🧠 AWARe: Mitigating Catastrophic Forgetting via Activation-Weighted Adaptive REtention

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) exhibit strong generalization and reasoning abilities due to large-scale multimodal pre-training. However, fine-tuning these models on downstream tasks often leads to catastrophic forgetting, where newly learned task-specific knowledge degrades previously acquired capabilities. This issue arises because gradient updates for new tasks overwrite parameters critical to prior knowledge, limiting the practical deployment of MLLMs. To address this challenge, we propose **Activation-Weighted Adaptive REtention (AWARe)**, a fine-tuning method that mitigates catastrophic forgetting by dynamically controlling parameter updates based on activation patterns. AWARe assigns activation-based importance scores to parameters, selectively freezing those essential for preserving prior capabilities while allowing less important parameters to adapt to new tasks. Importantly, AWARe operates *without modifying model architectures*, ensuring compatibility with existing inference engines. Extensive experiments demonstrate that AWARe effectively preserves upstream capabilities while achieving superior downstream performance compared to existing methods. Code is available at https://anonymous.4open.science/r/AWARe-206B/.

1 Introduction

Multimodal Large Language Models (MLLMs) have emerged as powerful versatile agents, capable of solving complex tasks that require understanding and reasoning across both inputs of vision and language (Zhu et al., 2023; Dai et al., 2023; Bai et al., 2023; Liu et al., 2023; Lin et al., 2023; Chen et al., 2024b; Wang et al., 2024; Zhu et al., 2025a; Wang et al., 2025; Bai et al., 2025b,a). Typical MLLMs are composed of a pre-trained visual encoder, a large language model (LLM), and a connector module (Li et al., 2022; Dai et al., 2023; Liu et al., 2023).

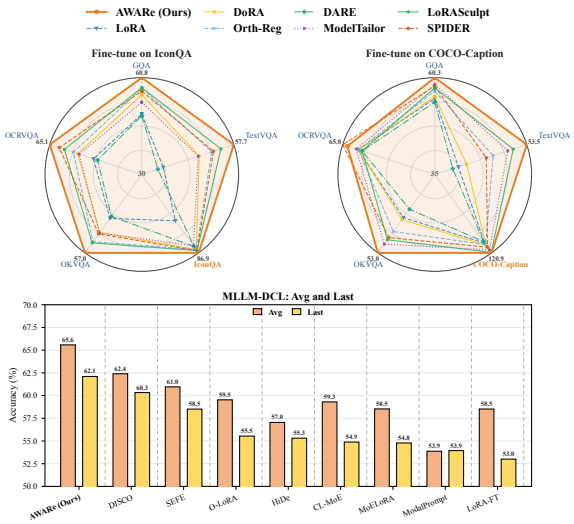


Figure 1: **Overall Performance of AWARe.** Top: radar plots showing comparison between our approach and various baselines under Single-Task Downstream Adaptation. Bottom: results on the MLLM-DCL continuous learning benchmark.

These models facilitate strong generalization abilities acquired from large-scale pre-training. Despite their impressive zero-shot capabilities, fine-tuning MLLMs on downstream tasks remains a standard practice to tailor models for specific domains or improve instruction-following performance (Dai et al., 2023; Bai et al., 2023; Liu et al., 2023). However, this adaptation process often comes at a cost: *catastrophic forgetting* (CF), where the model's proficiency in previously learned upstream tasks degrades significantly as it optimizes for new objectives (McCloskey and Cohen, 1989; McClelland et al., 1995; Kirkpatrick et al., 2017; Li et al., 2024b). Nonetheless, this form of forgetting remains particularly severe in MLLMs due to the complex interplay between modalities and the high dimensionality of the parameter space (Sung et al., 2023; Luo et al., 2025, 2023; Zhai et al., 2024a; Shen et al., 2024; Jiang et al., 2024; Zhai et al., 2024b).

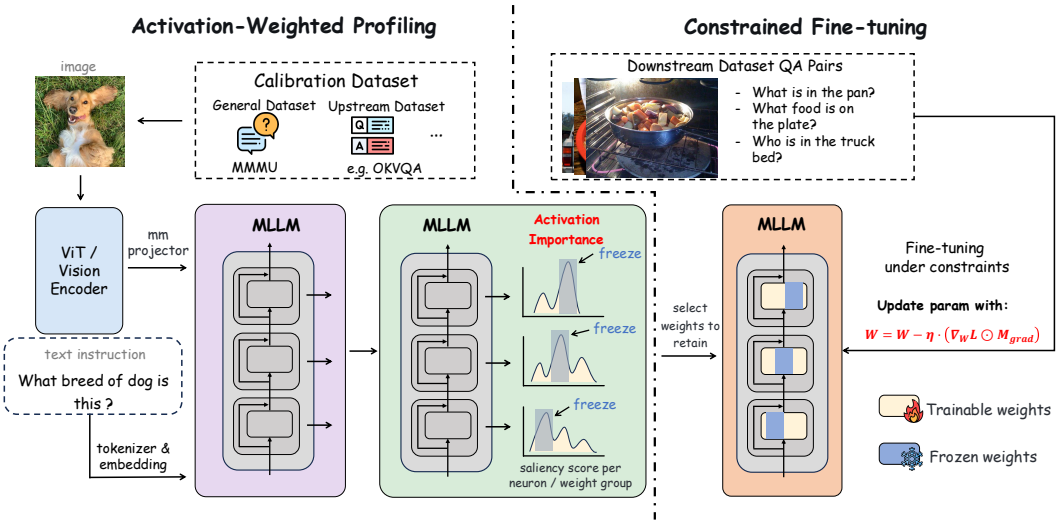


Figure 2: **Overview of AWARe.** Given a multimodal base model, AWARe identifies important neurons or weight groups using activation-based saliency on calibration samples, then freezes high-saliency weights and updates only low-saliency weights during fine-tuning.

Existing approaches to mitigate catastrophic forgetting, such as experience replay (Riemer et al., 2018; Chaudhry et al., 2019) or regularization-based methods (Kirkpatrick et al., 2017), often incur high computational overhead or struggle to scale to the billions of parameters in MLLMs. Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA (Hu et al., 2022) reduce the trainable parameter count but do not inherently prevent the erosion of pre-trained knowledge, as recent studies show they can still disrupt reliable upstream features (Biderman et al., 2024; Zhu et al., 2024). Consequently, a critical challenge arises: *How can we enable MLLMs to adapt effectively to downstream tasks while robustly preserving their pre-trained capabilities without prohibitive costs?*

Our work is primarily inspired by the principle of activation-aware parameter saliency, most notably demonstrated by Activation-aware Weight Quantization (AWQ) (Lin et al., 2024). AWQ shows that a small fraction of weights associated with large activation magnitudes are disproportionately important for preserving model behavior under quantization. Wanda reaches a complementary conclusion in pruning: static weight magnitude alone is insufficient for LLMs, while combining weights with input activation statistics gives a simple and effective estimate of functional importance (Sun et al., 2024). Recent model merging studies further suggest that activation-guided consensus can reduce parameter interference when combining specialized models (Nobari et al., 2025; Yao et al., 2025). Together, these findings indicate that activations provide a dynamic, data-dependent view of which internal pathways support stable behavior, beyond what weight-centric criteria such as magnitude pruning can capture (Han et al., 2015; Frankle and Carbin, 2019). We therefore repurpose activation saliency from compression and merging to forgetting mitigation.

Building on this motivation, we introduce **AWARe**, short for **A**ctivation-**W**eighted **A**daptive **RE**taining, a novel framework designed to surgically mitigate catastrophic forgetting. AWARe has two phases: *Knowledge Profiling*, which estimates neuron saliency from a small calibration set, and *Constrained Fine-tuning*, which freezes the most salient parameters during downstream training. When upstream data are unavailable, we can use a general-purpose calibration set such as MMMU (Yue et al., 2024); as shown by the AWARe (MMMU) results in Table 1, this remains highly effective while adding only a negligible profiling cost.

We validate our approach through comprehensive experiments on IconQA (Lu et al., 2021), COCO-Caption (Lin et al., 2014), and MLLM-DCL (Zhao et al., 2025) (Continuous learning) benchmarks. AWARe achieves strong stability-plasticity trade-offs, and freezing the top 30% of self-attention parameters is often sufficient to preserve upstream knowledge while keeping downstream performance competitive.

Our main contributions are summarized as follows:

❶ **AWARe Framework.** We introduce Activation-

Weighted Adaptive REtention (AWARe), a simple activation-based method for preserving critical upstream knowledge during downstream adaptation in MLLMs.

❷ **Efficiency and Simplicity.** Our approach requires no architectural changes and updates only a small subset of parameters (e.g., $\sim 17.5\%$; see Section C), avoiding replay buffers and extra modules.

❸ **Comprehensive Validation.** We show through ablations and benchmarks that activation-based saliency is effective and that AWARe preserves generalization across diverse settings.

2 Related Work

2.1 Catastrophic Forgetting and Continual Learning

Fine-tuning serves as a pivotal mechanism for adapting Multimodal Large Language Models (MLLMs) to downstream tasks. However, deep learning models often suffer from catastrophic forgetting (McCloskey and Cohen, 1989; McClelland et al., 1995), where previously learned knowledge is lost when acquiring new skills. Various continual learning algorithms have been proposed to address this issue, generally encompassing rehearsal-based, regularization-based (Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017), and architecture-based approaches (Houlsby et al., 2019; Lester et al., 2021). While established, many traditional methods rely on full-model fine-tuning or second-order statistics, making them computationally prohibitive for large-scale models.

2.2 Parameter-Efficient Fine-Tuning for Mitigation

In the era of large foundation models, the challenge shifts towards maintaining the model's strong generalization and "Open-World Stabilization" after downstream adaptation (Ni et al., 2023; Zhai et al., 2024a). Recent efforts have explored parameter-efficient solutions to mitigate forgetting. Reparameterization methods such as DoRA (Liu et al., 2024) and LoRA-based extensions like LoRAMoE (Dou et al., 2024) aim to enhance learning capacity while addressing generalization loss, whereas LoRASculpt (Liang et al., 2025) prunes redundant parameters via sparse updates. Regularization techniques have also been adapted; for instance, Orth-Reg (Hu et al., 2024) encourages fine-tuned features to remain orthogonal to pre-trained representations. CorDA (Yang et al., 2024) builds task-aware adapters through context-guided weight decomposition to preserve world knowledge. Selective tuning approaches, such as SPIDER (Huang et al., 2025) and Model Tailor (Zhu et al., 2024), leverage gradient or saliency analysis to construct sparse updates, aiming to protect critical pre-trained features. However, these criteria often rely on static information or gradient magnitudes, overlooking the dynamic nature of neuronal activations during forward propagation.

Recognizing these limitations, we propose AWARe. Unlike additive methods that introduce extra parameters or selective tuning that relies on static priors, AWARe utilizes task-induced activation distributions to measure parameter importance. By adaptively constraining high-activation neurons based on upstream statistics, AWARe effectively balances the preservation of core capabilities with the flexibility needed for downstream task adaptation.

3 Methodology

3.1 Problem Formulation

A Multimodal Large Language Model (MLLM) $\mathcal{M}_\theta$ generally comprises three core components: a vision encoder $\mathcal{V}$ (e.g., CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023)), a large language model $\mathcal{L}$ (e.g., LLaMA (Touvron et al., 2023b,a) or Vicuna (Chiang et al., 2023)), and a connector module ρ that bridges the visual and textual modalities. The standard paradigm involves aligning the pre-trained vision encoder representations with the LLM's embedding space via the connector, enabling the model to process multimodal inputs effectively.

Consider a Multimodal Large Language Model (MLLM) $\mathcal{M}_\theta$ parameterized by θ. We assume the model has been initially trained on a foundational set of upstream tasks $\mathcal{T}_{base}$. In a continual instruction tuning setting, the model is sequentially exposed to a stream of downstream tasks, denoted as $\mathcal{T}_{seq} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$. Our objective is to incrementally learn each new task $\mathcal{T}_t$ at step t by updating θ, while mitigating the *catastrophic forgetting* (McCloskey and Cohen, 1989; McClelland et al., 1995) of both the foundational capabilities from $\mathcal{T}_{base}$ and the previously acquired knowledge from tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_{t-1}\}$. As illustrated in Figure 3, we achieve this by partitioning the parameters into a

fixed set θ_{fixed}, which captures and preserves core established features, and a plastic set θ_{active} used for acquiring new task-specific knowledge.

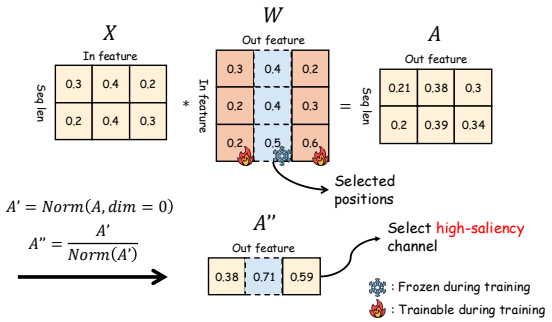


Figure 3: Overview of AWARe. For each target linear layer, we profile the activation matrix produced by calibration samples, normalize the activation statistics across sequence and feature dimensions, and obtain a saliency score for each output channel. High-saliency channels are selected and frozen to preserve pretrained knowledge, while the remaining channels stay trainable for downstream adaptation.

3.2 Activation-Based Saliency Estimation

To identify important parameters for retention, we utilize a calibration set $\mathcal{S}_{cal}$ composed of representative samples from each upstream task's train datasets ($\mathcal{T}_{up}$). Specifically, we randomly sample several instances from each upstream task. In scenarios where access to upstream data is restricted, we demonstrate that utilizing a comprehensive general benchmark, such as MMMU (Yue et al., 2024), as a proxy for $\mathcal{S}_{cal}$ is equally effective. We empirically demonstrate the robustness of model performance to the calibration set size in our ablation studies at Section 4.6. Crucially, the profiling incurs negligible overhead, requiring only a single forward pass on a small sample set. For a target linear layer with weight matrix $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, we analyze its activation tensor $\mathbf{A} \in \mathbb{R}^{B \times L \times d_{out}}$, where $i \in \{1, \ldots, B\}$, $j \in \{1, \ldots, L\}$, and $k \in \{1, \ldots, d_{out}\}$ index the batch size, context length and hidden dimension, respectively.

The saliency estimation follows a three-step aggregation and normalization process. First, we compute the L_2-norm of activations along the sequence length dimension for each sample j and neuron k:

$$a'_{i,k} = \sqrt{\sum_{j=1}^{L} a_{i,j,k}^2} \tag{1}$$

Second, to ensure comparability across samples and prevent those with outlier-scale activations from biasing the estimations, we apply per-sample L_2-normalization across the hidden dimension:

$$a''_{i,k} = \frac{a'_{i,k}}{\|\mathbf{a}'_i\|_2} = \frac{a'_{i,k}}{\sqrt{\sum_{k'=1}^{d_{out}} {a'_{i,k'}}^2}} \tag{2}$$

where $\mathbf{a}'_i = [a'_{i,1}, \ldots, a'_{i,d_{out}}]^\top$ is the vector of sequence-aggregated activations for sample i. This normalization ensures that the saliency is determined by the *relative* importance of neurons within each context rather than absolute activation scale, which is crucial for identifying neurons that consistently capture structural features across diverse upstream tasks. This refinement distinguishes our approach from simple magnitude-based methods and explains the superior performance observed in Table 3. Finally, the saliency score for each neuron k is obtained by averaging across the batch:

$$s_k = \frac{1}{B} \sum_{i=1}^{B} a''_{i,k} \tag{3}$$

The resulting saliency vector $\mathbf{s} \in \mathbb{R}^{d_{out}}$ reflects the relative importance of each output neuron in preserving upstream task capabilities.

3.3 Adaptive Parameter Retaining

Guided by the saliency scores, we introduce a retention ratio $\rho \in (0, 1)$ as a hyperparameter to control the fraction of parameters to be frozen. We rank output neurons across all target linear layers by their saliency scores and identify the global index set $\mathcal{I}$ containing the top-ρ fraction of neurons (e.g., top 30% or 10%).

Since these salient neurons characterize the core upstream knowledge, we freeze their corresponding parameters in the weight matrix. Specifically, we construct a binary gradient mask $\mathbf{M}_{grad} \in \{0, 1\}^{d_{out} \times d_{in}}$ such that:

$$\begin{aligned} (M_{grad})_{k,:} &= \mathbf{0} \quad \text{if } k \in \mathcal{I}, \\ (M_{grad})_{k,:} &= \mathbf{1} \quad \text{otherwise} \\ M_{ret} &= \mathbf{1} - M_{grad} \end{aligned} \tag{4}$$

where $\mathbf{0}$ and $\mathbf{1}$ are row vectors of size d_{in}. During downstream optimization, the parameters associated with salient regions are held constant via masked gradient updates:

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \cdot (\nabla_{\mathbf{W}} \mathcal{L}_{down} \odot \mathbf{M}_{grad}) \tag{5}$$

where η is the learning rate and $\odot$ denotes the Hadamard product. In practice, the PyTorch implementation achieves this by utilizing a training-time wrapper, `AwareLinear`, which decomposes

the linear layer into active and frozen components to efficiently manage gradient updates without altering the underlying operator's logic. This wrapper is removed post-training, leaving the original model configuration intact. This row-wise freezing mechanism ensures that the most influential output features of the pre-trained LLaVA model remain intact while allowing less critical regions to adapt to new tasks. As shown in Figure 4, the distribution of these salient neurons varies across transformer layers and projection types.

3.4 Selection of Target Linear

We specifically apply AWARe to the linear projection layers within the self-attention mechanism (i.e., `q_proj`, `k_proj`, `v_proj`) and the linear layers within the multimodal projector (`mm_projector`). This selection is informed by recent research (Zhu et al., 2025b) which demonstrates that updating self-attention projections tends to cause significantly less catastrophic forgetting of pre-existing knowledge compared to the multilayer perceptron (MLP) blocks. By focusing our retention strategy on those linear, we balance the preservation of core cross-modal reasoning capabilities with the flexibility needed for downstream task adaptation. An ablation analysis is provided in Section 4.5. Consistent with this finding, all other components of the LLM, including MLP blocks and the output projection ($\mathbf{W}_o$), are kept entirely frozen during adaptation. This configuration not only maximizes parameter efficiency but also minimizes the risk of distorting the pre-trained feature space in regions less critical for downstream alignment. The complete procedure for AWARe is summarized in Algorithm 1.

Algorithm 1 AWARe

Require: Pre-trained model $\mathcal{M}_\theta$, Upstream task $\mathcal{T}_{up}$, Downstream task $\mathcal{T}_{down}$, Retention ratio ρ

Ensure: Fine-tuned model $\mathcal{M}_\theta$

Phase 1: Knowledge Profiling
Sample calibration set $\mathcal{S}_{cal} \sim \mathcal{T}_{up}$
for each target layer l in $\mathcal{M}_\theta$ **do**
  Perform forward pass on $\mathcal{S}_{cal}$ to obtain activations $\mathbf{H}^{(l)}$
  Calculate neuron saliency $\mathbf{s}^{(l)}$ using Equations (1) to (3)
end for
$\mathcal{I} \leftarrow \text{GlobalTopIndices}(\{\mathbf{s}^{(l)}\}_l, \rho)$
for each target layer l in $\mathcal{M}_\theta$ **do**
  Extract layer-specific retained indices $\mathcal{I}^{(l)}$ from $\mathcal{I}$
  Generate row-wise gradient mask $\mathbf{M}_{grad}^{(l)}$ via Equation (4)
end for
Phase 2: Constrained Fine-tuning
while not converged on $\mathcal{T}_{down}$ **do**
  $\mathcal{L} \leftarrow \text{ComputeLoss}(\mathcal{M}_\theta, \mathcal{T}_{down})$
  Update parameters: $\theta \leftarrow \theta - \eta \cdot (\nabla_\theta \mathcal{L} \odot \mathbf{M}_{grad})$
end while

4 Experiments

In this section, we evaluate the effectiveness of our proposed method by comparing it with several state-of-the-art baselines on multimodal downstream tasks. We also conduct extensive ablation studies to analyze the impact of key components and hyperparameters.

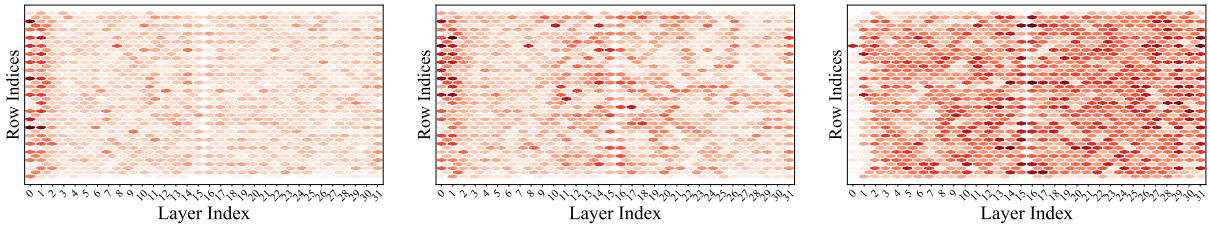
4.1 Experimental Setup

Datasets and Tasks. We use LLaVA-v1.5-7b (Liu et al., 2023) as the base model for training. The model has been pre-trained on several upstream tasks, including OKVQA (Marino et al., 2019), OCRVQA (Mishra et al., 2019), GQA (Hudson and Manning, 2019), and TextVQA (Singh et al., 2019). For downstream evaluation, we use IconQA (Lu et al., 2021) for VQA and COCO-Caption (Lin et al., 2014) for image captioning, where captioning performance is measured by CIDEr. For continual instruction tuning, we evaluate LLaVA-v1.5-7b on MLLM-DCL (Zhao et al., 2025), a five-task sequence covering remote sensing (RS), medical VQA (Med), autonomous driving (AD), science reasoning (Sci), and financial understanding (Fin). We provide additional experiments on Qwen2.5-VL-7B (Bai et al., 2025b) in Section B.

Implementation Details. During training, we freeze the vision tower to preserve the pre-trained visual representations. The complete training configuration and hyperparameter settings are provided in Section D.

4.2 Baselines

We evaluate AWARe under two complementary settings. First, in single-task downstream adaptation, the model learns one new downstream task and is evaluated for both target-task learning and forgetting on upstream tasks. In this setting, we compare against established parameter-efficient



(a) q_proj (b) k_proj (c) v_proj

Figure 4: Layer-wise distribution of high-activation neurons across different projection types. The heatmaps visualize indices identified for freezing (top 30% global activation). Notably, q_proj and k_proj show high concentration in early layers, whereas v_proj exhibits a more uniform distribution, suggesting value information is distributed more evenly across the depth of the network.

Methods	IconQA								COCO-Caption							
	OKVQA	OCRVQA	GQA	TextVQA	*Target*	$\mathcal{R}$	$\mathcal{E}$	$\mathcal{H}$	OKVQA	OCRVQA	GQA	TextVQA	*Target*	$\mathcal{R}$	$\mathcal{E}$	$\mathcal{H}$
Zero-shot	57.99	66.20	61.93	58.23	23.18	100.0	28.9	44.9	57.99	66.20	61.93	58.23	40.40	100.0	42.3	59.4
Full-FT †	0.04	0.00	0.00	0.46	80.15	0.2	100.0	0.4	0.00	0.00	0.00	0.04	95.59	0.0	100.0	0.0
LoRA	45.06	48.40	49.22	36.47	63.65	73.3	79.4	76.2	44.92	59.90	54.88	39.75	110.27	81.6	115.4	95.6
DoRA	49.94	54.40	55.39	47.26	84.48	84.7	105.4	93.9	45.38	59.75	55.40	41.38	112.60	82.6	117.8	97.1
Orth-Reg	53.12	56.10	57.43	51.00	84.52	89.1	105.5	96.6	48.11	59.85	56.83	46.69	111.87	86.5	117.0	99.5
DARE	44.39	46.85	48.75	34.84	82.28	71.5	102.7	84.3	42.98	58.65	53.92	38.61	108.57	79.5	113.6	93.5
Model Tailor	50.07	53.85	53.04	46.92	81.76	83.4	102.0	91.8	50.99	60.60	58.54	49.65	117.64	**89.9**	123.1	103.9
LoRASculpt	53.52	59.50	57.63	53.76	85.26	91.8	106.4	98.6	49.99	58.65	57.63	50.73	120.35	88.8	125.9	104.2
SPIDER	50.41	61.49	56.54	51.54	85.29	90.0	106.4	97.5	49.50	**65.00**	58.14	45.33	114.74	89.2	120.0	102.3
AWARe	**56.95**	**65.10**	**60.83**	**57.66**	**86.92**	**98.4**	**108.4**	**103.2**$_{\uparrow 4.6}$	**53.01**	63.45	**60.32**	**53.46**	**120.94**	**94.2**	**126.5**	**108.0**$_{\uparrow 3.8}$
AWARe (MMMU)	56.86	63.35	60.93	56.77	85.23	97.4	106.3	101.7$_{\uparrow 3.1}$	53.25	63.12	60.38	52.74	121.22	93.9	126.8	107.9$_{\uparrow 3.7}$

Table 1: **Comparison with State-of-the-Art Fine-Tuning Solutions on LLaVA-v1.5-7B** for IconQA and COCO-Caption. The optimal and sub-optimal results are denoted by boldface and underlining. ↑ means improved accuracy compared with the sub-optimal results. †Specifically, the near-zero performance of Full-FT on upstream tasks reflects extreme catastrophic forgetting due to overfitting to the target task. More details about the metrics are provided in Section 4.3. Training setting are detailed in Section D.

fine-tuning and anti-forgetting baselines, including **Full-FT**, **LoRA** (Hu et al., 2022), **DoRA** (Liu et al., 2024), **DARE** (Yu et al., 2024), **Orth-Reg** (Hu et al., 2024), **Model Tailor** (Zhu et al., 2024), **LoRASculpt** (Liang et al., 2025), and **SPIDER** (Huang et al., 2025). Second, in continual instruction tuning, the model learns a sequence of subtasks and is evaluated after each training stage. For this setting, we compare against continual learning baselines including **ModalPrompt** (Zeng et al., 2025), **SEFE** (Chen et al., 2025), **HiDe-LLaVA** (Guo et al., 2025a), **CL-MoE** (Huai et al., 2025), **O-LoRA** (Wang et al., 2023), **MoELoRA** (Chen et al., 2024a), and **DISCO** (Guo et al., 2025b). Detailed descriptions for all baselines are provided in Section E.

4.3 Single-Task Downstream Adaptation

We evaluate the performance of all methods by decomposing continual learning capabilities into two core components: **Knowledge Retention** ($\mathcal{R}$) and **Learning Efficiency** ($\mathcal{E}$). To provide a balanced assessment, we report their Harmonic Mean ($\mathcal{H}$):

$$\mathcal{H} = \frac{2 \cdot \mathcal{R} \cdot \mathcal{E}}{\mathcal{R} + \mathcal{E}} \tag{6}$$

The components are defined as follows:

- $\mathcal{R}$**:** A stability metric measuring the preservation of performance on upstream tasks after fine-tuning. It is calculated as the average ratio of the fine-tuned accuracy to the base (zero-shot) accuracy:

$$\mathcal{R} = \frac{1}{|\mathcal{T}_{up}|} \sum_{i \in \mathcal{T}_{up}} \frac{Acc_{i,\text{final}}}{Acc_{i,\text{base}}} \tag{7}$$

  where $Acc_{i,\text{base}}$ is the model's initial performance before fine-tuning. A sharp drop in $\mathcal{R}$ indicates catastrophic forgetting.

- $\mathcal{E}$**:** A plasticity metric measuring the model's adaptation to the new downstream task. To normalize against model capacity, it is defined relative to the Full Fine-Tuning (Full-FT) baseline:

$$\mathcal{E} = \frac{Acc_{down,\text{Method}}}{Acc_{down,\text{Full-FT}}} \tag{8}$$

  where $Acc_{down,\text{Full-FT}}$ represents the performance upper bound achievable by Full-FT.

As shown in Table 1, our method achieves a superior harmonic balance $\mathcal{H}$ between plasticity and stability. Compared to other methods, our approach maintains high retention rates $\mathcal{R}$ while achieving efficiency $\mathcal{E}$. Notably, when upstream task data is

Method	Avg						Last					
	RS ↑	Med ↑	AD ↑	Sci ↑	Fin ↑	Average ↑	RS ↑	Med ↑	AD ↑	Sci ↑	Fin ↑	Average ↑
ModalPrompt	53.19	45.73	40.78	41.82	87.82	53.87	53.63	45.68	40.77	41.81	87.82	53.94
HiDe	75.41	48.71	37.06	42.49	81.55	57.04	74.31	48.95	33.21	38.54	81.55	55.31
LoRA-FT	73.34	47.97	38.68	45.16	87.45	58.52	69.65	41.59	25.43	40.88	87.45	53.00
MoELoRA	77.19	47.02	37.30	44.40	86.75	58.53	77.54	41.85	27.62	40.13	86.75	54.78
CL-MoE	73.51	49.97	38.55	45.73	88.74	59.30	71.34	46.84	26.33	41.17	88.74	54.88
O-LoRA	76.12	47.38	40.97	46.02	87.15	59.53	74.64	44.42	30.02	41.47	87.15	55.54
SEFE	77.38	51.41	44.24	44.95	86.82	60.96	77.26	**50.37**	37.21	40.87	86.82	58.51
DISCO	77.41	48.04	48.70	48.65	89.22	62.40	76.49	44.48	44.84	**46.61**	89.22	60.33
AWARe	**79.88**	**55.42**	**51.38**	**49.79**	**91.45**	**65.58** $_{\uparrow 3.18}$	**78.96**	47.01	**46.94**	46.13	**91.45**	**62.10** $_{\uparrow 1.77}$

Table 2: **Detailed continuous instruction tuning performance on the MLLM-DCL benchmark** (Zhao et al., 2025). *Avg* reports the average score for each subtask across all later evaluation points after that subtask is learned, while *Last* reports the final score on each subtask after training on the last subtask.

restricted, we used the MMMU general dataset as a calibration dataset and still achieved relatively good results, outperforming both LoRASculpt and SPIDER. We further verify that AWARe is not limited to the LLaVA-v1.5 backbone by conducting an additional Qwen2.5-VL base-model experiment; these results are reported in Section B.

4.4 Continual Instruction Tuning on MLLM-DCL

To further demonstrate the effectiveness of our proposed method in sequential learning scenarios, we evaluate AWARe on the MLLM-DCL benchmark (Zhao et al., 2025). This benchmark is specifically designed to assess continual instruction tuning for multimodal large language models. As shown in Table 2, AWARe achieves the highest average per-task performance (*Avg*) and the best final performance after the last task (*Last*), improving the overall averages by 3.18 and 1.77 points over the strongest baseline, respectively. The detailed CL results are provided in Section A. Overall, AWARe effectively balances plasticity and stability, confirming its superiority in continual instruction tuning settings.

4.5 Ablation Studies

We conduct ablation studies to investigate the contribution of different components and hyperparameter choices in our framework.

Effectiveness of Activation-based Selection. We compare our activation-based parameter selection strategy against weight norm-based selection, a hybrid strategy ($0.3 \times \text{Weight} + 0.7 \times \text{Activation}$), and a baseline using random parameter selection at the same ratio (30%). As shown in Table 3, random selection significantly fails to protect the upstream knowledge, with the average performance on upstream tasks (*Source Avg*) dropping to 49.94 on the IconQA task. In contrast, our activation-based selection consistently achieves the best balance, outperforming both random and weight-based selection in both *Source Avg* and *Target* scores.

Strategy	IconQA		COCO-Caption	
	Source Avg	*Target*	*Source Avg*	*Target*
Random Selection	49.94	86.32	49.46	116.31
Weight Norm	58.51	85.51	53.52	115.19
$0.3W + 0.7A$	58.69	84.66	53.90	115.24
Activation	**60.13**	**86.92**	**57.56**	**120.94**

Table 3: **Ablation Study on Parameter Selection Strategy.** We compare different methods for selecting parameters to retain: random selection (Random Selection), weight magnitude (Weight Norm), activation magnitude (Activation), and a hybrid approach ($0.3W + 0.7A$).

Impact of Selection Ratio and Strategy. We explore the effect of varying the selection ratio (from 1% to 90%) and comparing Layer-Balanced versus Global-Highest selection strategies. As shown in Table 4, the Global-Highest selection strategy at a 30% ratio achieves the optimal balance between stability and plasticity, yielding the highest harmonic mean ($\mathcal{H} = 103.2$) and superior knowledge retention ($\mathcal{R} = 98.4$).

Ratio	OKVQA	OCRVQA	GQA	TextVQA	*Target*	$\mathcal{R}$	$\mathcal{E}$	$\mathcal{H}$
Layer balanced selection								
1%	55.33	64.00	60.17	54.39	82.44	95.7	102.9	99.2
10%	55.33	65.20	59.64	55.16	86.84	96.3	108.3	102.0
30%	56.28	62.75	60.62	56.13	85.29	96.5	106.4	101.2
40%	56.47	63.85	60.46	56.94	86.13	97.3	107.5	102.1
50%	57.33	63.95	60.80	57.19	86.21	97.9	107.6	102.5
70%	57.69	64.20	61.77	57.54	85.67	98.7	106.9	102.6
90%	57.65	63.70	62.01	57.92	84.07	98.7	104.9	101.7
Global highest selection								
1%	55.26	63.40	59.89	54.96	82.65	95.6	103.1	99.2
10%	55.88	63.25	59.59	55.39	86.13	95.8	107.5	101.3
30%	56.95	65.10	60.83	57.66	**86.92**	98.4	**108.4**	**103.2**
40%	57.66	64.35	61.47	57.77	86.64	98.7	108.1	**103.2**
50%	57.71	64.75	61.87	57.86	85.97	99.1	107.3	103.0
70%	57.78	64.35	**62.03**	57.84	80.64	99.0	100.6	99.8
90%	**58.13**	**65.55**	61.89	**58.33**	54.62	**99.8**	68.1	81.0

Table 4: **Ablation Study on Selection Ratio.** We evaluate the performance of our method with different ratios of parameters selected for retention on the IconQA task.

Calibration Dataset Composition. We exam-

ine how the composition of the calibration dataset influences the model's resistance to catastrophic forgetting on the IconQA task. As shown in Table 5, using any individual upstream task (e.g., OKVQA, OCRVQA, GQA, or TextVQA) as the calibration source consistently preserves knowledge, with relatively small performance variations across different configurations. However, our comprehensive calibration approach (ALL), which integrates samples from all upstream sources, yields the most balanced and superior overall performance ($\mathcal{H} = 103.2$). This demonstrates that while AWARe is robust to the specific choice of calibration task, a diverse mixture of upstream data ensures the most effective protection across the model's entire functional landscape.

Calibration Task	OKVQA	OCRVQA	GQA	TextVQA	*Target*	$\mathcal{R}$	$\mathcal{E}$	$\mathcal{H}$
OKVQA	55.17	63.96	58.56	54.72	**86.97**	95.1	**108.5**	101.4
OCRVQA	54.76	**65.16**	58.28	54.45	85.85	95.2	107.1	100.8
GQA	55.92	64.22	60.12	54.97	86.21	96.3	107.6	101.6
TextVQA	55.56	64.40	59.78	56.11	85.78	96.5	107.0	101.5
ALL	**56.95**	65.10	**60.83**	**57.66**	86.92	**98.4**	108.4	**103.2**

Table 5: **Ablation Study on Calibration Dataset Composition.** We investigate how using different subsets of upstream data for activation statistics calculation affects performance on the IconQA task.

Training Targets. We investigate the impact of applying our method to different components within the model architecture. As illustrated in Figure 5, targeting both the self-attention projections and the multimodal (`mm_projector`) yields the most favorable balance between plasticity and stability. Specifically, the `mm_projector` is essential for learning new downstream concepts; Locking it limits adaptation and causes a roughly 10-point decline in target performance. Furthermore, we observe that fine-tuning `MLP` blocks results in a severe performance collapse on upstream tasks (dropping below 30-point), even with our adaptive protection. This reinforces the hypothesis that MLP weights house the bulk of the model's fundamental general knowledge, whereas self-attention projections provide a more flexible landscape for specialized task adaptation.

4.6 Sensitivity to Calibration Datasets

To assess the statistical significance and robustness of our results, we evaluate the sensitivity of AWARe to the randomness of the calibration set. We repeat the Knowledge Profiling phase over three independent runs with different random seeds for sample selection. As shown in Table 6, the performance metrics exhibit minimal standard deviation across runs, demonstrating that the identified salient neurons are consistent and statistically stable artifacts of the model structure, rather than noise from specific data samples. Moreover, increasing the calibration size from 200 to 400 samples yields negligible performance differences, confirming that 200 samples are statistically sufficient to robustly estimate parameter importance.

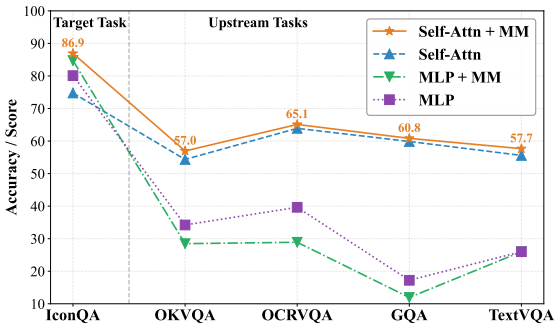


Figure 5: **Ablation Study on Training Target.** Performance comparison across different architecture components on the IconQA and upstream tasks. Applying AWARe to both `self-attn` and `mm_projector` achieves the best stable-plastic trade-off.

Calibration Size	OKVQA	OCRVQA	GQA	TextVQA	*Target*	$\mathcal{R}$	$\mathcal{E}$	$\mathcal{H}$
200	56.95	65.10	60.83	57.66	86.92	98.4	108.4	103.2
200	56.88	64.45	61.20	58.04	86.57	98.5	108.0	103.0
200	56.49	64.50	60.49	58.15	86.58	98.1	108.0	102.8
400	56.37	64.38	60.58	57.77	86.20	97.9	107.5	102.5
400	56.74	64.40	60.38	57.17	86.69	97.7	108.2	102.7
400	57.04	64.62	60.47	56.77	86.35	97.8	107.7	102.5
Std. Dev.	*0.24*	*0.25*	*0.28*	*0.48*	*0.23*	*0.30*	*0.30*	*0.30*

Table 6: **Sensitivity Analysis on Calibration Datasets.** We evaluate the performance of AWARe on IconQA task across different calibration set sizes (200 and 400) and random sample 3 times.

5 Conclusion

In this paper, we introduce AWARe, an activation-weighted framework for mitigating catastrophic forgetting in Multimodal Large Language Models. AWARe first profiles task-induced activations to identify salient neurons, then freezes these regions during downstream fine-tuning to preserve pre-trained capabilities. This design is architecture-agnostic, requires no additional adapter modules, and updates only a small fraction of parameters. Across downstream adaptation and MLLM-DCL experiments, AWARe consistently improves the stability-plasticity trade-off, and remains effective even when using a general-purpose calibration set such as MMMU.

6 Limitations

AWARe relies on a calibration set to estimate activation-based neuron importance. Although our experiments show that a general-purpose dataset such as MMMU can provide a practical substitute when upstream task data is unavailable, the quality and coverage of the calibration data may still affect the selected frozen regions. In addition, our evaluation focuses on representative MLLM backbones and multimodal tasks; broader validation on larger models, more diverse domains, and longer continual learning sequences would further clarify the generality of the approach. Finally, because AWARe explicitly freezes salient parameters to preserve prior capabilities, extremely large distribution shifts may still require tuning the retention ratio to balance stability and plasticity.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, and 1 others. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, and 1 others. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19552–19564.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Jingkuan Song, and Lianli Gao. 2024a. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in Neural Information Processing Systems*, 37:57817–57840.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. 2025. Sefe: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *ICML*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, and 1 others. 2024. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *ACL*, pages 1932–1945.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. 2025a. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13572–13586.

Haiyang Guo, Fanhu Zeng, Fei Zhu, Wenzhuo Liu, Da-Han Wang, Jian Xu, Xu-Yao Zhang, and Cheng-Lin Liu. 2025b. Federated continual instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1325–1335.

Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Jiajun Hu, Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. 2024. Learn to preserve and diversify: Parameter-efficient group with orthogonal regularization for domain generalization. In *European Conference on Computer Vision*, pages 198–216. Springer.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. Clmoe: 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*, pages 19608–19617.

Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. 2025. Learn from downstream and be yourself in multimodal large language models fine-tuning. In *ICML*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. 2024. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024b. Revisiting catastrophic forgetting in large language model tuning. *arXiv preprint arXiv:2406.04836*.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR.

Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. 2025. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26170–26180.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. In *CVPR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. In *CVPR*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *ICML*.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. 2018. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *NeurIPS*, 30.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2023. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. *arXiv preprint arXiv:2311.08011*.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. 2025. Activation-informed merging of large language models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of lora. *arXiv preprint arXiv:2402.15896*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. 2023. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. *arXiv preprint arXiv:2310.02998*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. 2024. Corda: Context-oriented decomposition adaptation of large language models. *arXiv preprint arXiv:2406.05223*.

Yuxuan Yao, Shuqi Liu, Zehua Liu, Qintong Li, Mingyang Liu, Xiongwei Han, Zhijiang Guo, Han Wu, and Linqi Song. 2025. Activation-guided consensus merging for large language models. *NeurIPS*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *ICML*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Modalprompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12137–12152.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024a. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. pages 202–227. PMLR.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024b. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *CPAL*, pages 202–227. PMLR.

Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*.

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

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. 2024. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *ICML*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhen Zhu, Yiming Gong, Yao Xiao, Yaoyao Liu, and Derek Hoiem. 2025b. How to teach large multimodal models new skills.

A MLLM-DCL Evaluation Details

We follow the MLLM-DCL benchmark (Zhao et al., 2025) and report four standard metrics for continuous instruction tuning. Let T be the total number of sequential tasks, and let $A_{t,i}$ denote the test accuracy on task i after the model has finished training on task t.

- **Mean Fine-tune Accuracy (MFT).** MFT measures plasticity, i.e., how well the model adapts to each task immediately after learning it:

$$\mathrm{MFT} = \frac{1}{T}\sum_{i=1}^{T} A_{i,i}. \tag{9}$$

  A higher MFT indicates stronger immediate adaptation to the downstream distribution of each new task.

- **Mean Final Accuracy (MFN).** MFN measures the model's average performance on all tasks after the entire training sequence:

$$\mathrm{MFN} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}. \tag{10}$$

  Unlike MFT, MFN accounts for later performance degradation, so a high MFN requires both learning new tasks and retaining previous ones.

- **Mean Average Accuracy (MAA).** MAA summarizes the accuracy trajectory throughout continual learning by averaging the cumulative accuracy over observed tasks at each training step:

$$\mathrm{MAA} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{t}\sum_{i=1}^{t} A_{t,i}\right). \tag{11}$$

  MAA rewards methods that maintain consistently high accuracy across all known tasks during the learning process.

- **Backward Transfer (BWT).** BWT quantifies the average final change in performance on previously learned tasks relative to their accuracy immediately after learning:

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(A_{T,i} - A_{i,i}\right). \tag{12}$$

  A less negative or positive BWT indicates less forgetting, while a negative BWT reflects degradation on earlier tasks.

Method	MFT ↑	MFN ↑	MAA ↑	BWT ↑
ModalPrompt	53.87	53.94	49.67	**+0.09**
HiDe	60.77	55.31	60.68	-6.82
LoRA-FT	64.98	53.00	61.13	-14.97
MoELoRA	64.94	54.78	61.76	-12.70
CL-MoE	66.06	54.88	61.79	-13.96
O-LoRA	65.16	55.54	62.12	-12.03
SEFE	65.01	58.51	63.63	-8.13
DISCO	64.78	60.33	63.93	-5.57
AWARe	**67.49**	**62.10**	**67.28**	-6.74

Table 7: Detailed CL metric summary on MLLM-DCL.

Training Stage	RS	Med	AD	Sci	Fin
RS	78.87	–	–	–	–
Med	80.76	59.44	–	–	–
AD	80.80	59.03	54.25	–	–
Sci	80.02	56.20	52.94	53.45	–
Fin	78.96	47.01	46.94	46.13	91.45

Table 8: Detailed AWARe accuracy results on MLLM-DCL.

For the detailed AWARe run in Table 8, we use 200 MMMU samples as the calibration dataset when training the first RS task. For each subsequent task, the calibration dataset consists of MMMU together with 100 samples from each previous task, which are used to recompute the activation distributions. Rows indicate the latest task after which the model is fine-tuned, and columns indicate evaluation tasks.

B Additional Model Qwen2.5-VL

To examine whether AWARe remains effective on a stronger recent backbone, we additionally evaluate Qwen2.5-VL (Bai et al., 2025b) on two downstream tasks from MLLM-DCL: Med and AD. We use HallusionBench (Guan et al., 2024), MicroVQA (Burgess et al., 2025), DocVQA (Mathew et al., 2021), VStarBench (Wu and Xie, 2024), and AI2D (Kembhavi et al., 2016) as upstream evaluation tasks to measure retained general multimodal capabilities after task-specific tuning. The zero-shot row reports the original Qwen2.5-VL performance before downstream adaptation.

Across both downstream tasks, SFT reaches the highest target-task accuracy but causes severe degradation on upstream capabilities, especially on DocVQA, VStarBench, and AI2D. In contrast, AWARe consistently preserves the strongest upstream performance among fine-tuned methods while retaining competitive downstream adaptation, indicating that activation-weighted parameter freezing transfers to Qwen2.5-VL and provides a

Method	DCL-Med ↑	HallusionBench ↑	Micro-VQA ↑	DocVQA ↑	VStarBench ↑	AI2D ↑
Zero-shot	0.357	0.770	0.480	0.947	0.620	0.740
SFT	**0.521**	0.330	0.070	0.599	0.010	0.040
AWARe	0.477	**0.760**	**0.480**	**0.913**	**0.650**	**0.730**
SPIDER	0.452	0.670	0.440	0.875	0.580	0.710
LoRASculpt	0.465	0.720	0.460	0.863	0.560	0.690

Table 9: **Qwen2.5-VL results on the MLLM-DCL Med task.** The downstream column reports DCL-Med adaptation performance, while the remaining columns report upstream task retention. Boldface marks the best downstream result and the best upstream retention among fine-tuned methods.

Method	DCL-AD ↑	HallusionBench ↑	Micro-VQA ↑	DocVQA ↑	VStarBench ↑	AI2D ↑
Zero-shot	0.213	0.770	0.480	0.947	0.620	0.740
SFT	**0.5595**	0.500	0.390	0.190	0.240	0.370
AWARe	0.4629	**0.760**	**0.440**	**0.901**	**0.600**	**0.710**
SPIDER	0.432	0.650	0.390	0.856	0.540	0.650
LoRASculpt	0.445	0.690	0.420	0.847	0.520	0.690

Table 10: **Qwen2.5-VL results on the MLLM-DCL AD task.** The downstream column reports DCL-AD adaptation performance, while the remaining columns report upstream task retention. Boldface marks the best downstream result and the best upstream retention among fine-tuned methods.

robust stability-plasticity trade-off.

C Parameter Efficiency Analysis

In this section, we provide a detailed calculation of the trainable parameter ratio in our optimal setting. Consider a standard Transformer block in LLaMA architecture with hidden dimension h. The parameters can be categorized into Self-Attention and Feed-Forward Network (FFN) modules:

- **Self-Attention:** Consists of four projection matrices: $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o \in \mathbb{R}^{h\times h}$. The total parameter count is $4h^2$.
- **FFN (SwiGLU):** Comprises three matrices: $\mathbf{W}_{up}, \mathbf{W}_{gate} \in \mathbb{R}^{h_{inter}\times h}$ and $\mathbf{W}_{down} \in \mathbb{R}^{h\times h_{inter}}$. In LLaMA, the intermediate dimension is typically set to $h_{inter} \approx \frac{8}{3}h$. The total parameter count is approximately $3 \times \frac{8}{3}h \times h = 8h^2$.

Thus, the total number of parameters per block is approximately $12h^2$.

In our best experimental setting, we apply the AWARe strategy specifically to the query, key, and value projection layers ($\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$) with a retention ratio of $\rho = 30\%$, while freezing the remaining components. The number of trainable parameters is calculated as:

$$N_{train} = 3 \times (1-\rho) \times h^2 = 3 \times 0.7 \times h^2 = 2.1h^2 \tag{13}$$

Consequently, the ratio of trainable parameters to the total parameter count is:

$$\text{Ratio} = \frac{2.1h^2}{12h^2} \approx 17.5\% \tag{14}$$

This analysis demonstrates that our method effectively updates only a quarter of the model parameters, significantly reducing computational overhead while preserving generalization capabilities.

D Experimental Configuration

We summarize the hyperparameter configuration that achieves the optimal performance reported in our main results in Table 11, specifically under the **Global-Highest 30%** setting. All experimental results are averaged over three random runs.

E Baseline Details

Single-task downstream adaptation. We calculate baselines using the following configurations:

- **Full-FT**: Full fine-tuning of the model parameters.
- **LoRA**[ICLR'22](Hu et al., 2022): Low-Rank Adaptation, which injects trainable rank decomposition matrices.
- **DoRA**[ICML'24](Liu et al., 2024): Enhances LoRA's learning capacity and training stability by decomposing weights into magnitude and direction components.
- **DARE**[ICML'24](Yu et al., 2024): Parameters from the fine-tuned model are randomly selected and rescaled to preserve both generalization and specialization capabilities.
- **Orth-Reg**[ECCV'24](Hu et al., 2024): Encourages fine-tuned features to remain orthogonal to pretrained features, thereby preserving model generalization.
- **Model Tailor**[ICML'24](Zhu et al., 2024): Pretrained parameters are preserved while a small fraction (e.g., 10%) of fine-tuned parameters is replaced based on salience and sensitivity analysis.
- **LoRASculpt**[CVPR'25](Liang et al., 2025): The method prunes redundant LoRA parameters via sparse updates guided by weight importance and conflict-aware regularization.
- **SPIDER**[ICML'25](Huang et al., 2025): Updates parameters only where specialization knowledge is more critical than generalization knowledge.

Continual instruction tuning. For the MLLM-DCL experiments, we compare against continual learning baselines that are designed for sequential task adaptation:

Hyperparameter	Value
Knowledge Profiling Phase	
Calibration Dataset Composition	OKVQA, OCRVQA, GQA, TextVQA (200 samples each)
When Upstream Dataset is restricted	MMMU (20 samples each category 600 samples in total)
Selection Strategy	Global Highest Activation
Training Target Components	`q_proj`, `k_proj`, `v_proj`, `mm_projector`
Retention Quota (ρ)	30% (Top 30% salient neurons frozen)
Downstream Fine-tuning Phase	
Epochs	3
Global Batch Size	16
Learning Rate (LLM Components)	2×10^{-5}
Learning Rate (MM Projector)	2×10^{-4}
LR Scheduler	Cosine Annealing (Warmup ratio: 0.03)
Optimizer	AdamW

Table 11: Detailed Hyperparameter Settings for the Optimal AWARe Configuration.

- **LoRA-FT**[ICLR'22](Hu et al., 2022): Sequential LoRA fine-tuning on each new subtask without an explicit forgetting-mitigation mechanism.
- **O-LoRA**[EMNLP'23](Wang et al., 2023): Learns new tasks in orthogonal subspaces to reduce interference with previously acquired knowledge.
- **MoELoRA**[NIPS'24](Chen et al., 2024a): Uses a mixture-of-LoRA-experts design for continual instruction tuning.
- **ModalPrompt**[EMNLP'25](Zeng et al., 2025): Introduces dual-modality guided prompts to preserve multimodal knowledge during continual learning.
- **CL-MoE**[CVPR'25](Huai et al., 2025): Uses a dual-momentum mixture-of-experts mechanism for continual multimodal visual question answering.
- **HiDe-LLaVA**[ACL'25](Guo et al., 2025a): Decouples multimodal continual instruction tuning hierarchically to mitigate forgetting.
- **SEFE**[ICML'25](Chen et al., 2025): Separates superficial and essential forgetting to preserve important multimodal capabilities.
- **DISCO**[ICCV'25](Guo et al., 2025b): Applies federated continual instruction tuning for sequential instruction adaptation.