KSS-MoE: Knowledge Space Synergy Framework in Mixture of Experts for Continual Visual Instruction Tuning

Lingyun Song[1, 2, 3*], Ziyao Chen[1], Kang Pan[4], Xiaolin Han[1], Xinbiao Gan[5*], Yudai Pan[1], Xiaofan Sun[1], Xiaoqi Wang[1], Xuequn Shang[3, 6*]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an
[2]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua
[3]Shenzhen Research Institute of Northwestern Polytechnical University, Shenzhen
[4]Independent Researcher
[5]School of Computer Science, National University of Defense Technology, Changsha
[6]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an
lysong@nwpu.edu.cn, xinbiaogan@nudt.edu.cn, shang@nwpu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) employing the Mixture-of-Experts (MoE) structure exhibit encouraging results in visual language tasks. However, they struggle with catastrophic forgetting due to a lack of effective collaboration among experts and negative transfer across tasks. This happens because the router typically employed in MoE for managing expert assignments is inadequate when there are significant shifts in data distribution across various tasks. A drop in the effectiveness of earlier tasks is caused by negative transfer, which occurs due to conflicts in shared knowledge between tasks, disturbing the knowledge already acquired. To address these issues, we propose the Knowledge Space Synergy Framework in Mixture of Experts (KSS-MoE) for Continual Visual Instruction Tuning (CVIT). It dynamically combines the knowledge subspaces of experts to improve the integration of fine-grained complementary knowledge and collaborative abilities of experts, thus addressing the limitations of the basic router. Furthermore, we introduce a general expert that maintains orthogonal subspaces for shared knowledge, enabling effective cross-task knowledge utilization while reducing negative transfer. Extensive experiments conducted on eight CVIT tasks confirm the excellence of KSS-MoE, showcasing its top-tier performance.

Code — https://github.com/sunlitsong/KSS-MoE

Introduction

Multimodal Large Language Models (MLLMs) (Achiam et al. 2023; Liu et al. 2023a; Team et al. 2024; Yang et al. 2025) exhibit exceptional capabilities in multiple visual–language tasks such as visual question answering and image captioning. The MLLM training process generally involves two distinct phases. During the pre-training phase, a large dataset of image-text pairs is utilized to ensure cross-modal alignment. Following this, the fine-tuning phase involves training the MLLMs with instruction datasets to adhere to human instructions. Although this offline training method works well for a variety of tasks, real-world data is continually changing and arrives sequentially, which can cause a gradual drop in performance. Furthermore, it often results in what is known as "catastrophic forgetting" of previously acquired knowledge (Zhai et al. 2023).

To this end, Continual Visual Instruction Tuning (CVIT) integrates continual learning techniques into fine-tuning pipeline (Zhu et al. 2024a; Zheng et al. 2024; Huai et al. 2025; Qiao et al. 2024; Wang et al. 2024), with the goal of assimilating new information while retaining competence on prior tasks. However, in multimodal settings, divergent cross-modal distributions and substantial domain gaps are frequently introduced, exacerbating the conflict between acquiring new knowledge and preserving previously learned representations. As a result, devising mechanisms that strike an optimal balance between adaptability and knowledge retention remains an open and pressing challenge.

Mixture-of-experts (MoE), comprising several specialized task experts and a trainable router, is extensively used in MLLMs because its sparse activation greatly lowers computational demands and boosts efficiency. Designing an advanced router (Nguyen et al. 2024; Zuo et al. 2022; Wang and Li 2024) to govern expert collaboration and incorporating a general expert (Gou et al. 2023; Dai et al. 2024) to gather knowledge across tasks are gaining significant attention. Moreover, MoE presents itself as an encouraging approach for continual learning (Gururangan et al. 2021; Huai et al. 2025), as it lessens interference in the presence of unbalanced task distributions by optimizing expert specialization.

Unfortunately, existing MoE-based MLLMs still struggle with several critical challenges in CVIT scenarios. Firstly, when there is a substantial shift in task data distribution, the router in MoE may experience forgetting, undermining the experts' collaboration. Secondly, because certain parameters of the general expert can be reused for novel tasks, task

*Corresponding authors.

discrepancies frequently cause knowledge conflicts. These conflicts significantly disturb the shared knowledge acquired from previous tasks, resulting in negative transfer (Zhang et al. 2022), where cross-task knowledge transfer leads to impaired performance.

To address these issues, we propose the Knowledge Space Synergy Framework in Mixture of Experts (KSS-MoE) for CVIT. We have crafted both a set of task-specific experts and a single general expert. The task-specific experts concentrate on acquiring knowledge for specific tasks, whereas the general expert is committed to capturing shared knowledge applicable across various tasks. Leveraging internal activations, we assess knowledge relevance across dimensions within each expert. The most relevant knowledge subspaces are then extracted and fused to construct a composite expert for inference, achieving fine-grained integration of all experts' capabilities. Furthermore, our KSS-MoE refreshes the shared knowledge for the general expert through an "orthogonal complement subspace" based on task vectors (Ilharco et al. 2022). This method prevents disruptions caused by task knowledge conflicts, and thus achieving a plasticity-stability balance. The key contributions of our work can be summarized as follows.

- We introduce a novel method aimed at enhancing knowledge complementarity among experts by dynamically collaborating knowledge subspaces within experts. This enhances MoE's ability to generalize, while minimizing the issue of catastrophic forgetting caused by distribution shifts and ensuring the long-term retention of prior knowledge.
- We propose an orthogonal fusion method to integrate shared knowledge subspaces, aiming to enhance the learning ability of the general expert in CVIT. Preserving orthogonality in shared knowledge across tasks results in the merging of subspaces that are both highly discriminative and minimally redundant. This aids in collecting shared knowledge while preventing negative transfer.
- We perform comprehensive comparisons against the previous state-of-the-art baselines, and the experimental results reveal that our work exceeds all competing baselines on the CoIN benchmark (Chen et al. 2024), encompassing eight datasets.

Background and Related Work

Multimodal Large Language Model

Due to their remarkable performance in multimodal tasks, MLLMs have gained significant research interest. The majority of MLLMs consist of three main parts: a Large Language Model (LLM) backbone(Touvron et al. 2023; Guo et al. 2025), a vision encoder (Dosovitskiy et al. 2020; Radford et al. 2021), and a vision-to-language projector(Li et al. 2023; Lin et al. 2022). However, full fine-tuning of MLLMs for each new task is prohibitively costly. To address this, parameter-efficient fine-tuning (PEFT) methods adapt MLLMs with minimal extra parameters. A group of approaches inject learnable parameters into input embeddings (Li and Liang 2021; Liu et al. 2021; Lester, Al-Rfou, and Constant 2021), whereas LoRA (Hu et al. 2022) employs low-rank matrices with a small number of extra parameters. Beyond these PEFT techniques, the Mixture-of-Experts (MoE) framework enables dynamic computational resource allocation, reducing training costs. In this study, we integrate MoE with LoRA-based MLLMs in the CVIT setting, allowing adaptation to new data while preserving prior knowledge.

Continual Learning for LLMs

Current approaches aim to address the issue of catastrophic forgetting in LLMs under a continual learning framework. Various studies (Tong et al. 2025; Wang et al. 2023; Zheng et al. 2024) either separate task learning into low-rank subspaces or use orthogonal gradient projection techniques to minimize interference between tasks. Recent studies (Liu et al. 2024; Zhu et al. 2024a; Cao et al. 2024) employ techniques like parameter isolation and selective updating to safeguard previously learned information from being erased. Capitalizing on MoE's outstanding performance in LLMs, a growing body of research (Wang et al. 2024; Zhao et al. 2025; Zhu et al. 2024b) examines more advanced experts and routers, aiming to localize updates for new tasks within specific substructures. However, existing methods focus primarily on task isolation but neglect cross-task knowledge sharing, impairing both knowledge retention and cross-task generalization. We therefore propose a cross-task synergy approach to mitigate forgetting issue triggered by distribution shift.

MoE in Continual Learning

The MoE framework has demonstrated potential in the realm of continual learning. In MoE, various studies (Gururangan et al. 2021; Chen et al. 2023a) assign task-specific units through the use of conditional activation and regularization, effectively minimizing task interference. Additional advances (Huai et al. 2025; Zhang et al. 2025; Wang and Li 2024) improve routing strategies for dynamically balancing existing and new knowledge. Nevertheless, MoE experiences constraints in terms of collaboration among experts, which hinders its ability to generalize effectively. Although some approaches (Cai et al. 2024; Chen et al. 2023b; Gou et al. 2023; Dai et al. 2024) introduce an always active general expert to foster knowledge exchange, they risk negative transfer due to inter-task conflicts. In contrast, our KSS-MoE enables cross-task complementary collaboration through fine-grained knowledge subspace fusion, while its general expert maintains the shared knowledge orthogonally, prevents negative transfer from knowledge conflicts.

Preliminaries

Task Definition

Let $\mathcal{D} = \{D_1, D_2 \cdots, D_T\}$ be a sequence of CVIT tasks over T multimodal domains. The t-th task D_t contains n_t inputs $X_t \in \{(I_i^v, I_i^q, A_i)\}_{i=1}^{n_t}$, where I_i^v, I_i^q, and A_i represent the image, question, and their associated answer for the i-th input, respectively. CVIT aims to learn new task

D_t while maintaining the knowledge acquired from previous tasks $D_{1:t-1}$, effectively preventing catastrophic forgetting.

We frame $\mathcal{D}$ as text generation tasks, and the MLLM is trained to maximize the conditional likelihood of the output sequence autoregressively by

$$P_\theta(A|I^v, I^q) = \prod_{k=1}^{K} P_\theta(A_k|A_{<k}, I^v, I^q), \tag{1}$$

where A_k is the k-th token in the answer, $A_{<k}$ denotes the first $k-1$ tokens, θ are the MLLM parameters and K is the total number of tokens in the answer. The MLLM is optimized with the negative log-likelihood loss

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K} logP_\theta(A_k|A_{<k}, I^v, I^q). \tag{2}$$

LoRA-based MoE

MoE consists of a set of experts E_i and a router $\mathcal{R}$. Each expert is trained to master specific knowledge or subtasks, allowing the ensemble to represent a broader distribution pattern. The router assigns gating weights to each expert, with the process expressible by

$$\mathcal{R}_i(\boldsymbol{x}) = softmax(\boldsymbol{x}W_g), \tag{3}$$

where $\mathcal{R}_i(\cdot)$ indicates the weight of expert E_i, W_g is the trainable weight of router $\mathcal{R}$, $softmax(\cdot)$ is used to normalize the weight. The output y of MoE is aggregated by the following expression:

$$\boldsymbol{y} = \sum_{i=1}^{N_e} \mathcal{R}_i(\boldsymbol{x}) E_i(\boldsymbol{x}), \tag{4}$$

where N_e is the number of experts and $E_i(\boldsymbol{x})$ indicates the output of the i-th expert.

LoRA, a well-established tool in PEFT, finds extensive application in LLMs. It incorporates two low-rank matrices, $A \in \mathbb{R}^{r\times M}$ and $B \in \mathbb{R}^{N\times r}$, into the bypass of the LLMs' backbone W_0, which remains frozen during the fine-tuning phase:

$$\boldsymbol{y} = W_0\boldsymbol{x} + \Delta W = W_0\boldsymbol{x} + \frac{\alpha}{r}BA\boldsymbol{x}, \tag{5}$$

where α is a hyperparameter employed to regulate the update magnitude, and $r \ll min(M, N)$ denotes the rank. Consider each expert to be composed of A and B: $E_i = \frac{\alpha}{r}B_iA_i$. Accordingly, the output of LoRA-based MoE is computed by

$$\boldsymbol{y} = W_0\boldsymbol{x} + \frac{\alpha}{r}\sum_{i=1}^{N_e} \mathcal{R}_i(\boldsymbol{x}) B_iA_i\boldsymbol{x}. \tag{6}$$

Methodology

Overview

In this section, we present the KSS-MoE, which fosters synergy among experts and prevents negative transfer of shared knowledge. As shown in Fig. 1, the primary elements include the Adaptive Collaboration of Expert Knowledge Subspaces and the Orthogonal Fusion of the Shared Knowledge Subspace.

Adaptive Collaboration of Expert Knowledge Subspaces

Upon encountering a new task D_t, we set up a task-specific expert $E_t = B_tA_t$, with $A_t \in \mathbb{R}^{r\times d}$ and $B_t \in \mathbb{R}^{d\times r}$ are low-rank matrices of LoRA, where $r \ll d$. It has the same shape as the prior expert $\{E_1, E_2, ...E_{t-1}\}$. In matrix A_t, each row vector $\boldsymbol{a}_t^i \in \mathbb{R}^{1\times d}$ and in matrix B_t, each column vector $\boldsymbol{b}_t^i \in \mathbb{R}^{d\times 1}$ encapsulate different pieces of knowledge. Based on the property of Permutation Invariance in LoRA(Zhao et al. 2024), the knowledge space of the expert E_t is represented as $S_t = span\{s_t^1, s_t^2, \cdots s_t^r\}$, where $s_t^j = \{A_t[j,:], B_t[:,j]\}$ is a basis vector of S_t.

Inspired by Wang et al., we incorporate regularization throughout the training process to distinguish the knowledge of the new expert from that of previous experts, as illustrated in Fig. 2, which in turn augments their distinct specialization. The row vectors of $A_t = [(\boldsymbol{a}_t^1)^\top, (\boldsymbol{a}_t^2)^\top, \cdots, (\boldsymbol{a}_t^r)^\top]^\top$ are treated as a collection of basis vectors, whereas each column $\boldsymbol{b}_t^i$ in $B_t = [\boldsymbol{b}_t^1, \boldsymbol{b}_t^2, \cdots, \boldsymbol{b}_t^r]$ serves as a linear weighting factor for the corresponding $\boldsymbol{a}_t^i$. To enforce mutual orthogonality between A_t for the current task and $A_{<t}$ for prior tasks, we minimize the loss as follows:

$$\begin{aligned}\mathcal{L}'(\theta) &= \mathcal{L}(\theta) + \lambda \cdot \mathcal{L}_{orth} \\ &= \mathcal{L}(\theta) + \frac{\lambda}{(t-1)\times r^2}\sum_{i=1}^{t-1} \|A_t \cdot A_i^\top\|,\end{aligned} \tag{7}$$

where λ represents the hyperparameter, and $\|\cdot\|$ signifies the L1 norm.

During the inference phase, we dynamically choose the subspaces from each expert that are most pertinent to the input. This reduces redundancy while preventing interference from irrelevant knowledge. We start by eliminating irrelevant subspaces characterized by low-rank activations to avoid interference, which can be formulated by

$$\mathcal{U}_t = span\{s_t^k | \boldsymbol{z}_t[k] > \epsilon, \boldsymbol{z}_t \in \mathbb{R}^r, s_t^* \in S_t\}, \tag{8}$$

where $\mathcal{U}_t$ denotes the subspace of E_t that is relevant to input $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{z}_t = A_t\boldsymbol{x}$ is the low-rank activation of E_t, $* \in \{1, 2, \cdots, r\}$, and ϵ is a constant positive threshold.

Redundant knowledge can generate noise and misinformation, causing instability and variations in inference. To eliminate weakly related basis vectors in $\mathcal{U}_t$, we assess the importance of knowledge and preserve only the most significant parts (i.e., basis vectors):

$$\mathcal{V}_t = span\{u_t^l | l = argtop_k(\boldsymbol{z}_t),\ u_t^* \in \mathcal{U}_t\}, \tag{9}$$

where $k = min(len(\mathcal{U}_t), round(\frac{r}{t}))$, r is the rank of E_t, u_t is basis vectors of $\mathcal{U}_t$, $* \in \{1, 2, \cdots, n_{u_t}\}$, n_{u_t} is the count of basis vectors of $\mathcal{U}_t$, and t is task number. $\mathcal{V}_t = span\{v_t^1, v_t^2, \cdots v_t^{n_t}\}$ represents the subspace of the expert E_t that is closely related to the input $\boldsymbol{x}$, where n_t denotes the count of basis vectors. By this method, we can obtain the task-relevant subspaces from all t task-specific experts, and their collection can be denoted as $\mathcal{W}^s = \{\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_t\}$.

It can be demonstrated that the output of the concatenated LoRA is equivalent to the sum of their individual outputs (Zhao et al. 2024). Therefore, we combine the subspaces

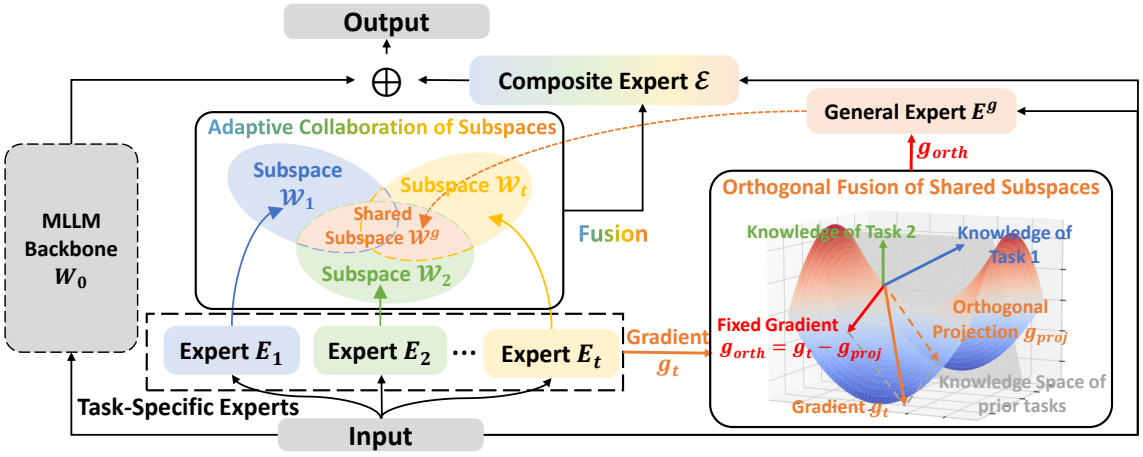

Figure 1: The framework of our KSS-MoE. It mainly contains two parts: Adaptive Collaboration of Expert Knowledge Subspaces and Orthogonal Fusion of the Shared Knowledge Subspace.

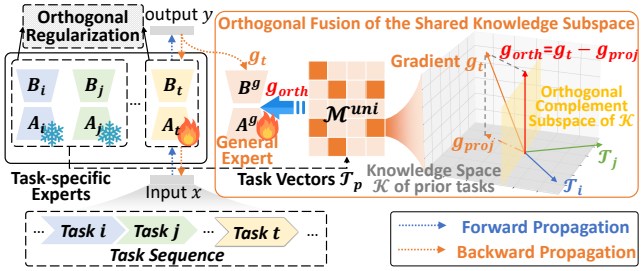

Figure 2: The orthogonal fusion of shared Knowledge subspaces in the general expert. It avoids the negative transfer caused by knowledge conflicts by projecting the gradients into the orthogonal complement subspace of the preceding tasks.

within $\mathcal{W}^s$ to create a composite expert $\mathcal{E}^s$, which inherits the strengths of all individual task-specific experts, thus realizing complementary benefits. Specifically, each $\mathcal{V}_i$ comprises the pair $\{A_{\mathcal{V}_i}, B_{\mathcal{V}_i}\}$, where $A_{\mathcal{V}_i} \in \mathbb{R}^{n_i \times d}$ and $B_{\mathcal{V}_i} \in \mathbb{R}^{d \times n_i}$, with n_i representing the number of basis vectors. We stack all matrices $A_{\mathcal{V}_*}$ vertically and combine all $B_{\mathcal{V}_*}$ horizontally to form $\mathcal{A}^s = [A_{\mathcal{V}_1}^\top, A_{\mathcal{V}_2}^\top, \cdots, A_{\mathcal{V}_t}^\top]^\top \in \mathbb{R}^{r_s \times d}$ and $\mathcal{B}^s = [B_{\mathcal{V}_1}, B_{\mathcal{V}_2}, \cdots, B_{\mathcal{V}_t}] \in \mathbb{R}^{d \times r_s}$, respectively. Here, $\top$ signifies the transpose operation, while r_s represents the count of basis in $\mathcal{W}^s$, and $* \in \{1, 2, \cdots, t\}$. The results from the expert $\mathcal{E}^s = \{\mathcal{A}^s, \mathcal{B}^s\}$ can be determined by

$$\boldsymbol{y} = W_0 \boldsymbol{x} + \frac{\alpha}{r} \mathcal{B}^s \mathcal{A}^s \boldsymbol{x}. \tag{10}$$

Orthogonal Fusion of Shared Knowledge Subspaces

By utilizing shared representations and exploiting the relationships between tasks, transferring knowledge between them enhances overall performance. Thus, we establish a general expert $E^g = B^g A^g$, where $A^g \in \mathbb{R}^{r \times d}$ and $B^g \in \mathbb{R}^{d \times r}$, with $r \ll d$, matching the dimensions of the task-specific experts. Its knowledge domain can likewise be depicted as $S_g = \{s_g^1, s_g^2, \cdots, s_g^r\}$, with each s_g^j defined by $s_g^j = \{A^g[j,:], B^g[:,j]\}$. As illustrated in Fig. 2, the update incorporates the gradients g_t derived from E_t within task D_t. A general expert strives to acquire shared knowledge that is adaptable for various tasks, avoids conflicts in knowledge across diverse tasks and negative knowledge transfer.

To fully capture historical knowledge, we maintain a task vector (Ilharco et al. 2022) $\mathcal{T}_p = \{\mathcal{T}_p^A, \mathcal{T}_p^B\}$ for each expert after training, where $p \in \{1, 2, \cdots, t-1\}$, $\mathcal{T}_p^A = A_p^{ft} - A_p^{pre}$ and $\mathcal{T}_p^B = B_p^{ft} - B_p^{pre}$. The initial weights for A_p and B_p are represented as A_p^{pre} and B_p^{pre}, respectively, whereas A_p^{ft} and B_p^{ft} indicate the weights following the fine-tuning process. $\mathcal{T}_p$ serves as a linear combination of learning gradients, so it can characterize the knowledge representation of the corresponding task D_p macroscopically.

In addition, we build a mask $\mathcal{M}_p = \{\mathcal{M}_p^A, \mathcal{M}_p^B\}$ linked to $\mathcal{T}_p$ to pinpoint the crucial parameters of expert E_p:

$$\mathcal{M}_p^*[i,j] = \begin{cases} 1, & \mathcal{T}_p^*[i,j] > quantile(\mathcal{T}_p^*, \frac{t-1}{t}), \\ 0, & \text{else}, \end{cases} \tag{11}$$

where t is the task number and $quantile(\mathcal{T}, q)$ represents the value in $\mathcal{M}_p$ corresponding to the quantile q, and $* \in \{A, B\}$. By uniting all masks $\{\mathcal{M}_i\}_{i=1}^t$, we derive the **conflict knowledge mask** $\mathcal{M}^{uni} = \{\mathcal{M}^{uni,A}, \mathcal{M}^{uni,B}\}$ via Eq.(12), which divides the subspaces into those exhibiting knowledge conflicts (indicated by 1) and those that do not (indicated by 0).

$$\mathcal{M}^{uni} = \bigcup_{i=1,2,\cdots,t-1} \mathcal{M}_i. \tag{12}$$

To prevent knowledge interference in subspaces where there's a risk of conflict, we project the gradient g_t of the expert E_t into the orthogonal complement of the space spanned by previous experts $E_{<t}$. By normalizing and stacking the prior task vectors $\mathcal{T}_{<t}$ after flattening, we obtain

$\mathcal{K} = \{\mathcal{K}^A, \mathcal{K}^B\}$, where $\mathcal{K}^A, \mathcal{K}^B \in \mathbb{R}^{(t-1)\times(r\times d)}$. Since $\mathcal{K}$ represents the essential knowledge space of $E_{<t}$, the gradient g_t is projected onto its orthogonal complement subspace. Given that experts $E_{<t}$ must maintain orthogonality throughout training, the equation can be formulated by

$$g^*_{orth} = g_t - \mathcal{K}^{*\top}\mathcal{K}^* g_t. \tag{13}$$

where $* \in \{A, B\}$.

For subspaces without knowledge conflicts, we simply update the general expert using the original gradient g_t. The update of the general expert during training can be formulated by

$$E^g_* = E^g_* - \eta \cdot (\mathcal{M}^{uni,*} \odot g^*_{orth} + (1 - \mathcal{M}^{uni,*}) \odot g_t), \tag{14}$$

where $* \in \{A, B\}$, $E^g_* \in \{A^g, B^g\}$, η represents the learning rate, and $\odot$ signifies the element-wise product.

Finally, we dynamically retrieve strongly correlated subspaces from E^g to effectively extract shared knowledge, then integrate these into the expert $\mathcal{E}^s$ for joint inference. Similarly to task-specific experts, we evaluate the contribution of each basis of the shared space by

$$\mathcal{W}^g = span\{s^k_g | \boldsymbol{z}^g_t[k] > \gamma,\ \boldsymbol{z}^g_t \in \mathbb{R}^r, s^*_g \in S_g\}. \tag{15}$$

$\mathcal{W}^g = \{\mathcal{A}^g, \mathcal{B}^g\}$ denotes the extracted shared knowledge subspace, $\boldsymbol{z}^g = A^g\boldsymbol{x}$ is inner activation, γ is the hyperparameter we set, and $* \in \{1, 2, \cdots, r\}$. By incorporating the shared knowledge into $\mathcal{W}^s$ following the method outlined in Eq.(10), we form the expert $\mathcal{E} = \{\mathcal{A}, \mathcal{B}\}$ that integrate specific-task knowledge and shared knowledge cross-tasks. $\mathcal{A} = [(\mathcal{A}^s)^\top, (\mathcal{A}^g)^\top]^\top \in \mathbb{R}^{(r_s+r_g)\times d}, \mathcal{B} = [\mathcal{B}^s, \mathcal{B}^g] \in \mathbb{R}^{d\times(r_s+r_g)}$, r_s and r_g denote number of basis in $\mathcal{W}^s$ and $\mathcal{W}^g$, respectively. The final result can be derived by

$$\boldsymbol{y}' = W_0\boldsymbol{x} + \frac{\alpha}{r}\mathcal{B}\mathcal{A}\boldsymbol{x}. \tag{16}$$

Experiment

Experimental Setup

Datasets and Metrics. KSS-MoE is trained and assessed using the CoIN benchmark (Chen et al. 2024), which comprehensively encompasses eight vision instruction tuning tasks: ScienceQA (Lu et al. 2022), TextVQA (Singh et al. 2019), ImageNet (Deng et al. 2009), GQA (Hudson and Manning 2019), VizWiz (Gurari et al. 2018), Grounding (Kazemzadeh et al. 2014; Mao et al. 2016), VQAV2 (Goyal et al. 2017), and OCR-VQA (Mishra et al. 2019). These tasks cover a broad array of scenarios relevant to CVIT.

Following CoIN, we utilize metrics frequently employed in continual learning, specifically, Mean Average Accuracy (MAA) and Backward Transfer (BWT). The former evaluates the overall performance, while the latter indicates its capacity to resist catastrophic forgetting. A model is regarded as exhibiting superior performance if it possesses higher MAA and BWT.

Implementation Details. In our experiments, we utilize LLaVA-1.5-7B (Liu et al. 2023a) as the base MLLM. It is built upon Vicuna as the LLM backbone and incorporates pre-trained CLIP ViT-L/14 (Radford et al. 2021) to serve as the visual encoder. KSS-MoE is attached to the Multi-Head Self-Attention (MHSA) layers and Feed-Forward Networks (FFNs). We set the number of task-specific experts to 8, assign each expert a rank of 16, and set λ to 0.5. In order to avoid knowledge redundancy while exploring relevant subspaces, ϵ and γ are both set to 0.1. Throughout the training process, every task undergoes training for a single epoch. The Adam optimizer is utilized with a learning rate set at 2e-4, in conjunction with the cosine learning rate scheduler.

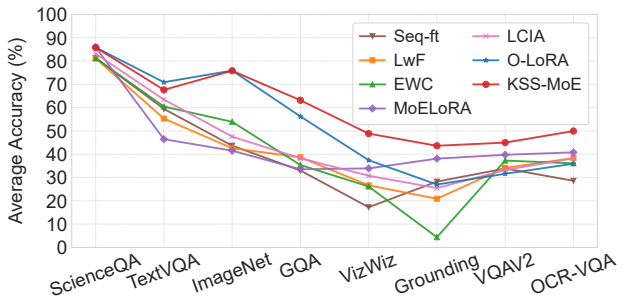

Figure 3: The overall performance of various methods during the CVIT process. KSS-MoE achieves state-of-the-art and shows strong resistance against catastrophic forgetting.

Main Results

To assess the performance of KSS-MoE, we evaluate it against both well-known and former top-performing methods, such as sequential LoRA fine-tuning (Seq-ft), LwF (Li and Hoiem 2017), EWC (Kirkpatrick et al. 2017), MoELoRA (Liu et al. 2023b), LCIA (Qiao et al. 2025), and O-LoRA (Wang et al. 2023). As indicated in Table 1, we present the experimental outcomes for our model alongside the baselines across eight tasks. From these results, three key observations can be made.

Firstly, KSS-MoE achieves state-of-the-art overall performance regarding MAA and BWT across all datasets, outperforming the second-best approach by 7.38% and 7.6%, respectively. As illustrated in Fig. 3, KSS-MoE maintains superior performance consistently throughout continual learning. This is ascribed to the joint influence of both task-specific experts and the general expert. By synergistically enhancing subspace complementarity, we mitigate the catastrophic forgetting that results from significant distribution shifts. For instance, transferring multimodal text recognition capabilities from OCR-VQA to TextVQA yields a 1.38% MAA improvement on TextVQA, demonstrating the general expert's ability to capture cross-task knowledge and enable positive transfer.

Secondly, KSS-MoE is a more advanced and effective MoE framework. Traditional MoE models are more prone to suffer severe forgetting due to expert incoordination and routing inconsistency. KSS-MoE fundamentally tackles the problems of load imbalance and routing-level forgetting by fusing the knowledge subspaces of individual experts to serve as the routing mechanism, instead of using the tra-

Methods	Accuracy on each dataset								Overall performance	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Seq-ft	81.35	46.53	96.28	39.58	43.32	32.39	51.75	58.28	40.69	-27.59
	52.11	31.06	4.16	27.29	20.79	0.21	34.87	58.28		
LwF	81.16	49.53	97.03	51.96	52.60	10.50	66.26	63.47	42.23	-20.71
	66.73	42.41	9.60	39.19	34.52	3.08	47.80	63.47		
EWC	81.21	49.31	96.83	44.91	49.18	10.59	66.09	62.31	41.86	-21.46
	70.81	41.76	9.13	38.17	18.73	1.32	46.55	62.31		
MoELoRA	85.90	51.81	97.19	61.25	44.20	30.03	66.81	63.87	45.00	-21.83
	72.91	48.82	9.58	42.55	36.70	3.40	48.62	63.87		
LCIA	83.12	50.84	97.03	37.26	43.64	16.63	62.91	63.24	45.02	-18.85
	55.53	44.18	12.85	38.79	44.71	0.13	45.48	63.24		
O-LoRA	85.85	58.28	91.31	55.41	50.89	0.37	57.87	44.42	52.61	-19.75
	13.65	41.62	51.72	49.98	30.01	0.00	56.01	44.42		
KSS-MoE	85.85	53.41	96.95	54.65	56.54	33.06	61.44	47.68	**59.99**	**-11.25**
	68.73	54.79	57.33	52.11	42.86	15.37	60.69	47.68		

Table 1: Comparison of performances(%) between KSS-MoE and other advanced methods. For each method, the first line shows the results obtained by evaluating on the task immediately after learning it, while the second line shows the results obtained by evaluating after training on the final task.

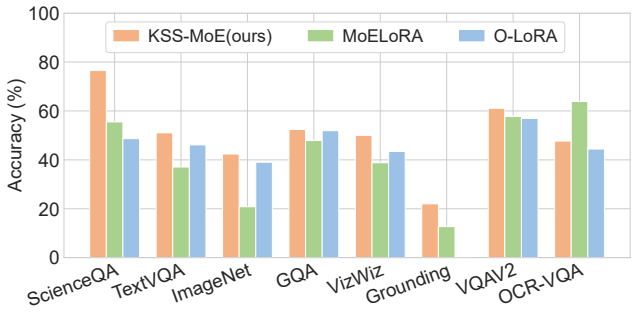

Figure 4: Comparison of the average accuracy of KSS-MoE, MoELoRA and O-LoRA on each dataset throughout the entire learning process.

ditional router, as illustrated in Fig. 4. KSS-MoE showcases improved efficiency in its use of knowledge. Unlike the vanilla router that broadly activates whole experts, KSS-MoE's subspace-level extraction allows for more precise localization and retrieval of knowledge.

Thirdly, KSS-MoE effectively acquires complementary knowledge while avoiding interference. In tasks involving classification and grounding, baseline models generally show subpar results. This is due to the substantial disparity between these tasks and others, which leads to increased knowledge interference and complicates the retention of previously acquired knowledge. During inference, O-LoRA uniformly combines all experts, which can result in overlapping or conflicting subspaces. In contrast, as depicted in Fig. 4, our KSS-MoE achieves significant advantages in tasks with substantial discrepancy by eliminating potentially conflicting information and performing fine-grained merging of the remaining subspaces.

Ablation Studies

We perform ablation studies to assess the contribution of each element in KSS-MoE, and the results from these experiments are presented in Table 2.

The effect of Orthogonal Fusion of Shared Knowledge Subspace. The general expert is employed to learn the shared knowledge among tasks, aiding in positive transfer. In particular, in experiment (b), the entire general expert is eliminated from KSS-MoE (d). By comparing the results of (b) and (d), we notice different levels of decrease in both MAA and BWT, indicating that the general expert is proficient in consistently acquiring cross-task knowledge and effectively mitigating catastrophic forgetting.

The role of the orthogonal shared space. We further examine the significance of acquiring shared knowledge within orthogonal spaces through the general expert. This is implemented by utilizing a general expert without learning the orthogonal shared space, as shown in (c). When comparing (c) with (d), a notable decline in performance is observed, and unexpectedly, (c) performs worse than the baseline lacking the general expert as seen in (b). This suggests that substantial task discrepancy and conflicting knowledge frequently occur within the CVIT environment. Vanilla general expert fails to consistently integrate shared knowledge across various tasks, causing disorganized knowledge and negative transfer. In contrast, our orthogonal shared space shows an advanced capability in avoiding interference between different tasks' knowledge.

The impact of the expert subspace collaboration strategy. The adaptive collaboration strategy is designed to address the limitations in cooperation among experts in MoE, improving generalization capabilities. Considering that the general expert exists depending on task-specific experts, we

Methods	Accuracy on each dataset								Overall performance	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
(a)w/o m1,m2	55.50	37.07	20.83	47.90	38.82	12.70	57.72	63.87	45.00	-21.83
(b)w/o m2	73.27	50.12	37.01	50.10	53.62	13.15	62.28	46.47	57.64	-13.54
(c)w/o m2*	69.58	46.31	40.06	51.12	46.58	17.63	61.04	45.52	52.61	-12.78
(d)full	76.51	51.05	42.38	52.42	50.04	22.02	61.07	47.68	59.99	-11.25

Table 2: Ablation studies of KSS-MoE. m1, m2, and m2* denote expert subspace collaboration strategy, the general expert, and the orthogonal shared space, respectively. The accuracy represents the average performance throughout the entire process.

Rank	Accuracy on each dataset								Overall performance	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
4	69.30	49.32	31.75	53.41	48.08	8.09	62.04	45.07	53.42	-11.89
8	75.78	48.93	38.95	52.87	48.47	14.47	61.78	45.27	58.50	-15.05
16	76.51	51.05	42.38	52.42	50.04	22.02	61.07	47.68	59.99	-11.25
32	79.39	49.86	47.97	51.57	49.52	16.91	62.68	45.82	60.74	-11.23

Table 3: The impact of capacity of knowledge space. The accuracy in the table represents the average performance throughout the entire process.

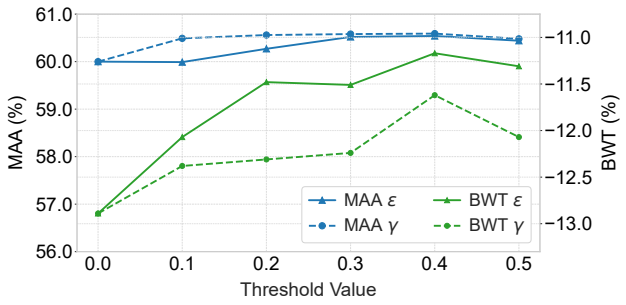

Figure 5: Results of the parameter sensitivity analysis. Initially, performance improves as ϵ and γ grow, peaking at 0.4, after which it begins to decline steadily.

replace expert subspace collaboration strategy with naive MoE, degenerating KSS-MoE to MoELoRA. Compare (a) to (b), it is clear that this strategy plays a crucial role in the entire model and determines how to effectively retrieve and integrate the experts' knowledge, thereby avoiding catastrophic forgetting caused by distribution drift.

Further Analysis

The impact of capacity of knowledge space. We established the expert ranks at 4, 8, 16, and 32 to investigate the impact of varying subspace capacities on performance. The experimental results are presented in Table 3. While enhancing subspace capacity typically boosts model performance for complex tasks, increasing the rank from 4 to 8 surprisingly diminishes BWT. This happens as smaller knowledge spaces ($r = 4$) limit learning that focuses on a particular task while reducing interference between different tasks. In contrast, larger spaces lead to subspace overlap, which intensifies the forgetting issue. This is demonstrated by considerably lower MAA values observed in smaller subspaces.

Sensitivity analysis of the threshold ϵ and γ. As depicted in Fig. 5, we examine how the hyper-parameters ϵ and γ influence the removal of irrelevant and weakly relevant knowledge. When adjusting one of them, the other remains fixed at zero. Given that they need to be more than zero yet close to it, we assign them values of 0, 0.1, 0.2, 0.3, 0.4 and 0.5, respectively. Initially, the BWT rises, but as ϵ and γ grow, it subsequently declines. In contrast, MAA, which serves as an indicator of overall performance, exhibits relatively minor variations. It suggests that setting thresholds too low introduces marginally relevant knowledge, disrupting the knowledge space, while setting them too high excludes useful knowledge, leading to diminished performance. KSS-MoE achieves optimal performance when ϵ and γ are both set to a balanced value of 0.4.

Conclusion

In this study, we introduce the KSS-MoE model tailored for MLLMs within the CVIT framework, which retrieves knowledge subspaces from task-specific experts and adaptively combines them to form a composite expert. Our KSS-MoE first facilitates fine-grained knowledge complementarity and collaboration among MoE experts, alleviating catastrophic forgetting caused by distribution drift. In addition, a general expert projects shared knowledge onto orthogonal complement subspace of prior experts, thus maintaining the integrity of cross-task knowledge and avoiding negative transfer due to inter-task conflicts. Comprehensive experiments conducted on eight benchmark datasets reveal that our KSS-MoE achieves state-of-the-art performance, introducing a new paradigm for MoE-based MLLMs in CVIT.

Acknowledgments

The research was supported in part by Open Research Fund of Zhejiang Key Laboratory of Intelligent Education Technology and Application (No. 2025ZNJYKF012), National Nature Science Foundation of China under Grant No. 62576283, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2025A1515012995, 2024A1515011715, and CCF-Zhipu Large Model Innovation Fund under Grant No. CCF-Zhipu202413.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cai, W.; Jiang, J.; Qin, L.; Cui, J.; Kim, S.; and Huang, J. 2024. Shortcut-connected expert parallelism for accelerating mixture-of-experts. *arXiv preprint arXiv:2404.05019*.

Cao, M.; Liu, Y.; Liu, Y.; Wang, T.; Dong, J.; Ding, H.; Zhang, X.; Reid, I.; and Liang, X. 2024. Continual llava: Continual instruction tuning in large vision-language models. *arXiv preprint arXiv:2411.02564*.

Chen, C.; Zhu, J.; Luo, X.; Shen, H. T.; Song, J.; and Gao, L. 2024. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in Neural Information Processing Systems*, 37: 57817–57840.

Chen, W.; Zhou, Y.; Du, N.; Huang, Y.; Laudon, J.; Chen, Z.; and Cui, C. 2023a. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, 5383–5395. PMLR.

Chen, Z.; Shen, Y.; Ding, M.; Chen, Z.; Zhao, H.; Learned-Miller, E. G.; and Gan, C. 2023b. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11828–11837.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

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.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gou, Y.; Liu, Z.; Chen, K.; Hong, L.; Xu, H.; Li, A.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Gururangan, S.; Lewis, M.; Holtzman, A.; Smith, N. A.; and Zettlemoyer, L. 2021. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*.

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.

Huai, T.; Zhou, J.; Wu, X.; Chen, Q.; Bai, Q.; Zhou, Z.; and He, L. 2025. CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19608–19617.

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

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Gururangan, S.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Lin, H.; Cheng, X.; Wu, X.; and Shen, D. 2022. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Wu, J.; Liu, J.; and Duan, Y. 2024. Learning attentional mixture of loras for language model continual learning. *arXiv preprint arXiv:2409.19611*.

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Xu, D.; Tian, F.; and Zheng, Y. 2023b. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *CoRR*.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Nguyen, H.; Akbarian, P.; Pham, T.; Nguyen, T.; Zhang, S.; and Ho, N. 2024. Statistical advantages of perturbing cosine router in mixture of experts. *arXiv preprint arXiv:2405.14131*.

Qiao, J.; Tan, X.; Qu, Y.; Ding, S.; Xie, Y.; et al. 2024. Llaca: Multimodal large language continual assistant.

Qiao, J.; Zhang, Z.; Tan, X.; Qu, Y.; Ding, S.; and Xie, Y. 2025. Large Continual Instruction Assistant. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

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.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tong, K.; Pan, K.; Zhang, X.; Meng, E.; He, R.; Cui, Y.; Guo, N.; and Zhuang, H. 2025. Analytic subspace routing: How recursive least squares works in continual learning of large language model. *arXiv preprint arXiv:2503.13575*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, R.; and Li, P. 2024. LEMoE: Advanced Mixture of Experts Adaptor for Lifelong Model Editing of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2551–2575.

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*.

Wang, Z.; Che, C.; Wang, Q.; Li, Y.; Shi, Z.; and Wang, M. 2024. Separable mixture of low-rank adaptation for continual visual instruction tuning. *arXiv preprint arXiv:2411.13949*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhai, Y.; Tong, S.; Li, X.; Cai, M.; Qu, Q.; Lee, Y. J.; and Ma, Y. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Zhang, W.; Deng, L.; Zhang, L.; and Wu, D. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2): 305–329.

Zhang, X.; Bai, L.; Yang, X.; and Liang, J. 2025. C-lora: Continual low-rank adaptation for pre-trained models. *arXiv preprint arXiv:2502.17920*.

Zhao, H.; Wang, Z.; Sun, Q.; Song, K.; Li, Y.; Hu, X.; Guo, Q.; and Liu, S. 2025. LLaVA-CMoE: Towards Continual Mixture of Experts for Large Vision-Language Models. *arXiv preprint arXiv:2503.21227*.

Zhao, Z.; Shen, T.; Zhu, D.; Li, Z.; Su, J.; Wang, X.; Kuang, K.; and Wu, F. 2024. Merging loras like playing lego: Pushing the modularity of lora to extremes through rank-wise clustering. *arXiv preprint arXiv:2409.16167*.

Zheng, J.; Ma, Q.; Liu, Z.; Wu, B.; and Feng, H. 2024. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*.

Zhu, D.; Sun, Z.; Li, Z.; Shen, T.; Yan, K.; Ding, S.; Kuang, K.; and Wu, C. 2024a. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*.

Zhu, T.; Qu, X.; Dong, D.; Ruan, J.; Tong, J.; He, C.; and Cheng, Y. 2024b. Llama-moe: Building mixture-of-experts from llama with continual pre-training. *arXiv preprint arXiv:2406.16554*.

Zuo, S.; Liu, X.; Jiao, J.; Kim, Y. J.; Hassan, H.; Zhang, R.; Gao, J.; and Zhao, T. 2022. Taming Sparsely Activated Transformer with Stochastic Experts. In *International Conference on Learning Representations*.