Multi-Modality Expansion and Retention for LLMs through Parameter Merging and Decoupling

Junlin Li[1*], **Guodong Du**[1*], **Jing Li**[1✉], **Sim Kuan Goh**[2], **Wenya Wang**[3],
Yequan Wang[4], **Fangming Liu**[5], **Ho-Kin Tang**[1], **Saleh Alharbi**[6], **Daojing He**[1], **Min Zhang**[1]
[1]Harbin Institute of Technology, Shenzhen, China [2]Xiamen University Malaysia
[3]Nanyang Technological University [4]Beijing Academy of Artificial Intelligence, China
[5]Peng Cheng Laboratory, China [6]Shaqra University, Saudi Arabia
leejunlin27@gmail.com jingli.phd@hotmail.com

Abstract

Fine-tuning Large Language Models (LLMs) with multimodal encoders on modality-specific data expands the modalities that LLMs can handle, leading to the formation of Multimodal LLMs (MLLMs). However, this paradigm heavily relies on resource-intensive and inflexible fine-tuning from scratch with new multimodal data. In this paper, we propose *MMER (Multi-modality Expansion and Retention)*, a *training-free* approach that integrates existing MLLMs for effective multimodal expansion while retaining their original performance. Specifically, MMER reuses MLLMs' multimodal encoders while merging their LLM parameters. By comparing original and merged LLM parameters, MMER generates binary masks to approximately separate LLM parameters for each modality. These decoupled parameters can independently process modality-specific inputs, reducing parameter conflicts and preserving original MLLMs' fidelity. MMER can also mitigate catastrophic forgetting by applying a similar process to MLLMs fine-tuned on new tasks. Extensive experiments show significant improvements over baselines, proving that MMER effectively expands LLMs' multimodal capabilities while retaining 99% of the original performance, and also markedly mitigates catastrophic forgetting.

1 Introduction

Large Language Models (LLMs) (Guo et al., 2025; Wang et al., 2025; Zhang et al., 2024) have recently become a cornerstone in artificial intelligence due to their exceptional performance. Building on LLMs, researchers (Li et al., 2023a; Liu et al., 2023) integrate encoders for other modalities and use extensive modality-text data for alignment. These synthesis are then fine-tuned to develop Multimodal LLMs (MLLMs), which excel at processing multimodal inputs. This paradigm has led to the successful creation of numerous MLLMs across various modalities (Wu et al., 2024; Jiang et al., 2023).

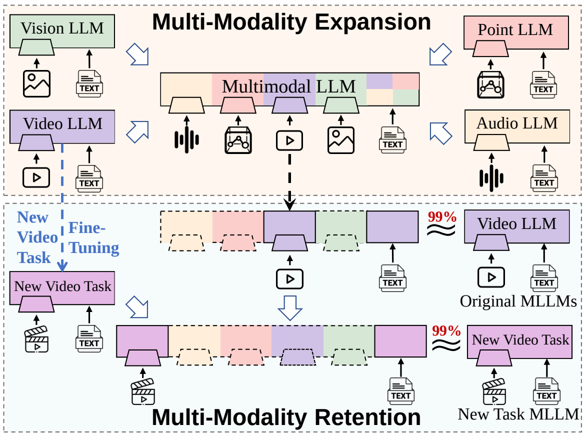


Figure 1: The key ideas of **MMER**. Multi-Modality Expansion creates a versatile model from existing MLLMs via a **training-free, extensible** process. Multi-Modality Retention reconstructs original or new task MLLMs to retain performance and mitigate catastrophic forgetting.

Most MLLMs specialize in dual modalities, including vision-oriented LLMs like LLaVA (Liu et al., 2023) and InternVL (Chen et al., 2024b), as well as video LLMs (Lin et al., 2023; Maaz et al., 2024) and audio LLMs (Chu et al., 2023; Deshmukh et al., 2023). Despite these advancements, there is a growing impetus to expand the modalities MLLMs can handle for diverse applications. A straightforward method involves adding multiple new modality encoders (Chen et al., 2023a; Lyu et al., 2023) or employing a unified multimodal encoder (Han et al., 2024), followed by re-fine-tuning the MLLMs with fresh modality-text data. However, this method is resource-intensive and lacks flexibility, as it requires generating or acquiring high-quality multimodal instruction data (Zhao et al., 2023, 2022) and fine-tuning from scratch.

To overcome the aforementioned limitations, re-

✉ Corresponding author. * Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30866–30887
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

searchers have explored model merging for multimodal expansion in MLLMs (Shukor et al., 2023; Panagopoulou et al., 2024). For instance, Chen et al. 2024a proposed NaiveMC, a basic, training-free framework that merges the LLMs of multiple MLLMs and combines their modality-specific encoders into the merged LLM. They further introduced the DAMC framework, which retrains MLLMs by separating modality parameters from language model parameters to mitigate parameter conflicts in the merged LLM. However, these two frameworks encounter a trade-off: NaiveMC is train-free but delivers lower performance, whereas DAMC requires training but yields better results.

In this paper, we propose a training-free approach named `MMER` (Multi-Modality Expansion and Retention), which enables multimodal expansion while bypassing the above trade-off and retains the original performance (See Figure 1). First, we merge the *task vectors* (Ilharco et al., 2023), which represent the difference between the fine-tuned and pre-train LLM parameters, into a merged task vector. Next, by comparing the *Directional Congruence* and *Dominant Significance* between the original and merged task vectors, we construct modality-specific binary masks. These masks can approximately identify and decouple the original modality-specific parameters retained in the merged task vector. This strategy allows the merged MLLM to independently process non-textual modality data, using its decoupled parameters, thereby significantly reducing interference from other modalities.

Furthermore, by re-adding a decoupled modality task vector into the base LLM parameters and integrating its corresponding encoder, we can reconstruct the near-original MLLMs. This strategy can retain the original modalities' performance while saving storage space. Remarkably, since our `MMER` approach is scalable, applying it to MLLMs fine-tuned on new tasks, along with multiple original MLLMs, yields a novel effect: effectively mitigating catastrophic forgetting. This approach enhances performance on new tasks without compromising previous ones by decoupling the new task's parameters from the original ones, thus preventing damage to the original parameters.

We demonstrated the effectiveness of `MMER` by composing four MLLMs (i.e., vision, audio, video, and point cloud) and conducted extensive experiments. In multimodal tasks like MCUB (Chen et al., 2024a), `MMER` significantly outperforms various baselines, confirming its ability to expand LLMs' multimodal capabilities without additional training. Moreover, we evaluated MLLMs reconstructed by `MMER` on fourteen dual-modal tasks spanning four modalities paired with text. The results reveal that they retain 99% of their original performance. Lastly, `MMER` proved resistant to catastrophic forgetting in single-task and cross-modal multi-tasks scenarios, effectively adapting to new tasks without undermining previous ones.

Our work makes several **contributions**:

- We propose `MMER`, a training-free approach for seamless multimodal expansion of LLMs through parameter merging and decoupling.
- We demonstrate two additional practical applications of the `MMER` approach: retaining the performance of original MLLMs and mitigating catastrophic forgetting in MLLMs.
- We conduct extensive and rigorous experiments on various multimodal tasks across three scenarios, with confirm the effectiveness of the `MMER` approach.

2 Related Work

Multimodal Large Language Models. Substantial researches (Dai et al., 2023; Achiam et al., 2023; Lee et al., 2024) is dedicated to developing LLMs for multimodal inputs. Vision LLMs (Alayrac et al., 2022; Li et al., 2023a) excel in vision-language tasks by connecting visual encoders to LLMs, sparking a surge in dual-modality MLLMs. Other modalities, like audio and video, quickly followed suit (Rubenstein et al., 2023; Lin et al., 2023). Meanwhile, researchers explored unifying multiple modalities into a single LLM. ImageBind-llm (Han et al., 2023) connects a multimodal encoder like ImageBind (Girdhar et al., 2023) to an LLM but relies solely on image-text data. OneLLM (Han et al., 2024) improves this by aligning all modalities with language. However, these methods cannot expand modalities due to the encoders have fixed input types. Other approaches connect multiple modality-specific encoders to an LLM, as seen in X-LLM (Chen et al., 2023a), MACAW-LLM (Lyu et al., 2023), which integrate encoders for vision, video, and audio. However, these methods require high-quality multimodal data for joint training and still struggle with modality expansion. In contrast, `MMER` provides an efficient, training-free solution for seamless multimodal expansion in LLMs.

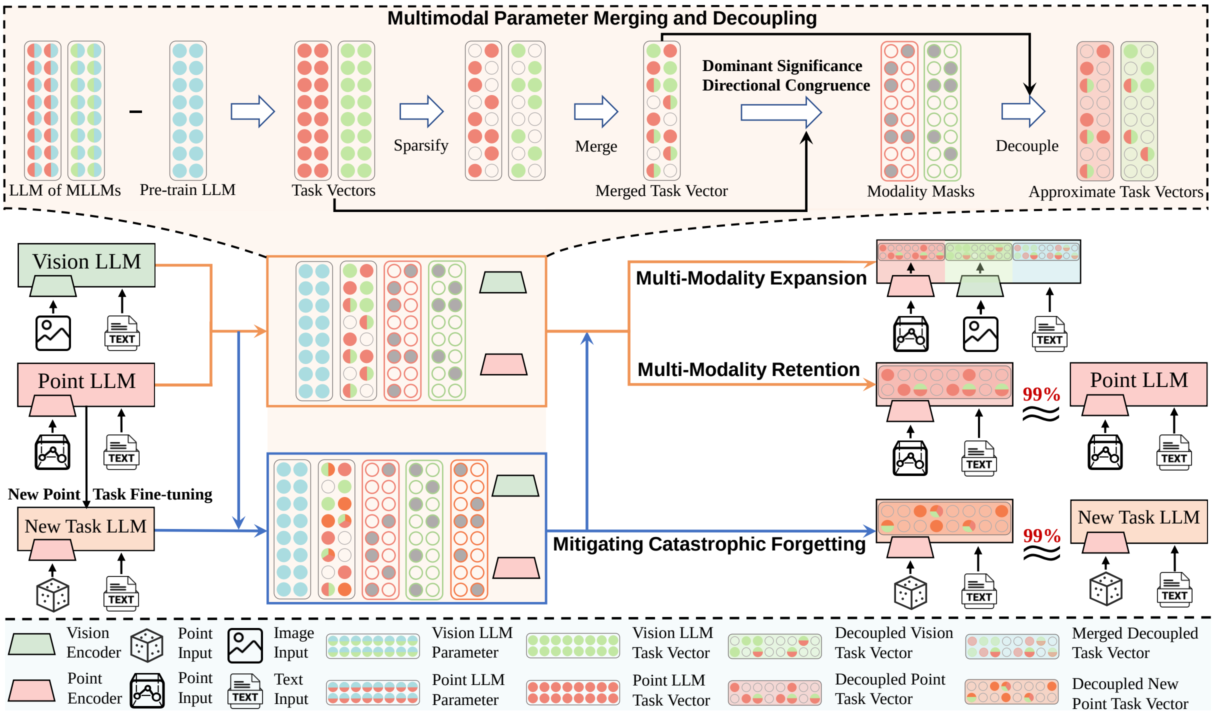


Figure 2: The overview of MMER, considering only the **Vision** and **Point Cloud** modalities **for clarity**. Each block corresponds to the same weight matrix, with empty blocks denoting zero value. "≈" signifies similar performance.

Model Merging and Model Composition. Model merging (Yang et al., 2024; Du et al., 2024b, 2025b,a) can improve single-task performance (Gupta et al., 2020), out-of-distribution generalization (Arpit et al., 2022), or combine the capabilities of multiple models (Wan et al., 2024). A basic method, TA (Ilharco et al., 2023) merges models by applying arithmetic operations on delta fine-tuned weights (i.e., task vectors), showing that such operations can yield comparable functional responses. Many subsequent methods have built upon this foundation. TIES (Yadav et al., 2023) mitigates interference during merging by pruning redundant parameters and resolving sign conflicts, while DARE (Yu et al., 2024) achieves the same by randomly dropping and scaling parameters in a preprocessing step. Moreover, Ortiz-Jimenez et al. 2023 established the theoretical foundation for TA, showing that weight disentanglement is key to its success. Model merging further applies to multimodal models. Sundar et al. 2024 explored multimodal transformers merging for specific tasks. Model Tailor (Zhu et al., 2024b) merges MLLMs to mitigate catastrophic forgetting. However, they do not explore the merging of MLLMs across modalities. To address this, the NaiveMC and DAMC frameworks (Chen et al., 2024a) merge models to create a unified MLLM that inherits multiple modality capabilities, enabling seamless expansion. However, one requires additional training, while the other delivers subpar performance. In contrast, MMER enhances the multimodal expansion capabilities of MLLMs without extra training while retaining original performance and demonstrating resistance to catastrophic forgetting. Detailed comparison with related methods is in Appendix A.

3 Methodology

In MMER, we first merge the LLM parameters $\{\theta_1, \theta_2, \ldots, \theta_n\}$ from multiple MLLMs, all fine-tuned from the same LLM θ_{pre}, into a unified LLM. However, such a merged model is prone to interference between modality-specific parameters, which can degrade the performance of representations. To handle this, we adopt a training-free parameter decoupling method that enhances the multimodal performance of the merged LLM while retaining the original performance. This is achieved by approximately decoupling modality-specific parameters within the merged parameter, ensuring independent processing of non-textual modality inputs. A visual workflow of MMER is depicted in Figure 2.

3.1 Multimodal Parameter Merging and Decoupling

Since TA (Ilharco et al., 2023) showed the effectiveness of arithmetic operations on task vectors, which is further theoretically supported by Ortiz-Jimenez et al. 2023, we apply these operations for parameter merging and decoupling. Specifically, we commence by employing the advanced model merging technique Ties (Yadav et al., 2023) to merge $\{\theta_1, \theta_2, \ldots, \theta_n\}$. Ties first extracts the task vectors for each MLLM as $\tau_{i,pre} = \theta_i - \theta_{\text{pre}}$, then refines them by selecting the Top$K\%$ absolute values to filter out non-essential parameters. This results in sparse task vectors τ_i, which are then merged base on sign consistency to generate the merged task vector $\tau_* = \mathit{merge}(\sum_{i=1}^{n} \tau_i)$. Finally, the final merged LLM parameter is $\theta_* = \theta_{\text{pre}} + \alpha \cdot \tau_*$, where $\alpha > 0$ is a scaling factor calibrated by the validation set from target tasks. If these sets are unavailable, α is determined based on the model's general performance across tasks of each modality.

Previous studies (Panigrahi et al., 2023; Wang et al., 2024) show that most of the information from the task vectors is retained and embedded in the merged task vector τ_*. By comparing the original task vectors τ_i with the merged task vector τ_*, we can identify relevant modality-specific parameter subsets from τ_*. This enables the construction of modality-specific binary masks m_i to decouple and approximate each original task vectors $m_i \circ \tau_*$. These masks filter out irrelevant parameters and reconstruct the original model parameters $\hat{\theta}_i$:

$$\hat{\theta}_i = \theta_{\text{pre}} + m_i \circ \tau_* \approx \theta_i \tag{1}$$

We construct the masks m_i by minimizing the Manhattan distance ℓ_1^* between the reconstructed model $\hat{\theta}_i$ and the LLM θ_i of original MLLMs:

$$\begin{aligned} \underset{m_i \in \{0,1\}^P}{\arg\min} \left|\hat{\theta}_i - \theta_i\right| &= \underset{m_i \in \{0,1\}^P}{\arg\min} \left|m_i \circ \tau_* - \tau_i\right| \\ &= \underset{m_i \in \{0,1\}^P}{\arg\min} \sum_{p=1}^{P} \left|m_i^{(p)} \circ \tau_*^{(p)} - \tau_i^{(p)}\right| \end{aligned} \tag{2}$$

where P represents the total number of parameters. The rationale for using the Manhattan distance is analyzed in Appendix D.1. If the sign of $\tau_i^{(p)}$ is inconsistent with that of $\tau_*^{(p)}$, the masks $m_i^{(p)}$ is set to 0 to avoid directional conflict. This step is referred to as **Directional Congruence**. Conversely, when the sign of $\tau_i^{(p)}$ aligns with $\tau_*^{(p)}$ and $\left|\tau_i^{(p)}\right| \geq \left|\tau_*^{(p)} - \tau_i^{(p)}\right|$, i.e., $\left|\tau_i^{(p)}\right| \geq 50\% \left|\tau_*^{(p)}\right|$, this indicates that $\tau_i^{(p)}$ is a dominant component of the merged parameter $\tau_*^{(p)}$. Thus, $\tau_*^{(p)}$ can be approximated as $\tau_i^{(p)}$, so $m_i^{(p)}$ is set to 1, which we refer to as **Dominant Significance**. We further introduce a scaling factor λ_i to refine this selection process, accommodating the varying numbers and modalities of original MLLMs, where a smaller λ_i selects more parameters. The selection of λ_i follows the same principle as α, enabling the modality-specific inputs to be processed in parallel and independently. The final mask m_i is constructed by the following formula:

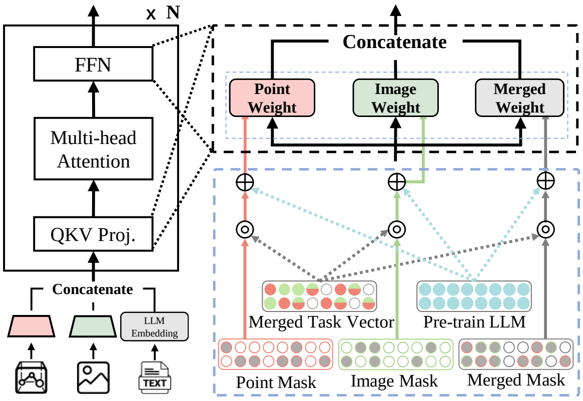


Figure 3: Details of MMER's dynamic processing. $\circledcirc$ and $\oplus$ represent the Hadamard product and addition.

$$m_i = \begin{cases} 1 & \text{if } |\tau_i^{(p)}| \geq \lambda_i \cdot 50\% |\tau_*^{(p)}| \text{ and} \\ & \quad \text{sign}(\tau_i^{(p)}) = \text{sign}(\tau_*^{(p)}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

3.2 The MMER Approach

We now comprehensively explain how the multimodal parameter merging and decoupling method enables multi-modality expansion, retention and addresses catastrophic forgetting in MLLMs.

3.2.1 Multi-Modality Expansion

Typical MLLMs consist of modality-specific components (i.e., multimodal encoders and alignment layers) and a base fine-tuned LLM. Our MMER approach disentangling these components, then applies the parameter merging and decoupling strategy to the fine-tuned LLMs of multiple MLLMs, producing a merged task vector τ_*, the pre-trained LLM parameter θ_{pre}, and n modality-specific binary masks m_i. The modality-specific components, including their weights, are reused directly, enabling the merged MLLM to seamlessly process all original modalities without losing functionality.

As depicted in Figure 3, upon receiving multimodal data, MMER respectively encodes them into representation inputs $X = [X_{M_1}, \ldots, X_{M_n}, X_t]$, where X_{M_i} and X_t represent the modality-specific sequences and text sequences. MMER then dynamically decouples the approximate modality-specific parameters $\theta_{\text{pre}} + m_i \circ \tau_*$. This ensures that non-textual modality representations are processed independently with their respective parameters. Text representations, on the other hand, are processed with the merged parameter $\theta_{\text{pre}} + \overline{m} \circ \tau_*$, where $\overline{m}$ is the average of all masks m_i. For example, when representations progress to the attention mechanism at the l-th layer, MMER decouples the modality-specific parameter from $W^{Q}_{*,l}$, the queries weights in the l-th layer from τ_*, then:

$$\begin{aligned}\mathbf{Q}_l = \Big[& X_{M_1,l}\left(m^{Q}_{1,l} \circ W^{Q}_{*,l} + W^{Q}_{pre,l}\right), \\ & \ldots,\ X_{t,l}\left(\overline{m}^{Q}_{l} \circ W^{Q}_{*,l} + W^{Q}_{pre,l}\right)\Big] \qquad (4)\end{aligned}$$

where $W^{Q}_{pre,l}$ denotes the queries weights in the l-th layer form θ_{pre}. Afterward, MMER sequentially decouples the modality-specific parameters for the keys and values in the l-th layer, and compute $\mathbf{K}_l$ and $\mathbf{V}_l$. Finally, we carry out attention operation:

$$X^{a}_{l} = \mathit{Attention}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l) \qquad (5)$$

$$[X^{a}_{M_1,l}, \ldots, X^{a}_{M_n,l}, X^{a}_{t,l}] = \mathit{Split}(X^{a}_{l}) \qquad (6)$$

Please note that the output representation should be partitioned by modality to match the input form. Consequently, the final output of the attention mechanism at the l-th layer is:

$$\begin{aligned}[X^{o}_{M_1,l}, \ldots, X^{o}_{t,l}] = \big[& X^{a}_{M_1,l}\left(m^{O}_{1,l} \circ W^{O}_{*,l} + W^{O}_{pre,l}\right) \\ & , \ldots, X^{a}_{t,l}\left(\overline{m}^{O}_{l} \circ W^{O}_{*,l} + W^{O}_{pre,l}\right)\big] \qquad (7)\end{aligned}$$

This procedure alleviates parameter conflicts across modalities, ensuring the merged MLLM retains fidelity when processing multimodal data.

3.2.2 Multi-modality Retention

Model merging and NaiveMC exhibit performance degradation (See Table 2) when handling modality-specific original tasks due to discrepancies between merged and original model parameters. However, MMER circumvents this issue by approximately reconstructing the original MLLMs. This process involves decoupling the modality-specific task vector $m_i \circ \tau_*$, adding it to the pre-trained LLM θ_{pre} to obtain the restored LLM $\hat{\theta}_i = \theta_{\text{pre}} + m_i \circ \tau_*$, and then integrating the corresponding modality-specific components to reconstruct the final MLLM. This strategy effectively mitigates parameter interference and retains original performance.

3.2.3 Mitigating Catastrophic Forgetting

Typically, fine-tuning MLLMs on new data improves performance on new tasks but often causes catastrophic forgetting on previous ones (Goodfellow et al., 2013). Drawing on the insight of Multi-modality Retention, MMER can additionally mitigate catastrophic forgetting. We first fine-tune the corresponding original MLLM on the new tasks. Next, we apply the parameter merging and decoupling method to the fine-tuned MLLM, alongside all original MLLMs, generating a new merged task vector and binary masks. Finally, we reconstruct the corresponding MLLM in a targeted manner to handle different tasks. This enables MMER to effectively adapt to new tasks without compromising previous ones, mitigating catastrophic forgetting.

4 Experiments Setup

4.1 Implementation

We explored MMER across four MLLMs: Vision, Audio, Video, and Point Cloud LLMs. To ensure fairness and comparability, we fine-tuned these four MLLMs in the same environment, each based on Vicuna-7B-v1.5 (Zheng et al., 2023), following previous works (Chen et al., 2024a; Panagopoulou et al., 2024). Details on experimental hyperparameters and fine-tuning can be found in Appendix B.2. We evaluated performance based on evaluation scores or accuracy and performance retention, the latter as defined in Appendix B.1.

4.2 Baseline Methods

We compared MMER with training-free methods: NaiveMC (Chen et al., 2024a), TA (Ilharco et al., 2023), TIES (Yadav et al., 2023), and PCB-Merging (Du et al., 2024a), where TA and TIES can substitute the merging strategy of NaiveMC for better performance. DARE (Yu et al., 2024) was integrated with these methods as it can complements them. For multi-modality expansion experiments, we included training-based baselines: ImageBind-LLM (Han et al., 2023) and X-InstructBLIP (Panagopoulou et al., 2024).

4.3 Datasets and Tasks

In multi-modality expansion experiments, we evaluated multimodal tasks, including MCUB (Chen et al., 2024a), MUSIC-AVQA (Li et al., 2022), and ModelNet40 (Wu et al., 2015) with images. For multi-modality retention experiments, we assessed fourteen dual-modal tasks spanning four

Task (→) Method (↓)	**ModelNet40** PI-T	**MUSCI-AVQA** IA-T	 VI-T	 VA-T	 AVI-T	**MCUB** AVI-T	 AVP-T	 AIP-T	 VIP-T	 AVIP-T	Avg.
–Training-based Multimodal Baselines											
ImageBind-LLM(Han et al., 2023)	39.86	36.54	38.76	39.72	38.16	35.20	31.40	33.40	31.80	32.93	35.51
X-InstructBLIP[ECCV24] (Panagopoulou et al., 2024)	57.93	40.71	41.23	48.34	47.39	41.40	25.20	21.20	29.40	27.94	37.04
–Training-free Model Merging Methods											
NaiveMC[ACL24] (Chen et al., 2024a)	60.53	39.31	47.65	47.40	49.64	53.64	56.28	60.53	54.60	59.16	53.23
TA[ICLR23] (Ilharco et al., 2023)	62.04	40.22	47.97	46.70	49.93	53.44	56.28	63.36	55.40	59.72	53.90
TIES[NeurIPS23] (Yadav et al., 2023)	61.74	43.27	49.27	48.60	51.19	53.64	55.47	61.74	54.60	58.55	54.10
PCB-Merging[NeurIPS24] (Du et al., 2024a)	62.15	44.32	50.24	49.67	51.54	54.54	56.68	63.97	55.60	60.48	54.92
NaiveMC (w/ DARE[ICML24] (Yu et al., 2024))	60.32	39.78	47.98	47.67	49.89	53.64	56.68	60.73	54.80	59.53	53.46
TA (w/ DARE)	**62.75**	40.46	47.98	46.92	50.27	54.25	56.48	64.17	55.40	60.08	54.27
TIES (w/ DARE)	61.96	43.78	49.54	48.98	51.36	54.25	55.87	62.55	55.20	59.06	54.57
MMER (ours)	62.15	**47.25**	**51.27**	**51.77**	**53.54**	**56.48**	**59.31**	**65.59**	**56.00**	**61.63**	**56.82**

Table 1: Accuracy (%) on multimodal tasks with various combinations of video (V), image (I), audio (A), point cloud (P), and text (T) inputs. Optimal results are in bold, while sub-optimal results are underlined.

Task (→) Method (↓)	**2 Point Tasks** Score (%) / Acc. (%)	**3 Audio Tasks** Score (%) / Acc. (%)	**2 Video Tasks** Acc. (%)	**7 Image Tasks** Acc. (%)	**Trimmed Avg.** Score (%) / Acc. (%)
Original MLLMs (Zero-shot)	23.15 / 21.27	25.30 / 24.71	39.79	62.23	24.23 / 51.01
NaiveMC [ACL2024] (Chen et al., 2024a)	22.65 (97.8) / 20.49 (96.3)	24.59 (97.2) / 30.65 (124.8)	36.92 (93.0)	52.56 (83.6)	23.62 (97.5) / 44.59 (88.3)
TA [ICLR23] (Ilharco et al., 2023)	22.96 (99.2) / 21.02 (98.8)	24.68 (97.5) / 31.88 (129.8)	37.57 (94.5)	54.89 (87.5)	23.82 (98.3) / 46.23 (91.0)
TIES [NeurIPS23] (Yadav et al., 2023)	22.82 (98.6) / 20.83 (97.9)	24.79 (98.0) / 32.15 (130.9)	37.81 (95.1)	54.10 (86.2)	23.80 (98.3) / 45.96 (90.6)
PCB-Merging [NeurIPS24] (Du et al., 2024a)	23.00 (99.4) / 21.16 (99.5)	25.03 (98.9) / 33.41 (135.2)	38.47 (96.7)	56.02 (90.0)	24.02 (99.1) / 47.24 (92.6)
NaiveMC (w/ DARE[ICML2024] (Yu et al., 2024))	22.83 (98.6) / 20.77 (97.6)	24.72 (97.7) / 31.62 (128.8)	37.63 (94.4)	53.61 (85.3)	23.78 (98.1) / 45.62 (89.8)
TA (w/ DARE)	23.04 (99.5) / 21.25 (99.9)	24.82 (98.1) / 32.44 (132.0)	37.52 (94.4)	55.47 (88.4)	23.95 (98.8) / 46.50 (91.4)
TIES (w/ DARE)	22.76 (98.3) / 20.98 (98.6)	24.92 (98.5) / 33.02 (134.4)	38.00 (95.6)	54.73 (87.2)	23.84 (98.4) / 46.37 (91.4)
MMER (ours)	**23.14 (99.9) / 22.49 (105.7)**	**25.20 (99.6) / 38.51 (155.6)**	**39.28 (98.5)**	**62.40 (100.3)**	**24.17 (99.8) / 50.84 (99.4)**

Table 2: Results of multi-modality retention experiments. The performance retention is shown in parentheses. "Trimmed Avg." represents the average result obtained after excluding three point or audio classification tasks.

modalities paired with text. Vision tasks include VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), TextVQA (Singh et al., 2019), VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), POPE (Li et al., 2023b), and OK-VQA (Marino et al., 2019). Audio tasks cover TUT (Mesaros et al., 2017), VocalSound (Gong et al., 2022), and Clotho (Drossos et al., 2020). Video tasks include MSRVTT (Xu et al., 2016) and MSVD (Chen and Dolan, 2011), and point tasks focus on ModelNet40 (Wu et al., 2015) and Objaverse (Deitke et al., 2023). We evaluated MMER's resilience to catastrophic forgetting on two new tasks, vision task Flickr30k (Young et al., 2014) and audio task Clotho-AQA (Lipping et al., 2022).

5 Main results

Results on Multi-Modality Expansion. As shown in Table 1 , we observe the following: **(i)** Advanced training-free model merging methods improve the NaiveMC framework's performance, suggesting their effective application to the merging of MLLMs–a previously unexplored area. This also suggests considerable parameter conflicts in the merged MLLM, as these methods primarily focus on mitigating conflicts among merging parameters. **(ii)** Our MMER approach significantly outperforms NaiveMC across all input combinations and tasks, demonstrating its effectiveness in extending multimodal capabilities and enhancing merged MLLMs' ability to manage modality combinations without additional training. **(iii)** Furthermore, MMER outperforms various baselines on nearly all tasks. This indicates that directly decoupling parameters after merging is more effective than merely reducing conflicts during the merging process. Lastly, the results for the original MLLMs are included in Appendix E.2.

Results on Multi-Modality Retention. The results in Table 2, reveal the following: **(i)** Interestingly, all methods show notable improvements on specific audio and point tasks. This likely due to these tasks are classification-based, whereas others involve captioning or QA tasks. The original audio and point LLMs, not fine-tuned for classification tasks, fail to follow instructions leading to poorer

Task (→)	Previous Tasks					New Tasks	
	2 Point tasks	3 Audio tasks	2 Video tasks	7 Image tasks	3 Multimodal tasks	Clotho-AQA	Flickr30k
Baseline (↓)	Score / Acc.	Score / Acc.	Acc.	Acc.	Acc.	Acc.	Score
Original MLLMs	23.15 / 21.27	25.30 / 24.71	39.79	62.23	-	49.40	51.26
Fine-tune on Clotho-AQA	-	19.82 / 12.31 (↓)	-	-	-	57.80 (↑)	-
Fine-tune on Flickr30k	-	-	-	57.25 (↓)	-	-	57.71 (↑)
MMER	23.14 / 22.49	25.20 / 38.51	39.28	62.40	56.82	49.28	51.00
MMER-Clotho-AQA	22.95 / 21.87	25.12 / 38.23 (∼)	39.17	62.20	56.53	57.71 (↑)	50.94
MMER-Flickr30k	23.05 / 22.03	24.96 / 37.68	38.90	62.27 (∼)	56.44	48.94	57.08 (↑)
MMER-Clotho-AQA+Flickr30k	22.82 / 21.56	24.88 / 37.69 (∼)	38.53	61.94 (∼)	55.89	57.52 (↑)	56.72 (↑)

Table 3: Results on previous and new tasks in both single-task and **cross-modal** multi-task scenario. MMER-xx refers to merging the MLLM fine-tuned on the new task xx into MMER. MMER-Clotho-AQA+Flickr30k denotes the merging of both the audio LLM fine-tuned on Clotho-AQA and the vision LLM fine-tuned on Flickr30k into MMER.

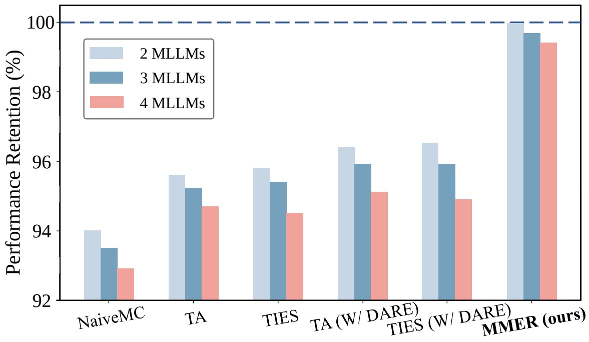


Figure 4: Performance retention vs. MLLMs quantity.

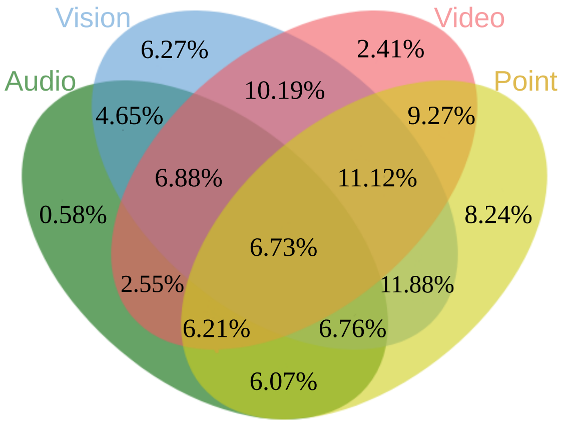


Figure 5: Parameters overlap across modalities.

performance. However, parameter merging may unlock their instruction-following ability, as the training data for other MLLMs included similar instructions. A detailed analysis is in Appendix D.3. For fairer comparison, we also provide average performance trimming these tasks. **(ii)** Although NaiveMC enables multimodal expansion for handling multimodal tasks, its performance on original tasks substantially lags behind the original MLLMs. While varied model merging methods can somewhat alleviate this decline, the gap remains notable. In contrast, MMER nearly retains the original performance. For instance, MMER achieves 99% performance retention in the trimmed average. Detailed performance for each task is in Appendix E.3.

Results on Mitigating Catastrophic Forgetting. The results for both single-task and cross-modal multi-tasks scenarios are shown in Table 3. **(i)** Fine-tuning MLLMs boosts performance on new tasks but often compromises on previous ones. In contrast, MMER, which additionally incorporates a fine-tuned MLLM (i.e., MMER-Clotho-AQA or MMER-Flickr30k), demonstrates strong robustness. It maintains nearly original performance on previous tasks and adapts effectively to new ones, achieving results comparable to fine-tuned MLLMs. **(ii)** We further integrated both fine-tuned MLLMs into MMER to showcase its performance in a cross-modal multi-tasks scenario. As more MLLMs are integrated, MMER continues to retain performance across new and previous tasks, though its ability to preserve performance slightly diminishes. Lastly, we compared MMER with LoRA in Appendix E.1. Detailed results for each task are provided in Appendix E.3.

6 Additional Results and Analysis

Performance & Storag vs. MLLM Quantity. Figure 4 presents the performance retention of merging different numbers of MLLMs in retention experiments. We can see that performance declines across all methods as more MLLMs are merged, indicating intensified parameter conflicts. Nevertheless, MMER consistently outperforms other methods with only minor degradation, while other methods exhibit a noticeable drop when dealing with multiple MLLMs. This highlights the robustness of parameter decoupling in mitigating conflicts. In terms of storage, MMER significantly reduces costs compared to maintaining individual MLLMs while

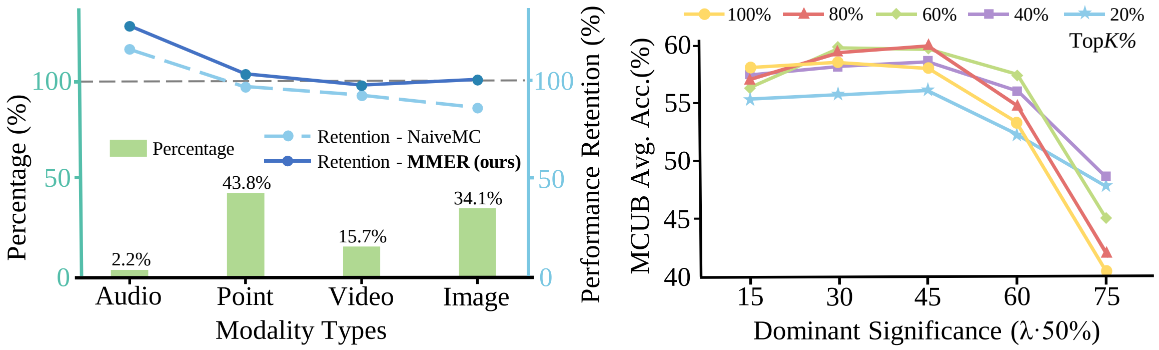


Figure 6: **(Left).** The bar plots illustrate the percentage of parameters selected by masks, while the lines show performance retention of NaiveMC and MMER across various dual-modal tasks. **(Right).** The lines depict the variations in MCUB average accuracy across different sparsity ratios (Top$K\%$) and Dominant Significance ($\lambda \cdot 50\%$).

Method	**Expansion**	**Retention**
	ACC.	**Score (%) / ACC. (%)**
MMER	**56.82**	**24.17 $_{(99.8)}$ / 50.84 $_{(99.4)}$**
− Directional Congruence	7.20	10.05 $_{(41.6)}$ / 8.34 $_{(16.7)}$
− Dominant Significance	33.87	14.71 $_{(60.5)}$ / 28.93 $_{(57.1)}$
− Scaling Factor λ	54.02	23.14 $_{(95.6)}$ / 47.78 $_{(93.9)}$

Table 4: Ablation study on parameter decoupling steps.

preserving similar performance and enabling multimodal expansion. Although it requires about twice the storage of model merging methods, it does not increase inference parameters and delivers notable performance improvements, striking an effective balance between the two approaches. Storage comparison details are in Appendix C.

Parameters Overlap in Merged Task Vector. Specifically, 40.43%, 55.36%, 64.49%, and 66.28% of audio, video, vision, and point parameters, are integrated into the merged task vector. The overlap between them shown in Figure 5, reveals a severe conflict between parameters across modalities. This underscores the need for MMER to decouple key parameters and effectively mitigate conflicts.

Modality-Specific Masks Analysis. Figure 6 (left) illustrates the percentage of parameters selected by different modality masks and compares the performance retention of MMER with NaiveMC. MMER achieves performance close to or even exceeding the original levels, indicating that crucial modality-specific information is preserved after merging. Surprisingly, we find that the audio mask, retaining only 2.2% of the parameters, still contributes to performance retention. This phenomenon aligns with previous research (Yu et al., 2024), which noted that "*Supervised fine-tuned language models tend to acquire excessively redundant delta parameters (i.e., task vectors).*" Our results further confirm that this holds true for MLLMs as well. A detailed analysis and explanation are provided in Appendix D.2.

Hyperparameters Analysis. Figure 6 (right) examines the effects of the Top$K\%$ hyperparameters and the scaling factor λ. Top$K\%$ controls the sparsity of the original task vectors. Excessive sparsity leads to marked performance degradation due to insufficient information in the sparse parameters. Conversely, insufficient sparsity fails to mitigate parameter conflicts, thereby hindering the decoupling of parameters. The effect of the scaling factor λ is akin to Top$K\%$. The scaling factor λ regulates the extent of information the mask extracts from the merged task vector. If λ is too high, the decoupled parameters lack effective information, leading to performance collapse. Conversely, if λ is too low, irrelevant parameters persist, resulting in poor performance. In summary, Top$K\%$ and λ work in tandem to regulate the amount of effective information in the decoupled parameters.

Ablation Study. In Table 4, we begin with the original parameter decoupling strategy and systematically remove components to evaluate their effectiveness. Removing Directional Congruence means selecting parameters based solely on Dominant Significance, i.e., $m_i = \mathbb{1}\{\,|\tau_i| \geq 50\% \cdot \lambda_i|\tau_*|\}$. Removing Dominant Significance retains parameters based only on the consistency of their signs, i.e., $m_i = \mathbb{1}\{sign(\tau_i) = sign(\tau_*)\}$. Table 4 shows these components are crucial for optimizing performance. Specifically, Directional Congruence is the most critical. Without it, the decoupled parameters lose all original modality information and become nearly meaningless. Next in importance is Dominant Significance. Without filtering out crucial

Method	One New Task		Two New Tasks		Storage
	Previous tasks	New task	Previous tasks	New tasks	
Model Tailor[ICML24] (Zhu et al., 2024b)	96.47 %	91.69 %	99.28 %	87.50 %	$32(P + P')$
MMER (ours)	**99.86 %**	**99.67 %**	**99.63 %**	**99.42 %**	$64P + 32P' + NP$

Table 5: Performance retention & Storage vs. Mitigating MLLMs' catastrophic forgetting methods in the **same modality**. Let N, P, and P' represent the number of new tasks, the total LLM parameters, and the modality-specific component parameters, assuming each float parameter occupies 32 bits.

parameters, irrelevant ones persist and disrupt the original parameters. Finally, the scaling factor λ also plays a role in further enhancing performance.

MMER vs. Model Tailor. In Table 5, we compare our MMER approach with the latest method for mitigating catastrophic forgetting in MLLMs within the same modality, since Model Tailor (Zhu et al., 2024b) is unable to accommodate new tasks across different modalities. The results show that MMER consistently outperforms Model Tailor in both single-task and multi-tasks scenarios, highlighting its effectiveness. Furthermore, as the number of new tasks increases, MMER maintains relatively stable performance, whereas Model Tailor exhibits a significant decline in performance on new tasks (i.e., from 91.69% to 87.50%), despite some improvement on previous tasks. However, a minor drawback of MMER is that its storage cost is approximately twice that of Model Tailor. Nonetheless, as the number of new tasks grows, MMER's practicality becomes more pronounced, making it a more viable solution in scenarios where balancing performance and storage efficiency is crucial.

7 Conclusion

In this paper, we propose MMER, a training-free method that resolves the dilemma of multimodal expansion for LLMs: costly retraining or suboptimal performance. MMER retains the multimodal encoders of existing MLLMs, merges their LLM parameters, and constructs binary masks to decouple modality-specific parameters. This mechanism enables independent handling of modality-specific inputs, reducing parameter conflicts. Besides, MMER can reconstruct original MLLMs, effectively retaining their performance and mitigating catastrophic forgetting. We conducted extensive experiments and analyses to validate the effectiveness and robustness of our MMER approach. We hope this work inspires further exploration of training-free multimodal expansion for LLMs.

Acknowledgements

This work was supported by National Science Foundation of China (62476070), Shenzhen Science and Technology Program (JCYJ20241202123503005, GXWD20231128103232001, ZDSYS20230626091203008, KQTD2024072910215406) and Department of Science and Technology of Guangdong (2024A1515011540). This work was also supported in part by the Major Key Project of PCL under Grant PCL2024A06 and PCL2022A05, and in part by the Shenzhen Science and Technology Program under Grant RCJC20231211085918010.

Limitations

We have focused exclusively on four commonly used modalities, leaving out a thorough analysis of the full range of potential modalities. Additionally, finding multiple existing MLLMs with the same architecture across modalities is currently challenging, and due to limited computational resources, experiments on larger-scale MLLMs are constrained. Finally, although our MMER approach does not increase inference parameters, the storage cost is twice that of the base model.

Ethical Considerations

Our research is conducted using publicly available and safe datasets and models. However, we explicitly acknowledge that the applicability of our MMER approach and findings may be limited to datasets or domains similar to those studied. The performance of our approach on other specific datasets or domains remains uncertain, and there may be potential risks when applying it to privacy-sensitive or high-risk scenarios. Furthermore, the generalizability of our findings to real-world applications may require further exploration and testing. Therefore, caution is advised, and thorough verification is necessary to ensure the method generates accurate and reliable results in such contexts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. 2022. Ensemble of averages: Improving model selection and boosting performance in domain generalization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024a. Model composition for multimodal large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 190–200.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023b. Beats: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 5178–5193.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 736–740.

Guodong Du, Zitao Fang, Jing Li, Junlin Li, Runhua Jiang, Shuyang Yu, Yifei Guo, Yangneng Chen, Sim Kuan Goh, Ho-Kin Tang, Daojing He, Honghai Liu, and Min Zhang. 2025a. Neural parameter search for slimmer fine-tuned models and better transfer. *arXiv preprint arXiv:2505.18713*.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, and Min Zhang. 2024a. Parameter competition balancing for model merging. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Guodong Du, Jing Li, Hanting Liu, Runhua Jiang, Shuyang Yu, Yifei Guo, Sim Kuan Goh, and Ho-Kin Tang. 2024b. Knowledge fusion by evolving weights of language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guodong Du, Xuanning Zhou, Junlin Li, Zhuo Li, Zesheng Shi, Wanyu Lin, Ho-Kin Tang, Xiucheng Li, Fangming Liu, Wenya Wang, Min Zhang, and Jing Li. 2025b. Knowledge grafting of large language models. *arXiv preprint arXiv:2505.18502*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuan Gong, Jin Yu, and James R. Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 151–155.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Weiyang Guo, Jing Li, Wenya Wang, YU LI, Daojing He, Jun Yu, and Min Zhang. 2025. Mtsa: Multi-turn safety alignment for llms through multi-round red-teaming. *arXiv preprint arXiv:2505.17147*.

Vipul Gupta, Santiago Akle Serrano, and Dennis DeCoste. 2020. Stochastic weight averaging in parallel: Large-batch training that generalizes well. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26584–26595.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

Drew A Hudson and Christopher D Manning. 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 (CVPR)*, pages 6700–6709.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2023. Dataless knowledge fusion by merging weights of language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19086–19096.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1140–1144.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 2507–2521.

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

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 32:3339–3354.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. DCASE2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pages 85–92.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2024. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 27011–27033.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21:140:1–140:67.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2646–2656.

Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2022. A simple but effective bidirectional framework for relational triple extraction. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, page 824–832.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Zesheng Shi and Yucheng Zhou. 2023. Topic-selective graph network for topic-focused summarization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 247–259.

Mustafa Shukor, Corentin Dancette, Alexandre Ramé, and Matthieu Cord. 2023. Unival: Unified model for image, video, audio and language tasks. *Transactions on Machine Learning Research (TMLR)*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Anirudh S. Sundar, Chao-Han Huck Yang, David M. Chan, Shalini Ghosh, Venkatesh Ravichandran, and Phani Sankar Nidadavolu. 2024. Multimodal attention merging for improved speech recognition and audio event classification. In *Proceedings of the Workshop on the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 655–659.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Feng Wang, Zesheng Shi, Bo Wang, Nan Wang, and Han Xiao. 2025. Readerlm-v2: Small language model for HTML to markdown and JSON. *CoRR*.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. 2024. Localizing task information for improved model merging and compression. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

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 *Proceedings of the International Conference on Machine Learning (ICML)*.

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2024. A two-stage adaptation of large language models for text ranking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lei Zhao, Junlin Li, Lianli Gao, Yunbo Rao, Jingkuan Song, and Heng Tao Shen. 2022. Heterogeneous knowledge network for visual dialog. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(2):861–871.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. 2024b. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

A Novelty and Contributions

Our research aims to achieve training-free multi-modality expansion and retention for LLMs through parameter merging and decoupling. We conduct a comparative analysis with existing relevant methods to demonstrate the innovation of our MMER approach.

Comparison with NaiveMC and DAMC frameworks. Our MMER approach is based on the NaiveMC framework (Chen et al., 2024a) and employs a parameter dynamic decoupling strategy similar to that of the DAMC framework (Chen et al., 2024a) to mitigate parameter conflicts in the merged MLLM. However, there are several key differences:

1. Compared to the NaiveMC framework, our MMER approach effectively enhances the multimodal performance of the merged MLLM.

2. Compared to the DAMC framework, our MMER approach employs a training-free parameter decoupling strategy instead of separating parameters during the initialization training of the MLLMs and achieves similar results. Additionally, MMER is additional compatible with full-parameter fine-tuned MLLMs, whereas DAMC is restricted to parameter-efficient fine-tuned MLLMs.

3. Compared to the NaiveMC and DAMC frameworks, our MMER approach retains the performance of the original MLLMs while also providing additional capabilities to mitigate catastrophic forgetting.

Our MMER approach integrates the strengths of the NaiveMC and DAMC frameworks, while additionally providing original performance retention capabilities.

Comparison with training-free model merging methods. Training-free model merging methods, such as TA (Ilharco et al., 2023), TIES (Yadav et al., 2023), PCB-Merging (Du et al., 2024a), and DARE (Yu et al., 2024), are primarily designed for merging models with identical architectures. Consequently, they must be combined with the NaiveMC framework to achieve multi-modality expansion for LLMs. These methods alleviate parameter conflicts in merged MLLMs to some extent, leading to performance enhancement. However, their overall effectiveness, both in terms of multimodal performance and retention of original performance, falls significantly short compared to our MMER approach.

Comparison with alignment and fine-tuning methods. Compared to methods (Chen et al., 2023a; Lyu et al., 2023; Han et al., 2024) that achieve multimodal expansion for LLMs by adding multiple new modality encoders or employing a unified multimodal encoder followed by alignment and fine-tuning, the advantages of our MMER approach are clear. MMER can effectively reuse a large number of MLLMs from the open-source community and merge them enabling multimodal expansion without the need for extensive resources and time spent on training models and constructing high-quality modality instruction data.

Comparison with TALL-masks. TALL-masks (Wang et al., 2024) is an information localization algorithm that, similar to our approach, compresses original parameters and subsequently approximates their restoration. However, there are several key differences:

1. From an algorithmic perspective, TALL-masks overlooks the Consistency of original and merged parameter signs. In contrast, we have addressed this aspect and demonstrated its effectiveness in our ablation experiments (See Table 4).

2. In terms of application scenarios, our MMER applies parameter merging and decoupling to the multimodal expansion for LLMs, enhancing their multimodal capabilities. Additionally, we utilize MMER to mitigate catastrophic forgetting. These aspects are not considered by TALL-masks.

3. Regarding the models utilized, the models used in our MMER approach are the 7B MLLMs across various modalities, while TALL-masks is applied to relatively smaller models within the same modality, such as T5 (Raffel et al., 2020) and ViT (Dosovitskiy et al., 2021).

B Implementation and Experimental Details

All our experiments are conducted on an NVIDIA 8×A800-SXM4-80GB machine.

Modality	Modality Encoder	Connector	Alignment Data	Fine-tuning Data	Referenced Work
Image	CLIP-ViT-L-336px (Dosovitskiy et al., 2021)	MLP	LCS 558K (Xu et al., 2024)	LLaVA-mixed 665K (Xu et al., 2024)	LLaVA-1.5 (Liu et al., 2024)
Audio	BEATs-Iter3+ (Chen et al., 2023b)	Q-Former	WaveCaps 400K (Mei et al., 2024)	OpenAQA filtered 350K (Gong et al., 2024)	X-InstructBLIP (Panagopoulou et al., 2024)
Video	LanguageBind (Zhu et al., 2024a)	MLP	LCS 558K, Valley 702K (Luo et al., 2023)	Video-ChatGPT 100K (Maaz et al., 2024), LLaVA-mixed sampled 140K	Video-LLaVA (Lin et al., 2023)
Point Cloud	Point Encoder (Xu et al., 2024)	MLP	PointLLM brief description 660K (Xu et al., 2024)	Point complex instruction 70K (Xu et al., 2024)	PointLLM (Xu et al., 2024)

Table 6: Training data and components of MLLMs for different modalities.

Stage	Hyperparameter	Image	Audio	Video	Point Cloud
Alignment-State	Batch size	256	256	256	128
	LR	1e-3	1e-3	1e-3	2e-3
	LR Schedule	cosine decay			
	Warmup Ratio	0.03			
	Epoch	1	1	1	3
Fine-tuning-Stage	Batch size	128	64	128	64
	LR	2e-5	1e-5	2e-5	2e-5
	LR Schedule	cosine decay			
	Warmup Ratio	0.03			
	Epoch	1	3	1	3

Table 7: Hyperparameters of different MLLMs.

B.1 Performance Retention

Considering the varying modalities of each original MLLM and the different evaluation metrics for distinct tasks, we provide performance retention in our results to validate the method's capacity to retain original performance. The definition is as follows:

$$\mathrm{PR} = \frac{1}{T}\sum_{t=1}^{T}\frac{\underset{x\sim\mu_t}{\text{metric}}\left[f_{\text{method}}(x)\right]}{\underset{x\sim\mu_t}{\text{metric}}\left[f_{\text{original}}(x)\right]} \tag{8}$$

where PR stands for performance retention and the "metric" refers to various evaluation metrics, such as accuracy and captioning scores(e.g., BLEU, ROUGE) (Ren et al., 2022, 2021; Shi and Zhou, 2023).

B.2 Implementation Details of Parameter Merging and Decoupling Process and Original Fine-tuned MLLMs

For the parameter merging and decoupling process, we set Top$K$ to 80%, while λ was calibrated according to the modality. We did not set the value of α as we did not use the merged MLLM merging by TIES in `MMER`. For fine-tuning the original MLLM, we used the same training data and components of each MLLM across the four modalities following NaiveMC (Chen et al., 2024a). More details are presented in Table 6. We adopted similar hyperparameters following previous works (Chen et al., 2024a; Liu et al., 2024; Panagopoulou et al., 2024; Lin et al., 2023; Xu et al., 2024). During the alignment stage, only the parameters in the connectors were trainable. In the fine-tuning stage, we tuned all connector parameters and base LLM parameters. For training efficiency, we utilized DeepSpeed Zero Optimization Stage 3. Detailed data are presented in the Table 7.

B.3 Baseline Details

In this section, we provide a detailed overview of the six baselines included in our experiments:

- **Original MLLMs** means that each MLLM is evaluated on its corresponding modality tasks to demonstrate its original performance, but they cannot perform cross-modal tasks simultaneously.

- **NaiveMC framework** (Chen et al., 2024a) combines modality-specific encoders from multiple MLLMs into the merged LLM, which is obtained by averaging the parameters of multiple LLMs from these MLLMs. The averaging merging strategy can be replaced by other model merging methods.

- **TA** (Ilharco et al., 2023) initially defines the concept of *task vector* and employs arithmetic operations for model merging, model forgetting, and support multi-tasks learning, etc. The final model is formed by scaling and adding task vectors to the initial model, represented mathematically as $\theta_m = \theta_{\text{init}} + \lambda \cdot \sum_{t=1}^{n} \tau_t$.

- **TIES** (Yadav et al., 2023) improves upon TA (Ilharco et al., 2023) by further mitigating parameter interference. It first prunes redundant parameters to retain the most important ones. When encountering conflicts in parameter signs during merging, it selects and merges parameters with the majority sign while ignoring those with minority signs.

- **DARE** (Yu et al., 2024) proposes a preprocessing step to address parameters conflict. This method randomly discards the majority of the delta parameters while scaling the remaining ones by $\theta' = \theta \cdot (1/(1-p))$ where p is the proportion of dropped delta parameters.

- **Model Tailor** (Zhu et al., 2024b) identifies the key parameters fine-tuned on the new tasks within the MLLM and integrates them into the original MLLM, thereby retaining the performance on previous tasks while adapting to new tasks.

C Storage Cost Calculation

As shown in Figure 7, although model merging methods maintain low storage costs that remain constant regardless of the number of merging MLLMs, their lower performance may constrain their practical applicability. In contrast, maintaining individual MLLMs preserves strong performance for their respective modalities but fails to achieve multimodal expansion and results in linear growth in storage costs. Our MMER approach strikes an effective balance between these approaches. It enables multimodal expansion while retaining nearly 100% of the original MLLMs' modality capabilities and provides additional resilience against catastrophic forgetting.

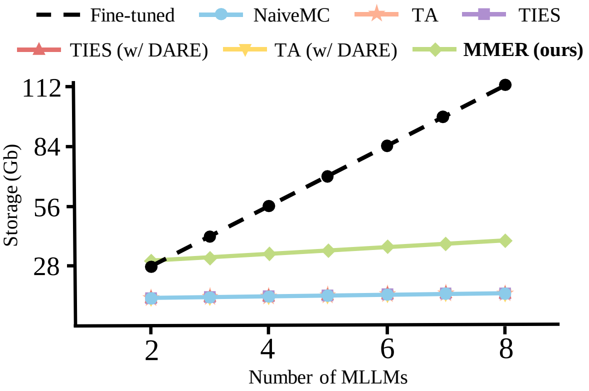


Figure 7: Storage cost vs. Number of MLLMs.

Additionally, we provide the calculation of storage costs for MMER approach and the relevant methods mentioned above. Let N, P, P', and P^* represent the number of original MLLMs, the total parameters of the LLMs, the number of the modality-specific component parameters, and the number of additional trainable parameters of parameter-efficient fine-tuning methods, respectively. Assuming each float parameter occupies 32 bits, the storage cost for these methods across N original MLLMs is calculated as follows:

- Original fine-tuned models: $32N(P + P')$. $32(P + P')$ represents the number of parameters contained in a single MLLM.

- NaiveMC framework: $32P + 32NP'$. Stores a merged LLM and N modality-specific components.

- DAMC framework: $32P + 32NP' + 2N(32P^*)$. Stores a merged LLM and N modality-specific components. $2N(32P^*)$ represents the need to store an additional $2N$ trainable parameters of parameter-efficient fine-tuning methods for parameter separation.

- NaiveMC wit TA / TIES / DARE: $32P + 32NP'$. Same as the NaiveMC framework.

- MMER: $64P + 32NP' + NP$. $64P$ is for storing the parameters of a base LLM and a merged task vector, while $32NP'$ indicates N modality-specific components. Additionally, NP denotes the storage for N binary masks.

	Expansion ACC.	Retention Score / ACC.
MMER (Manhattan)	56.82	24.17 / 50.84
MMER (Euclidean)	56.05	23.89 / 50.41

Table 8: Results of MMER with Manhattan distance or Euclidean distance

D More Analysis

D.1 Rationale for Using The Manhattan Distance

Firstly, we do not adopt methods like Fisher (Matena and Raffel, 2022) or Regmean (Jin et al., 2023), which require additional gradient-based computations to obtain the information matrix, as they demand substantial computational resources or data. Inspired by TIES (Yadav et al., 2023) and DARE (Yu et al., 2024), which propose that *"Supervised fine-tuned language models tend to acquire excessively redundant delta parameter"*, we aim to decouple the most critical parameter of each modality from the merged task vector so that the decoupled parameters are as close as possible to the original task vectors.

Based on the aforementioned concept, we decided to use a binary mask matrix to directly mask out irrelevant parameters in the merged task vector, retaining only the key information related to each modality. We chose to use the Manhattan distance to optimize the mask mainly due to its mathematical properties and its promotion of sparsity in high-dimensional parameter spaces.

In particular, since most of the delta parameters are redundant, this implies that most elements in the mask should be zero, with only a few elements set to 1. By minimizing the Manhattan distance, we can easily achieve this goal because the gradient of parameter updates with respect to the Manhattan distance is constant. This makes it more likely to penalize smaller non-zero parameters and drive them to zero, thus encouraging the sparsity of the mask. Moreover, these smaller non-zero parameters are often redundant (Yadav et al., 2023), which are the ones we wish to mask out.

Furthermore, Manhattan distance directly measures the element-wise difference between the merged task vector and the original task vectors. This comparison can precisely capture which parameters have undergone significant changes during fine-tuning and which parameters are irrelevant noise. Finally, We conducted both multi-modality expansion and retention experiments by replacing the Manhattan distance with the Euclidean distance. The results presented in the Table 8 validated the effectiveness of using Manhattan distance.

	Directional Alignment	Average Magnitude
Vision	69.20%	5e-4
Audio	50.62%	8e-5
Video	57.58%	2e-4
Point	70.09%	5e-4

Table 9: Percentage of parameters whose directions align with those in the merged task vector and the average magnitude of the parameters across the task vectors of the four modalities

D.2 Modality-Specific Masks Further Analysis

We construct the audio mask by comparing the merged task vector with the original audio MLLM task vector. Thus, the audio mask selecting only 2.2% of the parameters reflects the significant difference between these two task vectors. Next, we analyze why the remaining 97.8% of parameters were not selected. There are two possible reasons for the unselected parameters:

1. The signs of $\tau_*^{(p)}$ and $\tau_{audio}^{(p)}$ are opposite.

2. The signs of $\tau_*^{(p)}$ and $\tau_{audio}^{(p)}$ are the same, but the magnitude of $\tau_{audio}^{(p)}$ is too small.

We examined the percentage of $\tau_i^{(p)}$ whose signs align with those in the merged task vector and the average magnitude of $\tau_i^{(p)}$ across four modalities, the results are shown in Table 9.

It is evident that the direction mismatch is not the primary cause, as the percentage differences in directional alignment across the four modalities are relatively small. However, we found that the magnitude of the audio task vector is significantly smaller than those of the other modalities. This indicates that the original audio MLLM is highly similar to the pre-trained LLM. As a result, the merged model (97.8% of the parameters from the pre-trained LLM with 2.2% of the parameters activated by the audio mask from the merged task vector) only needs to activate 2.2% of the key parameters to retain its audio performance.

Task (→)	7 Original Image Tasks								New Tasks
	VQAv2	GQA	TextVQA	VizWiz	ScienceQA	POPE	OK-VQA	Avg.	Flickr30k
Method (↓)	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Score
Original MLLMs	78.11	61.52	55.89	51.51	71.12	86.17	31.33	62.23	51.26
Fine-tune on Flickr30k	72.27	54.19	46.10	52.88	70.22	76.28	28.31	57.25	57.71
Lora	75.72	58.24	52.87	52.64	70.63	85.08	29.21	60.63	54.85
MMER-Flickr30k (ours)	77.75	61.43	55.41	52.72	71.75	85.72	31.07	62.27	57.08

Table 10: The results of MMER and LoRA fine-tuning on original vision LLM for Flickr30k.

D.3 Analysis of Performance Improvement in Multi-Modality Retention Experiment

Firstly, the performance gain is not due to the removal of redundant parameters. In general, as more parameters are removed, performance tends to degrade (Yadav et al., 2023; Yu et al., 2024). This trend was also evident in our analysis (see Figure 6 (right)), where increasing the Dominant Significance λ·50% resulted in a reduction of selected parameters for each modality, leading to a gradual decline in performance.

So, what accounts for the performance improvement? We hypothesize that the parameters selected by the mask overlap with parameters from other modalities. To explore this further, we analyzed the overlap of the parameters selected by the audio mask with those from other modalities. We found that 41.7% of these parameters do not overlap with any other modality, while 23.2%, 21.1%, and 22.1% overlap with the video, vision, and point modalities, respectively.

It is possible that the model benefits from additional knowledge embedded in these overlapping parameters, such as prior knowledge or instruction-following capabilities. To validate this hypothesis, we replaced the overlapping parameters with the original audio task vector and conducted experiments on three audio tasks, yielding results of 24.71 (97.6%) / 24.32 (98.4%). Notably, the performance improvement was lost, which confirms the validity of our analysis.

E Detailed Results and Extended Experiments

E.1 Mitigating Catastrophic Forgetting Experiments

MMER vs. LoRA. We fine-tuned a LoRA adapter on original vision MLLM for Flickr30k, with the detailed results presented in Table 10. The results show that LoRA improves performance on target tasks but inevitably leads to a decline in performance on previous tasks, although this decline is less severe compared to full-parameter fine-tuning. In contrast, our MMER approach outperforms LoRA on target tasks, while causing almost no degradation in previous tasks. However, this comes at the cost of increased storage overhead. Both approaches have distinct advantages and disadvantages, enabling users to select the most suitable method based on their specific requirements.

Task (→) Model (↓)	ModelNet40	MUSCI-AVQA
Vision MLLM	51.94	44.06
Audio MLLM	-	30.63
Video MLLM	-	47.72
Point MLLM	21.27	-
MMER (ours)	62.15	53.54

Table 11: Accuracy (%) results of four original unimodal models on the multimodal tasks.

More importantly, our approach addresses an additional application scenario. In the open-source community, models are typically categorized into adapter-based models and full-parameter fine-tuned models. While the former can be easily integrated into existing models, the latter lacks such adaptability. Our approach bridges this gap by providing a solution to seamlessly incorporate full-parameter fine-tuned models.

E.2 Further Results of Multi-Modality Expansion Experiments

We supplemented the results of four original unimodal models on the multimodal tasks for a fairer comparison. Since MCUB cannot be evaluated using unimodal models, we excluded it from the analysis. As shown in Table 11, we observe that MMER consistently outperforms the unimodal models. This advantage arises from MMER's integration of additional modal information. This demonstrates MMER's ability to effectively decou-

Task (→)	7 Image Tasks						
	VQAv2	GQA	TextVQA	VizWiz	ScienceQA	POPE	OK-VQA
Method (↓)	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.
Original MLLMs	78.11	61.52	55.89	51.51	71.12	86.17	31.33
MMER (ours)	77.95	61.85	55.74	52.26	71.16	86.58	31.27
–Multi-Modality Retention							
NaiveMC [ACL2024] (Chen et al., 2024a)	59.73	45.83	42.29	47.87	68.52	79.41	24.28
TA [ICLR23] (Ilharco et al., 2023)	62.71	48.86	45.20	49.47	70.04	82.38	25.56
TIES [NeurIPS23] (Yadav et al., 2023)	61.78	48.23	44.60	48.67	69.05	81.21	25.13
NaiveMC (w/ DARE[ICML2024] (Yu et al., 2024))	60.91	46.62	42.88	49.04	70.09	81.08	24.62
TA (w/ DARE)	63.65	49.25	45.74	49.82	70.87	83.12	25.82
TIES (w/ DARE)	62.54	48.73	45.38	49.15	69.78	82.17	25.39
–Mitigating Catastrophic Forgetting							
Fine-tune on Flickr30k	72.27	54.19	46.10	52.88	70.22	76.78	28.31
MMER-Clotho-AQA	77.87	61.59	55.51	51.88	71.16	86.24	31.14
MMER-Flickr30k	77.75	61.43	55.41	52.72	71.75	85.72	31.07
MMER-Clotho-AQA+Flickr30k	77.32	61.33	55.23	52.33	71.02	85.43	30.94

Table 12: Results for each method on seven image tasks. All tasks are Question-Answering tasks.

Task (→)	2 Point Tasks					
	ModelNet40	Objavers-captioning				
Method (↓)	Acc.	BLEU-1	ROUGE-L	METEOR	Sentence-BERT	SimCSE
Original MLLMs	21.27	4.73	8.51	12.02	44.18	46.31
MMER (ours)	22.49	5.06	8.53	11.90	43.72	46.51
–Multi-Modality Retention						
NaiveMC [ACL2024] (Chen et al., 2024a)	20.49	4.43	8.24	11.37	43.22	45.97
TA [ICLR23] (Ilharco et al., 2023)	21.02	4.69	8.46	11.73	43.55	46.38
TIES [NeurIPS23] (Yadav et al., 2023)	20.83	4.55	8.39	11.60	43.29	46.27
NaiveMC (w/ DARE[ICML2024] (Yu et al., 2024))	20.77	4.41	8.38	11.59	43.47	46.28
TA (w/ DARE)	21.25	4.81	8.49	11.82	43.67	46.42
TIES (w/ DARE)	20.98	4.62	8.31	11.47	43.14	46.28
–Mitigating Catastrophic Forgetting						
MMER-Clotho-AQA	21.87	4.92	8.46	11.52	43.55	46.28
MMER-Flickr30k	22.03	5.08	8.55	11.63	43.61	46.36
MMER-Clotho-AQA+Flickr30k	21.56	4.98	8.39	11.38	43.34	46.02

Table 13: Results for each method on two point cloud tasks. Among them, ModelNet40 is a classification task, while Objavers is a captioning task.

ple modality parameters, enabling it to handle inputs from different modalities more efficiently, and highlights its strength in enhancing multimodal understanding.

E.3 Detailed Results

In this section, we present detailed results from the multi-modality retention and mitigating catastrophic forgetting experiments. The results of various baselines for seven vision tasks are shown in Table 12, two point cloud tasks in Table 13, three audio tasks and two video tasks in Table 14, three multimodal tasks in Table 15, and the last two new tasks in Table 16.

F Qualitative Results

We provide qualitative results in Figure 8. These results demonstrate the capability of the merged MLLM constructed by our MMER approach to understand and reason with multimodal inputs.

G Prompt for Evaluation

We present the evaluation prompts for each benchmark in Table 17. To denote the inputs for various modalities, we use “<image>”, “<audio>”, “<video>”, and “<point>” to represent vision, audio, video, and point cloud modalities, respectively.

Task (→)	3 Audio Tasks					2 Video Tasks	
	TUT	VocalSound	Clotho			MSVD	MSRVTT
Method (↓)	Acc.	Acc.	CIDEr	SPICE	SPIDEr	Acc.	Acc.
Original MLLMs	22.23	27.19	38.63	11.98	25.29	48.40	31.18
MMER (ours)	34.14	42.88	38.49	11.93	25.18	48.12	30.43
–Multi-Modality Retention							
NaiveMC [ACL2024] (Chen et al., 2024a)	29.50	31.80	37.56	11.61	24.61	44.53	29.31
TA [ICLR23] (Ilharco et al., 2023)	30.64	33.12	37.69	11.67	24.69	45.61	29.54
TIES [NeurIPS23] (Yadav et al., 2023)	30.87	33.42	37.89	11.72	24.78	45.88	29.74
NaiveMC (w/ DARE[ICML2024] (Yu et al., 2024))	30.50	32.75	37.75	11.66	24.74	45.69	29.58
TA (w/ DARE)	30.98	33.90	37.87	11.69	24.89	45.51	29.54
TIES (w/ DARE)	31.59	34.45	37.96	11.87	24.92	46.07	29.93
–Mitigating Catastrophic Forgetting							
Fine-tune on Clotho-AQA	6.98	17.65	30.02	9.40	20.04	-	-
MMER-Clotho-AQA	34.01	42.45	38.37	11.89	25.11	48.04	30.29
MMER-Flickr30k	33.41	41.94	38.10	11.81	24.98	47.74	30.05
MMER-Clotho-AQA+Flickr30k	33.54	41.83	37.97	11.76	24.92	47.38	29.67

Table 14: Results for each method on three audio tasks and two video tasks. Among them, TUT, VocalSound, MSVD, and MSRVTT are the classification tasks, while Clotho is a captioning task.

Task (→)	ModelNet40	MUSCI-AVQA			MCUB				
Method (↓)	PI-T	IA-T	VI-T	VA-T	AVI-T	AVP-T	AIP-T	VIP-T	AVIP-T
MMER-Clotho-AQA	61.98	47.01	51.22	51.43	56.08	59.11	65.08	55.80	61.08
MMER-Flickr30k	61.84	46.92	51.05	51.56	56.28	58.90	65.08	55.40	60.93
MMER-Clotho-AQA+Flickr30k	61.33	46.48	50.61	51.17	55.68	57.93	64.17	55.20	60.42

Table 15: Results of the mitigating catastrophic forgetting experiments for three MMER variants on multimodal tasks with different combinations of video (V), image (I), audio (A), point cloud (P), and text (T) inputs.

Task (→)	Clotho-AQA	Flickr30k						
Method (↓)	Acc.	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Original MLLMs	49.40	80.27	25.62	73.29	55.81	41.11	29.66	53.11
Fine-tune on Clotho-AQA	57.80	-	-	-	-	-	-	-
Fine-tune on Flickr30k	-	94.25	27.74	78.27	62.24	47.99	36.50	57.04
MMER (ours)	49.28	79.56	25.56	73.11	55.62	40.83	29.49	52.82
MMER-Clotho-AQA	57.71	79.72	25.51	73.04	55.48	40.72	29.33	52.77
MMER-Flickr30k	48.94	92.74	27.58	77.93	61.69	47.36	36.02	56.27
MMER-Clotho-AQA+Flickr30k	57.52	92.09	27.29	77.22	61.11	46.86	35.62	55.72

Table 16: Results of the mitigating catastrophic forgetting experiments for each method on two **new tasks**. Among them, Clotho-AQA is a Question-Answering task, while Flickr30k is a captioning task.

Describe this image in detail.

The image features a cat wearing a hat that is made of a pattern of orange and black stripes. The cat is looking directly at the camera with its green eyes, and its fur is a mix of orange and black. The cat's ears are perked up and its whiskers are visible. The hat is pulled down slightly over the cat's eyes, giving it a playful and mischievous look. The background is blurry, suggesting that the cat is the main focus of the image.

How many people are in this video?

There are two people in this video.

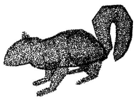
What are they doing?

The man is holding a baby in his arms and smiling while the baby is sleeping.

[Car horn sound]

Can the object in the point cloud produce the sound in the audio?

Yes, the object in the point cloud, which is a car, can produce the sound in the audio. The sound is likely a car horn or a police siren.



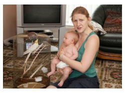
Are the specie of the animal in the image the same as that in the point cloud?

No, the animal in the image is a dog and the point cloud is a Squirrel.

[Motorcycle engine]

Can the object in the image produces the sound in the audio?

Yes, the object in the image is a motorcycle, and the sound in the audio is a motorcycle engine.



[Baby Laughs and Man Singing]

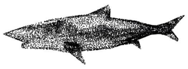
Describe the theme this image has in common with this video and this audio in detail.

The theme of this image, video, and audio is the bond between the parent and their child. The woman in the image is seen holding and caring for her baby, while the video and audio feature a man singing to his child, showcasing the love and affection he has for the baby. The theme of family and parenthood is prevalent in all three pieces of media, emphasizing the importance of nurturing and caring for one's child.

[Car Rev and Screech]

Describe this audio in detail.

The audio clip features a man speaking, followed by a car engine revving and tires screeching. The man's speech is difficult to understand due to the loud car noise. The car is accelerating and the engine is revving at a high RPM, creating a loud and intense sound. The tires screech as the car takes off, adding to the overall excitement of the scene.



What is this species?

This species is a blue shark.

Tell me more about this shark.

The shark in the image is a blue shark, which is a species of open ocean pelagic shark. It is a large shark, with a length of up to 21 feet and a weight of up to 1,100 pounds. The blue shark is known for its distinctive blue coloration, which is caused by a layer of reflective cells that help it blend into the surrounding water. It is a fast swimmer and can reach speeds of up to 35 miles per hour.

Describe this image in detail.

The image depicts a small boat, likely a gondola, carrying a large group of people on a river. The boat is covered in a canopy, and the passengers are holding umbrellas of various colors to protect themselves from the rain. The passengers appear to be mostly tourists, as they are dressed in casual clothing and some are taking photographs. The boat is being rowed by a man wearing a hat and a striped shirt. The river is lined with buildings on either side, including a bridge with a red roof in the background. The scene is peaceful and serene.

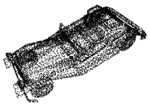
[Crowd Chatter]

Describe this audio in detail. Can this image produce this audio?

The audio is a recording of a group of people on a boat, with some of them speaking and others making noise. There are also sounds of water and wind. The audio seems to be from a different source than the image, which shows a boat with people on it, but it is possible that the audio was recorded at the same time as the image. The people on the boat are laughing and having a good time. The audio is not clear enough to make out what is being said, but it seems to be a group of people having a good time on a boat.

What is the object in the video? Could it be in the image?

The object in the video is a goldfish. It can be in the river of the image.



What is the object in the point? Could it be in the image?

The object in the point is a car. It is possible that the car is in the image, but it is not explicitly shown.

Figure 8: Qualitative results.

Benchmark	Modality	Prompt Template
MCUB	AVI-T	Based on four input entities:\nimage <image>\naudio <audio>\nvideo <video>\n {Question} {Options} Answer with the option's letter from the given choices directly.
	AVP-T	Based on four input entities:\naudio <audio>\nvideo <video>\npoint <point>\n {Question} {Options} Answer with the option's letter from the given choices directly.
	VIP-T	Based on four input entities:\nimage <image>\nvideo <video>\npoint <point>\n {Question} {Options} Answer with the option's letter from the given choices directly.
	AIP-T	Based on three input entities:\nimage <image>\naudio <audio>\npoint <point>\n {Question} {Options} Answer with the option's letter from the given choices directly.
	AVIP-T	Based on four input entities:\nimage <image>\naudio <audio>\nvideo <video>\npoint <point>\n {Question} {Options} Answer with the option's letter from the given choices directly.
MUSIC-AVQA	VI-T	Based on the video <video> and image <image>\n{Question} \nAnswer the question using a single word.
	VA-T	Based on the video <video> and audio <audio>\n{Question} \nAnswer the question using a single word.
	IA-T	Based on the image <image> and audio <audio>\n{Question} \nAnswer the question using a single word.
ModelNet40	PI-T	Based on rendered image <image> and point cloud <point>\nWhat is this? Select from these objects: {Options} Answer the question using a single word.
	I-T	<point>\nWhat is this? Select from these objects: {Options} Answer the question using a single word.
Objaverse	I-T	<point>\nOffer a clear and concise description of this point cloud object.
VocalSound & TUT	A-T	<audio>\nWhich of the following categories does this audio belong to? {Options} Answer the question using a single word.
Clotho	A-T	<audio>\nDescribe this audio in detail.
Clotho-AQA	A-T	<audio>\n{Question}\nAnswer the question using a single word or phrase.
MSRVTT & MSVD	V-T	<video>\n{Question}\nAnswer the question using a single word or phrase.
VQAv2 & GQA & POPE & OK-VQA	I-T	<image>\n{Question}\nAnswer the question using a single word or phrase.
Textvqa	I-T	<image>\n{Question}\nReference OCR token: {Options}\nAnswer the question using a single word or phrase.
VizWiz	I-T	<image>\n{Question}\nWhen the provided information is insufficient, respond with 'Unanswerable'.\nAnswer the question using a single word or phrase.
ScienceQA	I-T	<image>\n{Context}\n{Question}\nChoose the most likely ratio. {Options}
Flickr30k	I-T	<image>\nDescribe this image using one or more simple sentences.

Table 17: Prompt Template for different evaluation benchmarks.