DECOUPLED PROMPT-ADAPTER TUNING FOR CONTINUAL ACTIVITY RECOGNITION

Di Fu
School of Computing
National University of Singapore
Singapore
difu@u.nus.edu

Thanh Vinh Vo
School of Computing
National University of Singapore
Singapore
votv@comp.nus.edu.sg

Haozhe Ma
School of Computing
National University of Singapore
Singapore
haozhe.ma@u.nus.edu

Tze-Yun Leong
School of Computing
National University of Singapore
Singapore
leongty@comp.nus.edu.sg

ABSTRACT

Action recognition technology plays a vital role in enhancing security through surveillance systems, enabling better patient monitoring in healthcare, providing in-depth performance analysis in sports, and facilitating seamless human-AI collaboration in domains such as manufacturing and assistive technologies. The dynamic nature of data in these areas underscores the need for models that can continuously adapt to new video data without losing previously acquired knowledge, highlighting the critical role of advanced continual action recognition. To address these challenges, we propose Decoupled Prompt-Adapter Tuning (DPAT), a novel framework that integrates adapters for capturing spatial-temporal information and learnable prompts for mitigating catastrophic forgetting through a decoupled training strategy. DPAT uniquely balances the generalization benefits of prompt tuning with the plasticity provided by adapters in pretrained vision models, effectively addressing the challenge of maintaining model performance amidst continuous data evolution without necessitating extensive finetuning. DPAT consistently achieves state-of-the-art performance across several challenging action recognition benchmarks, thus demonstrating the effectiveness of our model in the domain of continual action recognition.

1 INTRODUCTION

The widespread deployment of cameras has significantly broadened the scope and influence of action recognition technology across multiple sectors. This technology is essential for boosting safety via security and surveillance, offering vital patient care in healthcare, and providing in-depth performance analyses in sports (Kong & Fu, 2022). It also enables robots to quickly perceive and respond to human actions during human-AI collaborations (Akkaladevi & Heindl, 2015), thereby enhancing the collaboration and efficiency between humans and AI in contexts such as manufacturing and assistive technologies. In this context, the importance of continual learning (CL) emerges, driven by the technological imperative to synchronize with the dynamically evolving nature of human activities and interactions. Defined as the ability of the model to incorporate new information from an ongoing data stream while retaining previously acquired knowledge, continual learning plays an essential role in this arena. It adeptly navigates the complex and heterogeneous spectrum of actions an action recognition algorithm encounters, enabling adaptation to emerging actions and contexts over time without compromising the recognition of previously observed actions. This flexibility is crucial for sustaining the efficacy and relevance of action recognition systems across their extensive applications, thereby establishing continual learning as a foundational element for the advancement and persistent applicability of action recognition technologies.

While continual learning has made significant progress in recent years, the specific challenge of continual action recognition, which involves learning from dynamic video data streams without forgetting previously acquired knowledge, remains a difficult problem. On the one hand, the majority of CL methodologies are primarily devised for static images (Wang et al., 2022b;a; Smith et al., 2023) , rendering them inadequately equipped to confront the unique challenges presented by video data. These challenges encompass the high-dimensional nature of video data, temporal

dependencies, and the significant variability across video sequences. Such challenges can either impede a model's ability to adapt to new tasks or induce rapid changes that lead to catastrophic forgetting. On the other hand, many existing methods tailored specifically for continual learning in the context of video (Villa et al., 2023; Pei et al., 2022) require the retention of data for new classes. These approaches inevitably lead to escalating memory costs, which become prohibitively expensive given the typically large size of video data. Consequently, the evolution of models in this domain is hampered by a delicate balance between adaptability and memory efficiency.

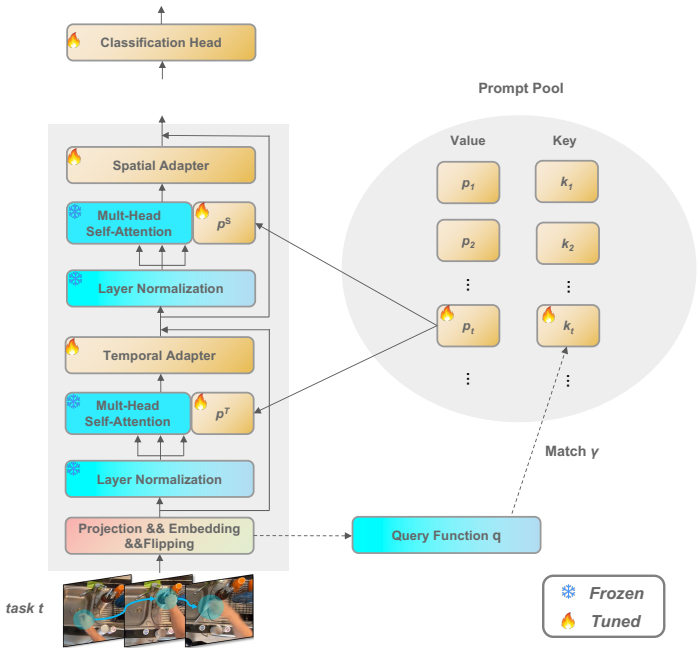
Recent advancements in large pre-trained models (Bao et al., 2021; Dosovitskiy et al., 2020; Radford et al., 2021) have spurred the development of lightweight tuning techniques, such as Adapters (Houlsby et al., 2019) and Prompt Tuning (Li & Liang, 2021; Lester et al., 2021), addressing the cost and time-intensive nature of fine-tuning large models. These methods enable the refinement of pre-trained models for new tasks with minimal increases in parameters, significantly reducing the computational and memory demands typically associated with training models from scratch. Furthermore, it has been observed that by tuning fewer parameters, these methods exhibit reduced susceptibility to forgetting and enhanced generalization capabilities (Vander Eeckt & Van Hamme, 2023), albeit with limited plasticity. This insight has catalyzed the trend of employing pre-trained models in continual learning scenarios (Wang et al., 2022a;b). Despite the distinct advantages of adapters and prompt tuning, where adapters mitigate catastrophic forgetting compared to conventional fine-tuning, and prompt tuning improves stability and generalizability, their standalone applications exhibit inherent limitations. While adapters have shown promise in continual learning, they struggle with rapid task specialization, as they require a certain amount of data to effectively adapt to new tasks. On the other hand, prompt tuning exhibits a slower rate of adaptation to new tasks, as the prompts need to capture task-specific information while maintaining the model's stability. This slow adaptation can lead to homogeneity in the learned representations, limiting the model's ability to distinguish between different tasks (Gao et al., 2023). DPAT addresses these limitations by combining the strengths of both adapters and prompt tuning in a decoupled training strategy, allowing for efficient task specialization while preserving the model's stability and generalization capabilities. By integrating adapters, DPAT exhibits enhanced adaptability to current tasks, achieving improved spatial-temporal adaptation to new tasks. Incorporation of learnable prompts further augments the model's stability, making it less susceptible to forgetting. Moreover, we employ an enhanced decoupled training strategy for prompts and adapters, leveraging the strengths of both components. This strategy streamlines the adaptation to new activities and fortifies generalization, while significantly lowers both computational and memory requirement.

Contributions: Our contributions are outlined as follows:

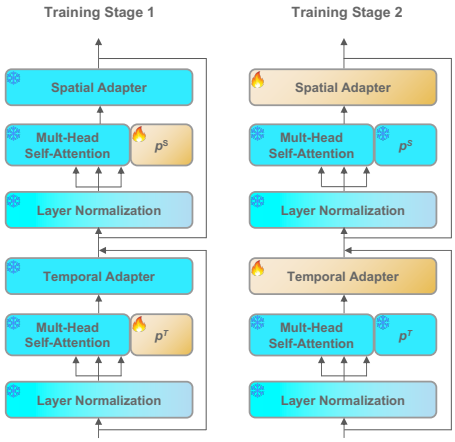
- **(i)** We introduce **D**ecouple **P**rompt-**A**dapter **T**uning (DPAT), a framework developed to improve the performance of pretrained image encoders in the context of Continual Action Recognition.
- **(ii)** By leveraging the intrinsic capabilities of pre-trained Vision Transformer (ViT) in conjunction with adapters and prompt tuning, we provide a robust solution that ensures the retention of previously acquired knowledge while seamlessly adapting to new spatial-temporal tasks.
- **(iii)** We substantiate that the designed dual-stage training strategy employed for adapters and spatial-temporal prompts plays a crucial role in harmonizing the model's plasticity and generalization capabilities, which safeguard against the adverse effects of new information on previously mastered tasks.
- **(iv)** In our extensive experiments across challenging datasets, DPAT consistently achieve state-of-the-art performance, showcasing its superiority in handling the fine-grained action recognition tasks.

2 Related Work

Continual Learning. Continual Learning endeavors to mitigate the issue of catastrophic forgetting. Inspired by the hippocampus's replay mechanism in the human brain, memory replay methods adopt either the conservation of real samples for future replay (Rolnick et al., 2019; Rebuffi et al., 2017; Buzzega et al., 2020) or the employment of generative models to create samples that emulate previous task distributions (Shin et al., 2017; Ostapenko et al., 2019). These methodologies, while effective, are constrained by substantial storage requirements and the complexity inherent in producing high-fidelity synthetic samples, especially when dealing with high-dimensional data such as video content. On the other hand, regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Zenke et al., 2017; Li & Hoiem, 2017) present an alternative strategy aimed at preserving essential weight configurations from preceding tasks. This is achieved through an array of analytical instruments, including the Fisher information matrix, gradient assessments, and uncertainty metrics, which are employed to evaluate and rank the significance of weights. These approaches, which obviate the need for the storage of additional data, offer a distinct advantage in terms of efficiency and privacy. However, empirical evidence suggests that excessive regularization can impair the model's ability to generalize effectively across complex tasks. This limitation underscores the delicate balance required in



(a) Depiction of the model architecture designed for continual video action recognition, leveraging spatial and temporal adapters in conjunction with two prompts ($\boldsymbol{p}^S$ and $\boldsymbol{p}^T$), all operating within a frozen Multi-Head Self-Attention (MSA) in pre-trained ViT. Meanwhile, the inclusion of a learnable key (k_t), utilized in the model's key-query matching mechanism, facilitates optimal prompt selection and mitigates forgetting during inference.



(b) Demonstration of our decoupled training strategy: we firstly freeze the adapter to establish a stable and generalizable foundation through Prefix tuning, then refine task-specific capabilities by focusing on adapter tuning while preserving the initial prompts for balanced adaptation and generalization.

Figure 1: Overview of the proposed Decoupled Prompt-Adapter Tuning (DPAT) approach: (a) Model architecture integrating adapters and prefix prompts to facilitate adaptation to new tasks; (b) Decoupled training paradigm designed to bolster knowledge preservation through phase-separated optimization of the model components

the application of regularization techniques, ensuring that the preservation of prior knowledge does not come at the expense of the model's adaptability and learning capacity for future tasks.

Continual Learning for Large Pre-trained Model. Recently, the emergence of large pre-trained models has catalyzed the development of prompt-based approaches (Wang et al., 2022a;b; Smith et al., 2023) in the realm of continual learning (CL). Pioneering works such as DualPrompt (Wang et al., 2022a) and L2P (Wang et al., 2022b) leverage a minimal set of trainable prompts instead of directly modifying the model's encoder parameters. These methods align input data with appropriate prompts via a task-agnostic, local clustering-like optimization process by assembling a pool of prompts from which selections are made. On the other hand, in the sphere of prompt-based learning for Vision-Language Models (VLMs), significant strides have been made with contributions from works like ProGrad (Zhu et al., 2023), CoCoOp (Zhou et al., 2022a), and CoOp (Zhou et al., 2022b). These efforts have introduced strategies to mitigate forgetting and enhance the adaptability of prompts for downstream tasks. ProGrad, for example, unveils a progressive prompt training strategy that gradually escalates prompt length. CoCoOp employs a conditional framework for generating input-dependent prompts, while CoOp embraces a cooperative learning methodology to simultaneously optimize image and text prompts. These techniques provide valuable perspectives on employing prompt learning to boost VLMs' performance within continual learning scenarios. However, it is worth noting that the application of these prompt tuning methods has predominantly been explored within the context of static image learning. The extension of these methods to video processing remains an area ripe for exploration.

Continual Learning for Activity Recognition. Several studies have addressed the challenges of temporal dynamics and complexity in video data to improve continual learning frameworks' knowledge retention across video tasks. Park et al. (2021) leveraged time-channel information for weighted knowledge distillation, aiming to better encode temporal dynamics and combat forgetting. Villa et al. (2023) introduced the PIVOT model, which enhances temporal modeling and mitigates forgetting through the integration of spatial prompts, memory replay, and a transformer encoder, significantly boosting video classification accuracy. Meanwhile, Pei et al. (2022) developed a memory-efficient approach

by creating a condensed frame representation for each representative video from previously seen classes. While these methods demonstrate innovative techniques to mitigate forgetting and manage memory, they share a reliance on storing samples, indicating that their strategies are not rehearsal-free. Conversely, ST-Prompt (Pei et al., 2023) presented a novel rehearsal-free approach in video continual learning by leveraging pre-trained vision-language models with temporal prompts for temporal information encoding. While this approach, necessitating additional inputs from a text encoder in CLIP (Radford et al., 2021) for distribution alignment, may struggle to capture videos with intricate temporal dynamics solely through prompts, our methods circumvents the need for a text encoder and instead employs adapters to bolster temporal modeling capability

3 PRELIMINARY

3.1 CONTINUAL ACTION RECOGNITION

In the domain of action recognition, continual learning frameworks are tailored to incrementally adapt to a series of data streams, denoted as $\{D_1, D_2, \ldots, D_T\}$. Each stream D_t at stage t comprises N_t labeled video clips $\{(v_t^b, y_t^b)\}_{b=1}^{N_t}$, where v_t^b, a video clip, is characterized within $\mathbb{R}^{T\times H\times W\times C}$, and y_t^b represents the label of the video clip v_t^b. In this context, v_t^b is defined in a four-dimensional space, with T representing the number of frames, H the height, W the width, and C the number of channels. The label y_t^b is a categorical variable that identifies the class of the video clip v_t^b. In the class incremental learning framework, distinct, non-overlapping class groups are introduced at each stage (Y_t), with the constraint that $Y_i \cap Y_j = \emptyset$ for any $i \neq j$. This setup mandates that the model, f_θ, not only assimilates new data from the current dataset D_t while retaining the ability to accurately classify across an expanding set of classes, encapsulated as $\tilde{Y}_t = \bigcup_{i=1}^{t} Y_i$. The primary objective is to optimize a singular model's average classification accuracy across all tasks.

3.2 CONTINUAL LEARNING WITH PREFIX-TUNING

In the realm of continual learning, Prefix-tuning (Li & Liang, 2021) has become a pivotal strategy for adapting Transformer models to new tasks with minimal retraining effort. Let the input to the Multi-Head Self-Attention (MSA) layer be $\boldsymbol{h} \in \mathbb{R}^{L\times D}$, and we further denote the input query, key, and values for the MSA layer to be $\boldsymbol{h}_Q$, $\boldsymbol{h}_K$, and $\boldsymbol{h}_V$, respectively. The prompt parameter $\boldsymbol{p} \in \mathbb{R}^{L_p\times D}$ is divided into $\{\boldsymbol{p}^K, \boldsymbol{p}^V\} \in \mathbb{R}^{\frac{L_p}{2}\times D}$ and prepended to $\boldsymbol{h}_K$, and $\boldsymbol{h}_V$ in the MSA as:

$$f_{\text{Pre-T}}\left(\boldsymbol{p}, \boldsymbol{h}\right) = \text{MSA}\left(\boldsymbol{h}_Q, \left[\boldsymbol{p}^k; \boldsymbol{h}_K\right], \left[\boldsymbol{p}^v; \boldsymbol{h}_V\right]\right). \tag{1}$$

To address catastrophic forgetting during Prefix tuning, state-of-the-art methods such as L2P (Wang et al., 2022b) and DualPrompt (Wang et al., 2022a) utilize a key-value pair query strategy for dynamically selecting instance-specific prompts from a pool. Each prompt p_m is linked to a learnable key k_m, with M denoting the pool's size, selected based on cosine similarity against an input-conditioned query $q(x)$, thereby identifying the key k_m with the highest similarity $\gamma(q(x), k_m)$. During the test, the prompt embedding task-specific information is selected by $\arg\min_m \gamma(q(x), k_m)$, ensuring that the keys, embedded with task-specific knowledge during training, are precisely matched to the input for inference. However, these approaches were primarily tested for dealing with static images and do not accommodate the temporal information crucial for video action recognition, rendering them inadequate for such applications, whereas our proposed model seeks to address such limitations by incorporating additional adapters.

4 METHOD

Our proposed approach is illustrated in Figure 1. We begin by explaining the configuration of the adapter and prompt in our model in Section 4.1, where we discuss how these components work with the MSA layer in the pretrained image encoder. In Section 4.2, we describe our decoupled training process. In Section 4.3, we outline the redesigned query-key matching loss. Finally, we detail the training objectives in Section 4.4.

4.1 POSITION OF ADAPTER AND PROMPT

Figure 1a illustrates our approach, diverging from recent methods that append an additional temporal model atop the ViT backbone to capture temporal dynamics. Instead, we deploy adapters to refine the model's inherent processing capabilities. Specifically, our model incorporates a spatial adapter, *Adapter-S*, with a spatial prompt $\boldsymbol{p}^S$, and a temporal adapter, *Adapter-T*, with a temporal prompt $\boldsymbol{p}^T$. This configuration maintains the original functionality of the MSA layer within the pre-trained image encoder, leveraging its inherent strengths while minimizing modifications.

Expanding upon Villa et al. (2023)'s demonstration of repurposing pre-trained image encoders for temporal analysis, our approach distinctively augments the pre-trained image model with adapters and prompts. This not only facilitates exhaustive extraction of spatial and temporal information with adapter but also leverages prompts to mitigate model forgetting and enhance stability. Following the dual-prompt approach, $\boldsymbol{p}^T$ includes a task-agnostic prompt, $\boldsymbol{g}^T$, positioned at a shallower layer to capture task-independent information, which is learned and shared across all tasks, and a task-specific prompt, $\boldsymbol{e}^T$), aimed at collecting task-related information. Similarly, $\boldsymbol{p}^S$ is structured with its task-agnostic component, $\boldsymbol{g}^S$, and task-specific component, $\boldsymbol{e}_S$. Furthermore, for any task t, the task-specific prompt $\boldsymbol{e}_t = \{\boldsymbol{e}_t^T, \boldsymbol{e}_t^S\}$ is associated with a task-specific key, $\boldsymbol{k}_t$, a learnable parameter designed to capture the distinctive features of task t. During inference, a pre-defined query function q is employed to perform key-query matching to facilitate the selection of the appropriate task-specific prompt to be prepended to the model.

Specifically, upon receiving a video patch embedding, $\boldsymbol{z} \in \mathbb{R}^{T\times(N+1)\times D}$, it is first reshaped into $\boldsymbol{z}^T \in \mathbb{R}^{(N+1)\times T\times D}$, where T represents the temporal dimension, or the number of frames. This reshaped embedding, $\boldsymbol{z}^T$, is subsequently fed into a pre-trained MSA layer, enabling it to learn the complex relationships among the T frames. Concurrently, we employ Prefix tuning by prepending a prompt parameter $\boldsymbol{p}^T$ to the MSA layer. Following the extraction of temporal features, the output from the temporal adapter is reshaped back to $\mathbb{R}^{T\times(N+1)\times D}$ to enable the extraction of spatial features. This reshaped output is then processed by the spatial adapter, which is specifically designed to enhance the model's capabilities in spatial analysis. Similar to the approach taken with the temporal domain, we prepend a learnable prompt $\boldsymbol{p}^S$ to the spatial adapter's output. Given the aforementioned model structure, the forward process in our proposed model can be written as:

$$\begin{aligned} h_T^l &= h^{l-1} + \textit{Adapter-T}\left(f_{\text{Pre-T}}(\boldsymbol{p}^T, \text{LN}(h^{l-1}))\right), \\ h_S^l &= h_T^l + \textit{Adapter-S}\left(f_{\text{Pre-T}}(\boldsymbol{p}^S, \text{LN}(h_T^l))\right), \end{aligned} \tag{2}$$

where h_T^l and h_S^l denote the output after temporal and spatial feature extraction, respectively. LN is the layer normalization. The function $f_{\text{Pre-T}}$ is defined as in Equation 1.

4.2 Decoupled Prompt-Adapter Tuning

Despite the theoretical advantage of adapters being less susceptible to catastrophic forgetting compared to traditional fine-tuning, their standalone application does not completely circumvent the challenges associated with rapid task specialization. Conversely, employing Prefix tuning directly, although it enhances model stability and generalizability, it has been observed that Prefix tuning exhibits a slower adaptation rate to new tasks and is prone to homogeneity (Gao et al., 2023), thereby constraining the adaptability required for varied and intricate tasks. These observations form the core motivation for our proposed strategy, Decoupled Prefix Prompt and Adapter Tuning, which aims to harness the complementary strengths of both adapter and Prefix tuning in a unified framework.

Figure 1b illustrates the proposed decoupled training strategy, the training process is delineated into two distinct phases, meticulously designed to balance adaptability with generalizability:

First Stage: Prefix Tuning. Initially, Prefix tuning is employed to provide the model with a stable and generalizable base. Learnable prompts encapsulate the task-specific information, reducing the immediate adaptation pressure for adapter and providing the model with a robust understanding of the task. This stage is crucial for setting the stage for specialized adaptation.

Second Stage: Adapter Tuning. Subsequently, the focus shifts to adapter tuning, emphasizing task-specific refinement and adaptation. By maintaining the prompt learned in the initial phase, we preserve the generalization and stability benefits of Prefix tuning, while the adapter module targets task-specific learning. This approach aims to strike a balance between rapid adaptation and maintaining the model's ability to generalize, addressing the limitations of employing either tuning strategy in isolation.

4.3 Redesigned Query-Key Matching loss

In the DualPrompt (Wang et al., 2022a), the matching loss is formalized as $\mathcal{L}_{\text{match}}\left(\boldsymbol{x}, \boldsymbol{k}_t\right) = \gamma\left(q(\boldsymbol{x}), \boldsymbol{k}_t\right)$, with γ acting as a distance metric and $q(\boldsymbol{x})$ as a query function. This design aims to minimize the distance between $\boldsymbol{k}_t$ and the query representation of $\boldsymbol{x}$, thus enhancing the affinity between task-specific keys and inputs from corresponding tasks. Nonetheless, this initial formulation overlooks the inter-task relationships, focusing solely on the proximity to a single task key without considering the influence of other task keys. To address this limitation, we propose an enhancement through normalization of similarity scores using a softmax function to ensure that the model's predictions are influenced not only by the nearest task key but also by the relative similarity to all task keys. The revised matching

loss incorporating softmax normalization is given by:

$$\mathcal{L}_{\text{match}}\left(\boldsymbol{x}, \boldsymbol{k}_t\right) = -\log\left(\frac{e^{\frac{-\gamma\left(q(\boldsymbol{x}), \boldsymbol{k}_t\right)}{\tau}}}{\sum_{i=1}^{t} e^{\frac{-\gamma\left(q(\boldsymbol{x}), \boldsymbol{k}_i\right)}{\tau}}}\right). \tag{3}$$

In this updated formula, the temperature factor τ is introduced to adjust the model's sensitivity to variations in the distance metric, thereby refining the perception of distances between query and key. Furthermore, the modification includes a normalization component in its denominator, aggregating the probabilities that the input $\boldsymbol{x}$ aligns with each of the task-specific keys $\boldsymbol{k}_i$. This enhancement facilitates a more balanced and thorough evaluation of the input's affiliation with all tasks, optimizing the alignment between task-specific inputs and their corresponding keys and addressing the initial formulation's limitation by considering the relative similarity to all task keys.

Similar to the DualPrompt approach, for evaluating a test example x, we select the most appropriate task key index by identifying the minimum distance between $q(\boldsymbol{x})$ and $\boldsymbol{k}_t$, as determined by the criterion $\text{argmin}_t \gamma(q(\boldsymbol{x}), \boldsymbol{k}_t)$.

4.4 Training Objective

The comprehensive training and testing processes are outlined in Algorithm 1 and Algorithm 2, respectively. The objective for the two-stage decoupled training is formulated as follows:

$$\begin{aligned} &\text{Stage 1:} \quad \min_{\boldsymbol{p}^S, \boldsymbol{p}^T, \phi} \mathcal{L}\left(f_\phi\left(f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(\boldsymbol{x})\right), y\right), \\ &\text{Stage 2:} \quad \min_{\theta_T, \theta_S, \boldsymbol{k}_t, \phi} \mathcal{L}\left(f_\phi\left(f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(\boldsymbol{x})\right), y\right) + \lambda \mathcal{L}_{\text{match}}\left(\boldsymbol{x}, \boldsymbol{k}_t\right), \end{aligned} \tag{4}$$

where f_ϕ denotes the classification head parametrized by ϕ, $f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(\boldsymbol{x})$ represents the forward process of the adapted, frozen pretrained vision model. It integrates spatial-temporal adapters and prompts with parameters $\boldsymbol{\theta}_T, \boldsymbol{\theta}_S$ and $\boldsymbol{p}^S, \boldsymbol{p}^T$, respectively, as described in Equation 2. $\mathcal{L}$ is the cross-entropy loss, $\mathcal{L}_{\text{match}}$ is the matching loss defined in Equation 3, and λ is a scalar balancing factor.

5 Experiments

We first compare the performance of DPAT against baseline models on several benchmarks for action recognition in Section 5.2. We then conduct ablation studies to highlight the significance of each component of DPAT in Section 5.3. Furthermore, Furthermore, additional experiments on extra datasets and ablation studies on hyperparameters, such as prompt positioning and bottleneck ratio, are detailed in Appendix A.3.

5.1 Experiments Settings

Datasets. We evaluate our method across three public datasets: Kinetics-400 (Kay et al., 2017) and ActivityNet (Caba Heilbron et al., 2015) for standard action recognition, alongside EPIC-Kitchens-100 (Damen et al., 2021) for fine-grained action recognition. Following a strategy akin to the vCLIMB (Villa et al., 2022) benchmark guidelines, we organize the data into a class-incremental setting. In this arrangement, each dataset's classes are introduced sequentially across a series of 10 tasks, with careful measures taken to prevent class overlap. This structure rigorously evaluates our method's adaptability and learning evolution in a systematic, incremental manner. Further details are provided in Appendix A.1.

Evluation Metrics. We use the widely recognized evaluation metrics of average accuracy (Acc) and backward forgetting (BWF) to measure the model's overall performance across tasks and the extent to which it retains knowledge from previous tasks after learning new ones, respectively, as defined below:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} R_{N,i}, \qquad \text{BWF} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left(R_{i,i} - R_{N,i}\right), \tag{5}$$

where N represents the total number of tasks, $R_{N,i}$ denotes the accuracy of the model on task i after learning all N tasks, and $R_{i,i}$ represents the accuracy of the model on task i immediately after learning it.

Implementation. We leverage the ViT-B/16 architecture (Dosovitskiy et al., 2020), pre-trained on the ImageNet-21K (IN-21K) dataset, as the backbone for DPAT. To adapt to task-specific requirements while maintaining the integrity of the pre-trained features, we freeze the backbone and sequentially fine-tune adapters and prompts. Optimization is

Algorithm 1 DPAT at Training Stage with Decoupled Prompt and Adapter Optimization

Require: Pre-trained ViT backbone f, classification head f_ϕ, number of tasks N, training set $\boldsymbol{D} = \{D_t\}_{t=1}^{T}$.
Require: Temporal prompt $\boldsymbol{p}^T = \{\boldsymbol{g}^T, \boldsymbol{e}^T\}$ that contains task-agnostic prompt g^T and task-specific prompt $\{\boldsymbol{e}_t^T\}_{t=1}^{N}$, spatial prompt $\boldsymbol{p}^S = \{\boldsymbol{g}^S, \boldsymbol{e}^S\}$, a pool of task keys $\mathbf{K} = \{\boldsymbol{k}_t\}_{t=1}^{N}$.
Require: Adapter-S parameterized with θ_S, Adapter-T parameterized with θ_T.
Require: Initialize $\phi, \boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S, \mathbf{K}$.
for $t = 1$ **to** N **do**
  Select the task-specific prompt $\boldsymbol{e}_t = \{\boldsymbol{e}_t^T, \boldsymbol{e}_t^S\}$ and corresponding task key $\boldsymbol{k}_t$.
  Generate the prompted architecture $f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}$: attach temporal prompt $\boldsymbol{p}^T = \{\boldsymbol{g}^T, \boldsymbol{e}_t^T\}$ and spatial $\boldsymbol{p}^S = \{\boldsymbol{g}^S, \boldsymbol{e}_t^S\}$ at specified locations within the backbone model f, adapting the structure to incorporate spatial and temporal prompts effectively.
  for $e = 1$ **to** M_t **do**
    Draw a mini-batch $B = \{(v_t^b, y_t^b)\}_{b=1}^{|B|}$ from D_t.
    for all $(\boldsymbol{x}, y)$ in B **do**
      Calculate the adapted feature by $f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(x)$ ▷ Stage 1 Optimization
      Calculate the per sample loss $\mathcal{L}_{\S}$ via Equation. 4
    end for
    Update $\phi, \boldsymbol{p}^S, \boldsymbol{p}^T$ by backpropagation.
  end for
  for $e = 1$ **to** M_t **do**
    Draw a mini-batch $B = \{(v_t^b, y_t^b)\}_{b=1}^{|B|}$ from D_t
    for all $(\boldsymbol{x}, y)$ in B **do**
      Calculate the adapted feature by $f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(\boldsymbol{x})$ ▷ Stage 2 Optimization
      Calculate the per sample loss via Equation. 4
    end for
    Update $\boldsymbol{\theta}_T, \boldsymbol{\theta}_S, \boldsymbol{k}_t, \phi$ by backpropagation.
  end for
end for

Algorithm 2 DPAT at Testing Stage

Require: Pre-trained ViT backbone f, classification head f_ϕ
Require: Temporal prompt $\boldsymbol{p}^T = \{\boldsymbol{g}^T, \boldsymbol{e}^T\}$, spatial prompt $\boldsymbol{p}^S = \{\boldsymbol{g}^S, \boldsymbol{e}^S\}$, a pool of task keys $\mathbf{K} = \{\boldsymbol{k}_t\}_{t=1}^{N}$.
Require: Adapter-S parameterized with θ_S, Adapter-T parameterized with θ_T.
Require: Test sample $\boldsymbol{x}$
Generate query feature $q(\boldsymbol{x})$
$t_x = \arg\min_t \gamma(q(\boldsymbol{x}), \boldsymbol{k}_t)$ ▷ Matching for the index of task-specific prompt
Select the task-specific Prompt $\boldsymbol{e}_{t_x} = \{\boldsymbol{e}_{t_x}^T, \boldsymbol{e}_{t_x}^S\}$
Generate the prompted architecture $f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}$: attach temporal prompt $\boldsymbol{p}^T = \{\boldsymbol{g}^T, \boldsymbol{e}_{t_x}^T\}$ and spatial $\boldsymbol{p}^S = \{\boldsymbol{g}^S, \boldsymbol{e}_{t_x}^S\}$ at specified locations within the backbone model f, adapting the structure to incorporate spatial and temporal prompts effectively.
Predict with $f_\phi\left(f_{\boldsymbol{p}^T, \boldsymbol{p}^S, \boldsymbol{\theta}_T, \boldsymbol{\theta}_S}(\boldsymbol{x})\right)$

performed using the Adam (Kingma & Ba, 2014) optimizer, with a batch size of 64 across 50 epochs for each task. We set the bottleneck ratios for spatial and temporal adapters at 0.25 and the scaling factor, λ, to 1. To ensure robustness and generalizability of our findings, we select benchmark methods that operate under analogous conditions, facilitating a rigorous and fair comparison.

5.2 Comparison with baseline

The comparative analysis includes EWC (Kirkpatrick et al., 2017), MAS (Aljundi et al., 2018), iCaRL (Rebuffi et al., 2017), and PIVOT (Villa et al., 2023). EWC penalizes changes to important weights to preserve knowledge, requiring significant computational resources. MAS uses synaptic plasticity to balance knowledge preservation and adaptation. iCaRL retains examples of previous classes for replay, improving accuracy but increasing memory usage. PIVOT leverages prompts for video continual learning, achieving high performance with reduced forgetting. To ensure a fair

comparison, we enhanced EWC, MAS, and iCaRL with a tunable adapter and used the CLIP (Radford et al., 2021) ViT-B/16 model for PIVOT. Our method, DPAT, combines adapters for efficient task adaptation and prompt tuning to mitigate forgetting, resulting in superior performance with lower computational and memory costs.

Results on Kinetics-400 and ActivityNet. As illustrated in Table 1, DPAT (IN-21K), employing a ViT-B/16 model pre-trained on ImageNet-21K, substantially surpasses traditional rehearsal-free methodologies MAS and EWC. Further analysis reveals that against the state-of-the-art PIVOT, DPAT (CLIP), which also utilizes the same CLIP ViT-B/16 backbone as PIVOT, secures higher prediction accuracy and demonstrates reduced backward forgetting on both Kinetics-400 and ActivityNet datasets, all achieved without relying on replayed video instances. This enhanced performance not only highlights DPAT's adeptness at preserving previously acquired knowledge amidst new data assimilation but also signifies its profound impact on advancing continual learning paradigms.

Results on EPIC-Kitchens-100. Table 2 showcases the comparative results on the EPIC-Kitchens-100 dataset. DPAT notably excels in verb prediction against baseline models, underscoring the significant capability of the temporal adapter design to capture temporal information effectively. Furthermore, the reduced backward forgetting observed across models highlights how our innovative prompt design, incorporating a key-query matching strategy, effectively mitigates forgetting in temporal modeling contexts. Although the prediction accuracy for nouns is slightly lower than that of PIVOT, our model still demonstrates the best overall performance in action prediction. This nuanced balance between temporal and spatial knowledge preservation, despite the slight trade-off, underscores DPAT's robustness and adaptability, making it a competitive choice for continual activity recognition in dynamic environments such as those presented by the EPIC-Kitchens-100 dataset.

Table 1: Comparative Evaluation Metrics for Models on Kinetics-400 and ActivityNet Datasets: Memory Usage (Mem. (RI), the number of Replayed Instances), Accuracy (Acc ↑, higher is better), and Backward Forgetting (BWF ↓, lower is better). DPAT outperforms existing rehearsal-free methods EWC and MAS, replay-based iCaRL, and the PIVOT when employing the same CLIP ViT-B/16 backbone in terms of accuracy and backward forgetting.

Model	**Kinetics-400**			**ActivityNet**		
	Mem. (RI)	Acc ↑	BWF ↓	Mem. (RI)	Acc ↑	BWF ↓
iCaRL	8000	48.7%	30.3%	4000	60.9%	15.72%
PIVOT	4000	56.1%	25.7%	2000	73.6%	11.1%
EWC	0	25.2%	15.1%	0	11.3%	13.1%
MAS	0	23.8%	11.6%	0	10.4%	5.5%
DPAT (IN-21K)	0	58.5%	24.7%	0	71.9%	12.6%
DPAT (CLIP)	0	**61.3%**	22.3%	0	**74.5%**	11.2%
Upper Bound	0	84.1%	-	0	89.3%	-

Table 2: Performance Analysis on the EPIC-Kitchens-100 Dataset: Accuracy (Acc) and Backward Forgetting (BWF) Metrics for Verb, Noun, and Action Predictions, where an action is deemed correctly predicted only if both the verb and noun components are accurately identified.

Model	**Mem. (RI)**	**Verb**		**Noun**		**Act**	
		Acc ↑	BWF ↓	Acc ↑	BWF ↓	Acc ↑	BWF ↓
iCaRL	5940	39.7%	25.2%	30.8%	19.6%	17.9%	16.4%
PIVOT	2970	46.1%	20.2%	**43.5%**	16.8%	29.8%	13.2%
EWC	0	18.1%	10.1%	11.1%	8.3%	5.6%	1.1%
MAS	0	19.6%	9.3%	13.3%	7.7%	6.7%	0.9%
DPAT (IN-21K)	0	51.1%	19.1%	38.1%	17.7%	30.3%	12.6%
DPAT (CLIP)	0	**53.8%**	18.6%	42.3%	17.1%	**32.3%**	11.5%
Upper Bound	0	65.3%	-	56.1%	-	42.8%	-

5.3 Ablation Studies

In this section, we conduct ablation studies to scrutinize the characteristics and efficacy of our fundamental design elements.

Effect of Model Component. In the ablation studies summarized in Table 3, we systematically assess the contribution of each component to our model's proficiency in continual action recognition, utilizing the Kinetics-400 dataset for our experiments. The absence of the Temporal Adapter leads to a substantial decrease in accuracy by 38.2%, underscoring its critical role in capturing temporal dynamics crucial for precise action recognition. The complete removal of all adapters results in a marked decrease in accuracy to 25%, indicating a further decline in learning performance. This underscores the critical role of adapters in adapting to new tasks. Interestingly, this configuration leads to minimal forgetting, suggesting that a combination of prompt learning with a frozen pre-trained model provides a stable foundation, even as it points to the necessity of adapters for effective task adaptation. In contrast, the exclusion of the Prefix Prompt detrimentally impacts accuracy and leads to the highest BWF rate among the configurations tested, emphasizing the prompt's vital importance in bolstering the model's learning stability and its ability to mitigate forgetting efficiently. Derived from experiments on the Kinetics-400 dataset, these findings highlight the intricate balance between leveraging adapters for generalization and prompt tuning for stability. This balance underscores the necessity of designing both components within our model to navigate the challenges of continual learning effectively.

Table 3: Results of ablation studies on model components

Training Method	Acc ↑	BWF ↓
DPAT	**61.3**%	22.3%
Ablate Temporal Adapter	33.1%	19.5%
Ablate all Adapter	25.0%	5.1%
Ablate Task-agnostic Prefix	58.7%	24.1%
Ablate All Prefix Prompt	48.2%	33.6%

Effect of Decoupled Training. To illustrate the efficacy of our decoupled training strategy, we present the learning curve comparison between decoupled and joint training in Figure 2. The x-axis represents the current task, while the y-axis measures the average accuracy (Acc). Conducting experiments on the Kinetics-400 dataset, our findings reveal that although both strategies commence with comparable accuracy levels, the decoupled training demonstrates a trend of slower degradation in mean accuracy alongside reduced backward forgetting. This highlights our model's enhanced capability to preserve previously learned knowledge over time.

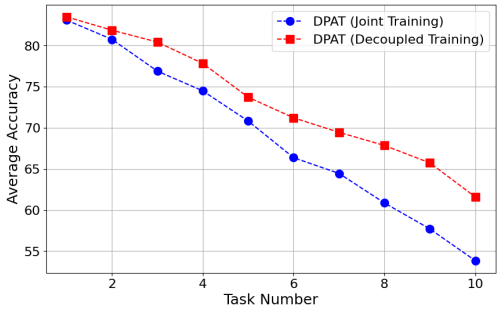


Figure 2: Comparative Result of DPAT with Joint and Decoupled Training Strategies on Kinetics-400

Effect of Query Matching Loss. Table 4 highlights DPAT's superiority over the DualPrompt model, with DPAT achieving a Matching Accuracy of 45.6% compared to DualPrompt's 32.8% on Kinetics-400. This improvement underscores our optimized matching loss's efficacy in enhancing task-specific contrast, allowing for more precise task-specific key embeddings. Notably, the elevation in Matching Accuracy corresponds with a significant boost in Prediction Accuracy and a reduction in Backward Forgetting, showcasing the direct impact of improved alignment on model performance and memory retention.

Table 4: Effect of Different Query Matching Loss

Query Loss	Matching Acc	Acc ↑	BWF ↓
DualPrompt	32.8%	57.2%	25.5%
DPAT	**45.6**%	61.3%	22.3%

6 CONCLUSION

In this paper, we present a simple yet effective rehearsal-free approach for continual activity recognition. Eliminating the necessity for integrating auxiliary temporal architectures, relying on external modal inputs, or undertaking extensive fine-tuning, our architecture employs adapters in conjunction with prompt tuning to excel in the realm of continual action recognition, leveraging a frozen pre-trained model. Furthermore, We introduce a decoupled training strategy that capitalizes on the adapter's generalization capabilities and the stability provided by prompt tuning, effectively mitigating the issue of forgetting. Evaluations on various challenging benchmarks for continual action recognition indicate that our model performs well in comparison to existing methods.

Despite the considerable advantages offered by our methodology, it is not without its limitations. The reliance of our task-specific prompt design on predefined task boundaries precludes its application in an online learning context. Our approach currently operates within a closed-set classification paradigm, transitioning to a more realistic open-set continual action recognition setting stands as a promising avenue for future research.

ACKNOWLEDGMENTS

This research is supported by an Academic Research Grant No. MOE-T2EP20121-0015 from the Ministry of Education in Singapore.

REFERENCES

Sharath Chandra Akkaladevi and Christoph Heindl. Action recognition for human robot interaction in industrial applications. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pp. 94–99. IEEE, 2015.

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)*, pp. 139–154, 2018.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

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.

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.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021. doi: 10.1109/TPAMI.2020.2991965.

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.

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.

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.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, 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*, pp. 11321–11329, 2019.

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*, pp. 13698–13707, 2021.

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.

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*, pp. 11932–11942, 2023.

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.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

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, 2023.

Steven Vander Eeckt and Hugo Van Hamme. Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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*, pp. 19035–19044, 2022.

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*, pp. 24214–24223, 2023.

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, 2022a.

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, 2022b.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

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*, pp. 16816–16825, 2022a.

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, 2022b.

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*, pp. 15659–15669, 2023.

A Appendix

A.1 Datasets

Kinetics-400. Kinetics-400 stands as an extensive dataset tailored for action recognition, encompassing roughly 300k video snippets classified into 400 distinct human action categories. Originating from a wide variety of YouTube videos, each clip is meticulously trimmed to approximately 10 seconds to ensure uniformity. For the purposes of this work, the dataset is partitioned into 10 separate tasks, with each task consists of 40 unique action classes, arranged in a sequential, class-incremental manner.

ActivityNet. ActivityNet is a large-scale dataset designed for action recognition, featuring over 20,000 video clips spread across 200 activity classes. Sourced from YouTube, it offers a wide-ranging and diverse collection of real-world scenarios that capture a broad spectrum of everyday human activities. Similar to Kinetics-400, we organize the dataset into a class-incremental setting with 10 tasks, ensuring a neat division of 20 activities per task.

Epic-Kitchen-100. Epic-Kitchen-100 is a large-scale egocentric video dataset that records over 100 hours of kitchen unscripted activities. It consists of 90K action segments, which are split into train/val/test sets of 67K/10K/13K. Differ from preceding two datasets, Epic-Kitchen-100 defines an action as a combination of a verb and a noun. Given the necessity to match both verbs and nouns for action recognition, this task presents a higher level of complexity compared to action recognition in previous datasets characterized by single label prediction. We employ a class-incremental strategy for verb classification, dividing 97 verb categories into 10 non-overlapping tasks to systematically introduce new classes. Concurrently, noun prediction is approached with a task-incremental strategy, wherein all tasks share a consistent set of total 300 noun classes. This deliberate division, prioritizing verbs for the class-incremental learning setting, stems from our intent to scrutinize the model's ability to discern and learn from the nuanced temporal dynamics across different tasks.

A.2 Implementation Details

Following the ViT architecture guidelines, we sample videos to 16 frames. Each frame is then randomly cropped and resized to 224×224 pixels. Additionally, we enhance the diversity of the training data by applying data augmentation techniques, including mixup (Zhang et al., 2017), label smoothing, horizontal flipping, color jittering, and RandAugment (Cubuk et al., 2020). Specifically, for mixup, an alpha of 0.2 was utilized. Label smoothing was implemented using a factor of 0.1. In the color jittering process, brightness, contrast, and saturation adjustments were uniformly

set to 0.4, with hue adjustments at 0.1. We deployed RandAugment with a configuration of 2 transformations at a magnitude level of 10. Finally, to ensure consistency across diverse datasets, we normalize each frame to the range $[0, 1]$

As depicted in Figure 1a, our method, DPAT, integrates spatial and temporal adapters into every ViT block. The architecture of the adapter adheres to a bottleneck design, incorporating two fully connected (FC) layers separated by an activation layer. The design involves diminishing the input dimensionality through the initial FC layer, and subsequently restoring it via the second FC layer. The extent of dimensionality reduction is determined by a bottleneck ratio—defined as the quotient of the bottleneck to the input dimension—with a consistent ratio of 0.25 applied throughout the experiment.

Task-agnostic prompts, denoted as $\boldsymbol{g}^T$ and $\boldsymbol{g}^S$, are introduced in the first two blocks of the pre-trained ViT model. Conversely, task-specific prompts, represented by $\boldsymbol{e}^T$ and $\boldsymbol{e}^S$, target the third through fifth ViT blocks. Our analysis revealed an optimal setup consisting of task-specific prompts with a length of 5, alongside task-agnostic prompts extended to a length of 20.

DPAT employs a sequential optimization strategy, optimizing adapters and prompts using the Adam optimizer in two separate stages. We utilize a batch size of 64 and a total of 50 epochs for both training stage, starting with a base learning rate of 1×10^{-3} for prompt tuning and 3×10^{-4} for adapter tuning, followed by a cosine decay schedule. Moreover, the balancing scaling factor λ was selected from the set $\{0.01, 0.1, 1, 10\}$. Following a similar parameter search for the temperature scaling factor τ among $\{0.01, 0.05, 0.1, 0.5\}$, we determined that $\lambda = 1$ and $\tau = 0.1$ provide the optimal outcomes.

A.3 Additional Experiments and Analysis

In this section, we present additional experiments and ablation studies to further validate the effectiveness of our proposed DPAT approach and investigate the impact of various components and hyperparameters on its performance.

A.3.1 Ablation Studies on ActivityNet

To further validate the effectiveness of our proposed DPAT approach and the individual contributions of its components, we conducted additional ablation studies on the ActivityNet dataset. The results are summarized in Table 5, highlighting the importance of both the temporal adapter and the task-agnostic prefixes within our framework.

Table 5: Ablation studies on the ActivityNet dataset.

Method	**Acc↑**	**BWF ↓**
DPAT	**74.5**	11.2
Ablate Temporal Adapter	41.2	15.7
Ablate All Adapters	32.6	**6.8**
Ablate Task-agnostic Prefix	68.1	13.3
Ablate All Prefixes	59.4	19.5

The findings on the ActivityNet dataset corroborate those observed on the Kinetics-400 dataset, underscoring the critical role that both adapters and prefixes play in our methodology. The removal of the temporal adapter or all adapters results in a significant decline in accuracy, whereas the elimination of task-agnostic prefixes or all prefixes exacerbates backward forgetting and diminishes overall performance.

A.3.2 Experiments on UCF-101

Further experiments were conducted on the UCF-101 dataset to offer a broader evaluation of our approach. The outcomes, detailed in Table 6, show our method's performance in comparison to existing benchmarks.

While our DPAT approach did not outperform PIVOT on the UCF-101 dataset, the results were closely matched. The superior performance of PIVOT might be attributed to its use of a replay buffer, which plays a pivotal role in mitigating forgetting and enhancing model performance on this relatively simpler dataset. Notably, our method requires no additional memory for data storage, presenting an advantage in complex continual learning scenarios for which it was designed.

Table 6: Performance comparison on the UCF-101 dataset.

Model	**Mem. (RI)**	**Acc** ↑	**BWF** ↓
iCaRL	2020	82.1	12.3
PIVOT	1010	**94.1**	**3.7**
EWC	0	15.2	30.1
MAS	0	16.8	8.6
DPAT (IN-21K)	0	89.2	4.3
DPAT (CLIP)	0	92.8	3.9
Upperbound	–	96.8	–

A.3.3 Ablation Studies on Prompt Position

A parameter search was conducted to fine-tune the positioning of temporal and spatial prompts within the DPAT framework, with the starting position (start_g) set to 1 and the ending position (end_e) set to 5. We varied the ending position (end_g) and adjusted the starting position of the subsequent element (start_e) accordingly. The search results, provided in Table 7, validate our initial configuration choices, though it is acknowledged that due to the expansive search space, these findings may not represent the optimal configuration.

Table 7: Ablation studies on prompt position.

end$_g$	**Acc**
1	59.2
2	**61.3**
3	60.9
4	58.8

A.3.4 Ablation studies on Bottleneck Ratio

The bottleneck ratio, a critical hyperparameter in our DPAT framework, influences the trade-off between model capacity and efficiency. Experiments were conducted on the Kinetics-400 dataset to identify the optimal bottleneck ratio. Various ratios were tested, and their impacts on average accuracy and backward forgetting were assessed. As Table 8 indicates, a bottleneck ratio of 0.25 optimally balances model capacity with efficiency, leading to the best performance on this dataset.

Table 8: Ablation studies on bottleneck ratio on Kinetics-400.

Bottleneck Ratio	**Acc** ↑	**BWF** ↓
0.05	59.7	21.8
0.1	60.5	22.1
0.25	**61.3**	22.3
0.5	60.9	22.7