

Decomposing and Composing: Towards Efficient Vision-Language Continual Learning via Rank-1 Expert Pool in a Single LoRA

Zhan Fa¹, Yue Duan¹, Jian Zhang¹, Lei Qi², Wanqi Yang³, Yinghuan Shi^{1*}

¹National Key Laboratory for Novel Software Technology, Nanjing University, China

²School of Computer Science and Engineering, Southeast University, China

³School of Computer and Electronic Information, Nanjing Normal University, China

{fazhan, yueduan, zhangjian7369}@smail.nju.edu.cn, qilei@seu.edu.cn, yangwq@nju.edu.cn, syh@nju.edu.cn

Abstract

Continual learning (CL) in vision-language models (VLMs) faces significant challenges in improving task adaptation and avoiding catastrophic forgetting. Existing methods usually have heavy inference burden or rely on external knowledge, while Low-Rank Adaptation (LoRA) has shown potential in reducing these issues by enabling parameter-efficient tuning. However, considering directly using LoRA to alleviate the catastrophic forgetting problem is non-trivial, we introduce a novel framework that restructures a single LoRA module as a **decomposable** Rank-1 Expert Pool. Our method learns to dynamically **compose** a sparse, task-specific update by selecting from this expert pool, guided by the semantics of the [CLS] token. In addition, we propose an **Activation-Guided Orthogonal (AGO) loss** that orthogonalizes critical parts of LoRA weights across tasks. This sparse composition and orthogonalization enable fewer parameter updates, resulting in domain-aware learning while minimizing inter-task interference and maintaining downstream task performance. Extensive experiments across multiple settings demonstrate state-of-the-art results in all metrics, surpassing zero-shot upper bounds in generalization. Notably, it reduces trainable parameters by 96.7% compared to the baseline method, eliminating reliance on external datasets or task-ID discriminators. The merged LoRAs retain less weights and incur no inference latency, making our method computationally lightweight.

Code — <https://github.com/Fazhan-cs/DAC>

Introduction

Vision-Language Models (VLMs) are vital for real-world multi-domain tasks as they possess the ability to address the challenging problem of aligning visual and language modalities (Parelli et al. 2023; Antol et al. 2015; Hong et al. 2024). In real-world scenarios, continual learning (CL) in VLMs is imperative due to the dynamic and evolving nature of real-world data, and systems need to adapt continuously to new information (Zheng et al. 2023; Jha, Gong, and Yao 2024). Leveraging pre-trained VLM like CLIP (Radford et al. 2021), recent CL advancements enhance downstream task performance via continual fine-tuning (Jha, Gong, and

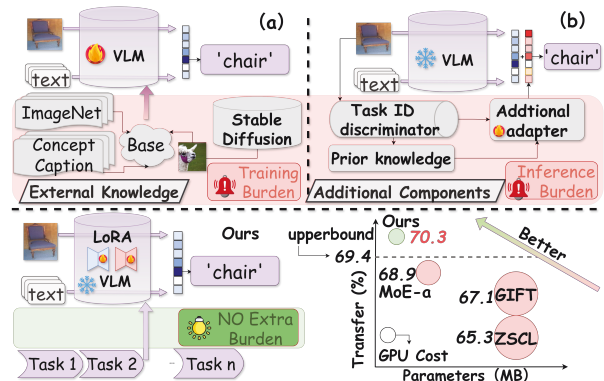


Figure 1: Advantages of our method over previous works: (a) ZSCL and GIFT introduce large external datasets or synthetic images from generative models for model regularization. (b) MoE-a and RAIL introduce additional components and utilize prior knowledge during testing. Our method uses **NO** external data knowledge and introduces **NO** additional burden. Meanwhile, it has fewer training parameters and lower GPU cost, with Transfer, a metric to measure the generalization ability of VLMs, exceeding the upper bound of the original CLIP zero-shot performance.

Yao 2024; Thengane et al. 2022; Lee, Zhong, and Wang 2023). In this paper, we center on multi-domain continual learning of VLMs, aiming to boost downstream tasks and maintain the transfer ability of pre-trained CLIP models. Many works have explored the learning and forgetting problems of VLMs (Wu et al. 2025a; Yu et al. 2024; Xu et al. 2024). Previous works, as depicted in Figure 1, have tangible limitations in the training and inference stages of continual learning: (a) **Training burden:** Works such as ZSCL (Zheng et al. 2023) and GIFT (Wu et al. 2025a) heavily rely on external data sources. ZSCL uses over 100K images in large-scale datasets like ImageNet (Deng et al. 2009) for model regularization, while GIFT utilizes Stable Diffusion (Rombach et al. 2021) to generate synthetic images for over 10 hours for data replay. (b) **Inference burden:** Methods like MoE-a (Yu et al. 2024) and RAIL (Xu et al. 2024) introduce additional complexity during inference. They add adapters to the pre-trained model, and the size of these addi-

*Corresponding author

tional components increases linearly with task numbers, i.e., 11 tasks here, bringing a great inference overhead. Moreover, they obtain prior knowledge such as task-id through discriminators during inference, which is difficult to obtain under real-world scenarios.

Moreover, some previous works have attempted to use LoRA to address continual learning problems (Wei, Li, and Marculescu 2024; Wang et al. 2023; Yang et al. 2024a), often by isolating training stage updates into multiple separate low-rank matrices. However, directly relying on LoRA to alleviate catastrophic forgetting is non-trivial. As previous studies have shown (Jiang et al. 2025; Gekhman et al. 2024), LoRA still suffers from redundant parameter updates despite its low-rank characteristics. While many LoRA-based studies have focused on static rank pruning to reduce this redundancy (Jiang et al. 2025; Zhang et al. 2023a; Valipour et al. 2022; Ding et al. 2023; Meng, Wang, and Zhang 2024), such approaches are not suitable for dynamic data injection and changing task domains in continual learning. To address this, we introduce a dynamic learning approach. We innovatively restructure a single LoRA module as a **decomposable Rank-1 Expert Pool**. The key insight behind our method is the equivalence that **a single LoRA with rank r is equivalent to using r LoRAs with rank 1**. This decomposition allows us to solve the parameter redundancy problem from a new perspective: instead of updating the entire low-rank matrix, we can dynamically **compose** a sparse, task-specific update by selecting only the most relevant experts from the pool. This process makes the parameter updates more targeted and efficient, significantly reducing unnecessary interference and leading to better performance. Our framework exploits this efficient update mechanism while retaining the core advantages of LoRA. The training process remains highly parameter-efficient, and after each task, the composed sparse LoRA weights are merged back into the original model. This ensures our method introduces no extra components or latency during inference, creating a truly lightweight and effective solution for continual learning. To implement the dynamic selection, we then employ the semantic-rich [CLS] token (Liang et al. 2022; Wang et al. 2024) to guide a lightweight router, enabling the selection of appropriate experts to adapt to different task domains.

In addition, considering that the orthogonality of parameter update directions is helpful for isolating task optimization objectives (Wang et al. 2023; Yang et al. 2024a; Feng et al. 2025), we innovatively propose an **Activation-Guided Orthogonal (AGO) loss**. Our approach uses the expert activation frequency recorded during our dynamic composition process. When updating for each task, we leverage this pre-recorded information to calculate the orthogonal loss with all previous trained LoRAs. Unlike previous works that use additional components to determine domain information, our orthogonal strategy isolates domains in the parameter space, considering that LoRA parameters are not merely numerical adjustments but encapsulate crucial model update directions (Wang et al. 2023). Therefore, by directly zeroing the weights of low-activation-frequency experts, the task can focus on being orthogonal to the crucial parts of previous tasks, without causing excessive interference to the cur-

rent task. This can avoid the problem of parameter collision (Yang et al. 2024a) to a certain extent.

In general, the contributions of our work are as follows:

- We systematically analyze the training and inference burdens in prior VLM continual learning works and design an efficient framework to address them.
- We propose a novel method that restructures a single LoRA module as a **decomposable Rank-1 Expert Pool**, allowing for the dynamic **composition** of sparse, task-specific subspaces and reducing critical burden.
- We design a synergistic **Activation-Guided Orthogonal (AGO) loss** that leverages expert activation frequencies to reduce inter-task interference precisely.

In diverse experimental settings, despite reducing training parameters by **96.7%** compared to the baseline (570.76 MB vs. 18.99 MB), our method achieves state-of-the-art (SOTA) results across all metrics, with the transfer metric even exceeding CLIP’s zero-shot performance by 0.9%.

Related Works

Existing continual learning methods primarily focus on solving Class Incremental Learning (CIL) or Task Incremental Learning (TIL) based on the domain changes of data (Rebuffi et al. 2017; Agarwal et al. 2022; Wortsman et al. 2022; Ding et al. 2022). Traditional continual learning methods encompass replay-based methods (Rebuffi et al. 2017; Lavda et al. 2018), distillation-based methods (Ding et al. 2022; Li and Hoiem 2017), regularization-based methods (Wortsman et al. 2022; Aljundi et al. 2018), and architecture-based methods (Li et al. 2019; Douillard et al. 2022). Here we focus on the setting of Multi-Task Incremental Learning (MTIL) (Zheng et al. 2023; Yu et al. 2024), which enables VLMs to sequentially learn in different domains while retaining the pre-trained generalization ability for previously seen tasks. Recent works have focused on the continual fine-tuning of pre-trained VLMs: ZSCL (Zheng et al. 2023) adds a regularization term to the cross-entropy loss function to penalize changes in model parameters or feature space and regularize the parameter space using a large-scale reference dataset. MoE-a (Yu et al. 2024) cooperates the pre-trained CLIP with a Mixture of Experts (MoE) adapters and uses the reference dataset to discriminate task-ID, enabling the model to distinguish between unseen and seen tasks. Moreover, RAIL (Xu et al. 2024) uses the principle of ridge regression to add an additional high-dimensional classification adapter to CLIP and consolidates the learned knowledge using task IDs. GIFT (Wu et al. 2025a) uses a pre-trained diffusion model to generate replay images for the text inputs of past tasks to enhance memory.

These methods either introduce additional model components and datasets or require a large amount of prior knowledge for inference, which brings great training and inference burden. However, with the rapid development of LoRA in the era of pre-trained models (Meng, Wang, and Zhang 2024; Mao et al. 2024; Zhang et al. 2023b; Yang et al. 2024b; Ren et al. 2024), we leverage its principles in continual learning scenarios. Considering that LoRA can fine-tune VLM in a low-rank characteristic and merge the weights

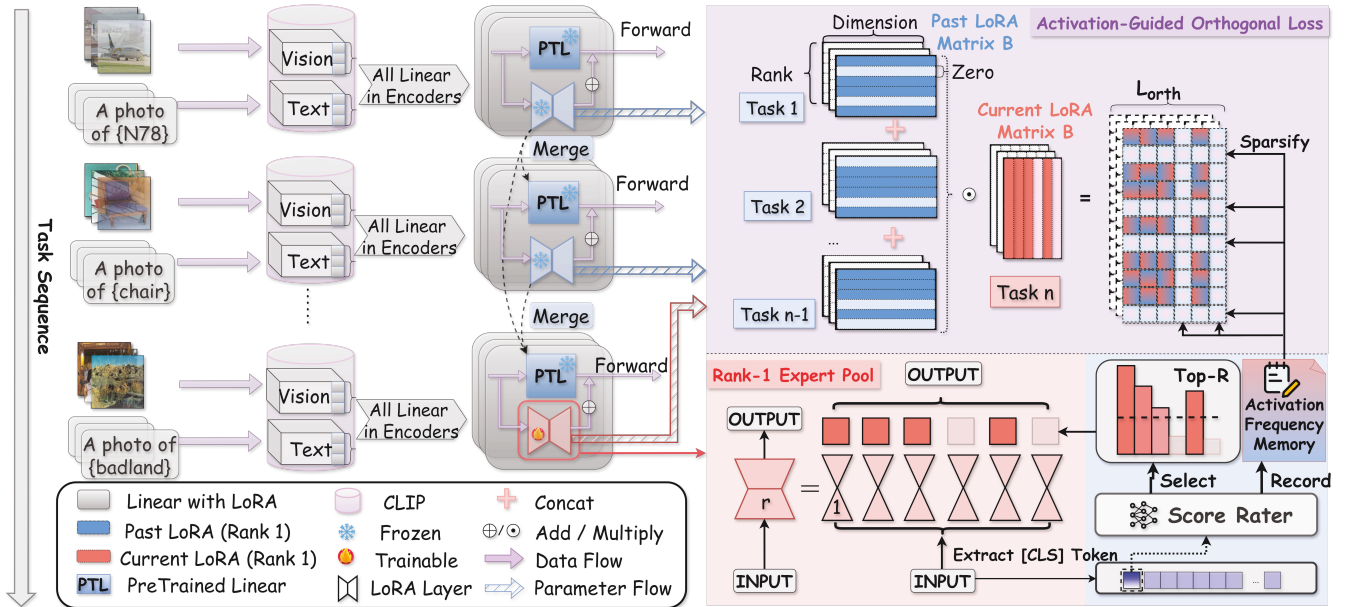


Figure 2: The overall framework of our proposed method. LoRA is configured for all linear layers of the CLIP text and image encoder transformers, with the original parameters frozen. Each LoRA module is treated as a **Rank-1 Expert Pool**. For each input, the [CLS] token is extracted to guide a router that composes a sparse update by selecting critical experts from this pool. During training, an **Activation-Guided Orthogonal loss** is calculated between the current LoRA and those from past tasks. After training, the composed LoRA weights are merged back into the original model for zero-overhead inference.

back into the original weights when training is completed, our method introduces no additional burdens during training and testing, making it more lightweight and more suitable for practical task scenarios. Furthermore, our method differs from previous work that treats multiple parallel LoRAs as individual experts in a Mixture of Experts (MoE) framework (Liu et al. 2024; Gao et al. 2024; Dou et al. 2023). Rather than adding more LoRA modules and increasing parameter counts, our approach restructures a single LoRA as a **decomposable Rank-1 Expert Pool**. We then use the semantics of the [CLS] token to dynamically **compose** sparse, task-specific subspaces from this internal expert pool. This provides a more fine-grained and parameter-efficient mechanism for continual learning.

Remarks. To sum up, while existing continual learning methods for VLMs are limited by their training and inference burdens, our work offers a significantly more efficient approach, by restructuring a single LoRA as a decomposable Rank-1 Expert Pool that are dynamically composed for each task. This avoids the parameter and computational overhead caused by combining multiple, separate LoRA modules, offering a truly lightweight solution.

Method

Problem Definition

Following the standard setup for Vision-Language Models (VLMs) like CLIP (Radford et al. 2021), our model consists of an image encoder f_θ and a text encoder g_ψ implemented as transformers. Classification is performed by calculating

the cosine similarity between the image embeddings $\mathbf{z}^V = f_\theta(\mathbf{x})$ and text embeddings $\mathbf{z}^T = g_\psi(\mathbf{t})$. During inference, the probability of classifying the image \mathbf{x} into class $y_i \in \{1, \dots, C\}$ is calculated as $p(y_i|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{z}^V, \mathbf{z}_i^T))}{\sum_{c=1}^C \exp(\text{sim}(\mathbf{z}^V, \mathbf{z}_c^T))}$, where $\text{sim}(\cdot)$ represents a cosine similarity metric.

In this work, we first evaluate our method on the multi-domain task-incremental learning (MTIL) scenario, a classic benchmark for multi-task and multi-domain continual learning (Zheng et al. 2023; Yu et al. 2024). In this setting, the model learns sequentially from a series of 11 tasks. For each task t , the dataset is represented as $\mathcal{D}^t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N^t}$, where $\mathbf{x}_i^t \in \mathbb{R}^{H \times W \times C}$ is the input image, $y_i^t \in \mathcal{C}^t$ is its corresponding class label, and N^t is the number of samples in task t , respectively. The class set $\mathcal{C}^t = \{y_j^t\}_{j=1}^{M^t}$ consists of the class names in task t , with a total of M^t classes. Additionally, to demonstrate the robustness and generalization of our approach, we adopt another cross-domain task-agnostic incremental learning (X-TAIL) setting (Xu et al. 2024). This benchmark is particularly challenging because it introduces task-agnostic settings during training, making it suitable for evaluating model generalization under realistic conditions.

Overview Framework

The core design of our framework is to achieve efficient continual learning by intelligently managing parameter updates within a single module, thereby avoiding common dependencies on external data or inference overhead. To this end, we restructure a single LoRA module as a **decompos-**

able Rank-1 Expert Pool. Guided by the semantics of the [CLS] token, our method learns to dynamically **compose** a sparse, task-specific update by selecting from this expert pool. To complement this sparse composition, we introduce an **Activation-Guided Orthogonal (AGO) loss** that minimizes parameter collision (Yang et al. 2024a) by isolating critical updates in the existing parameter space, ensuring efficient learning without complex architectural burdens.

The overall framework of our method is shown in Figure 2. We freeze the original weights of CLIP and add LoRA layers to all linear layers in the multi-head attention and MLP blocks of the Image and Text encoders. After the training for each task is completed, the composed sparsified LoRA weights are merged back into the original CLIP weights to ensure no additional inference burden is introduced. This process is repeated for each subsequent task, and for each task, only the LoRA weights and frequency information need to be saved. Except for the first task, which has no ‘‘Past LoRA’’ currently, the AGO loss is applied in subsequent training to isolate the parameter optimization direction between the current and previous tasks’ LoRAs, enhancing the model’s adaptability to new tasks without being overly affected by previous parameter updates.

Dynamic Composition from a Rank-1 Expert Pool

Many previous works fine-tuning with LoRA have introduced static metrics for sparsity and pruning (Meng, Wang, and Zhang 2024; Zhang et al. 2023b). For example, DoRA (Mao et al. 2024) uses the Frobenius norm to measure rank importance. However, we argue that such static analyses are ill-suited for the dynamic nature of continual learning. As shown in Figure 3, the static importance of a rank does not reliably predict its contribution to performance in a dynamic setting. This finding motivates our dynamic approach: instead of treating LoRA as a monolithic block to be pruned, we view it as a pool of fine-grained, rank-1 experts that can be dynamically composed based on different task demands.

Our approach begins by decomposing a rank- r LoRA matrix ΔW . Following (Hu et al. 2022), $\Delta W = \mathbf{B}\mathbf{A}$, where $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d}$. This can be expressed as the sum of r rank-1 matrices:

$$\Delta W = [\mathbf{b}_1 \mathbf{b}_2 \cdots \mathbf{b}_r] \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_r^\top \end{bmatrix} = \sum_{i=1}^r \mathbf{b}_i \mathbf{a}_i^\top, \quad (1)$$

where $\Delta W_{(1,i)} = \mathbf{b}_i \mathbf{a}_i^\top$ is a rank-1 matrix formed by vectors $\mathbf{b}_i \in \mathbb{R}^{d \times 1}$ and $\mathbf{a}_i^\top \in \mathbb{R}^{1 \times d}$. We treat this set of r rank-1 matrices, $\{\Delta W_{(1,i)}\}_{i=1}^r$, as our **Rank-1 Expert Pool**.

To dynamically compose an update from this pool, we employ a lightweight router guided by the input’s semantics. Inspired by the special role of the [CLS] token in capturing global information in Transformers (Liang et al. 2022; Wang et al. 2024), we use this feature representation $\phi(\mathbf{x})$ as the router’s input. The router, a linear layer $\mathbf{W}_{router} \in \mathbb{R}^{r \times d_{CLS}}$, produces expert selection scores $\mathbf{s}_n = \mathbf{W}_{router} \phi(\mathbf{x}_n)$ for each input \mathbf{x}_n . The output is then a com-

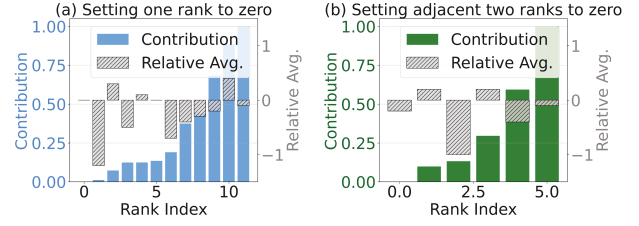


Figure 3: Comparison of the Average metric between setting no rank to zero and setting either one rank (a) or two adjacent ranks (b) to zero during the merging process. The rank is sorted in ascending order of Frobenius-norm: The contributions of each rank increase from left to right in sequence. The figure indicates that a higher contribution does not necessarily lead to performance improvement and shows ranks with low contribution may also play a crucial role in the task.

position of the original weights and the selected experts:

$$y = \sigma \left(W_0(x) + \sum_{i=1}^r \pi_i(\mathbf{W}_{router} \phi(x)) \mathbf{b}_i \mathbf{a}_i^\top x \right), \quad (2)$$

where W_0 is the original weight matrix and σ is an activation function. The term $\mathbf{W}_{router} \phi(x)$ calculates scores that determine the contribution of the i -th expert, $\mathbf{b}_i \mathbf{a}_i^\top$. These scores are also used for the subsequent sparsification of LoRA. Finally, π_i is a gating function that leverages these scores to determine the selection of the i -th expert. During training, we apply a two-stage composition to select the most relevant R experts for the current batch. First, for each sample, we identify the top- R experts based on its scores, denoted by the index set $\mathcal{S}_n = \text{Top}(\mathbf{s}_n, R)$. Second, we aggregate these selections across the batch to obtain an expert vote count \mathbf{v} , and select the overall top- R experts for activation based on \mathbf{v} in current batch, resulting in the final set $\mathcal{S}_{batch} = \text{Top}(\mathbf{v}, R)$. This ensures that only the most critical experts for the current data distribution are updated. We also maintain an activation frequency memory, \mathbf{C}_l , and for each sample \mathbf{x}_n , $\mathbf{C}_l(i) = \mathbf{C}_l(i) + 1$ if $i \in \mathcal{S}_n$. This dynamic selection mechanism allows the model to form sparse subspaces for different datasets, as discussed in ablation studies.

Activation-Guided Orthogonal (AGO) Loss

Previous work suggests that LoRA parameters encapsulate crucial model update directions (Wang et al. 2023; Yang et al. 2024a). Therefore, learning in a subspace orthogonal to those of previous tasks can alleviate catastrophic forgetting. A standard approach is to enforce orthogonality between the LoRA matrix \mathbf{B}_t of the current task t and the concatenated matrix of all past tasks, $\mathbf{B}_{past} = \text{Concat}(\mathbf{B}_1, \cdots, \mathbf{B}_{t-1})$. We formalize the orthogonality constraint as follows:

$$L_{orth} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \left| (\mathbf{b}_{past}^i)^\top \mathbf{b}_t^j \right|, \quad (3)$$

where we define $\mathbf{B}_{past} = [\mathbf{b}_{past}^1, \mathbf{b}_{past}^2, \cdots, \mathbf{b}_{past}^m]$ and $\mathbf{B}_t = [\mathbf{b}_t^1, \mathbf{b}_t^2, \cdots, \mathbf{b}_t^n]$, and m and n are the number of column vectors (basis vectors) in \mathbf{B}_{past} and \mathbf{B}_t , respectively.

However, this naive orthogonality suffers from two key issues. First, it can lead to “parameter collision” (Yang et al. 2024a), where parameters still interfere with each other despite the overall subspace orthogonality. Second, applying a strong orthogonal loss can compromise the optimization of the primary classification task, harming final performance.

To resolve this, we propose an **Activation-Guided Orthogonal (AGO) loss** that synergizes with our Rank-1 Expert Pool. Instead of enforcing orthogonality on the entire dense LoRA subspace, our approach uses the expert activation frequencies \mathbf{C}_l recorded during previous compositions. We identify the set of the top- R most critical experts from past tasks, denoted as $\mathcal{S}_{freq} = Top(\mathbf{C}_l, R)$. We then construct a sparse LoRA matrix, ΔW_{AGO} , where only these critical experts are retained for the loss calculation:

$$\Delta W_{AGO} = \sum_{i \in \mathcal{S}_{freq}} \mathbf{b}_i \mathbf{a}_i^\top + \sum_{i \notin \mathcal{S}_{freq}} \mathbf{0}. \quad (4)$$

This means that ranks deemed unimportant for the current task, based on historical activation, are temporarily discarded here. By applying L_{orth} only to these sparse, task-critical subspaces, we can effectively reduce parameter collision and isolate optimization directions without overly impacting the learning of the downstream classification task.

Training and Inference

Our training objective combines a standard supervised classification loss, L_{sup} (using the alignment loss from CLIP), with our proposed AGO loss, L_{orth} , described in previous subsection. The final loss is:

$$L = L_{sup} + \lambda L_{orth}, \quad (5)$$

where the hyper-parameter λ balances the two terms.

Upon completion of training for each task, the learned LoRA weights are merged back into the main model’s original weights, resulting in zero additional inference overhead. The decomposition and composition of the expert pool allows for flexible merging strategies at inference time. By default, to strike a balance between performance on the new task and retention of past knowledge, we merge only the top- $R/2$ experts as determined by their activation frequency. However, in scenarios where task-id is known (Yu et al. 2024), our method can adaptively merge a more complete set of experts specific to that task, offering flexibility for practical situations. *We provide a detailed analysis of our method’s computational advantages in **Computation Cost** section.*

Experiments

Experimental Setting

In this part, we detail the experimental setup employed to evaluate our proposed method. Our experiments mainly focus on classic and challenging learning scenarios for CLIP.

Datasets In the Multi-domain Task Incremental Learning (MTIL) setting and the Task-Agnostic Incremental Learning (X-TAIL) setting, we utilize a total of 11 datasets: Aircraft (Maji et al. 2013), Caltech101 (Fei-Fei, Fergus, and Perona 2004), CIFAR 100 (Krizhevsky, Hinton et al. 2009),

DTD (Cimpoi et al. 2014), EuroSAT (Helber et al. 2019), Flowers (Nilsback and Zisserman 2008), Food (Bossard, Guillaumin, and Van Gool 2014), MNIST (Deng 2012), OxfordPet (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), and SUN397 (Xiao et al. 2010). Each dataset is treated as an individual task for continual learning. Specifically, we adopt a 5-shot split for both MTIL and X-TAIL (Yu et al. 2024; Xu et al. 2024).

Metrics Following previous works (Zheng et al. 2023), we use three standard metrics: “Transfer” (generalization to unseen data), “Average” (overall performance across all tasks), and “Last” (knowledge retention from past tasks).

Implementation Details We follow the overall setups in (Zheng et al. 2023; Yu et al. 2024), using a CLIP ViT-B/16 backbone with our technique applied to all linear layers. Key hyperparameters are: LoRA rank set to 12, with 8 ranks retained after sparsification and the top 4 experts merged after training. The orthogonal loss weight λ is 0.1. We train for 500 iterations per task using the AdamW optimizer with a batch size of 32 and a learning rate of $2e-3$. All experiments were conducted on RTX A6000 GPUs.

Comparisons with State-Of-The-Arts

We compare our method with traditional continual learning baselines (LWF-VR (Ding et al. 2022), WISE-FT (Wortman et al. 2022)) and recent VLM-specific methods (ZSCL (Zheng et al. 2023), MoE-a (Yu et al. 2024), RAIL (Xu et al. 2024), GIFT (Wu et al. 2025a)).

As presented in Table 1, here 65.3% is the CLIP’s average across all 11 tasks. It can be seen that our method achieves the best performance regardless of whether task-id prior knowledge is available or not. Specifically, when task-id is not given, our method outperforms the previous state-of-the-art (SOTA) methods by 3.2%, 3.7%, 6.0% in the Transfer, Average, Last metric, indicating that the method not only improves the performance of downstream tasks, but also better preserves the ability of pre-trained weights. It is worth noting that here our method even does not introduce any additional external large-scale datasets or synthetic data replay. Moreover, when task-id is given, our method still has an edge over previous methods, with improvements of 0.9%, 1.6%, and 2.0% respectively in the corresponding metrics. Notably, our method exceeds the zero-shot upper bound performance of CLIP in the Transfer metric, and consistently shows a steady improvement over all datasets during continual learning, which demonstrates the excellent generalization ability of our method.

Computation Cost

As shown in Table 2, our framework is remarkably computationally efficient. Compared to full fine-tuning baselines (Zheng et al. 2023; Wu et al. 2025a; Yu et al. 2024), our approach reduces trainable parameters by **96.7%** and peak GPU memory consumption by **66.7%**. Moreover, Our method’s training/inference speeds are 1.84/2.98 it/s, compared to MoE-a’s 0.78/1.04 it/s for our method’s adapter-free style, introducing zero inference overhead. Crucially, these efficiency gains are achieved without requiring any

Method	Task-ID	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	Pets	Cars	SUN397	Average
CLIP													
Zero-shot													
5-shot Full Fine-tune	TIF	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.6	89.0	64.7	65.2	65.3
		30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
Transfer													
Continual-FT		-	72.8	53.0	36.4	35.4	43.3	68.4	47.4	72.6	30.0	52.7	51.2
LwF-VR [Arxiv'22]		-	82.2	62.5	40.1	40.1	56.3	80.0	60.9	77.6	40.5	60.8	60.1
WiSE-FT [CVPR'22]		-	77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
ZSCL [ICCV'23]	TIF	-	84.0	68.1	44.8	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
GIFT [CVPR'25]	-	-	88.6	65.0	45.9	49.8	64.1	84.9	61.7	88.5	55.5	67.5	67.1
Ours	-	-	88.3	68.8	45.4	58.6	71.3	87.7	61.0	90.0	64.1	67.4	70.3 (+3.2)
MoE-a [CVPR'24]	TIK	-	87.9	68.2	44.1	48.1	64.7	88.8	69.0	89.1	64.5	65.1	68.9
RAIL [NeurIPS'24]	-	-	88.4	68.2	44.6	54.9	71.0	88.5	59.6	89.0	64.7	65.2	69.4
Ours	-	-	88.3	68.8	45.4	58.6	71.3	87.7	61.0	90.0	64.1	67.4	70.3 (+0.9)
Average													
Continual-FT		28.1	86.4	59.1	52.8	55.8	62.0	70.2	64.7	75.5	35.0	54.0	58.5
LwF-VR [Arxiv'22]		24.9	89.1	64.2	53.4	54.3	70.8	79.2	66.5	79.2	44.1	61.6	62.5
WiSE-FT [CVPR'22]		32.0	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
ZSCL [ICCV'23]	TIF	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
GIFT [CVPR'25]	-	29.2	90.4	65.7	57.6	69.9	78.0	84.8	73.0	88.2	57.8	67.9	69.3
Ours	-	33.4	91.4	74.5	59.6	77.8	80.0	88.0	73.4	90.9	65.4	67.9	73.0 (+3.7)
MoE-a [CVPR'24]	TIK	30.0	89.6	73.9	58.7	69.3	79.3	88.1	76.5	89.1	65.3	65.8	71.4
RAIL [NeurIPS'24]	-	32.9	94.5	69.9	58.1	71.8	84.4	88.5	70.4	89.0	66.1	65.7	71.9
Ours	-	36.5	90.7	76.3	59.9	78.6	80.6	88.0	73.4	91.0	65.4	67.9	73.5 (+1.6)
Last													
Continual-FT		27.8	86.9	60.1	58.4	56.6	75.7	73.8	93.1	82.5	57.0	66.8	67.1
LwF-VR [Arxiv'22]		22.9	89.8	59.3	57.1	57.6	79.2	78.3	77.7	83.6	60.1	69.8	66.9
WiSE-FT [CVPR'22]		30.8	88.9	59.6	60.3	80.9	81.7	77.1	94.9	83.2	62.8	70.0	71.9
ZSCL [ICCV'23]	TIF	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
GIFT [CVPR'25]	-	27.9	89.8	46.0	62.0	71.9	87.8	83.4	93.0	86.1	67.6	71.8	71.6
Ours	-	31.1	93.0	75.0	64.1	87.4	85.4	88.5	95.1	93.5	71.2	72.8	77.9 (+6.0)
MoE-a [CVPR'24]	TIK	30.1	89.3	74.9	64.0	82.3	89.4	87.1	89.0	89.1	69.5	72.5	76.1
RAIL [NeurIPS'24]	-	32.9	95.1	70.3	63.2	81.5	95.6	88.5	89.7	89.0	72.5	71.0	77.2
Ours	-	36.5	90.9	77.9	65.3	90.1	88.4	88.4	95.2	93.7	71.4	72.8	79.2 (+2.0)

Table 1: Comparisons with state-of-the-art methods on few-shot MTIL Order 1 benchmark in terms of ‘‘Transfer’’, ‘‘Average’’, and ‘‘Last’’ scores (%). **TIK** means Task-Id Known, and **TIF** means Task-Id Free. We label the best method with bold style.

Method	Training Param. (MB)	Extra Data & Components	GPU Mem Cost (MB)
ZSCL	570.76 (-0)	IN	28454 (-0)
MoE-a	194.78 (-65.5%)	AN & TN	14040 (-50.7%)
GIFT	570.76 (-0)	SD	22990 (-19.2%)
Ours	18.99 (-96.7%)	-	9490 (-66.7%)

Table 2: Computational Cost Comparison (IN: ImageNet1K, AN: AlexNet, TN: TinyImageNet, SD: Stable Diffusion).

extra datasets or additional components (data replay techniques (Zheng et al. 2023) or additional adapters (Xu et al. 2024)), which have been commonly used in these previous works, showing our method’s practicality for realistic continual learning scenarios.

Discussion

Analysis of Method Components

We ablate the effects of our method’s key components in Table 3. Both the dynamic composition from our Rank-1 Expert Pool and the **Activation-Guided Orthogonal (AGO) loss** individually contribute positively to performance when compared to baselines. Specifically, the dynamic composition of rank-1 experts prevents redundant weight updates, enhancing transfer ability and improving overall classification performance. The AGO loss, in turn, focuses the model on critical parameter updates, boosting downstream task accuracy. When the two are combined, they achieve the best results, demonstrating a synergistic effect.

Notably, all metrics in this section are presented as (%), and all ablation studies are conducted on the same experimental setting as Table 1, unless otherwise stated.

Method	Trans.	Δ	Avg.	Δ	Last	Δ
Baselines						
Zero-shot	69.4	0.0	65.3	0.0	65.3	0.0
ZSCL	65.3	-4.1	64.4	-0.9	67.4	<u>+2.1</u>
SD-LoRA	67.4	-2.0	70.1	<u>+4.8</u>	74.2	<u>+8.9</u>
LoRAMoE	68.0	-1.4	69.9	<u>+4.6</u>	70.3	<u>+5.0</u>
Ablation						
Vanilla LoRA	65.0	-4.4	68.5	<u>+3.2</u>	73.8	<u>+8.5</u>
+ Expert Pool	69.7	<u>+0.3</u>	72.4	<u>+7.1</u>	77.1	<u>+11.8</u>
+ AGO loss	65.3	-4.1	69.2	<u>+3.9</u>	74.5	<u>+9.2</u>
+Ours	70.3	<u>+0.9</u>	73.0	<u>+7.7</u>	77.9	<u>+12.6</u>

Table 3: Ablation study on method components, conducted on MTIL setting Order 1. Underline means improvement.

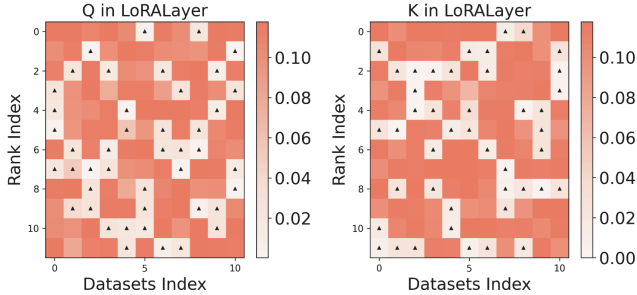


Figure 4: Visualization of the activation frequencies of LoRA under different tasks. The darker the color, the higher frequency. “▲” means disregarded experts. This figure shows frequency of the first LoRA layer in image encoder, indicating that the composed part of each task is exactly the experts with relatively higher activation.

Discussion on LoRA-based CL method

We also present other LoRA-based continual learning methods in Table 3: SD-LoRA (Wu et al. 2025b) and LoRAMoE (Dou et al. 2023), which are primarily designed to address simple static incremental problems. As shown, their performance is suboptimal when facing fine-grained dynamic domain discrepancies, where our approach performs better.

Analysis of Expert Composition

To verify that our framework selects domain-specific experts from the expert pool, we visualize the expert activation frequencies in Figure 4. The heatmaps show distinct activation patterns across different datasets (columns), confirming that the [CLS] token effectively guides the router to compose domain-aware subspaces.

We also study the impact of which experts are chosen for the final merged LoRA in Table 4. The results clearly indicate that merging the experts with the highest activation frequency (“Top”) yields the best performance. Conversely, using low-frequency experts (“Down”) critically degrades performance. Interestingly, a highly sparse composition (Top-2) further improves the Transfer metric, likely due to extreme fewer parameter updates enhancing generalization. Our default setting (Top-4) is chosen to strike an optimal balance

Strategy	Trans.	Δ	Avg.	Δ	Last	Δ
Great Experts						
Top-2 (2 E)	70.5	<u>+0.2</u>	71.5	-1.5	75.3	-2.6
Top-4 (Ours)	70.3	0.0	73.0	0.0	77.9	0.0
Top-8 (8 E)	67.2	-3.1	69.6	-3.4	73.7	-4.2
Good Experts						
Mid-4 (4 E)	70.3	0.0	72.8	-0.2	77.4	-0.5
Bad Experts						
Down-4 (4 E)	68.4	-1.9	64.0	-9.0	62.9	-15.0
Down-8 (8 E)	68.3	-2.0	70.2	-2.8	72.4	-5.5

Table 4: Ablation study on rank composing selection, involving expert number and frequency selection, 2 E means selecting 2 experts for sparsification.

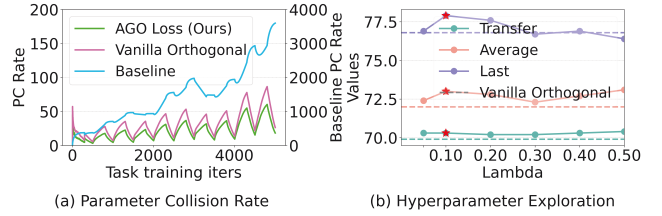


Figure 5: (a) Parameter collision rate comparison: Our method has fewer parameter collisions than vanilla orthogonal LoRA and non-orthogonal LoRA. (b) λ hyperparameter exploration: Our method is robust to λ , outperforming vanilla orthogonal LoRA in all metrics.

between downstream performance and generalization.

Analysis of Activation-Guided Orthogonal Loss

We analyze the advantage of AGO loss by measuring the parameter collision rate, where a higher L_{orth} value implies more collisions. As shown in Figure 5(a), our activation-guided approach maintains a lower collision rate than both a non-orthogonal baseline and a vanilla (dense) orthogonal loss, effectively isolating task-specific updates. Figure 5(b) explores the impact of the loss weight, λ , which defines the proportion of our AGO loss. The figure shows that as the proportion of the orthogonal loss increases, downstream task performance is affected while the generalization ability (Transfer) rises. The results also show our method is robust to the choice of λ , with $\lambda = 0.1$ providing the best balance.

Conclusion

In this paper, we introduced an efficient continual learning framework by restructuring a single LoRA module into a **decomposable** Rank-1 Expert Pool. Guided by a lightweight, semantics-aware router, our method dynamically **composes** sparse, task-specific subspaces from this pool. These composed subspaces are further decoupled by a synergistic **Activation-Guided Orthogonal (AGO) loss**. Extensive experiments demonstrate that our approach achieves state-of-the-art results across various continual learning settings while eliminating additional inference costs.

Acknowledgments

This work is supported by NSFC Project (62222604, 62192783, 624B2063, 62506162, 62476136), Jiangsu Science and Technology Project (BK20251241, BK20250142).

References

- Agarwal, A.; Banerjee, B.; Cuzzolin, F.; and Chaudhuri, S. 2022. Semantics-driven generative replay for few-shot class incremental learning. In *Proceedings of the 30th ACM international conference on multimedia*, 5246–5254.
- Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, 139–154.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, 446–461. Springer.
- Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.
- Ding, N.; Lv, X.; Wang, Q.; Chen, Y.; Zhou, B.; Liu, Z.; and Sun, M. 2023. Sparse low-rank adaptation of pre-trained language models. *arXiv preprint arXiv:2311.11696*.
- Ding, Y.; Liu, L.; Tian, C.; Yang, J.; and Ding, H. 2022. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*.
- Dou, S.; Zhou, E.; Liu, Y.; Gao, S.; Zhao, J.; Shen, W.; Zhou, Y.; Xi, Z.; Wang, X.; Fan, X.; et al. 2023. LoRAMoE: Alleviate world knowledge forgetting in large language models via MoE-style plugin. *arXiv preprint arXiv:2312.09979*.
- Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9285–9295.
- Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.
- Feng, J.; Pu, Z.; Hu, T.; Li, D.; Ai, X.; and Wang, H. 2025. OMoE: Diversifying Mixture of Low-Rank Adaptation by Orthogonal Finetuning. *arXiv preprint arXiv:2501.10062*.
- Gao, C.; Chen, K.; Rao, J.; Sun, B.; Liu, R.; Peng, D.; Zhang, Y.; Guo, X.; Yang, J.; and Subrahmanian, V. 2024. Higher layers need more lora experts. *arXiv preprint arXiv:2402.08562*.
- Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; and Herzig, J. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.
- Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.
- Hong, H.; Wang, S.; Huang, Z.; Wu, Q.; and Liu, J. 2024. Navigating Beyond Instructions: Vision-and-Language Navigation in Obstructed Environments. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7639–7648.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Jha, S.; Gong, D.; and Yao, L. 2024. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *arXiv preprint arXiv:2403.19137*.
- Jiang, S.; Liao, Y.; Zhang, Y.; Wang, Y.; and Wang, Y. 2025. Fine-tuning with Reserved Majority for Noise Reduction. In *The Thirteenth International Conference on Learning Representations*.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Lavda, F.; Ramapuram, J.; Gregorova, M.; and Kalousis, A. 2018. Continual classification learning using generative models. *arXiv preprint arXiv:1810.10612*.
- Lee, K.-Y.; Zhong, Y.; and Wang, Y.-X. 2023. Do pre-trained models benefit equally in continual learning? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6485–6493.
- Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International conference on machine learning*, 3925–3934. PMLR.
- Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.
- Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*.

- Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Xu, D.; Tian, F.; and Zheng, Y. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1104–1114.
- Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.
- Mao, Y.; Huang, K.; Guan, C.; Bao, G.; Mo, F.; and Xu, J. 2024. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. *arXiv preprint arXiv:2405.17357*.
- Meng, F.; Wang, Z.; and Zhang, M. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072.
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.
- Parelli, M.; Delitzas, A.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5607–5612.
- Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmlR.
- Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.
- Ren, W.; Li, X.; Wang, L.; Zhao, T.; and Qin, W. 2024. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752*.
- Thengane, V.; Khan, S.; Hayat, M.; and Khan, F. 2022. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*.
- Valipour, M.; Rezagholizadeh, M.; Kobzyev, I.; and Ghodsi, A. 2022. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*.
- Wang, A.; Sun, F.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2024. [CLS] Token Tells Everything Needed for Training-free Efficient MLLMs. *arXiv preprint arXiv:2412.05819*.
- Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X. 2023. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*.
- Wei, X.; Li, G.; and Marculescu, R. 2024. Online-LoRA: Task-free Online Continual Learning via Low Rank Adaptation. *arXiv preprint arXiv:2411.05663*.
- Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7959–7971.
- Wu, B.; Shi, W.; Wang, J.; and Ye, M. 2025a. Synthetic Data is an Elegant GIFT for Continual Vision-Language Models. *arXiv preprint arXiv:2503.04229*.
- Wu, Y.; Piao, H.; Huang, L.-K.; Wang, R.; Li, W.; Pfister, H.; Meng, D.; Ma, K.; and Wei, Y. 2025b. SD-LoRA: Scalable Decoupled Low-Rank Adaptation for Class Incremental Learning. In *The Thirteenth International Conference on Learning Representations*.
- Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.
- Xu, Y.; Chen, Y.; Nie, J.; Wang, Y.; Zhuang, H.; and Okumura, M. 2024. Advancing Cross-domain Discriminability in Continual Learning of Vision-Language Models. *arXiv preprint arXiv:2406.18868*.
- Yang, S.; Ning, K.-P.; Liu, Y.-Y.; Yao, J.-Y.; Tian, Y.-H.; Song, Y.-B.; and Yuan, L. 2024a. Is Parameter Collision Hinderling Continual Learning in LLMs? *arXiv preprint arXiv:2410.10179*.
- Yang, Y.; Li, X.; Zhou, Z.; Song, S.; Wu, J.; Nie, L.; and Ghanem, B. 2024b. CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning. *Advances in Neural Information Processing Systems*, 37: 71768–71791.
- Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; and He, Y. 2024. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23219–23230.
- Zhang, F.; Li, L.; Chen, J.; Jiang, Z.; Wang, B.; and Qian, Y. 2023a. Incelora: Incremental parameter allocation method for parameter-efficient fine-tuning. *arXiv preprint arXiv:2308.12043*.
- Zhang, Q.; Chen, M.; Bukharin, A.; Karampatziakis, N.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023b. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.
- Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; and You, Y. 2023. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19125–19136.