendix 8 for details). These prompts are appended to the patch embedding and inputted into the frozen backbone. Different from them, we model the prompt space as a conditional distribution, defining a distribution $p(\mathbf{P}|\hat{\mathbf{E}})$ over the prompts, conditioned on the input video feature $\hat{\mathbf{E}}$. Details on obtaining $\hat{\mathbf{E}}$ will be discussed later. The learning objective for task \mathcal{T}^k is then formulated as

$$\ell_{\text{classification}} = \mathbb{E}_{V_i^k, y_i^k} \left[-\log \mathbb{E}_{\mathbf{P} \sim p(\mathbf{P}|\hat{\mathbf{E}})} \left[p(y_i^k | V_i^k, g_\phi, \mathbf{P}) \right] \right], \quad (1)$$

where g_ϕ is a trainable classifier parameterized by ϕ . In our method, $p(\mathbf{P}|\hat{\mathbf{E}})$ is parameterized by a diffusion model. Given the complexity of solving Equation 1, we use a dual-phase training approach for each task \mathcal{T}^k , as shown in Figure 2 left. Initially, we focus on optimizing a task-specific space-time prompt \mathbf{P}^k using a novel spatio-temporal modeling technique (Section 3.3). This prompt acts as a pseudo reconstruction target to regularize the training of the diffusion model in the second phase. During this phase, we train a diffusion model [18] to solve Equation 1 and model the distribution within the prompt space (Section 3.4). During inference, as illustrated in Figure 2 right, we iteratively generate instance-level prompts from random Gaussian noise, which then guide the pre-trained ViT to make the final prediction. We also detail the training and inference algorithm in Appendix 7.

3.3. Space-Time Prompt Learning

Pre-trained image models inherently lack the ability to process temporal context across frames. To address this limitation in optimizing \mathbf{P}^k , previous prompt-based VCIL methods [47, 61] have utilized mean pooling or a temporal transformer to process all global frame features. Yet, these approaches often neglect the interactions among local patches across frames, crucial for video transformers [37] to capture strong temporal contexts effectively. To overcome this, as illustrated on the right side of Figure 3, we introduce a frame grid to train prompts that effectively capture both spatial and temporal contexts. This method is straightforward yet effective, approaching temporal modeling in a

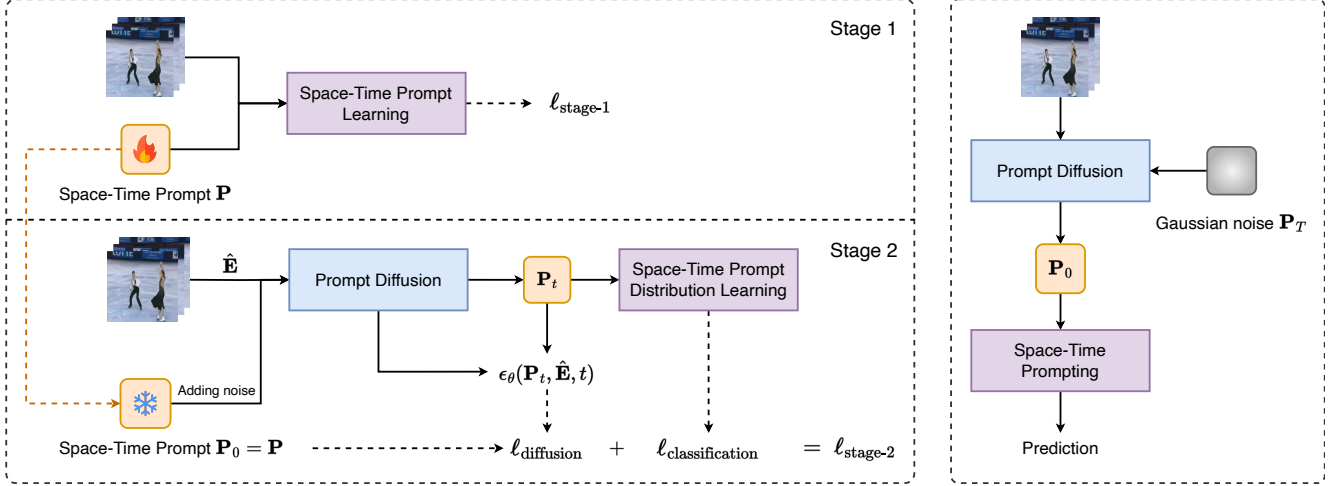


Figure 2. An overview of the CoSTEP training (left) and inference (right) pipelines.

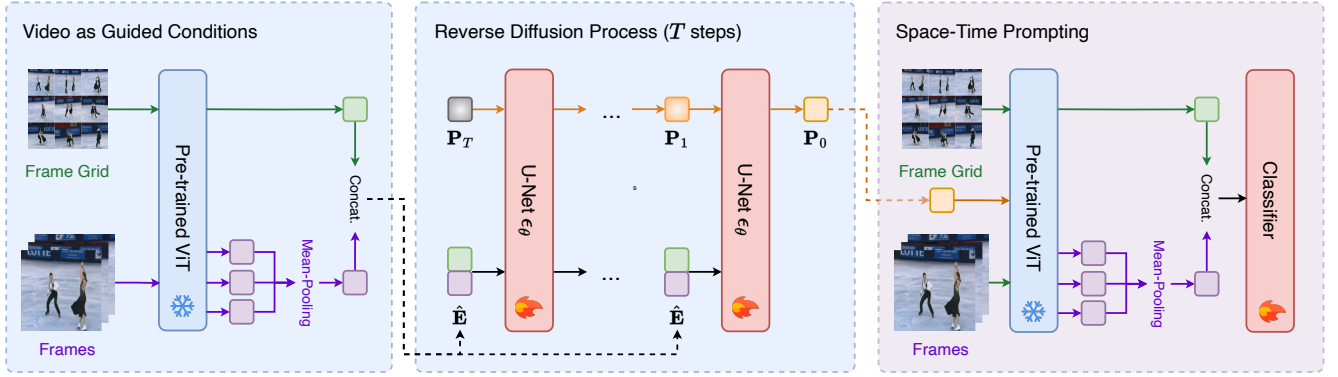


Figure 3. Illustration of the different components within CoSTEP. Initially, a pre-trained ViT is employed to encode the spatio-temporal information from a video clip into an embedding. This embedding subsequently serves as the condition for generating prompts. Finally, the generated prompts are integrated with the video input to assist in making predictions.

manner akin to spatial modeling and eliminating the need for additional complex modules for temporal analysis.

In our model’s prompt learning process for task \mathcal{T}^k , we start by sampling N_{sp} frames for spatial context and N_{tp} frames for temporal context from a video. Since temporal information requires modeling over consecutive frames, we typically sample more frames (N_{tp}) than for spatial context (N_{sp}). Also, pixel-level detail in each frame is less crucial for modeling temporal information, allowing us to down-sample temporal frames to reduce computational costs. Specifically, we down-sample each temporal frame by a scaling factor of $N_{tp}^{1/2}$, resulting in frames of size $[\frac{W}{N_{tp}^{1/2}} \times \frac{H}{N_{tp}^{1/2}}]$. These frames are then stacked in temporal order to form a grid matching the original frame size $[W \times H]$. Using \mathbf{P}^k , the pre-trained ViT extracts features from this frame grid, \mathbf{e}^{tp} , serving as the video’s temporal

representation. Similarly, \mathbf{P}^k extracts features from each spatial frame, generating a spatial representation \mathbf{e}_i^{sp} for each frame. The video representation $\mathbf{E} = [\mathbf{e}^{sp}, \mathbf{e}^{tp}]$ is then formed, where \mathbf{e}^{sp} is the average of all spatial features $\{\mathbf{e}_1^{sp}, \mathbf{e}_2^{sp}, \dots, \mathbf{e}_{N_{sp}}^{sp}\}$, and both \mathbf{e}_i^{sp} and \mathbf{e}^{tp} are vectors in $\mathbb{R}^{1 \times D_e}$, with D_e representing the dimension of the output feature. Finally, \mathbf{E} is used in a classifier to compute the classification loss $\ell_{\text{stage-1}}$.

3.4. Prompt Diffusion

We use a diffusion-based generative framework [18] to solve Equation 1 and model the distribution in the prompt space. Once the distribution is learned, we generate a prompt tailored to each input video from it. For clarity, we omit the symbol k in this section. Unless stated otherwise, this refers to scenarios within the incremental task \mathcal{T}^k .

Video features as guided conditions. We aim to model the probabilistic distribution based on the input video clip V_i . Accurate spatio-temporal encoding of the video is crucial for the diffusion model to effectively capture this distribution. Therefore, we follow the procedure in Section 3.3 to extract video embeddings $\hat{\mathbf{E}} = [\hat{\mathbf{e}}^{sp}, \hat{\mathbf{e}}^{tp}]$ using a frozen pre-trained ViT, but without prompting.

Training. We use a denoising diffusion probabilistic model to approximate the data distribution $p(\mathbf{P}_0|\hat{\mathbf{E}})$. This model aims to progressively reconstruct \mathbf{P}_0 to closely match \mathbf{P} by iteratively refining random Gaussian noise $\mathbf{P}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which has the same dimensions as \mathbf{P} . T denotes the number of diffusion steps. The forward diffusion process, $q(\mathbf{P}_t|\mathbf{P}_{t-1})$, is a Markov chain that incrementally adds Gaussian noise to each step t , starting from a clean tensor $\mathbf{P}_0 = \mathbf{P}$ drawn from $q(\mathbf{P}_0)$. The forward diffusion process is described as follows:

$$q(\mathbf{P}_t|\mathbf{P}_0) = \prod_{t=1}^T q(\mathbf{P}_t|\mathbf{P}_{t-1}), \quad (2)$$

where $q(\mathbf{P}_t|\mathbf{P}_{t-1}) = \mathcal{N}(\mathbf{P}_t; \sqrt{1 - \beta_t}\mathbf{P}_{t-1}, \beta_t\mathbf{I})$.

Here $\{\beta\}_{t=0}^T$ is a pre-defined variance schedule. By defining $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$, the forward diffused sample at time step t , denoted as \mathbf{P}_t , can be generated in a single step as follows:

$$\mathbf{P}_t = \sqrt{\bar{\alpha}_t}\mathbf{P}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \text{ where } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (3)$$

During training, the diffusion model selects a diffusion step t from the range $[1, T]$ and computes \mathbf{P}_t as per Equation 3. It then uses a trainable U-Net model to determine $\epsilon_\theta(\mathbf{P}_t, \hat{\mathbf{E}}, t)$ and computes the reconstruction loss for step t as follows:

$$\ell_{\text{diffusion}} = E_{\mathbf{P}_0, \hat{\mathbf{E}}, \epsilon, t} \left[\|\mathbf{P}_0 - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{P}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \hat{\mathbf{E}}, t)\|^2 \right]. \quad (4)$$

\mathbf{P}_t is treated as a sample from the distribution $p(\mathbf{P}|\hat{\mathbf{E}})$ and used to prompt the pre-trained ViT for calculating $\ell_{\text{classification}}$ in Equation 1. Although typically, the expectation in $\ell_{\text{classification}}$ requires multiple Monte Carlo samplings, we found that using just one sample can achieve SoTA performance. Thus, for efficiency, we only use one sample in this process. However, using more Monte Carlo samplings can further improve performance; see Appendix 12 for details. The final objective is defined as $\ell_{\text{stage-2}} = \ell_{\text{diffusion}} + \ell_{\text{classification}}$.

Given that the diffusion model is parameterized by θ , it could be prone to catastrophic forgetting during training. To

address this for the current k^{th} task, we retain the prompts $\{\mathbf{P}^i\}_{i=1}^{k-1}$ from previous tasks and a limited set of video representations $\{\hat{\mathbf{E}}^i\}_{i=1}^m$, where m is the number of stored representations. The model trains on the current task data alongside these stored elements to reduce forgetting. Notably, storing complete videos or specific frames isn't necessary; a single global representation per video is sufficient. This method significantly reduces memory use and privacy risks, as storing one vector per video occupies much less space than retaining raw video segments, and deep feature space outputs are less likely to expose private information than raw data.

Inference. After training, the diffusion model can generate \mathbf{P}_0 by iteratively denoising, starting from random Gaussian noise. At each step, to sample from $p_\theta(\mathbf{P}_{t-1}|\mathbf{P}_t)$, the model performs the following calculation:

$$\mathbf{P}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{P}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{P}_t, \hat{\mathbf{E}}, t) \right) + \sigma_t \epsilon. \quad (5)$$

For more details, refer to Appendix 9. Finally, we use \mathbf{P}_0 , which is intended to approximate \mathbf{P} , as a prompt for the pre-trained ViT.

4. Experiments

4.1. Experiment setup

Datasets and Evaluation Protocol. We assess our method using the standard action recognition datasets: UCF101 [55], HMDB51 [28], and Something-Something V2 [16], following the VCIL benchmarks from TCD [44] and vCLIMB [60]. For UCF101, we adopt two different setups: one where classes are introduced sequentially in 10 or 20 tasks, and another where the model first learns 51 classes and then trains on the remaining 50 classes divided into 5, 10, or 25 tasks. For HMDB51, the base model is trained on 26 classes, with the remaining classes divided into 5 or 25 tasks. For Something-Something V2, training begins with 84 classes, followed by 9 incremental tasks of 10 classes or 18 tasks of 5 classes. Performance is evaluated across all seen classes, measuring average accuracy (Acc) and backward forgetting (BWF), with metric details in Appendix 10.

Implementation Details. Our approach utilizes CLIP ViT-B/16 [48], a visual-language model pre-trained on numerous image-text pairs. We train our model with a batch size of 50, through 20 epochs in the first stage and an additional 20 in the second stage. From each video, we uniformly sample $N_{sp} = 8$ and $N_{tp} = 25$ frames and use a prompt length of 6. The number m of stored video representation vectors from previous tasks is set at 20 vectors per

Table 1. Comparison with existing VCIL approaches on UCF101, HMDB51 and Something-Something v2 (using TCD [44]’s split) with average accuracy (Acc) reported.

Methods	UCF101			HMDB51		SSv2	
	5 tasks	10 tasks	25 tasks	5 tasks	25 tasks	9 tasks	18 tasks
iCaRL [49]	70.58	69.51	67.28	43.90	37.15	20.41	16.62
UCIR [19]	77.55	74.59	71.77	48.20	39.42	24.32	19.31
PODNet [10]	76.50	76.17	73.83	48.38	43.35	27.63	20.14
TCD [44]	78.13	76.87	75.74	50.29	44.04	29.32	24.69
FrameMaker [46]	79.37	79.55	79.32	51.43	46.37	31.41	26.57
L2P [67]	81.24	80.09	78.58	49.98	45.87	26.02	21.33
HCE [31]	80.01	78.81	77.62	52.01	48.94	36.88	32.82
CLIP ViT-B/16							
CLIP (Zero Shot) [48]	69.68	69.61	69.87	40.19	40.42	3.56	3.57
S-Prompts [65]	80.60	80.27	80.43	53.11	53.89	33.69	30.84
ST-Prompt [47]	84.75	85.54	85.67	60.14	60.54	39.98	35.44
CoSTEP (Ours)	86.05	86.71	86.95	61.70	61.84	41.44	36.60

Table 2. Comparison with existing VCIL approaches on UCF101 (using vCLIMB [60]’s split).

Methods	10 tasks		20 tasks	
	Acc (↑)	BWF (↓)	Acc (↑)	BWF (↓)
EWC [27]	9.51	98.94	4.71	92.12
MAS [1]	10.89	11.11	5.90	5.31
BiC [68]	78.16	18.49	70.69	24.90
iCaRL [49]	80.97	18.11	76.59	21.83
CLIP ViT-B/16				
CLIP (ZS) [48]	65.73	11.4	65.38	12.35
PIVOT [61]	93.36	4.47	93.07	3.89
CoSTEP (Ours)	96.51	1.99	95.47	2.55

class. For further details on implementation and efficiency discussions, refer to Appendices 11 and 12.

4.2. Main Results

Our novel approach, CoSTEP, was benchmarked against existing VCIL approaches across four distinct datasets, as detailed in Tables 1 and 2. CoSTEP outperforms other prompt-based VCIL methods like ST-Prompt [47] and PIVOT [61] that use CLIP, achieving higher average accuracy across all datasets. This success demonstrates CoSTEP’s effective capture of the complex space-time prompt distribution and its ability to generate useful instance-level prompts. Furthermore, CoSTEP excels in minimizing forgetting. This advantage can be at-

Table 3. Benefit of prompt diffusion over other prompt learning methods.

	UCF101	HMDB51
Task-agnostic	74.98	48.77
L2P [67]	89.61	56.49
CODA-Prompt [54]	92.07	55.58
DAP [22]	85.38	52.58
$\mathcal{N}(0, I)$	58.37	34.93
$\mathcal{N}(\mu(\mathcal{P}), \Sigma(\mathcal{P}))$	75.92	44.45
BPT [7]	93.93	58.83
VAE [26]	95.18	58.38
GAN [14]	92.32	58.05
CoSTEP (Ours)	96.51	61.70

tributed to CoSTEP’s strategy of encoding previous knowledge within a diffusion model rather than a prompt pool, which enhances capacity and stability. We highlight that ST-Prompt uses soft prompts for both CLIP vision and text encoders, significantly boosting performance through text prompts [47]. Although we follow the standard CIL pipeline [54, 67] of adding prompts only to the vision encoder, we still outperform ST-Prompt.

4.3. Ablation Studies

In this section, we perform ablation studies to assess the characteristics and performance of our core design elements, using the UCF101 dataset with 10 tasks and

Table 4. Ablation study on the objective function in the second training stage.

$\ell_{\text{diffusion}}$	$\ell_{\text{classification}}$	UCF101	HMDB51
✓	✗	49.54	33.53
✗	✓	92.80	54.34
✓	✓	96.51	61.70

HMDB51 with 5 tasks, unless noted otherwise. For additional ablation studies, please see Appendix 12.

Benefit of prompt diffusion. To assess the effectiveness of our diffusion-based prompt distribution learning, we compare two categories of methods: (1) *Deterministic methods*, including *Task-agnostic*, which tunes a single set of prompts sequentially for all tasks; L2P [67], which selects prompts from a pool for each input using key-value pairs; CODA-Prompt [54], creating prompts by combining prompt components based on input-conditioned weights; and DAP [22], generating prompts from a network based on input features. (2) *Probabilistic methods*, involving sampling a prompt from a normal distribution $\mathcal{N}(0, I)$ or $\mathcal{N}(\mu(\mathcal{P}), \Sigma(\mathcal{P}))$, where $\mu(\mathcal{P})$ and $\Sigma(\mathcal{P})$ are the mean and covariance matrix computed from $\mathcal{P} = \{\mathbf{P}^k\}_{k=1}^K$. Additionally, we explore replacing our diffusion model with Bayesian Prompt Learning (BPT) [7], Variational Autoencoder (VAE) [26], and Generative Adversarial Network (GAN) [14]. Note that for methods unable to utilize stored video representation vectors to counter forgetting, we instead store an equivalent number of raw videos for them. We present the average accuracies in Table 3.

The results reveal that *Task-agnostic* performs poorly due to significant forgetting. Deterministic methods like L2P, CODA-Prompt, and DAP generally underperform compared to probabilistic methods such as BPT, VAE, and GAN, highlighting the benefits of modeling the space-time prompt distribution. Sampling a prompt directly from $\mathcal{N}(0, I)$ or $\mathcal{N}(\mu(\mathcal{P}), \Sigma(\mathcal{P}))$ yields poor results, demonstrating that treating the prompt space as a simple normal distribution is inadequate and that capturing its complex distribution is crucial. Our CoSTEP, which employs a diffusion model, surpasses BPT, VAE, and GAN, suggesting it more effectively models complex distributions and generates more informative prompts.

Loss terms. In the second training stage, we optimize the total loss $\ell_{\text{diffusion}} + \ell_{\text{classification}}$, where $\ell_{\text{classification}}$ is the main objective and $\ell_{\text{diffusion}}$ regulates the prompt space with a pseudo reconstruction target. Results from different loss

Table 5. Ablation study on different ways of learning prompts for videos.

Spatial	Temporal	UCF101	HMDB51	SSv2
✓	✗	92.46	59.12	37.52
✗	✓	70.37	49.64	30.58
✓	✓	96.51	61.70	41.44

Table 6. Ablation study on different ways of modeling temporal context.

	UCF101	HMDB51
Mean-pooling	94.41	58.37
Max-pooling	93.86	57.79
Frame grid	96.51	61.70

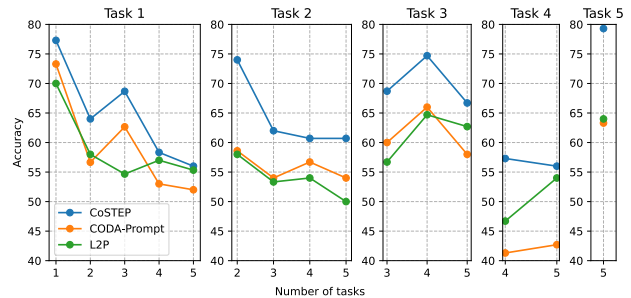


Figure 4. Changes in accuracy for each incremental task on HMDB51 (5 tasks).

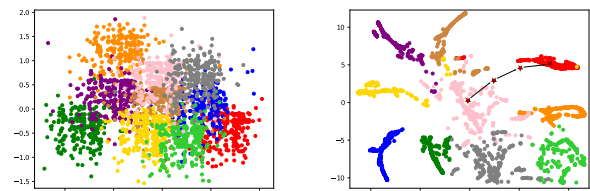


Figure 5. t-SNE visualization of **cross-task** prompts produced by CODA-Prompt (left) and CoSTEP (right) on UCF101 (10 tasks). Each point represents a D -dimensional prompt vector, with distinct colors denoting different tasks. Black arrows and red stars (right) show how a diffused prompt adapts.

combinations are shown in Table 4, indicating that using both losses together achieves the most effective learning. Using only $\ell_{\text{diffusion}}$ performs poorly, as expected, because it

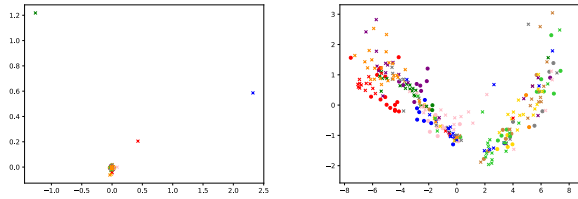


Figure 6. t-SNE visualization of prompts within the **same task** produced by CODA-Prompt (left) and CoSTEP (right) on UCF101 (10 tasks). Different colors indicate different classes and different shapes of the same color represent various clusters within the same class.

neglects the marginal likelihood in Equation 1 that we aim to optimize.

Space-time prompts. We conducted tests focusing solely on spatial or temporal dimensions, as shown in Table 5. In the spatial-only approach, we excluded the frame grid, whereas in the temporal-only approach, we relied exclusively on frame grid features. Our space-time prompting outperforms these methods, even on Something-Something v2 (9 tasks) which places a greater emphasis on temporal reasoning. Omitting the frame grid leads to lower performance, underscoring the importance of using the ViT’s innate capabilities for temporal modeling in our prompt design. Using only the frame grid also fails to achieve satisfactory accuracy, likely due to resizing frames for our grid, which reduces spatial detail. This emphasizes the benefit of using original-size frames to better capture and utilize spatial information.

Is using a frame grid necessary? We explored alternative methods for temporal modeling, specifically Mean-pooling, which averages the RGB values of all frames, and Max-pooling, which selects the maximum RGB values across frames. Our results, as shown in Table 6, indicate that our Frame grid method surpasses the others. This is expected since Mean-pooling and Max-pooling significantly reduce temporal information by condensing all frames into one. In contrast, Frame grid leverages the ViTs’ capability to analyze relationships between frames, providing a distinct advantage over the other methods.

4.4. Why does CoSTEP outperform other methods?

Analysis of per-task accuracy. To demonstrate how CoSTEP outperforms others, we visualize accuracy changes for each of the 5 incremental tasks on HMDB51, following pre-training on 51 classes, in Figure 4. CoSTEP consistently achieves the highest initial accuracy for each

task. Furthermore, the initial accuracy gap between CoSTEP and other methods widens with more tasks, indicating CoSTEP’s superior generalization ability to transfer knowledge from pre-trained and earlier tasks to future tasks. CODA-Prompt and L2P fail to achieve high initial accuracy, indicating their generalizability and plasticity are insufficient for video domains.

Visualization of prompts. We use t-SNE [59] to visualize the differences in prompts generated by CoSTEP and CODA-Prompt on UCF101 (10 tasks). Figure 5 shows the distribution of prompts across different tasks, while Figure 6 focuses on the distribution within the first task. To verify whether generated prompts correspond to input videos, we apply k -means to cluster the video features (averaged CLIP frame features per video) of each class into two centroids. These centroids presumably reflect visual variation within the class. We observe that prompts generated by CoSTEP are more effectively clustered by tasks and exhibit greater diversity than those from CODA-Prompt. Moreover, CODA-Prompt-generated prompts for the same task show little variation across classes or instances, while CoSTEP-generated prompts display more diversity within the same task. Prompts for different classes and even different clusters within the same class are well clustered. This explains why CoSTEP performs better: it generates prompts that more effectively guide predictions for each instance. Additionally, we illustrate CoSTEP’s diffusion process during inference. Starting with a prompt generated from a Gaussian distribution, which is initially far from the optimal area, it gradually moves closer during the diffusion process, demonstrating CoSTEP’s ability to generate meaningful prompts from the learned distribution. More visualizations are available in Appendix 13.

5. Conclusion

In this paper, we introduce CoSTEP, an approach that enhances prompt learning for Video Class-Incremental Learning by modeling the overall distribution within the space-time prompt space. This approach enables the generation of diverse and informative prompts, minimizes overfitting to current tasks, and improves generalization to future tasks, thus boosting overall performance. Additionally, we have developed a straightforward yet effective space-time prompting strategy to better transfer spatial modeling capabilities from pre-trained image models to temporal modeling. This strategy allows these models to analyze relationships across frames similarly to how they analyze within-frame relationships, leveraging their intrinsic capabilities. Our experimental results demonstrate that CoSTEP achieves superior performance over current state-of-the-art methods.

References

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 1, 6
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1
- [3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 1
- [4] Sungmin Cha, Sungjun Cho, Dasol Hwang, Sunwon Hong, Moontae Lee, and Taesup Moon. Rebalancing batch normalization for exemplar-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20127–20136, 2023. 1
- [5] Geng Chen, Wendong Zhang, Han Lu, Siyu Gao, Yunbo Wang, Mingsheng Long, and Xiaokang Yang. Continual predictive learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10728–10737, 2022. 3
- [6] Hao Cheng, Siyuan Yang, Chong Wang, Joey Tianyi Zhou, Alex C Kot, and Bihan Wen. Stsp: Spatial-temporal subspace projection for video class-incremental learning. In *European Conference on Computer Vision*, pages 374–391. Springer, 2025. 3
- [7] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Bayesian prompt learning for image-language model generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15237–15246, 2023. 1, 3, 6, 7
- [8] Jiahua Dong, Wenqi Liang, Yang Cong, and Gan Sun. Heterogeneous forgetting compensation for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11742–11751, 2023. 1
- [9] 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. 2
- [10] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020. 6
- [11] Yingjun Du, Zehao Xiao, Shengcai Liao, and Cees Snoek. Protodiff: Learning to learn prototypical networks by task-guided diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [12] Di Fu, Thanh Vinh Vo, Haozhe Ma, and Tze-Yun Leong. Decoupled prompt-adapter tuning for continual activity recognition. *arXiv preprint arXiv:2407.14811*, 2024. 3
- [13] Yunhao Ge, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, and Laurent Itti. Clr: Channel-wise lightweight reprogramming for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18798–18808, 2023. 1
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 6, 7
- [15] Dipam Goswami, René Schuster, Joost van de Weijer, and Didier Stricker. Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3195–3204, 2023. 1
- [16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Freund, Peter Yianilos, Moritz Mueller-Freitag, et al. The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5
- [17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 4, 1
- [19] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 6
- [20] Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11858–11867, 2023. 1
- [21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3, 1
- [22] 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*, pages 11847–11857, 2023. 1, 3, 6, 7
- [23] Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *arXiv preprint arXiv:1710.10368*, 2017. 1
- [24] Haeyong Kang, Jaehong Yoon, Sung Ju Hwang, and Chang D Yoo. Continual learning: Forget-free winning subnetworks for video representations. *arXiv preprint arXiv:2312.11973*, 2023. 3

- [25] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 2
- [26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6, 7
- [27] 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. 1, 6
- [28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5
- [29] Hyeongjun Kwon, Taeyong Song, Somi Jeong, Jin Kim, Jinhun Jang, and Kwanghoon Sohn. Probabilistic prompt learning for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6768–6777, 2023. 3
- [30] Xiaochen Li, Jian Cheng, Ziyang Xia, Zichong Chen, Junhao Shi, Zhicheng Dong, and Nyima Tashi. Ts-ilm: Class incremental learning for online action detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1158–1167, 2024. 3
- [31] Sen Liang, Kai Zhu, Wei Zhai, Zhiheng Liu, and Yang Cao. Hypercorrelation evolution for video class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3315–3323, 2024. 3, 6
- [32] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24246–24255, 2023. 1
- [33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 1
- [34] 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. 3
- [35] Xinyang Liu, Dongsheng Wang, Miaoge Li, Zhibin Duan, Yishi Xu, Bo Chen, and Mingyuan Zhou. Patch-token aligned bayesian prompt learning for vision-language models. *arXiv preprint arXiv:2303.09100*, 2023. 3
- [36] Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, and Joost van de Weijer. Augmented box replay: Overcoming foreground shift for incremental object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11367–11377, 2023. 1
- [37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1, 3
- [38] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 1, 3
- [39] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022. 1
- [40] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022. 2, 1
- [41] Zichong Meng, Jie Zhang, Changdi Yang, Zheng Zhan, Pu Zhao, and Yanzhi Wang. Diffclass: Diffusion-based class incremental learning. In *European Conference on Computer Vision*, pages 142–159. Springer, 2024. 1
- [42] Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 35:14516–14528, 2022. 1
- [43] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jah-nichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019. 1
- [44] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13698–13707, 2021. 1, 3, 5, 6
- [45] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2203–2213, 2023. 1
- [46] Yixuan Pei, Zhiwu Qing, Jun Cen, Xiang Wang, Shiwei Zhang, Yaxiong Wang, Mingqian Tang, Nong Sang, and Xueming Qian. Learning a condensed frame for memory-efficient video class-incremental learning. *Advances in Neural Information Processing Systems*, 35:31002–31016, 2022. 1, 2, 3, 6
- [47] Yixuan Pei, Zhiwu Qing, Shiwei Zhang, Xiang Wang, Yingya Zhang, Deli Zhao, and Xueming Qian. Space-time prompting for video class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11932–11942, 2023. 1, 2, 3, 6
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5, 6, 4
- [49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE con-*

- ference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 1, 6
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [51] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 1
- [52] Dawid Rymarczyk, Joost van de Weijer, Bartosz Zielinski, and Bartłomiej Twardowski. Icicle: Interpretable class incremental continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1887–1898, 2023. 1
- [53] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 1
- [54] 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*, pages 11909–11919, 2023. 1, 2, 6, 7, 4, 5
- [55] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- [56] Qing Sun, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Exploring example influence in continual learning. *Advances in Neural Information Processing Systems*, 35: 27075–27086, 2022. 1
- [57] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. When prompt-based incremental learning does not meet strong pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1706–1716, 2023. 1, 3
- [58] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 2, 1
- [59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8
- [60] Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, and Bernard Ghanem. vclimb: A novel video class incremental learning benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19035–19044, 2022. 1, 2, 3, 5, 6
- [61] Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, and Bernard Ghanem. Pivot: Prompting for video continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24214–24223, 2023. 1, 2, 3, 6, 4, 5
- [62] Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F Grewe. Continual learning with hypernetworks. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020. 1
- [63] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14836–14845, 2023. 2
- [64] Wenjin Wang, Yunqing Hu, Qianglong Chen, and Yin Zhang. Task difficulty aware parameter allocation & regularization for lifelong learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7776–7785, 2023. 1
- [65] 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, 2022. 1, 2, 6
- [66] 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*, pages 631–648. Springer, 2022. 2
- [67] 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*, pages 139–149, 2022. 1, 2, 6, 7, 4, 5
- [68] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 6
- [69] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. *Advances in Neural Information Processing Systems*, 35:14771–14783, 2022. 1
- [70] Hanbin Zhao, Xin Qin, Shihao Su, Yongjian Fu, Zibo Lin, and Xi Li. When video classification meets incremental classes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 880–889, 2021. 3
- [71] Shu Zhao, Tan Yu, Xiaoshuai Hao, Wenchao Ma, and Vijaykrishnan Narayanan. Kalahash: Knowledge-anchored low-resource adaptation for deep hashing. *arXiv preprint arXiv:2412.19417*, 2024. 1
- [72] Shu Zhao, Xiaohan Zou, Tan Yu, and Huijuan Xu. Reconstruct before query: Continual missing modality learning with decomposed prompt collaboration. *arXiv preprint arXiv:2403.11373*, 2024. 2
- [73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 3

- [74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. [1](#), [3](#)
- [75] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. [1](#)
- [76] Xiaohan Zou and Tong Lin. Efficient meta-learning for continual learning with taylor expansion approximation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. [1](#)
- [77] Xiaohan Zou, Changqiao Wu, Lele Cheng, and Zhongyuan Wang. Tokenflow: Rethinking fine-grained cross-modal alignment in vision-language retrieval. *arXiv preprint arXiv:2209.13822*, 2022. [2](#)