Empowering Large Language Model for Continual Video Question Answering with Collaborative Prompting

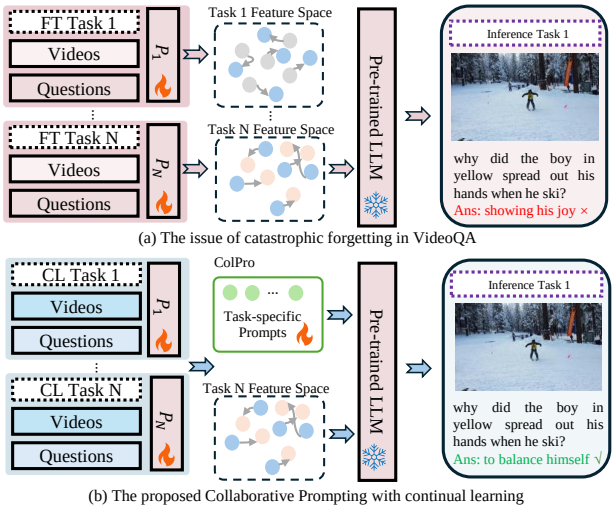
Chen Cai*, Zheng Wang*, Jianjun Gao, Wenyang Liu
Ye Lu, Runzhong Zhang, Kim-Hui Yap†
Nanyang Technological University
{e190210, zheng011, ekhyap}@ntu.edu.sg

Abstract

In recent years, the rapid increase in online video content has underscored the limitations of static Video Question Answering (VideoQA) models trained on fixed datasets, as they struggle to adapt to new questions or tasks posed by newly available content. In this paper, we explore the novel challenge of VideoQA within a continual learning framework, and empirically identify a critical issue: fine-tuning a large language model (LLM) for a sequence of tasks often results in catastrophic forgetting. To address this, we propose Collaborative Prompting (ColPro), which integrates specific question constraint prompting, knowledge acquisition prompting, and visual temporal awareness prompting. These prompts aim to capture textual question context, visual content, and video temporal dynamics in VideoQA, a perspective underexplored in prior research. Experimental results on the NExT-QA and DramaQA datasets show that ColPro achieves superior performance compared to existing approaches, achieving 55.14% accuracy on NExT-QA and 71.24% accuracy on DramaQA, highlighting its practical relevance and effectiveness.

1 Introduction

Video Question Answering (VideoQA) is critical for video understanding, involving training machine learning models to accurately respond to questions across various tasks (e.g., finding specific information Choi et al., 2021, counting objects Xiao et al., 2021, recalling actions Zhang et al., 2022) based on given video content. However, existing VideoQA models are typically trained on fixed datasets in static environments. With a continual increase in the number of videos on the internet every day, these static models may face challenges in answering new questions posed by the newly available content. One straightforward solution to overcome this challenge is to fine-tune the models when new data is introduced. However, this approach can lead to higher computational costs when retraining on all the data. Alternatively, fine-tuning only on newly available video question-and-answer pairs may lead to the catastrophic forgetting issue (McCloskey and Cohen, 1989), as shown in Figure 1(a).



(a) The issue of catastrophic forgetting in VideoQA

(b) The proposed Collaborative Prompting with continual learning

Figure 1: (a) Existing fine-tuning techniques train for different VideoQA tasks, which could lead to catastrophic forgetting, and generate inferior results. (b) We introduce the Collaborative Prompting (ColPro) within a continual learning framework, which retains task-specific knowledge to generate accurate answers, where $\mathbf{P}_N$ denotes a projection layer.

This motivates us to explore continual learning techniques (Rebuffi et al., 2017; Rolnick et al., 2019; Wang et al., 2022b) for VideoQA, facilitating ongoing fine-tuning of models across a sequence of data while mitigating catastrophic forgetting of previous tasks (e.g., finding information, or counting objects mentioned earlier), thereby addressing the needs of real-world dynamic environments. Recent continual learning techniques (Wang et al.,

*Equal contributions.
†Corresponding author.

2022b,a) have achieved good performance by employing rehearsal-free methods, such as learnable prompting (Jia et al., 2022) and prefix-tuning (Li and Liang, 2021). These approaches eliminate the need for memory-intensive stored experiences from previous tasks (Rolnick et al., 2019; Cha et al., 2021), reduce computation costs, and minimize forgetting. Specifically, L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), DBI (Qian et al., 2023), and ProgPrompt (Razdaibiedina et al., 2023) employ task-aware prompting techniques to fine-tune pre-trained models for downstream tasks using fewer learnable parameters. While these methods have improved performance in vision and language continual learning tasks, they often transfer either single-modal (text only) or multimodal (text and images) information from task to task. In terms of VideoQA, it is crucial to incorporate textual question context (**A1**), visual content (**A2**), and video temporal dynamics (**A3**) in the continual training setting. In this paper, we introduce Collaborative Prompting (ColPro), which explores these aspects for the VideoQA problem, and represents an area that has not been fully explored in prior research, as shown in Figure 1(b).

The core idea of ColPro is to empower a base model to achieve enhanced performance when transferring across a sequence of tasks. Inspired by the robust reasoning abilities of recent Large Language Models (LLMs), we instantiate the base model as a LLM (e.g., LLaMA Touvron et al., 2023) to generate accurate answers from textual questions and video inputs. Specifically, ColPro integrates three types of prompting techniques: task-specific question constraint prompting (TQCP), knowledge acquisition prompting (KAP), and visual temporal awareness prompting (VTAP), aimed at enhancing accuracy in answer prediction while minimizing forgetting. TQCP enables the model to gain awareness of the task type and select the correct prompt representation using a negative guiding approach (Jiang et al., 2024; Li et al., 2024). This method allows the prompt representation to positively correlate with the current task-specific question and negatively correlate with negative question samples. Additionally, KAP acquires task-specific question and video information to enhance accurate answer prediction (for **A1**). Furthermore, VTAP integrates visual information with the LLM by continuously incorporating video dynamics into prompts through autoregressive temporal dynamics and video distillation loss (for **A2** and **A3**). With these prompting strategies, ColPro encapsulates multimodal information to enhance task-specific question answering and mitigate catastrophic forgetting during inference.

Our main contributions to this paper are as follows: (1) We explore the novel problem of video question answering (VideoQA) in a continual learning context, and demonstrate a critical issue: efficiently fine-tuning a LLM for a sequence of tasks leads to catastrophic forgetting. This motivates us to conduct empirical studies to mitigate this issue. (2) We propose Collaborative Prompting (ColPro), which utilizes three distinct aspects: textual question context, visual content, and video temporal dynamics, in VideoQA to facilitate knowledge transfer to future tasks. (3) We conduct extensive experiments on the split VideoQA dataset (NExT-QA (Xiao et al., 2021) and DramaQA (Choi et al., 2021)) for continual task-specific answer prediction. Our findings show that ColPro achieves state-of-the-art results, with 55.14% accuracy on NExT-QA and 71.24% accuracy on DramaQA.

2 Related Works

2.1 Video Question Answering

VideoQA is a fundamental task in video understanding, aiming to answer questions based on video content (Xiao et al., 2023; Gao et al., 2023a; Choi et al., 2021). Many recent works have explored LLM-based VideoQA (Yu et al., 2024; Luo et al., 2023; Ko et al., 2023), which requires a LLM to predict the correct answer given a video and question pair. Flipped-VQA (Ko et al., 2023) uses the prompting technique to fine-tune a LLM to learn the specific VideoQA task. SeViLA (Yu et al., 2024) is built based on a pre-trained large image-language model (Li et al., 2023), extending its capabilities to perform reasoning on video inputs. However, most existing methods are trained on fixed datasets to handle reasoning in static environments, which struggle to answer new questions or tasks posed by newly available content. In contrast, we study a continual video question answering problem, and address its inherent challenges caused by catastrophic forgetting.

2.2 Continual Learning for Visual Question Answering

Over the past few years, various continual learning approaches have been proposed to address the is-

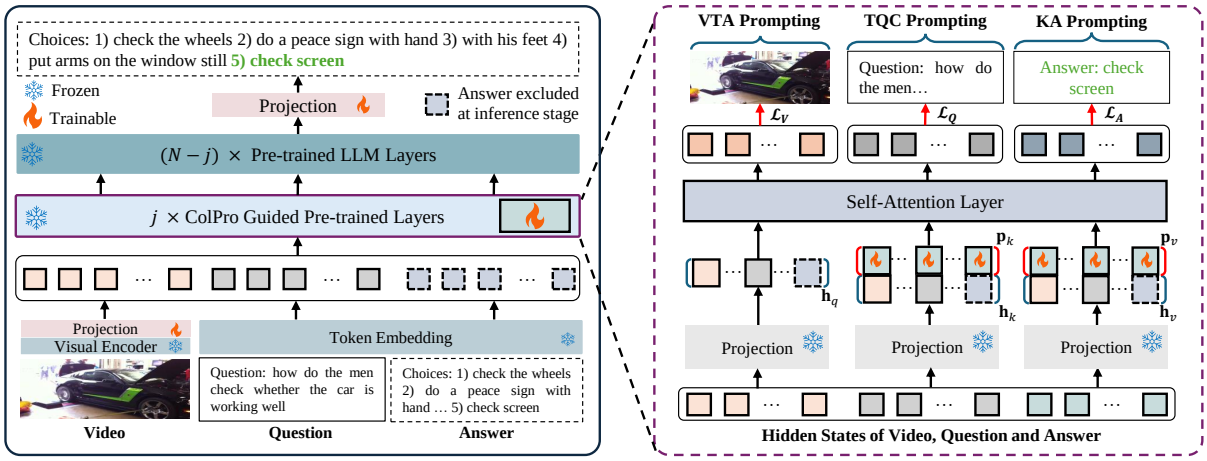


Figure 2: Illustration of the Collaborative Prompting (ColPro) framework. **Left**: The training process incorporates ColPro into the first j ColPro Guided Pre-trained Layers to enhance answer prediction accuracy while minimizing forgetting. **Right**: Three detailed prompting techniques within ColPro are demonstrated: task-specific question constraint prompting (TQCP), knowledge acquisition prompting (KAP), and visual temporal awareness prompting (VTAP). Together, these techniques encapsulate the textual question context, visual content, and video temporal dynamics for each VideoQA task.

sue of catastrophic forgetting (Li and Hoiem, 2017; Rebuffi et al., 2017; Rolnick et al., 2019). Existing methods can be summarized into three categories: rehearsal-based (Buzzega et al., 2020; Rolnick et al., 2019), architecture-based (Li et al., 2019; Ke et al., 2020), and regularization-based (Aljundi et al., 2018; Paik et al., 2020). Rehearsal-based approaches involve constructing a subset of learned samples in a memory buffer and replaying them when learning a new task. Architecture-based approaches allocate separate sets of dedicated parameters for each different task. Regularization-based approaches preserve changes to weights associated with older tasks and selectively stabilize parameter changes. Recent studies (Wang et al., 2022b,a; Gao et al., 2023b) draw inspiration from learnable prompting (Lester et al., 2021; Zhang et al., 2023a) in natural language processing to address catastrophic forgetting by learning a small number parameter that is attached to a pre-trained model. Specifically, L2P (Wang et al., 2022b) utilizes a set of task-specific learnable prompts to dynamically instruct a pre-trained model for continual learning. ProgPrompt (Razdaibiedina et al., 2023) adopts progressive networks with a pre-trained language model to learn prompts for different tasks and sequentially concatenates the task-specific learned prompts for text classification. Learning-Accumulation-Ensemble (LAE) (Gao et al., 2023b) utilizes different Parameter-Efficient Fine-Tuning (PEFT) methods such as adaptor (Houlsby et al., 2019), Lora (Hu et al., 2021), and prompting (Lester et al., 2021) for image classification.

Recent visual question answering models (Lei et al., 2023; Qian et al., 2023) have been exploring the continual learning techniques to answer new questions with given images without experiencing catastrophic forgetting. Zhang et al., 2023b and Lei et al., 2023 introduce replay-based method to address image-based question answering tasks. Qian et al., 2023 use multimodal decoupled prompts to interact with a pre-trained vision-language model, capturing the intricate relationships between modalities. Similar to adapter-based LAE (Gao et al., 2023b), Dynamic Adapter Merging (DAM) (Cheng et al., 2024) utilizes an adaptor-based framework (Houlsby et al., 2019) for video question answering. Unlike DAM, which addresses domain shift in datasets using an adaptor, our work aims to guide a LLM to comprehend multimodal information, including question context, visual content, and temporal dynamics, through a novel prompting technique called ColPro. To the best of our knowledge, this approach is the first of its kind.

3 Methodology

3.1 Motivation and Objective

In this section, we provide an overview of our approach. First, we discuss continual learning scenarios and their applications for VideoQA. Next, we explain our motivation for utilizing prompting strategies with LLM to achieve our goals. Finally,

we present the overall architecture of the proposed method and its training objective.

Continual Learning Scenarios. In continual learning scenarios, a model is trained sequentially through various stages using a dataset $D = <d_1, d_2, ..., d_T>$, where d_t $(1 \leq t \leq T)$ denotes the t-th training task, and data from previous tasks is not accessible during the training of stage t. In this paper, we study the problem of rehearsal-free continual learning on video question answering tasks, where the data $d_t = <V^t, Q^t, A^t>$ consists of video V^t, question Q^t, and answer A^t pairs. For our experiments, we segment the types of questions into T tasks followed by (Zhang et al., 2023b; Lei et al., 2023) to benchmark our proposed approach on the NExT-QA (Xiao et al., 2021) and DramaQA (Choi et al., 2021) datasets. Following the settings in existing rehearsal-free works (Wang et al., 2022b; Razdaibiedina et al., 2023), we assume a pre-trained LLM model (e.g., LLaMA (Touvron et al., 2023)) is available for our experiments.

Prompting for LLM-based Video Question Answering. Prompting, a learnable prompt-based learning technique (Zhang et al., 2023a) has been introduced as a streamlined fine-tuning approach, transforming large language models (e.g., LLaMA (Touvron et al., 2023)) into highly efficient instruction-following models. The core concept of prompting is to incorporate additional instructions into pre-trained LLMs, enabling them to perform downstream tasks in both NLP and multimodal reasoning contexts (Liu et al., 2024; Zhu et al., 2023). In this work, we leverage the efficient instruction-following capabilities and outstanding reasoning abilities of LLMs to achieve accurate multimodal question answering in a continual learning scenario. We illustrate the prompting for LLM-based continual VideoQA as follows.

Our primary focus is on leveraging LLMs for continual learning in VideoQA tasks, with LLaMA-Adapter (Zhang et al., 2023a) serving as our baseline. We adhere to their approach of adaptation through prompt tuning, which is central to our methodology. We use prompt to incorporate task-specific information by learning through LLaMA layers, given the input of task-specific questions, videos, and answers. At the inference stage, we keep the model frozen and utilize the learned task-specific prompt knowledge to predict the task-specific answer. In our framework, given the N-layers LLaMA, we inject prompts for the first j-layers LLaMA transformer layers, named ColPro Guided Pre-trained Layers $\theta(.)$. We maintain the pre-trained model frozen while tuning a select few additional learnable prompts. Rather than appending prompts directly to the input tokens, our approach involves adding prompts to the keys and values within the Multihead Self-Attention (MSA) layer, following the structure described in Transformer architectures (Vaswani et al., 2017). With the split sets of learnable prompts denoted as $\mathbf{P}_k$ and $\mathbf{P}_v \in \mathbb{R}^{l \times d}$, integrated into the key $\mathbf{H}_k$ and values $\mathbf{H}_v$ within the LLaMA model, where $\mathbf{H}_q$ represents the query, the attention module is adapted as follows:

$$\mathbf{H}_i = \text{Attention}(\mathbf{H}_q, [\mathbf{P}_k; \mathbf{H}_k], [\mathbf{P}_v; \mathbf{H}_v]) \quad (1)$$

$$\text{MSA} = \text{Concat}(\mathbf{H}_1, ..., \mathbf{H}_m)\mathbf{W}_o \quad (2)$$

where $[;]$ denotes concatenation, $\mathbf{W}_o$ is projection matrices, $\mathbf{H}_i$ denotes i-th head and m is the number of total heads. In this paper, we adopt the complementary learning principle (Wang et al., 2022a), incorporating learnable General prompts $\mathbf{P}_g$ (G-Prompt) and Expert prompts $\mathbf{P}_e$ (E-Prompt) into the first j layers of the LLaMA model, where the G-Prompt is applied to the first i layers to capture task-invariant knowledge, whereas the E-Prompt is applied to the subsequent layers from $i+1$ to j for task-specific knowledge adaptation. Through this prompting approach, we effectively train a small number of parameters while retaining the knowledge of existing tasks, all without the need for external memory.

Overall Architecture. Our proposed method, termed collaborative prompting (ColPro) for continual VideoQA, is illustrated in Figure 2. Leveraging LLM-based VideoQA models as a foundation, our goal is to establish a cohesive set of collaborative and interactive prompts. This approach aims to mitigate the issue of catastrophic forgetting often associated with straightforward sequential fine-tuning methods.

Each training task set consists of video V^t, question Q^t, and answers A^t in pairs. We extract a sequence of visual tokens $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{N_v}\} \in \mathbb{R}^{N_v \times D}$ from the raw video using a frozen visual encoder (Radford et al., 2021), and utilize a tokenizer to process the raw question and answer into tokens, i.e., $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_{N_q}\} \in \mathbb{R}^{N_q \times D}$ and $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{N_a}\} \in \mathbb{R}^{N_a \times D}$, where N_v, N_q and N_a denote the number of video frames,

lengths of question and answer tokens, respectively. During the training stage, the task-specific token sequences $\mathbf{q}^t$, $\mathbf{v}^t$, and $\mathbf{a}^t$ are concatenated and inputted into LLaMA along with an additional prompts $\mathbf{P} = <\mathbf{P}_e, \mathbf{P}_g> = \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_p}\} \in \mathbb{R}^{N_p \times D}$, where N_p denotes length of prompts. This setup allows the output feature to be calculated as follows:

$$\mathbf{X}^t, \mathbf{P} = \theta(< \mathbf{Q}^t, \mathbf{V}^t, \mathbf{A}^t >, \mathbf{P}), \tag{3}$$

where $\mathbf{X}^t = <\mathbf{X}_q^t, \mathbf{X}_v^t, \mathbf{X}_a^t>$ denotes the sequence of output features for question, video and answer for task t. In our framework, $\mathbf{P}_g$ is trained alone using the global cross-entropy loss similar to existing methods (Wang et al., 2022a,b), while we focus on optimizing the E-Prompt $\mathbf{P}_e$ to effectively capture and preserve task-specific knowledge, thereby reducing catastrophic forgetting. During the inference stage, LLaMA takes $\mathbf{V}^t$, $\mathbf{Q}^t$, and the learned prompts $\mathbf{P}$ to predict task-specific answers.

3.2 Collaborative Prompting

We systematically explore continual learning, focusing on integrating multimodal distributions into a unified set of prompts. This approach provides a comprehensive framework for continuous improvement in LLM-based VideoQA. Our methodology includes collaboratively incorporating prompts designed for task-specific question constraints, visual temporal awareness, and knowledge acquisition.

Task-specific Question Constraint Prompting (TQCP). TQCP extracts question-specific knowledge from learned prompt representations, enhancing task awareness during the inference stage. Different from existing methods (Wang et al., 2022a,b) that rely on a known task identity to select and train specific sets of prompts alongside the classifier, we directly utilize the question type to guide the learning of a single set of prompts for question awareness. Drawing inspiration from a recent negative label guided algorithm (Jiang et al., 2024; Li et al., 2024), we enable $\mathbf{P}_e$ to be positively correlated with the current task-specific question type (e.g., how many) and negatively correlated with negative question samples (e.g., negative question types: what, where, etc). This facilitates task-type awareness and links input features to E-prompts during the inference stage. To achieve this, we optimize $\mathbf{X}_q^t$ and $\mathbf{P}_e$ with question generation loss and negative questions ($\mathbf{Q}^-$) guided loss, which allows the given prompt to learn question-specific representation for the current task. It can be formulated as follows:

$$\mathcal{L}_q = -(\mathcal{L}_q^{gen} + \mathcal{L}_q^{neg}) \tag{4}$$

$$\mathcal{L}_q^{gen} = \sum_{n=0}^{N_q-1} \log P(\mathbf{q}_{n+1}^* | \mathbf{V}, \mathbf{A}, \mathbf{P}, \mathbf{q}^+ \leq n, \mathbf{Q}^-), \tag{5}$$

$$\mathcal{L}_q^{neg} = \frac{1}{\mathcal{B}} \sum_{i \in \mathcal{B}} \left(\frac{e^{sim(\mathbf{P}_e, \mathbf{Q}_i^+)/\tau}}{\sum_{j \in \mathcal{B}} (e^{sim(\mathbf{P}_e, \mathbf{Q}_j^+)/\tau} + e^{sim(\mathbf{P}_e, \mathbf{Q}_j^-)/\tau})} \right), \tag{6}$$

where $P(\mathbf{q}_{n+1}^*) = \text{Softmax}(\text{Linear}([\mathbf{X}_q; \mathbf{P}_e]))$ for task t, and $[;]$ denotes concatenation. $\mathbf{P} = <\mathbf{P}_e, \mathbf{P}_g>$ and τ is a temperature parameter. We employ cross-entropy loss $\mathcal{L}_q^{gen}$ locally to generate task-specific questions based on learned $\mathbf{X}_q$ and $\mathbf{P}_e$. $\mathcal{L}_q^{neg}$ to correlate the given question $\mathbf{Q}^+$ with $\mathbf{P}_e$, where $sim(\cdot, \cdot)$ computes the cosine similarities between the $\mathbf{P}_e$ and the i-th positive question $\mathbf{Q}_i^+$ (resp. j-th negative question $\mathbf{Q}_j^-$) samples in the batch $\mathcal{B}$.

Visual Temporal Awareness Prompting (VTAP). VTAP aims to bridge the gap between video features and the LLM, allowing E-prompts to incorporate visual information with temporal dynamics. This improves the video understanding abilities of the LLM and enhances its answer prediction capabilities with given questions and videos. However, modeling both the visual content of videos and their temporal dynamics simultaneously presents a challenge. To overcome this, we guide the E-prompt in learning video with temporal dynamics by using the question and answer choices as prior knowledge and leverage the autoregressive sequential abilities of the LLM to model and predict the order of video frames based on preceding frames. Furthermore, we distill video information extracted from an image encoder (Radford et al., 2021) into an E-prompt (Zhong et al., 2024; Li et al., 2023), enabling the LLM to understand visual features. In this work, we use contrastive loss ($\mathcal{L}_q^{con}$) to facilitate this process, which is formulated as follows:

$$\mathcal{L}_v = -(\mathcal{L}_v^{dyn} + \mathcal{L}_v^{con}) \tag{7}$$

$$\mathcal{L}_v^{dyn} = \sum_{n=0}^{N_v-1} \log P(\mathbf{v}_{n+1}^* | \mathbf{Q}, \mathbf{A}, \mathbf{P}, \mathbf{v} \leq n), \tag{8}$$

$$\mathcal{L}_v^{con} = \frac{1}{\mathcal{B}} \sum_{i \in \mathcal{B}} \log \left(\frac{e^{sim(\mathbf{P}_e, \mathbf{V}_i)/\tau}}{\sum_{j \in \mathcal{B}} e^{sim(\mathbf{P}_e, \mathbf{V}_j)/\tau}} \right), \tag{9}$$

where $P(\mathbf{v}^*_{n+1}) = \text{Softmax}(\text{Linear}([\mathbf{X}_v; \mathbf{P}_e]))$ for task t, $\mathcal{L}_v^{dyn}$ is the optimization function for video temporal dynamic modelling, and $sim(\cdot,\cdot)$ computes the cosine similarities between the $\mathbf{P}_e$ and the i-th video $\mathbf{V}_i$ (resp. j-th video $\mathbf{V}_j$) in the batch $\mathcal{B}$ for current task.

Knowledge Acquisition Prompting (KAP). KAP injects task-specific multimodal information from the question and video into the E-prompts to accurately predict answers for the current task. To achieve this, at the training stage, $\mathbf{P}_e$ leverages the autoregressive abilities of the LLM to encapsulate the task-specific context information of $\mathbf{V}$, $\mathbf{Q}$, and all answer choices $\mathbf{A}$ as prior knowledge to predict the specific answer. The objective function is formulated as:

$$\mathcal{L}_a = -\sum_{n=0}^{N_a-1} \log P(\mathbf{a}^*_{n+1}|\mathbf{Q}, \mathbf{V}, \mathbf{P}, \mathbf{a} \leq n), \quad (10)$$

where $P(\mathbf{a}^*_{n+1})$ = Softmax(Linear($[\mathbf{X}_a; \mathbf{P}_e]$)). At the inference phase, the continual VideoQA model predicts the task-specific answer with $\mathbf{V}$, $\mathbf{Q}$, $\mathbf{P}$ as:

$$\bar{a} = \arg\max_{a \in \mathcal{A}^t} P(\mathbf{a}|\mathbf{V}, \mathbf{Q}, \mathbf{P}), \quad (11)$$

where $\mathcal{A}^t$ is a set of answer choice for task t.

4 Experiments

4.1 Datasets

We use the multi-choice NExT-QA dataset (Xiao et al., 2021), which includes various types of questions. These include causal questions, such as why (CW) and how (CH), that ask for the intentions or reasons behind earlier actions; temporal questions, which determine the relationships between actions like what are (TC), what did (TN), and what was (TP); and descriptive questions, like how many (DC), where (DL) and other types of question (DO), which focus on visible contents such as places and attributes. We split (Lei et al., 2023; Zhang et al., 2023b) the NExT-QA dataset into eight distinct tasks based on question types in the NExT-QA dataset. In CL, the order of task learning impacts the learning outcome. Therefore, we conducted experiments and set our tests in the sequence that resulted in the highest forgetting rate (suffers more in catastrophic forgetting) using the baseline method. The sequence of the training order follows this sequence: <TP, CW, DC, TC, DL, DO, TN, CH>. DramaQA dataset (Choi et al., 2021) features a video story understanding with hierarchical difficulty levels. We split it into 5 distinct tasks according to the question types, and experimentally chose the order that has the worst forgetting result to be our learning order for the CL VideoQA task. The order of CL learning follows <what, who, where, how, why>. Similar to the existing VideoQA methods (Yu et al., 2024), we report the performance of the validation set.

Table 1: The results on the NExT-QA dataset which are divided into 8 tasks, where the Avg. Acc denotes average accuracy across tasks and Avg. Fog is the average forgetting rate. The symbols ↑ and ↓ indicate whether a higher or lower value is preferable for a given metric, respectively.

Method	Avg. Acc (↑)	Avg. Fog (↓)
LLaMA	46.58	13.83
L2p	48.82	12.25
DualPrompt	50.62	11.74
LAE	49.38	11.47
L2p+	52.26	11.61
DualPrompt+	53.97	10.26
LAE+	53.75	9.74
DAM	53.88	9.99
ProgPrompt	53.95	10.69
ColPro	**55.14**	**7.43**

4.2 Evaluation Metrics

We evaluate methods using two metrics: the average final accuracy (Avg. Acc), where higher values are better and represent the final accuracy averaged over N tasks for all previous classes, and the average forgetting (Avg. Fog) that is widely used in existing works (Wang et al., 2022b,a; Qian et al., 2023), where the lower the better which indicates the tasks experienced less forgetting averaged over N tasks.

4.3 Implementation Details

We train CL VideoQA for five epochs on both datasets with a batch size of 8 for the NExT-QA dataset (4 for DramaQA) with the gradient accumulation technique. We fine-tuned the LLaMA-7B model in this paper. AdamW optimizer is used with β = (0.9, 0.95). We search learning rate and weight decay in [0.05, 0.1] and [0.15, 0.25], respectively. The number of video frames V is set to 10. Each frame is resized by 224 × 224 and fed into CLIP VIT-L/14 to extract frame features. The total sequence length of the concatenated visual, question,

Table 2: The results on the DramaQA dataset which are divided into 5 tasks.

Method	Avg. Acc (↑)	Avg. Fog (↓)
LLaMA	60.99	24.39
L2p	62.50	20.67
DualPrompt	65.89	17.93
LAE	65.82	17.35
L2p+	66.75	16.73
DualPrompt+	67.44	15.09
LAE+	67.03	14.82
DAM	67.37	15.19
ProgPrompt	67.92	14.95
ColPro	**71.24**	**12.64**

and answer tokens is 128 for NExT-QA and 280 for DramaQA. Temperature parameter τ is set to 1. The prompt tokens are empirically set to 10 for $\mathbf{p}_k$, $\mathbf{p}_v$. The positing of G-prompt and E-prompt are set to 0-6 and 7-18 LLaMA layers, respectively, for the best performance. The prompts are not attached to the remaining LLaMA layers.

4.4 Comparison with Continual Learning Methods

Table 1 compares the performance of the Collaborative Prompting (ColPro) on the split NExT-QA benchmark with existing CL approaches, including the fine-tuned LLaMA (Touvron et al., 2023) with addtional projection layer, L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), ProPrompt (Razdaibiedina et al., 2023), DAM (Cheng et al., 2024), and LAE (Gao et al., 2023b). Additionally, we report deep CL implementations of the L2P+, DualPrompt+, and LAE+ methods, which activate more layers of LLaMA for CL tasks by applying prompts to 18 layers. In the comparisons, our proposed ColPro achieved better average prediction accuracy and significantly lower average forgetting compared to existing methods. This improvement in average forgetting can be attributed to the fact that the ColPro method experiences less forgetting and allows better forward transfer of different tasks, which is beneficial in CL VideoQA. Similarly, we compare the performance of the ColPro on the split DramaQA benchmark with existing CL approaches in Table 2, further validating the effectiveness of our proposed method in addressing catastrophic forgetting issues. These tables indicate that the models experience catastrophic forgetting, with the Avg. Fog score up to 24%. This underscores the need to address catastrophic forgetting in video QA, and we have minimized the forgetting with ColPro.

Table 3: The results on both NExT-QA and DramaQA datasets with PEFT and our methods.

Methods	Dataset	Avg. Acc (↑)	Avg. Fog (↓)
Prefix	NExT-QA	47.76	13.21
Lora		52.00	11.07
L-Adaptor		51.83	12.42
Ours		**55.14**	**7.43**
Prefix	DramaQA	60.93	21.18
Lora		62.11	19.73
L-Adaptor		61.50	19.77
ColPro		**71.24**	**12.64**

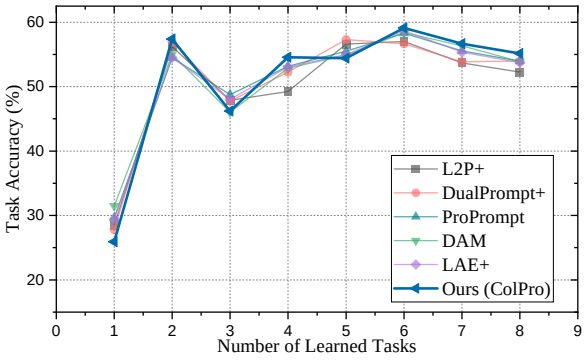


Figure 3: The results of the average accuracy for each task, which following the training order within the CL setting.

4.5 Comparison with Parameter-Efficient Fine Tuning Methods

Table 3 demonstrates the performance of ColPro with other Parameter-Efficient Fine-Tuning (PEFT) methods that can be used to address catastrophic forgetting in CL settings, such as LLaMA-Adapter (L-Adapter) (Zhang et al., 2023a), Lora (Hu et al., 2021), and Prefix (Li and Liang, 2021). Our proposed method shows a significant improvement in minimizing forgetting compared to our baseline LLaMA-Adapter method, as evidenced by a lower Avg. Fog and a higher Avg. ACC when evaluated with the NExT-QA and DramaQA datasets. ColPro also outperforms Lora and Prefix, demonstrating the effectiveness of our specially designed strategy for LLM-based CL VideoQA settings.

4.6 Task-by-Task Average Accuracy

Continuous learning (CL) in real-world scenarios is an ongoing process, making the performance of

Table 4: The ablation results for the three proposed multimodal interaction prompting strategies.

$\mathcal{L}_a$	$\mathcal{L}_q$	$\mathcal{L}_v$	Avg. Acc (↑)	Avg. Fog (↓)
✓	✗	✗	52.60	10.62
✓	✓	✗	53.09	9.09
✓	✗	✓	54.38	10.27
✓	✓	✓	**55.14**	**7.43**

Table 5: The ablation results for different prompting optimization functions.

$\mathcal{L}_q^{neg}$	$\mathcal{L}_v^{dyn}$	$\mathcal{L}_v^{con}$	Avg. Acc (↑)	Avg. Fog (↓)
✗	✗	✗	52.87	10.15
✓	✗	✗	53.09	9.09
✓	✓	✗	52.80	9.95
✓	✗	✓	54.20	8.71
✓	✓	✓	**55.14**	**7.43**

each learning phase critical for the VideoQA model. Consequently, we plotted the task-by-task continual learning average accuracy in Figures 3 for the NeXT-QA dataset with respect to the order of training. We accumulated with the previous tasks, and tested on the current task to report the average accuracy. Our results show that ColPro achieves better performance in most of the perdition accuracy than existing methods in the learning phase.

4.7 Ablation Study

The Effectiveness of Multimodal Prompting. Our proposed method includes three primary multimodal interaction prompting strategies: question constraint (TQCP), visual temporal alignment (VTAP), and knowledge acquisition (KAP). Each strategy is optimized with its respective optimization function, $\mathcal{L}_q$, $\mathcal{L}_v$, and $\mathcal{L}_a$. We performed ablation studies on these strategies and reported the results in Table 4. A tick indicates that the respective prompting strategies was used during model training, while a cross means not included. Notably, $\mathcal{L}_a$ is always included to optimize the model with the ground truth answer. We can observe that the inclusion of each optimization function significantly impacts the model's performance. Specifically, models trained with all three optimization functions consistently achieve higher accuracy and lower forgetting. The ablation results demonstrated show that incorporating VTAP enhances the accuracy of the LLM, while utilizing TQCP helps the model suffer less from forgetting. This underscores the importance of question constraint and visual temporal alignment prompting, which help the LLM gain awareness of the task type to reduce catastrophic forgetting and understand the visual information with temporal dynamics for better answer reasoning. The combination of three strategies is crucial for optimizing multimodal interaction in CL VideoQA scenarios.

Table 6: The ablation results for prompt positioning in LLaMA layers.

$\mathbf{p}_g^{end}$	$\mathbf{p}_e^{end}$	Avg. Acc (↑)	Avg. Fog (↓)
18	0	52.53	10.45
0	18	53.66	8.81
6	18	53.74	9.01
10	18	54.78	8.11
8	16	54.45	8.59
8	18	**55.14**	**7.43**
8	20	54.75	9.02

Table 7: The ablation results for using different lengths of the prompt in the model.

Lenght of $\mathbf{p}$	Avg. Acc (↑)	Avg. Fog (↓)
5	53.91	8.91
10	**55.14**	**7.43**
15	54.55	8.35
20	53.77	9.63

Furthermore, in Table 5, we break down each prompting optimization to evaluate the effectiveness of using negative contrast ($\mathcal{L}_q^{neg}$), visual distillation ($\mathcal{L}_v^{con}$), and temporal dynamic understanding ($\mathcal{L}_v^{dyn}$) techniques. The optimization functions $\mathcal{L}_a$ and $\mathcal{L}_q^{gen}$ are always included to optimize the model with the ground truth answer and question during training. We can observe that $\mathcal{L}_q^{neg}$ tends to reduce forgetting in large margin, indicating that the negative contrast technique allows question constraint prompts to gain task-specific knowledge, making the model better aware of the task type. Furthermore, we can see that the stand along temporal dynamic ($\mathcal{L}_v^{dyn}$) does not benefit the model, but the visual distillation ($\mathcal{L}_v^{con}$) is able to improve the average accuracy. The method achieves excellent performance when the model the combination of $\mathcal{L}_v^{dyn}$ and $\mathcal{L}_v^{con}$. This emphasis the importance of the proposed temporal dynamic understanding that bridges LLM with video information for VideoQA to achieve better performance.

Number of layers for G-Prompts and E-

Prompts. In our method, we utilize both E-prompt and G-prompt. In Table 6, we empirically evaluate the effectiveness of placing the G-prompt and E-prompt within the LLaMA layers, following existing methods, to achieve the best performance. $\mathbf{p}_g$ and $\mathbf{p}_e$ denote the last layer attached to LLaMA-7B, the pre-trained network with 32 layers. The table shows that performance drops when either the $\mathbf{p}_g^{end}$ = 0 or E-prompt $\mathbf{p}_e^{end}$ = 0 is excluded. We found that the optimal performance was achieved by attaching the G-Prompt from layers 0 to 8 and the E-Prompt from layers 9 to 18.

Number of prompt. For prompt-based learning, the length of the prompt is a crucial parameter that can significantly affect learning performance (Gao et al., 2023b; Wang et al., 2022b). Table 7 illustrates the impact of varying prompt lengths on model accuracy and forgetting rates. Our results indicate that shorter prompts may not provide sufficient context for the model, leading to lower accuracy and higher forgetting. Conversely, excessively long prompts can introduce noise and unnecessary information, which also negatively impacts performance in our experiment. We found an optimal prompt length of 10 to balance the amount of information provided to the model and maintain high performance.

5 Conclusion

In this work, we explore the novel problem of VideoQA, which efficiently fine-tunes the LLM to answer new questions with video in a continual learning context. We propose Collaborative Prompting (ColPro) to integrate textual question context, visual content, and video temporal dynamics in each learning phase, facilitating knowledge transfer to future tasks while minimizing catastrophic forgetting. We achieves state-of-the-art results on split NExT-QA and DramaQA datasets.

6 Limitations

We propose the efficient Collaborative Prompting (ColPro), which integrates task-specific question constraint prompting, knowledge acquisition prompting, and visual temporal awareness prompting with a large language model (LLM) to enhance the performance of continual VideoQA. However, catastrophic forgetting remains high for the DramaQA dataset using our method, indicating a substantial decline in performance for VideoQA prediction when using LLaMA-7B. Furthermore, we did not experiment with larger models (e.g., 33B LLM) due to memory constraints, which limits our ability to explore catastrophic forgetting that may arise when fine-tuning other LLMs for CL VideoQA.

Acknowledgments: This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-08-024[T]). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930.

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

Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. 2024. Dam: Dynamic adapter merging for continual video qa learning. *arXiv*, 2403.08755.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1166–1174.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023a. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14773–14783.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. 2023b. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11483–11493.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea

Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2024. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*.

Zixuan Ke, Bing Liu, and Xingchang Huang. 2020. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in neural information processing systems*, 33:18493–18504.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*.

Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. 2023. Symbolic replay: Scene graph as prompt for continual learning on vqa task. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1250–1259.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. 2024. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17584–17594.

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

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International conference on machine learning*, pages 3925–3934. PMLR.

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

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

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

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

Inyoung Paik, Sangjun Oh, Taeyeong Kwak, and Injung Kim. 2020. Overcoming catastrophic forgetting by neuron-level plasticity control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5339–5346.

Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. 2023. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2953–2962.

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

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*.

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

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems*, 30.

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

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. 2023. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13265–13280.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *In NeurIPS*.

Jipeng Zhang, Jie Shao, Rui Cao, Lianli Gao, Xing Xu, and Heng Tao Shen. 2022. Action-centric relation transformer network for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):63–74.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

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

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Critical Continual Learning Order

In Table 8, we outline the sequence of continual learning tasks in VideoQA, enabling us to identify and select the most critical tasks that are particularly susceptible to catastrophic forgetting. By understanding which learning order is most affected by this phenomenon, we can prioritize and implement targeted strategies to mitigate forgetting, thereby enhancing the overall robustness and effectiveness of the continual learning system. We use baseline model (Touvron et al., 2023) with additional linear layer for this experiment. We can see that the learning sequences <TP, CW, DC, TC, DL, DO, TN, CH> have higher Avg. Fog for NExT-QA dataset.

Table 8: The results of the task learning sequences for continual learning for the NExT-QA datasets.

Orders	Avg. Fog (↓)
CH, DL, TP, TC, DC, DO, TN, CW	7.26
TP, TN, CH, TC, DL, DO, CW, DC	8.41
DO, CW, DC, CH, TP, TC, TN, DL	9.83
CW, DO, TN, DL, TC, TP, DC, CH	11.86
TP, CW, DC, TC, DL, DO, TN, CH	13.83

A.2 Learning Parameter Analysis with Full Model Fine-tuning

Since most parameters are fixed at the inference stage, the performance of a fine-tuned prompt-based model may be worse than that of a fully fine-tuned model for each specific task. However, during the training stage, fine-tuning the entire LLM incurs high computational costs. Here, we provided an analysis of this aspect to better understand the trade-offs between the effectiveness and computation cost of these two approaches using the score we get from the DramaQA dataset as an example. According to the training parameters indicated in LLaMa (Touvron et al., 2023) and Lora (Hu et al., 2021), we can assume that to fully fine-tune an LLM requires more than 500M parameters, whereas our prompt-based method only requires around 33.5M parameters. Although the Avg. Acc (assumed to be >71.24) of full LLaMA fine-turning may be higher than our score, but it requires a much higher computation cost. Our method can efficiently and effectively fine-tune LLaMA-7B model for CL in VideoQA using a single 24GB

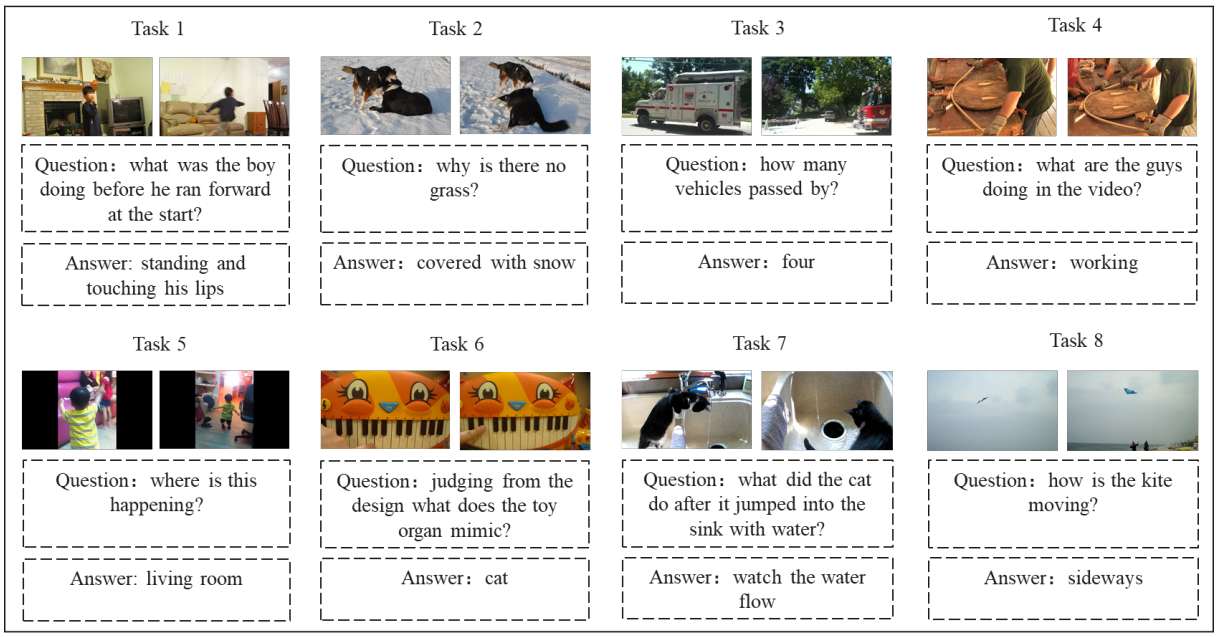


Figure 4: The video examples with their corresponding questions and answers for each task.

Table 9: Analysis of Model Parameters and Average Accuracy

Model	Parameters	Avg. Acc
LLaMA fine-tuning	> 500 M	> 71.24
Prompt-tuning	33.5 M	71.24

GPU. Furthermore, ColPro achieved better performance compared to existing prompt-based methods, as illustrated in Tables 1 and 2. It's worth noting that due to limited computational power, we were unable to provide the results for full LLaMA fine-tuning.

A.3 Continual Learning Setting and Examples

In this paper, we split the dataset towards the function-incremental setting of continuous learning, similar to existing CL ImageQA works (Lei et al., 2023; Qian et al., 2023), to better evaluate the CL VideoQA task. We split the dataset according to different functions. For instance, we split NExT-QA into causal reasoning function, which includes logic understanding of asking why (CW) and how (CH), temporal reasoning function that involves the relationship understanding of objects or attributes recognition in what are (TC), what did (TN), and what was (TP), and descriptive reasoning function encompasses knowledge understanding of how many (DC), where (DL), and other types of questions (DO), as illustrated in Section 4.1. Similar for DramaQA, we split the dataset according to the function of each question type. The raw video examples for CL VideoQA with their corresponding question type and answer are illustrated in Figure 4. The figure showing the differences between them for NExT-QA (Xiao et al., 2021) dataset.