

Transferring Math Reasoning Abilities from LLMs to MLLMs via Causal-Aware Model Merging

Yijie Hu^{1,2}, Guanyu Yang^{1,2}, Zihao Zhou^{1,2}, Bin Dong⁴,
Xiaowei Huang², Kaizhu Huang³, Qiufeng Wang^{1,2*}

¹School of Advanced Technology, Xi'an Jiaotong-Liverpool University,

²University of Liverpool, ³Digital Innovation Research Center, Duke Kunshan University,

⁴Ricoh Software Research Center Beijing Co., Ltd

Yijie.Hu20@student.xjtlu.edu.cn, Qiufeng.Wang@xjtlu.edu.cn

Abstract

To build a math MLLM, current research efforts predominantly rely on elaborate training schemes or high-quality multi-modal math reasoning data. On the other hand, model merging has attracted much attention recently due to its low-cost and tune-free property, which has shown great potential for constructing math MLLM. However, most recent attempts either assume all parameters are task-relevant or adopt a less interpretable parameter selection strategy, thus merging parameters on a coarse granularity, leading to limited or even degrading performance. In this paper, we propose a causal-analysis-guided parameter merging framework, which identifies the math reasoning-related parameters in a finer granularity and interpretable manner, then merges the math LLMs with MLLM. Specifically, we first construct targeted arithmetic and logical tasks, generate reference and counterfactual samples, and then apply intervention-based causal attribution to localize parameters that are causally responsible for math reasoning. Once we identify the crucial reasoning parameters in the math LLM that exhibit higher causal effects, we can more precisely align and merge only the corresponding parameters from the math LLM into the base MLLM. The procedure is tuning-free and enhances both math reasoning and general reasoning abilities. Experimental results of three model families on four math reasoning benchmarks and two general reasoning benchmarks validate the effectiveness of the proposed approach.

1 Introduction

Mathematical reasoning is a core capability of large language models (LLMs) and plays a crucial role in solving complex analytical and symbolic tasks (Minaee et al., 2024). Progress in this area has been accelerated by the development of high-quality

*Corresponding author.

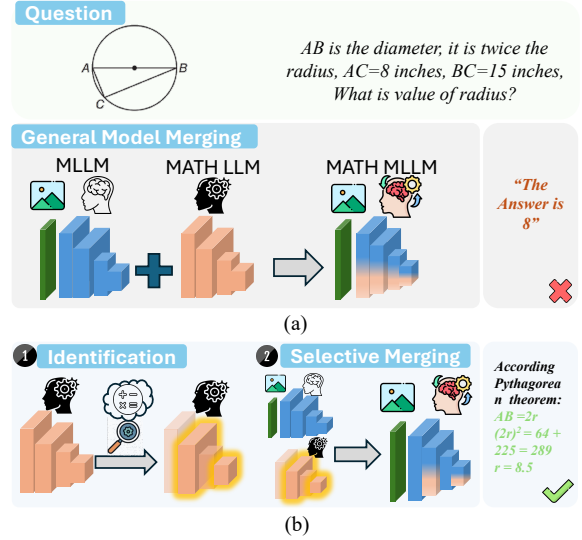


Figure 1: (a) General model merging methods merge all the parameters without selection, which may hinder reasoning performance after merging. (b) We first identify the crucial parameters through casual analysis, then select to merge the crucial parameters in the math LLM with MLLM to obtain the math MLLM.

math reasoning datasets (Yu et al., 2024a) and advanced training paradigms, enabling models such as DeepSeek-R1 and Qwen to deliver strong performance across a wide range of text-based math benchmarks (Minaee et al., 2024; Bai et al., 2023; Gou et al., 2024). However, math reasoning remains difficult partly due to the high cost of collecting and annotating high-quality multimodal reasoning data for multimodal large language models (MLLMs) (Gao et al., 2025; Zheng et al., 2024).

Serving as a low-cost and tuning-free alternative, model merging has recently been employed to enhance multi-modal math reasoning by directly merging a math LLM with a base MLLM, enabling the transfer of specialized reasoning capabilities without additional training (Chen et al., 2025; Hu et al., 2026). An MLLM is usually finetuned from a base LLM, consisting of a visual encoder and a

foundation LLM as its core components (Liu et al., 2023): the visual encoder extracts visual features, while the LLM is responsible for the reasoning process, which often shares an identical architecture with math LLMs. General model merging approaches (Ilharco et al., 2022; Yadav et al., 2023) derive task vectors for each model by subtracting the pre-trained weights from the fine-tuned weights (i.e., $\Delta W = W_{ft} - W_{pre}$), and then combine them with MLLM via addition or weighted interpolation. However, as demonstrated in fig. 1, such methods implicitly assume that all parameters are task-relevant, ignoring the need to isolate task-specific parameters, which often results in sub-optimal or even detrimental to model performance. Recent advances, such as IP-Merging (Hu et al., 2026), attempt to select reasoning parameters via computing layer-wise subspace similarities of parameters. However, IP Merging requires an extra math model for comparing the similarities and operates at a coarse layer-wise granularity (fig. 2(a)), which lacks interpretability and limits their effectiveness (fig. 2(b)). This raises a fundamental question: *How can we accurately and explainably identify the parameters responsible for math reasoning?*

In this paper, inspired by recent progress in mechanistic interpretability (MI) (Stolfo et al., 2023; Zhang et al., 2024a; Nikankin et al., 2025; Chen et al., 2025; Yang et al., 2025), we propose a causal analysis-guided framework to perform fine-grained and interpretable parameter selection in a data-driven manner. Treating LLMs as causal graphs, MI studies (Lin et al., 2025) have shown that model capabilities are often localized within sparse functional circuits composed of specific attention heads and MLP blocks, which can be revealed via attribution patching. Motivated by this insight, we design two targeted reasoning tasks, i.e., arithmetical calculation (e.g., $0 + 1 = ?$) and logical operation calculation (e.g., $0 \& 1 = ?$). We generate approximately 200K samples per task and retain only those that the model answers correctly. We then construct reference-counterfactual input pairs (e.g., $0 + 1$ vs. $2 - 3$). For each block, we replace the reference sample’s activations with those of the counterfactual sample, measuring the resulting change in output logits as a causal score. As demonstrated in fig. 2 (a), blocks with a large causal score are identified as reasoning-critical parameters. Under this paradigm, we can alleviate the need for the extra math LLM and locate reasoning parameters

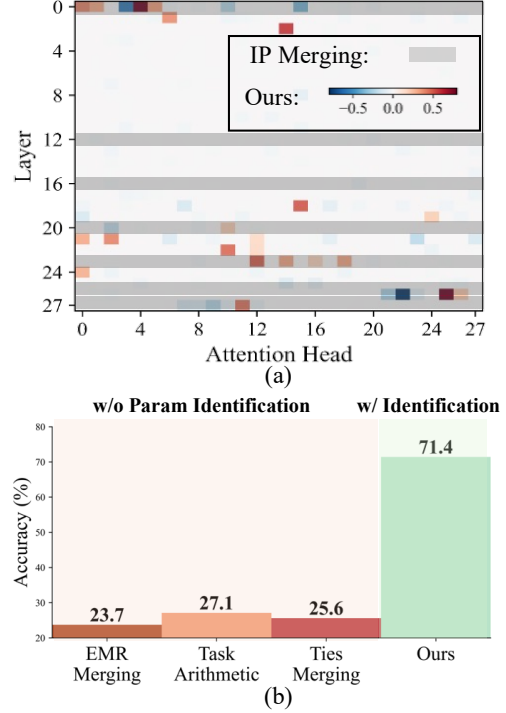


Figure 2: (a) We use attribution patching to identify the key reasoning-related parameters within Qwen2.5-Math-7B, darker color represents higher importance. IP merging identifies key parameters in a coarse manner. (b) We merge the top eight reasoning-related attention heads in the math LLM and compare our method with other model merging methods in MathVista.

accurately and explainably in finer granularity.

After identifying the crucial parameters, we further analyze their patterns and functional behaviors. We first analyze the identified parameters in the math LLM. To better understand how reasoning capabilities are encoded into model weights, we further investigate the evolution of parameter importance before and after math post-training. Our analysis reveals a distinct subset of parameters whose causal influence shifts significantly after post-training, indicating that these parameters form strengthened or newly emerged reasoning abilities. We interpret these parameters as encoding the core reasoning abilities acquired during math specialization. Based on this insight, we propose a causal-aware parameter selection strategy that prioritizes parameters with increased causal contribution after post-training, enabling more effective and interpretable knowledge transfer during merging. By comparing parameters activated by arithmetic and logical reasoning, we observe that logical reasoning parameters consistently contribute more to model merging performance, suggesting that

logical circuits encode more transferable and structurally compositional reasoning abilities. Compared to other model merging methods, our method obtains higher performance, as illustrated in fig. 2 (b). We then conduct experiments among three different model families and six datasets to validate the effectiveness of the proposed method. We summarize our contributions as follows:

- We propose a causal analysis-guided parameter selection strategy for merging MLLM with math LLM to enhance the reasoning performance.
- We design two targeted tasks for locating the reasoning parameter in an explainable manner and analyzing the validity of identified parameters.
- We propose a parameter selection strategy that further enhances the model merging performance.
- Extensive experimental results of three model families on six reasoning benchmarks validate the effectiveness of the proposed approach.

2 Related Works

2.1 Math Reasoning of MLLMs

The community has made substantial efforts to enhance mathematical reasoning, which is widely regarded as a fundamental capability of MLLMs. In the pre-training stage, previous works strengthen the reasoning ability of base models by collecting large-scale mathematical corpora (Wang et al., 2024c; Han et al., 2025; Deitke et al., 2025), generating synthetic data (Linger et al., 2025), and optimizing training strategies (Lin et al., 2024). During post-training, researchers further improve the scale and quality of mathematical instruction data through data augmentation techniques (Gao et al., 2025; Han et al., 2023; Shi et al., 2024), while the introduction of variant fine-tuning paradigms (Zhang et al., 2024b), reinforcement learning algorithms (Yu et al., 2024b; Wang et al., 2024d), and self-evolving frameworks (Zelikman et al., 2022; Liu et al., 2024) enhances the efficiency of data utilization. Recently, O1-like models have further advanced reasoning capabilities by leveraging deep-thinking chains of thought and tool-assisted inference scaling (Dong et al., 2025; Deng et al., 2025; Hu et al., 2024; Zhang et al., 2025a).

2.2 Mechanistic Interpretability

Mechanistic Interpretability (MI) methods aim to interpret what or how the model learns via causal analysis or sparse autoencoders (Rai et al., 2024). For math reasoning, recent causal-mediation studies (Stolfo et al., 2023) show arithmetic information is carried by early or mid attention to the final position where late MLPs write result features, identifying the specific components mediating predictions. Recent findings (Zhang et al., 2024a) also show that a small, sparse circuit that often under five percent of heads plus associated MLPs suffices for calculation, and selectively tuning these parts improves math without harming other skills. Recent evidence indicates models combined activity explains most arithmetic behavior (Nikankin et al., 2025). Emergent symbolic mechanisms describe a three-stage organization of abstraction, induction, and retrieval across layers (Yang et al., 2025).

2.3 Model Merging

Model merging is a low-cost solution to elevate large language models (LLMs) by fusing multiple task-specialized models into a unified one (Yang et al., 2026). Existing approaches can be broadly categorized into pre-merging and during-merging strategies. Pre-merging methods operate at the parameter or representation level, including weight arithmetic (Ilharco et al., 2022; Yadav et al., 2023). During-merging methods instead address task interference online via averaging, adaptive reweighting, orthogonal or subspace projection techniques, dynamic routing, or post-merging calibration (Yang et al., 2024c,b). To extend merging to low-rank adaptation models, recent works adapt SVD-based formulations (Gargiulo et al., 2025; Panariello et al., 2026; Hu et al., 2026), aligning LoRA updates into a shared basis to retain the efficiency of low-rank representations while significantly improving accuracy across tasks. However, these efforts predominantly focus on unimodal language models (Yang et al., 2026; Lim et al., 2024) without precise or explainable parameter identification, whereas our work locates the parameters clearly and then merges the parameters.

3 Methodology

3.1 Model Merging for MLLM and Math LLM

Model merging seeks to fuse multiple task-specialized models into a single multi-task model

without additional training. Consider a shared base model W_0 , from which both the MLLM W_{MLLM} and the math LLM W_{Math} are fine-tuned. We define their fine-tuning deltas as $\Delta W_{\text{MLLM}} = W_{\text{MLLM}} - W_0$ and $\Delta W_{\text{Math}} = W_{\text{Math}} - W_0$. With these deltas, we can derive our math-reasoning MLLM W^* by selectively merging the math-specific deltas into the MLLM. A general formulation of model merging can be written as

$$W^* = F(W_0, \Delta W_{\text{MLLM}}, \Delta W_{\text{Math}}; \mathbf{M}), \quad (1)$$

where $\mathbf{M}$ represents the parameter selection or alignment operator. Under this formulation, prior model merging approaches can be further decomposed as

$$W^* = W_0 + \alpha_1 f_1(\Delta W_{\text{MLLM}}, \mathbf{M}_1) + \alpha_2 f_2(\Delta W_{\text{Math}}, \mathbf{M}_2), \quad (2)$$

where α_1, α_2 are (global or layer-wise) scaling coefficients, and f_1, f_2 denote mapping functions such as identity, masking, normalization, or projection. Under this formulation, Task Arithmetic (Ilharco et al., 2022) sets identity mask and f_1, f_2 as identity, yielding a weighted sum of deltas. EMR-Merging (Huang et al., 2024) introduces magnitude-based sparse masks with data-dependent rescaling, while IP-Merging (Hu et al., 2026) constructs $\mathbf{M}_1, \mathbf{M}_2$ via subspace similarity and calculates scaling factors using trace ratios, with f_1, f_2 implemented as projection matrices.

Following the same framework in (2), our method defines f_1 and f_2 using causal attribution scores to identify reasoning-critical attention heads and MLP modules, from which we build selection masks $\mathbf{M}_1, \mathbf{M}_2$ by retaining parameters with high causal importance. For notation clarity, we denote the selected parameters used for merging as $\Delta W_{\text{MLLM}}^{\text{select}}, \Delta W_{\text{Math}}^{\text{select}} \in \mathbb{R}^{m \times n}$, and the remaining unmodified MLLM parameters as $\Delta W_{\text{MLLM}}^{\text{retain}}$. To preserve general abilities, we follow prior work (Hu et al., 2026) and fix $\alpha_1 = 1$. To prevent the math delta from dominating the merge, we compute α_2 via SVD-based normalization (Marczak et al.,

Algorithm 1: Parameter Identification

Input: Set D_Ω of reference and counterfactual pairs (x_r, x_c) ; frozen model $\mathcal{M}$ with components N

Output: Causal effect score E_n for each $n \in N$

```

1 foreach  $(x_r, x_c)^{(i)} \in D_\Omega$  do
2   Run  $\mathcal{M}(x_r)$  and cache activations  $A_r$ ;
   run  $\mathcal{M}(x_c)$  and cache activations  $A_c$ 
3   Let  $y_r, y_c$  be correct answers for  $x_r, x_c$ 
4   foreach  $n \in N$  do
5      $A'_r \leftarrow A_r, A'_r[n] \leftarrow A_c[n]$ 
6      $A'_c \leftarrow A_c, A'_c[n] \leftarrow A_r[n]$ 
7      $P(y_r) \leftarrow \mathcal{M}(x_r, A_r)$     original  $X_r$ 
      logits
8      $P'(y_r) \leftarrow \mathcal{M}(x_r, A'_r)$     patched
       $X_r$  logits
9      $P(y_c) \leftarrow \mathcal{M}(x_c, A_c)$     original
       $X_c$  logits
10     $P'(y_c) \leftarrow \mathcal{M}(x_c, A'_c)$     patched
       $X_c$  logits
11     $s_n^{(i)} \leftarrow \frac{1}{2} \left[ \frac{P'(y_c) - P(y_c)}{P(y_c)} + \frac{P(y_r) - P'(y_r)}{P(y_r)} \right]$ 
12 return  $E_n = \frac{1}{|D_\Omega|} \sum_{i=1}^{|D_\Omega|} s_n^{(i)}$ 

```

2025):

$$\begin{aligned} \Delta W_{\text{Math}}^{\text{select}} &= \sum_{i=1}^r \sigma_M^i \mathbf{u}_M^i \mathbf{v}_M^{i\top}, \\ \Delta W_{\text{MLLM}}^{\text{select}} &= \sum_{i=1}^r \sigma_V^i \mathbf{u}_V^i \mathbf{v}_V^{i\top}, \\ \bar{\sigma}_V &= \frac{1}{r} \sum_{i=1}^r \sigma_V^i, \alpha_2 = \frac{\bar{\sigma}_V}{\sum_{i=1}^r \sigma_M^i}, \end{aligned} \quad (3)$$

where r is the number of the singular values. This formulation scales the math parameters according to their relative spectral magnitude, enforcing balanced contribution from reasoning parameters. The final merged model preserves general MLLM parameters, merges selected reasoning parameters with controlled strength:

$$W^* = W_0 + \Delta W_{\text{MLLM}}^{\text{retain}} + \Delta W_{\text{MLLM}}^{\text{select}} + \alpha_2 \Delta W_{\text{Math}}^{\text{select}}. \quad (4)$$

3.2 Parameter Identification

3.2.1 Task Data Formulation

We design two controlled reasoning tasks, *i.e.*, arithmetic computation and logical reasoning, to con-

Arithmetic Calculations	Logical Calculations
$\{A\} + \{B\} = \{C\}$	$\{A\} \text{ AND } \{B\} = \{C\}$
$\{A\} - \{B\} = \{C\}$	$\{A\} \text{ OR } \{B\} = \{C\}$
$\{A\} * \{B\} = \{C\}$	$\{A\} \text{ XOR } \{B\} = \{C\}$
$\{A\} / \{B\} = \{C\}$	NOT $\{A\} = \{C\}$

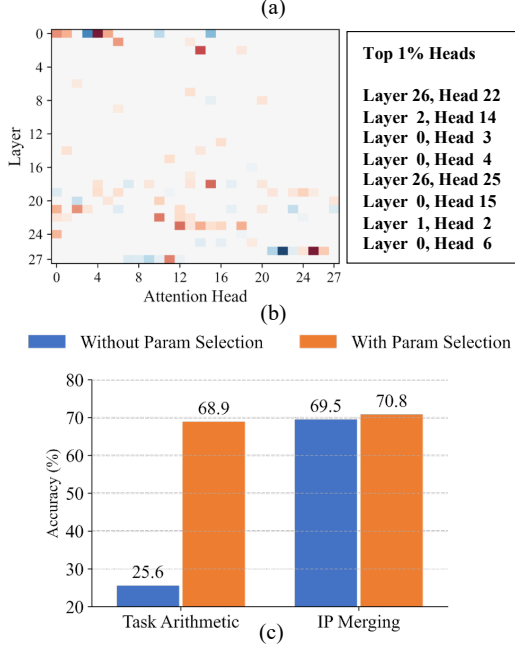


Figure 3: (a) Prompt generation template for arithmetic and logical operation tasks. (b) We demonstrate the causal map for attention heads involved in math calculation and the location of the top 1% attention heads. (c) We select the top 1% of attention heads ranked by causal importance and apply existing model merging methods both with and without this parameter selection.

struct reference-counterfactual prompt pairs (templates shown in fig. 3(a)). The arithmetic task, formatted as $\{A\} + \{B\} =$, evaluates direct numerical computation by instantiating operands $\{A\}, \{B\} \in [0, 300]$, yielding inputs such as “0+1=” and “2-3=”. The logical task is formulated as $\{A\} \text{ AND } \{B\} =$ to probe multi-step symbolic reasoning, which requires intermediate representation conversion. For instance, solving “2 AND 3” involves converting decimals to binary ($2 \rightarrow 10, 3 \rightarrow 11$), performing bitwise AND ($10 \wedge 11 = 10$), and converting the result back to decimal ($10 \rightarrow 2$), making its reasoning trajectory structurally deeper than arithmetic. Sampling operands from the same numerical range ensures that both tasks produce prompts that are semantically homogeneous yet causally distinct in their reasoning paths, a property essential for isolating functional components via attribution patching. We generate 200K samples for each task. To ensure intervention signals reflect faithful model behavior,

we filter the generated pool and retain only instances that the model answers correctly, which are then used to construct reference-counterfactual pairs for causal analysis.

3.2.2 Method

The computation of the LLM can be reorganized as a graph with N nodes (Stolfo et al., 2023), where each node is a computation component, including attention heads, MLP blocks, the residual connections, and each edge represents the data flow that the output of the previous node will be transposed to the input of the later node. To identify the key reasoning parameters, we adopt the attribution patching approach (Wang et al., 2024a). Concretely, given i -th reference prompt $x_r^{(i)}$, simply x_r (e.g., “0+1=”) that the model answers correctly and a counterfactual prompt x_c (e.g., “2-3=”) with a different correct answer, we run the frozen model on both inputs and cache intermediate activations for every component n , denoted as A_r and A_c respectively. We then perform a targeted intervention patching: we run x_r again, but we perturb exactly one component n with the activation taken from A_c , while keeping all other activations unchanged. Let P and P' denote the output token distributions before and after patching, and let y_r and y_c be the correct answers for x_r and x_c , respectively. The causal effect score (Wang et al., 2022) for component n at data pair i is defined as

$$s_n^{(i)} = \frac{1}{2} \left[\frac{P'(y_c) - P(y_c)}{P(y_c)} + \frac{P(y_r) - P'(y_r)}{P'(y_r)} \right], \quad (5)$$

which captures both the relative increase in the probability of the counterfactual answer y_c and the relative decrease in the probability of the reference answer y_r caused by patching. We repeat this procedure over (x_r, x_c) pairs in the generated target set D_Ω across all model components, and average $s_n^{(i)}$ to get E_n , obtaining the causal score for each component. See detailed algorithm in algorithm 1.

3.3 Results

We first take the arithmetic task as an example and locate the crucial attention heads in Qwen2.5-Math-7B-Instruct, as illustrated in fig. 3(b). We then rank all heads by their causal scores and retain only the top 1% of attention heads that exhibit the strongest causal influence on reasoning performance, then apply the model merging methods using the selected parameters (fig. 3(c)). Instead of merging all parameters or selected layers as in prior approaches,

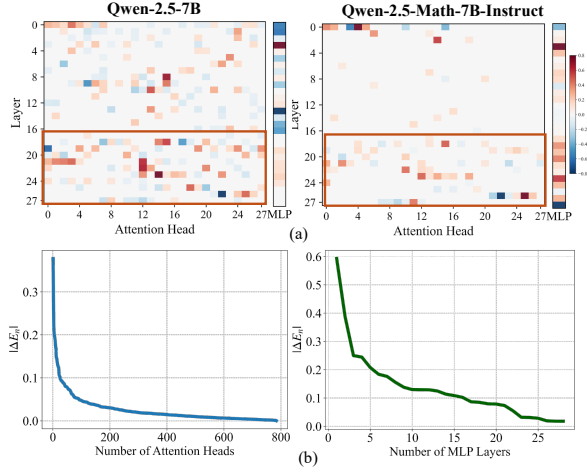


Figure 4: (a) Causal map of base model and its post-trained math model. The math model has a more sparse causal map. For attention heads, the post-training process up-weights middle to later layers of parameters and down-weights the front layers. For MLP layers, the post-training process upweights later layers. (b) Causal shift distribution in attention heads and MLP layers. A small part of the parameter has a high causal shift.

we restrict the merge operation to the sparse set of causally verified attention heads. Experimental results demonstrate that merging on this fine-grained parameter subset consistently improves math reasoning accuracy across benchmarks. These findings validate that the located attention heads capture critical reasoning pathways, and elective merging is more effective than merging all parameters.

3.4 Parameter Selection Strategy

In this section, we analyze which parameters encode enhanced math reasoning capabilities after post-training by comparing the base LLM and its math-specialized variant in section 3.4.1. We observe that a small subset of parameters exhibits significant growth in causal attribution scores, indicating strengthened or newly acquired reasoning functions. Based on this, we propose a parameter selection strategy that prioritizes parameters with the largest causal influence shifts in section 3.4.2.

3.4.1 Which Parameters Encode Enhanced Reasoning Abilities?

We further investigate how reasoning abilities are encoded during post-training, i.e., how the math LLM structurally modifies the parameter space of the base model to enhance the math reasoning abilities. Since the math LLM is a post-trained variant

of the base LLM, we perform attribution patching on both models and compute the causal effect score of every attention head and every MLP block. By comparing the causal contributions before and after training, we can quantify whether the causal effect of each parameter increases, decreases, or remains unchanged. Parameters that show an increase in causal effect are interpreted as the components responsible for the enhanced reasoning ability gained from math post-training.

We present the causal maps of the base LLM (Qwen-2.5-7B) and the post-trained math LLM (Qwen-2.5-Math-7B-Instruct) in fig. 4(a). The causal heatmap of the math LLM becomes significantly more sparse, indicating that math post-training does not uniformly modify the entire model but instead reallocates causal influence to a compact set of reasoning-critical components. Specifically, for attention heads, we observe that post-training consistently up-weights middle to late layers (layers 16–27) while down-weighting early layers. This suggests that math reasoning shifts the burden of computation toward deeper attention layers, where more abstract symbolic representations are formed. For MLP blocks, the effect is even more concentrated: post-training strongly increases causal influence in the final layers, indicating that mathematical reasoning relies on late-stage feature transformation and decision refinement. We further plot the causal shift distribution in fig. 4(b) further confirms this structural specialization. Only a small fraction of attention heads and MLP blocks exhibit large causal shifts, while the majority undergo minimal change. These observations collectively reveal that math LLMs gain reasoning ability not by uniformly increasing capacity, but by specializing a sparse circuit within deeper layers of the model.

3.4.2 Selection Strategy

Based on the above analysis, we propose a parameter selection strategy to identify reasoning-enhanced parameters in the math LLM. Let $\mathcal{M}_{\text{BaseLLM}}$ and $\mathcal{M}_{\text{MathLLM}}$ denote the base and math post-trained models, respectively, where both consist of N computational nodes. From the causal attribution described in section 3.3, each node n is assigned a causal influence score, denoted as E_n^{base} for the base model and E_n^{math} for the math model. To quantify how much reasoning capability each node gains after math post-training, we measure

Table 1: We merge LLaVA 3 with Deepseek-R1-Distilled-LLaMA, Qwen2.5-VL-7B-Instruct with Qwen-2.5-Math-7B-Instruct models, InternVL3-8B-Instruct with DeepSeek-R1-distilled-Qwen-7B, respectively. **Bold** represents the best. Underline represents the second best.

Approach	Math Reasoning				General Reasoning	
	MathVista	MathVerse	DynaMath	MathVision	MMstar	MMMU Pro
<i>LLaMA3-LLaVA-Next-8B as Base Model</i>						
Base Model	37.8	18.4	23.9	14.1	43.0	12.1
Task Arithmetic	35.5	17.6	23.4	13.6	39.3	9.8
Ties Merging	33.1	18.0	23.4	13.3	39.1	8.2
EMR Merging	38.1	19.3	15.4	13.1	42.0	12.3
IP Merging	38.1	18.4	23.8	14.6	42.9	12.4
Ours	38.7	<u>18.6</u>	24.1	14.7	43.4	12.5
<i>InternVL3-8B-Instruct as Base Model</i>						
Base Model	66.1	38.5	50.7	24.9	66.5	27.5
Task Arithmetic	32.0	16.9	20.1	6.2	22.3	9.4
Ties Merging	25.0	15.3	16.1	9.6	25.1	11.5
EMR Merging	49.2	30.1	39.3	18.8	51.3	21.1
IP Merging	67.6	39.0	51.4	25.2	67.3	27.7
Ours	68.4	<u>38.8</u>	52.0	25.6	68.3	28.7
<i>Qwen2.5-VL-7B-Instruct as Base Model</i>						
Base Model	68.2	46.5	51.9	25.1	63.9	34.7
Task Arithmetic	25.6	9.7	14.1	8.2	23.3	6.9
Ties Merging	25.6	3.4	14.1	10.2	14.7	1.3
EMR Merging	23.7	10.7	15.7	22.3	9.9	5.6
IP Merging	69.5	46.2	52.4	26.3	64.8	35.3
Ours	71.4	46.8	53.4	27.1	65.3	35.8

the causal influence change as

$$\Delta E_n = E_n^{\text{math}} - E_n^{\text{base}}. \quad (6)$$

We then rank all nodes in descending order based on $|\Delta E_n|$ and select the top- k nodes with the largest shifts, thereby constructing the selection masks M_1, M_2 in Eq. 2 that define the set of reasoning-critical parameters. During merging, we merge the selected reasoning-enhanced parameters into the MLLM. In this way, our strategy prioritizes parameters that gain the reasoning abilities during post-training, enabling a more effective and interpretable transfer of reasoning abilities while avoiding interference from irrelevant or unchanged parameters.

4 Experiments

4.1 Experiments Setup

We test our models on six benchmarks, i.e., MathVista (Lu et al., 2024), MathVerse (Zhang et al.,

2025b), DynaMath (Zou et al., 2025), MathVision (Wang et al., 2024b) and general QA benchmarks MMMU Pro (Yue et al., 2024), and MMStar (Chen et al., 2024). We compare our method with general model merging techniques such as Task Arithmetic (Ilharco et al., 2022), Ties Merging (Yadav et al., 2023), EMR Merging (Huang et al., 2024) and IP Merging (Hu et al., 2026). We conduct experiments using three model families, i.e., Qwen series, LLaVA series, and InternVL3 series. See more details in section B in the appendix.

4.2 Benchmark Performance

We conduct experiments and report the results on four math reasoning benchmarks and two general reasoning benchmarks using three different types of models in table 1. We use the logical calculation task to identify the reasoning parameters in the base and math LLM (section 4.3), then adopt the selection strategy and merge the selected parameters us-

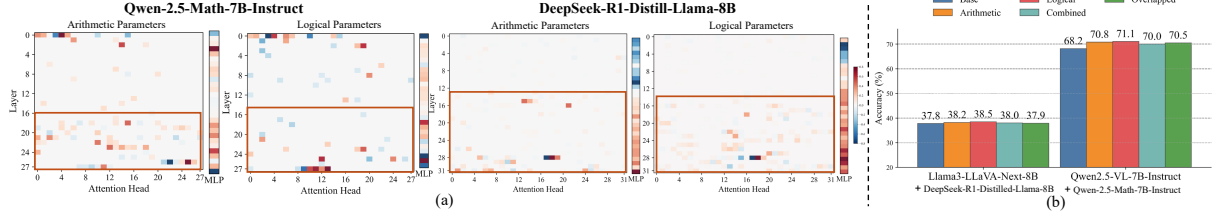


Figure 5: (a) Parameter location map for arithmetical and logical calculations in Qwen2.5-7B-Math-Instruct. (b) We compare the performance of merging different parameters in the math LLM with MLLM using different merging strategies.

ing eq. (3). Across all three base architectures and six benchmarks, our causal-analysis-guided merging consistently outperforms existing approaches while avoiding the severe degradation frequently induced by coarse-grained methods such as Task Arithmetic and Ties Merging. By selectively integrating parameters with the largest causal-influence shifts, our method stabilizes or improves math-reasoning performance. For instance, preserving or slightly enhancing MathVista, MathVerse, and DynaMath, whereas prior approaches often collapse accuracy (e.g., Task Arithmetic reducing MathVista by 30% to 40% on InternVL3-8B and Qwen2.5-VL-7B). Importantly, our strategy also maintains strong general-reasoning performance: unlike Ties Merging and EMR, which markedly degrade MMStar and MMMU-Pro across multiple backbones, our method either matches or surpasses the base models, such as achieving 68.3% on MMStar with InternVL3-8B and 35.8% on MMMU-Pro with Qwen2.5-VL-7B. These results substantiate our core claim that reasoning-critical parameters form a sparse, transferable subset and that aligning these causally identified components enables principled and effective integration of math-reasoning capabilities into MLLMs without compromising their general reasoning performance. See more experimental results in section A.

4.3 Parameter Analysis

We further analyze the parameter circuits identified via arithmetic and logical reasoning tasks. Using Qwen2.5-Math-7B-Instruct and DeepSeek-R1-Distill-Llama-8B, we visualize the spatial distribution of parameters in fig. 5(a). The visualization shows that only a sparse subset of attention heads and MLP blocks contributes significantly to the model’s reasoning capability. This observation aligns with prior findings that reasoning emerges from a compact set of specialized components,

rather than the full parameter space (Wang et al., 2024a; Nikankin et al., 2025). Across both models, reasoning-related attention heads are located in the middle and later layers, whereas important MLP blocks are more concentrated in the later layers. To evaluate whether these parameters transfer crucial reasoning ability, we perform selective parameter merging in fig. 5(b). Specifically, we merge only the top 1% parameters ranked by causal importance for each task (arithmetic vs. logical). We further compare two merging strategies: merging the union of both parameter sets (i.e., “Combined”) and merging only their intersection (i.e., “Overlapped”). As shown in fig. 5(b), merging logical reasoning parameters leads to larger performance gains than merging arithmetic parameters, indicating that logical parameters encode more complex reasoning knowledge and can be better transferred to MLLM for multi-modal math reasoning. Moreover, merging the overlapped subset yields performance comparable to merging the full combined set, despite involving significantly fewer parameters. This result confirms that the reasoning capability of a math-specialized LLM is governed by a minimal, highly informative circuit, and causality-guided parameter merging can enhance reasoning performance while avoiding unnecessary parameter interference.

5 Case Studies

We further provide several user cases in fig. 6 to illustrate the improved reasoning consistency after merging the math LLM. Across numerical, graphical, and pattern-recognition tasks, the merged model corrects multiple systematic errors made by the base MLLM: it identifies the correct local maximum in a function plot, accurately counts values below a threshold, pinpoints the year of the steepest increase in a time-series graph, and infers the missing shape in a visual pattern puzzle. These examples highlight that the merged model



Figure 6: More user cases

not only improves quantitative reasoning but also shows stronger visual and numerical alignment and pattern abstraction. Such gains demonstrate that integrating reasoning-enhanced parameters enables the MLLM to generalize more reliably across heterogeneous multimodal tasks.

6 Conclusion

In this work, we present a causal-analysis-guided parameter merging framework to strengthen multimodal math reasoning without model fine-tuning. Through causal attribution, we find that only a sparse subset of parameters is truly responsible for reasoning. Logical reasoning parameters transfer more effectively than arithmetic ones. Parameters whose causal influence increases after math post-training encode the most valuable reasoning abilities for merging. Guided by these findings, we selectively align and merge only the identified parameters into the MLLM, preserving general multi-modal capabilities while improving reasoning performance. This work demonstrates that fine-grained, interpretable parameter selection is crucial for effective reasoning enhancement and provides a practical, training-free solution for building stronger reasoning-aware MLLMs.

7 Limitation

Our framework primarily addresses models that share the same architecture, which enables direct alignment and merging of parameters. We leave architecture-agnostic parameter mapping and cross-structure model merging as future work.

8 Acknowledgment

This work was supported by National Natural Science Foundation of China under No. 62436009, 62276258, 92370119, 62376113 and Jiangsu Science and Technology Programme BK20251812, and Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems. This work was also supported by Research Development Fund in XJTLU under no. RDF-24-01-111.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang,

- Yu Qiao, Dahua Lin, and Zhao Feng. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favien Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.
- Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.
- Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkan Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9062–9072. IEEE.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2025. G-llava: Solving geometric problem with multi-modal large language model. In *International Conference on Learning Representations*, volume 2025, pages 3490–3511.
- Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. 2025. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *International Conference on Learning Representations*, volume 2024, pages 48362–48395.
- Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and Quanzeng You. 2025. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. In *EMNLP (Findings)*, pages 14221–14231.
- Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.
- Yijie Hu, Zihao Zhou, Kaizhu Huang, Xiaowei Huang, and Qiufeng Wang. 2026. Can mllms absorb math reasoning abilities from llms as free lunch? *Advances in Neural Information Processing Systems*, 38:130912–130940.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.
- Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. 2024. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.
- Derek Lim, Theo Putterman, Robin Walters, Haggai Maron, and Stefanie Jegelka. 2024. The empirical impact of neural parameter symmetries, or lack thereof. *Advances in Neural Information Processing Systems*, 37:28322–28358.
- Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024. Not all tokens are what you need for pretraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, Soheil Feizi, and Lifu Huang. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*.
- Deng Linger, Linghao Zhu, Yuliang Liu, Yu Wang, Qunyi Xie, Jingjing Wu, Gang Zhang, Yingying Zhu,

- and Xiang Bai. 2025. Theorem-validated reverse chain-of-thought problem generation for geometric reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 718–735.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. 2024. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.
- Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. 2025. No task left behind: Isotropic model merging with common and task-specific subspaces. *arXiv preprint arXiv:2502.04959*.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations*, volume 2025, pages 55939–55965.
- Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew Bagdanov, Simone Calderara, and Joost van de Weijer. 2026. Accurate and efficient low-rank model merging in core space. *Advances in Neural Information Processing Systems*, 38:61793–61825.
- Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.
- Haoliang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2026. V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Weiyun Wang, Zhe Chen, Wenhui Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024d. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 1(3).
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024a. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2026. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41.
- Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. 2024b. Representation surgery for multi-task model merging. *arXiv preprint arXiv:2402.02705*.

- Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2024c. Adamerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, volume 2024, pages 22743–22763.
- Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. 2025. Emergent symbolic mechanisms support abstract reasoning in large language models. *arXiv preprint arXiv:2502.20332*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024a. Meta-math: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, volume 2024, pages 45040–45061.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-seng Chua. 2024b. RLhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- 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, Huan Sun, Yu Su, and Wenhui Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1859–1869. IEEE.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2025b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024a. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*.
- Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2024b. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingting Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pages 48337–48383.

A More Experimental Results

A.1 Merging Math MLLM with Math LLM

We further evaluate the effectiveness of our approach by merging two math reasoning MLLMs with their corresponding math LLMs, as reported in table 2. We select two representative models, i.e., R1-VL-7B (Zhang et al., 2025a) and VL-Rethinker-7B (Wang et al., 2026), both of which are enhanced through SFT and RL from a base MLLM to strengthen visual reasoning. For R1-VL-7B, the merged model yields consistent improvements on MathVista, MMStar, and MMMU-Pro, indicating that causally guided parameter selection effectively amplifies its reasoning ability. Likewise, for VL-Rethinker-7B, our merging strategy achieves higher accuracy across all evaluated benchmarks, demonstrating strong generalization across architectures and task formats. These results collectively demonstrate that reasoning-critical parameters identified through causal analysis can be successfully transferred to further enhance mathematical reasoning, while maintaining overall general reasoning ability.

Table 2: We merge Math MLLM with the math LLM model. **Bold** represents the best performance.

Benchmarks	R1-VL-7B		VL-Rethinker-7B	
	Base Model	Ours	Base Model	Ours
MathVista	60.2	61.1	72.7	74.6
MathVerse	37.6	38.5	51.7	52.1
DynaMath	38.9	39.2	55.3	55.5
MathVision	24.1	24.3	28.1	29.3
MMStar	57.8	58.3	63.5	64.7
MMMU Pro	23.2	24.6	36.8	37.8

A.2 Number of Merged Parameters

We compare the number of merging parameters in fig. 7 using Qwen2.5-VL-7B-Instruct as an example. General methods such as Task Arithmetic, Ties Merging, and EMR Merging merge nearly the entire model (6.5B parameters), while IP-Merging still involves 1.8B parameters. In contrast, our method merges only 0.63B parameters as it selectively integrates only the reasoning critical circuits identified by our causal analysis. This compact merge footprint highlights the efficiency and precision of our approach.

A.3 Ablation Studies

We conduct ablations to analyze how the low-rank merging dimension and the percent of causally se-

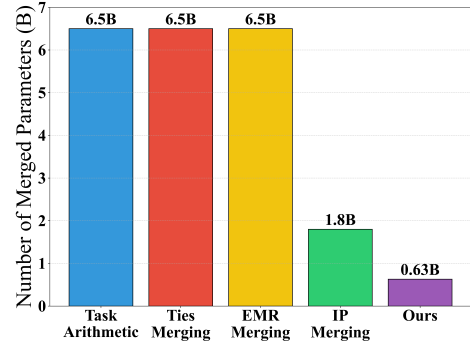


Figure 7: Comparison of the number of merging parameters in different methods.

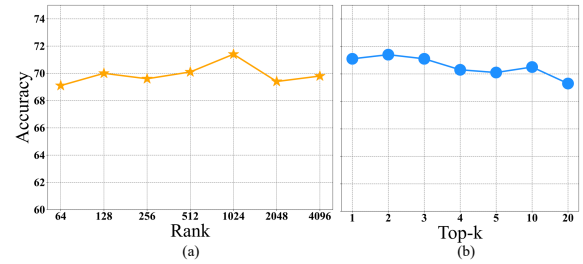


Figure 8: (a) Ablation results of rank. (b) Ablation results of the percentage of selected parameters.

lected parameters affect overall performance using Qwen series models. As shown in fig. 8 (a), the model is relatively stable across a wide range of ranks/ Performance gradually improves from rank 64 to rank 1024, reaching the highest accuracy at rank 1024, suggesting that moderate ranks provide a better balance between information preservation and noise suppression. When the rank becomes too large (e.g., 2048 or 4096), accuracy slightly drops, implying that overly large subspaces may reintroduce redundant or task-irrelevant components. In fig. 8 (b), varying the top-k percentage reveals that using a small subset of high-influence parameters is most effective. Accuracy peaks at 2 percent, after which including more parameters causes a small decline, confirming that reasoning-critical parameters form a sparse parameter subset.

B Detailed Experimental Settings

B.1 Datasets

We test our models on four math reasoning benchmarks, i.e., MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2025b), DynaMath (Zou et al., 2025) and MathVision (Wang et al., 2024b), two general reasoning benchmarks, i.e., MMMU Pro (Yue et al., 2025) and MMStar (Chen et al., 2024). All experiments are conducted using two

Table 3: Results of different hyperparameters of baseline methods in different models.

Methods	LLaMA3-LLaVA-Next-8B				InternVL3-8B-Instruct				Qwen2.5-VL-7B-Instruct			
	MathVista	MathVerse	MathVision	MMMU Pro	MathVista	MathVerse	MathVision	MMMU Pro	MathVista	MathVerse	MathVision	MMMU Pro
Task Vector												
scale=0.1	31.8	17.6	13.6	7.9	25.6	1.7	5.0	0.8	32.0	16.9	15.3	9.4
scale=0.3	35.5	16.0	11.9	9.1	24.7	6.1	8.1	0.5	26.9	9.7	9.2	4.7
scale=0.5	32.8	10.6	11.0	9.8	27.1	9.7	8.2	1.8	23.9	1.5	8.7	3.8
scale=0.7	35.0	7.4	9.8	7.9	23.8	3.8	1.4	6.9	22.3	0.6	9	3.5
scale=0.9	31.4	3.2	9.9	8.5	22.5	1.5	6.9	2.1	21.7	0	6.2	1.9
Ties Merging												
R=0.3, P=0.1	29.6	17.3	12.7	7.1	24.0	1.7	9.3	1.3	25	10.5	8.3	6.5
R=0.3, P=0.3	31.4	16.1	11.9	8.2	23.3	3.4	6.2	0.7	22.9	10.6	8.5	4.6
R=0.3, P=0.5	30.2	14.5	10.8	8.0	25.6	3.3	8.6	0.9	21.5	4.5	7.1	3.9
R=0.3, P=0.7	32.1	13.5	9.6	7.8	22.1	0.7	4.1	0.9	20.7	1.3	7.7	0.8
R=0.3, P=0.9	29.6	9.4	9.3	6.7	24.0	0.2	6.8	0.3	21.8	0.4	7.1	2.1
R=0.2, P=0.1	30.2	17.7	13.3	7.0	23.5	1.7	9.4	1.2	21.9	13.1	8.4	3.4
R=0.2, P=0.3	30.5	17.2	12.6	7.7	24.5	3.0	6.8	1.0	23.5	3.8	7.3	2.9
R=0.1, P=0.5	31.7	16.1	10.2	7.7	23.7	2.3	7.5	0.8	24.6	3.8	9.3	6.2
R=0.2, P=0.7	33.1	14.6	9.9	7.7	21.8	1.6	10.2	0.8	19.9	1.6	5.5	8.2
R=0.2, P=0.9	30.8	11.3	9.6	7.6	23.8	0.3	10.1	1.2	19.9	0	4.6	6.9
R=0.1, P=0.1	28.6	18.0	12.6	7.4	23.4	0.9	1.4	1.3	22.6	15.3	8.9	6.1
R=0.1, P=0.3	29.3	17.2	12.6	7.7	24.2	2.8	6.6	0.3	22.5	14.8	9.6	11.5
R=0.1, P=0.5	31.4	16.4	10.7	7.7	22.5	2.4	7.4	1.2	21.4	2.9	4.7	0.1
R=0.1, P=0.7	32.5	13.6	10.0	7.7	21.9	0.5	6.2	0.9	20.7	0	6.5	0.1
R=0.1, P=0.9	31.5	11.2	9.1	7.8	24.1	0.3	9.4	1.3	19.8	0	1.9	0.1
EMR Merging	38.1	19.3	13.1	12.3	23.7	10.7	9.9	5.6	49.2	30.1	18.8	21.1
IP Merging												
sim=0.1	37.6	18.4	13.8	12.4	69.0	46.2	25.0	35.3	66.1	38.2	23.6	27.5
sim=0.3	36.4	18.1	14.6	12.4	69.3	46.1	26.3	35.3	66.9	38.7	25.7	27.6
sim=0.5	36.2	18.3	14.6	12.3	69.5	46.1	26.2	35.2	67.0	39.0	25.0	27.7
sim=0.7	36.8	18.4	14.2	12.2	69.4	45.8	25.8	34.9	67.6	38.4	24.0	27.5
sim=0.9	38.1	18.4	14.1	12.1	69.1	45.8	26.2	34.7	67.2	38.0	24.1	27.4

RTX 4090 gpus.

- **MathVista** assesses the MLLMs’ multimodal mathematical skills, the testing data can be divided into five subsets: Figure Question Answering (FQA), Geometry Problem Solving (GPS), Math Word Problems (MWP), Textbook Question Answering (TQA), and Visual Question Answering (VQA). For evaluation, following (Lu et al., 2024; Shi et al., 2024), we first employ GPT-4 to extract the predicted choices or answers from responses, then report the answer accuracy, which determines whether the final answer matches the ground truth.
- **MathVerse** includes a diverse set of math problems that require understanding and reasoning over both textual and visual information, such as charts, diagrams, and equations. The testing data can be divided into five subsets, i.e., Text Dominant, Text Lite, Vision Intensive, Vision Dominant and Vision Only.
- **DynaMath** is a dynamic visual mathematics benchmark developed for comprehensive evaluation of MLLM. It consists of 501 carefully curated seed problems, each implemented as a Python program, enabling the assessment of a model’s generalization capability by measuring its robustness across multiple input variations derived from the same underlying seed question.
- **MathVision** is a carefully constructed dataset comprising 3,040 high-quality visual mathematics problems collected from real-world math competitions. Covering 16 mathematical domains and organized into 5 difficulty levels, it offers a broad and balanced benchmark for assessing the visual mathematical reasoning capabilities of MLLMs.
- **MMMU Pro** is an upgraded version of the MMMU benchmark that provides a more rigorous evaluation of advanced AI models across multiple modalities. Compared with the original MMMU, it introduces several enhancements that make the tasks more challenging and realistic, allowing us to better measure a model’s true ability to understand and integrate visual and textual information.
- **MMStar** is a carefully curated multimodal benchmark with 1,500 vision-dependent challenge samples designed to test six core capabilities and eighteen finer axes of LVLMs. Samples are selected through an automated pipeline and strict human review to ensure

true visual dependency, minimal data leakage, and the need for advanced multimodal reasoning.

B.2 Details of Hyperparameter Selection

Following previous works in model merging (Yadav et al., 2023; Huang et al., 2024), we adopt the grid search for hyperparameters for the baseline methods, specifically, we set the hyperparameters based on the following range:

- **Task Arithmetic (Ilharco et al., 2022)** involves the scaling coefficients for merged task vectors, which are set ranging from [0.1, 0.3, 0.5, 0.7, 0.9].
- **Ties Merging (Yadav et al., 2023)** involves the scaling coefficient and ratio to retain large parameters, the scaling coefficients are set ranging from [0.1, 0.3, 0.5, 0.7, 0.8, 0.9], ratio to retain parameters with largest-magnitude values: [0.1, 0.2, 0.3].
- **EMR Merging (Huang et al., 2024)** does not involve specific hyperparameters.
- **IP Merging** involves the similarity threshold to determine whether the layer should be selected for merging, the similarity thresholds are set ranging from [0.1, 0.3, 0.5, 0.7, 0.8, 0.9].

We show the results of different hyperparameters of baseline methods in table 3. For Task Arithmetic, the scale in the table represents the scaling factor. For Ties Merging, “R” represents the retain rate and “P” is the rescaling factor. EMR merging does not involve hyperparameters. The “sim” in IP merging represents the similarity threshold.

B.3 Models and Comparison Methods

We employ the LLaVA series (Liu et al., 2023), Qwen 2.5 series (Wang et al., 2024c) and InternVL3 (Zhu et al., 2025) series as our base MLLM. We use other fine-tuned Math LLMs such as Deepseek-R1-distilled-Llama3 model, Qwen2.5-Math models (Yang et al., 2024a) and Deepseek-R1-distilled-Qwen model. The pretrained foundation of Qwen series models is the Qwen-2.5 model (Wang et al., 2024c). The InternVL model is trained based on the Qwen2.5 models. We use RTX 4090 GPUs for all of our experiments. We compare our proposed methods with prevailing model merging techniques:

- **Base Model** The performance of base MLLM on three tasks, we reproduce the results with the officially released code.
- **Task Arithmetic (Ilharco et al., 2022)** combines all the task vectors extracted from the models into one multi-task model.
- **Ties Merging (Yadav et al., 2023)** addresses task inference by pruning redundant parameters. The process involves three steps: Trim, Elect Sign, and Disjoint Merge.
- **EMR Merging (Huang et al., 2024)** computes one unified task vector and computes task-specific masks based on a unified task vector. The final task vector is computed by combining all the masked task vectors and weighting all the task vectors by rescaling parameters.
- **IP Merging (Hu et al., 2026)** uses svd based approach to first select layers that are crucial for model merging, then use rescaling and projection to merge the model parameters selectively.

We list the models used in the experiments in table 4.

Table 4: List of MLLMs and LLMs.

Models	Type	Pretrained Base Model	Source LLM
LLaVA-Next-Llama3-8B	MLLM	Llama-3-8B	Llama-3-8B
InternVL3-8B-Instruct	MLLM	Qwen2.5-7B	Qwen2.5-7B
VL-Rethinker	MLLM	Qwen2.5-VL-7B-Instruct	Qwen2.5-7B
R1-VL-7B	MLLM	Qwen2-VL-7B-Instruct	Qwen2-7B
Qwen2.5-VL-7B-Instruct	MLLM	Qwen2.5-7B	Qwen2.5-7B
Qwen2.5-Math-7B-Instruct	LLM	Qwen2.5-7B	Qwen2.5-7B
DeepSeek-R1-distilled-Qwen-7B	LLM	Qwen2.5-Math-7B	Qwen2.5-7B
DeepSeek-R1-distilled-Llama-3-8B	LLM	Llama-3-8B	Llama-3-8B