DRIFT: Transferring Reasoning Priors for Efficient MLLM Fine-Tuning

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have made rapid progress, yet their reasoning ability often lags behind strong text-only LLMs. Bridging this gap typically requires large-scale multimodal reasoning data or reinforcement learning, incurring substantial cost. An appealing alternative is parameter-space model merging between reasoning-enhanced LLMs and MLLMs, but we show that naive merging is fragile: its effectiveness varies widely across model families and can significantly degrade performance (e.g., for Qwen-based MLLMs). We propose **Directional Reasoning Injection for Fine-Tuning (DRIFT)**, a lightweight method that transfers reasoning knowledge in the gradient space while preserving multimodal alignment. DRIFT precomputes a reasoning prior from the parameter differences between text-only reasoning experts and multimodal models, and uses it to bias gradients during supervised fine-tuning. This design retains the simplicity of standard SFT pipelines while enabling efficient and stable reasoning transfer. Experiments on multimodal reasoning benchmarks, including MathVista and MathVerse, show that DRIFT consistently outperforms naive merging and standard SFT, and matches or surpasses training-intensive methods with substantially lower data and compute.

1 Introduction

Multimodal large language models (MLLMs) (Bai et al., 2025; Team et al., 2023; Li et al., 2024b) have recently achieved impressive progress in perception and alignment, enabling them to answer questions about images, analyze charts, and engage in grounded dialogue. However, despite these advances, their reasoning ability remains substantially weaker than that of text-only large language models (LLMs). Across benchmarks in mathematical reasoning (Pan Lu et al., 2024), logical inference (Xiao et al., 2024), and multi-hop question answering (Xiang Yue et al., 2025), a persistent gap emerges: MLLMs can perceive correctly but struggle to chain information into coherent reasoning steps. Bridging this gap is essential for applications that demand not only multimodal understanding but also structured, reliable reasoning.

A mainstream approach to improving reasoning in MLLMs is multimodal supervised fine-tuning (SFT) or reinforcement learning (RL) on reasoning-intensive datasets. Yet both are resource-heavy: collecting multimodal CoT-style data is costly, and reinforcement learning adds instability and computational overhead. In contrast, text-only reasoning models (DeepSeek-AI, 2025) are far easier to obtain due to the growing availability of large-scale text-only CoT data. This naturally raises a research question: *Can we transfer reasoning capability from text-only experts into MLLMs efficiently?*

A promising direction is parameter-space model merging, where the weights of a reasoning model are interpolated with those of an MLLM (Chen et al., 2025a). While exciting in its simplicity, our experiments reveal that naive merging is fragile (as shown in Sec. 3.2). It often disrupts perception and alignment, and in many cases even reduces reasoning performance. Learning merge coefficients during fine-tuning partly alleviates this issue, but at the cost of huge training overhead and instability.

To address these limitations, we propose DRIFT, *Directional Reasoning Injection for Fine-Tuning*, a lightweight gradient-based method that transfers reasoning knowledge without destabilizing multimodal training. Rather than interpolating weights in parameter space, DRIFT operates in gradient space: it computes a **reasoning vector**, defined as the parameter difference between a reasoning-rich text model and its multimodal counterpart, and uses this as a directional prior to guide updates during multimodal SFT. By injecting this guidance selectively into transformer modules (e.g., atten-

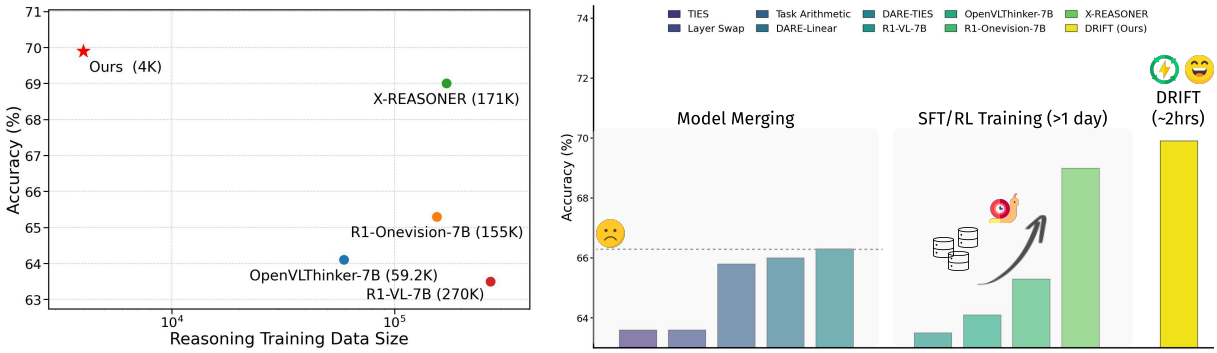

Figure 1: **DRIFT enables efficient reasoning transfer for MLLMs.** *Left:* Compared with reasoning-oriented training methods, DRIFT achieves comparable performance while requiring dramatically less multimodal SFT data (4K vs. >59K examples). *Right:* Simple parameter merging performs poorly on MathVista (Pan Lu et al., 2024). Training-based methods improve performance but rely on costly data curation and multi-day training. In contrast, DRIFT reaches competitive results within ∼2 hours of training, making it **both data- and compute-efficient**.

tion projections or MLP layers), DRIFT biases optimization toward reasoning while preserving perception. Essentially, DRIFT introduces no additional parameters, requires only a small amount of multimodal reasoning data (as shown in Fig. 1), and integrates seamlessly into existing fine-tuning pipelines. Our contributions are summarized as follows: (1) We show that parameter-space model merging for reasoning injection is brittle and often degrades performance when models are misaligned. (2) We introduce DRIFT, a lightweight gradient-based method that injects reasoning by using text-only expert–to–multimodal parameter differences as a directional prior during SFT. (3) Experiments on multimodal reasoning benchmarks show that DRIFT consistently outperforms SFT and merging baselines, matching training-heavy methods with far less data and compute.

2 Related Work

2.1 Multimodal Reasoning in Large Language Models

Recent work has investigated whether the strong reasoning abilities of text-only LLMs can be transferred to multimodal large language models (MLLMs). Mathematical and logical reasoning have become standard evaluation settings. Benchmarks such as MathVista (Lu et al., 2023), LogicVista (Xiao et al., 2024), MathVision (Wang et al., 2024a), MathVerse (Renrui Zhang et al., 2024), and WeMath (Qiao et al., 2024) assess MLLMs on visually grounded reasoning tasks of varying complexity. To improve multimodal reasoning, prior work has explored instruction tuning and supervised fine-tuning (SFT) (Ratzlaff et al., 2025; Li et al., 2024d; Ranaldi and Freitas, 2024; Subramaniam et al., 2025; Huang et al., 2024b; Dong et al., 2025), as well as reinforcement learning (RL) (Wan et al., 2025; Liu et al., 2025b; Chen et al., 2025b). While SFT is relatively lightweight, its effectiveness depends on large-scale, high-quality multimodal data. RL methods can yield stronger and more robust improvements, but at substantially higher computational cost.

2.2 Efficient Fine-Tuning of LLMs

To mitigate the cost of full fine-tuning, prior work has proposed parameter-efficient and data-efficient approaches.

Parameter-Efficient Fine-Tuning. Methods such as LoRA (Hu et al., 2022) and its variants (Dettmers et al., 2023; Hayou et al., 2024; Pan et al., 2024) reduce trainable parameters via low-rank updates. Adapter-based approaches insert lightweight trainable modules while freezing backbone weights, including LLaMA-Adapter (Gao et al., 2023; Zhang et al., 2024b), AdaptM-LLM (Lankford et al., 2023), and Bt-Adapter (Liu et al., 2024).

Data-Efficient Fine-Tuning. Another line of work improves efficiency by reducing training data or computation, through data selection (Lin et al., 2024; He et al., 2024) or visual token compression (Shang et al., 2024; Cai et al., 2024).

Model Merging. Model merging combines independently trained models via parameter arithmetic without additional training (Ilharco et al.; Yadav et al., 2023; Yu et al., 2024). While effective in

vision models (Huang et al., 2024a; Gargiulo et al., 2025), its application to MLLMs remains limited. Recent work such as BR2V (Chen et al., 2025a) explores transferring reasoning via merging, but faces challenges due to parameter misalignment and cross-modal interference. Our work addresses these limitations by injecting reasoning priors from LLMs into MLLMs in the gradient space.

The complete list of related work is provided in the appendix.

3 Method

3.1 Task Formulation

Starting from a text-only base LLM ϕ, one can derive multiple variants such as instruction-tuned models or task-specific experts for domains like mathematics, programming, or chemistry. Reasoning can be injected into this base model through two primary approaches: (i) supervised fine-tuning (SFT) on chain-of-thought (CoT) datasets, or (ii) reinforcement learning (RL), incentivizing step-by-step reasoning behavior without explicit CoT labels. To equip the model with visual understanding, a standard strategy is to integrate a visual encoder that maps images into token representations processed jointly with text, then train the encoder and LLM backbone end-to-end.

Despite sharing the same base, reasoning and vision capabilities are often developed in isolation: multimodal large language models rarely inherit the reasoning ability of their text-only counterparts. Building an MLLM capable of reasoning typically requires SFT over costly multimodal CoT data. RL can further refine reasoning, but usually assumes a seed of reasoning ability or sufficient long-context capacity. In contrast, the growing availability of text-only CoT resources makes it often easier to first obtain a strong text-only reasoning model from ϕ. This imbalance naturally motivates our research question ($\mathcal{Q}$): *can we leverage a text-only reasoning model to guide the transformation of a non-reasoning multimodal LLM into a reasoning-capable one?*

Formally, let the base model be ϕ and its variant fine-tuned on a task T_i be denoted ϕ_{T_i}. Our objective is to efficiently learn a model $\phi_{T'}$ by leveraging M domain experts $\{\phi_{T_1}, \phi_{T_2}, \ldots, \phi_{T_M}\}$, where $T' = \{T_1, T_2, \ldots, T_M\}$. In this work, we focus on the case where $T_1 =$ text-only reasoning and $T_2 =$ visual understanding, and aim to combine them in a data- and compute-efficient manner to obtain a reasoning-capable multimodal model.

3.2 Is Model Merging Always a "Free Lunch"?

Model merging, which combines the weights of domain experts so that the resulting model inherits desirable properties from each, appears to offer a promising path toward addressing our research question. In particular, one can merge a text-only reasoning LLM with the backbone of a multimodal LLM (MLLM) to unify their complementary strengths. Recent work, such as BR2V (Chen et al., 2025a), has explored this direction by attempting to integrate reasoning into multimodal LLM.

To explore the potential of model merging, we apply BR2V to the LLM backbones of a text-only reasoning model and a multimodal LLM, both derived from the same base model. We explore a series of models. Concretely, we experiment with `Mistral-7B` (Jiang et al., 2023), `LLaMA3-8B`, `Qwen-2-7B` (Yang et al., 2024), and `Qwen-2.5-7B` (Bai et al., 2025) as base models; `Dart-Uniform` (Tong et al., 2024), `Meta-Math` (Yu et al., 2023), `Qwen2-Math-7B` (Yang et al., 2024), and `DeepSeek-R1-Distill-Qwen-7B` (DeepSeek-AI, 2025) as text-only reasoning experts; and `LLaVA-Next-LLaMA3-8B` (Li et al., 2024a), `Idefics-8B` (Laurençon et al., 2024), `Qwen2-VL-7B-Instruct` (Wang et al., 2024b), and `Qwen-2.5-VL-7B-Instruct` (Bai et al., 2025) as multimodal variants.

We evaluate the merged models on multimodal reasoning benchmarks, including MathVista (Pan Lu et al., 2024), MathVision (Ke Wang et al., 2024), and MathVerse (Renrui Zhang et al., 2024) Vision-Only subset (see Tab. 1). While BR2V enhances the reasoning ability of `LLaVA-Next` and `Idefics`, yielding up to a 2% improvement when merged with reasoning-augmented variants, it often causes performance degradation in the `Qwen` series.

To further investigate these mismatched behaviors across different models, we compute layer-wise $\mathcal{L}_2$ norm and cosine similarity between model backbones, quantifying both magnitude and directional shifts in parameter space. This analysis enables us to examine how reasoning and visual understanding are distributed in parameter space, thereby characterizing the relationships between post-trained variants derived from the same base LLM. As shown in Fig. 2, variants of `LLaMA` and `Mistral` remain relatively close in parameter

Table 1: **Effect of model merging on multimodal reasoning benchmarks.** Performance is reported on MathVista (Pan Lu et al., 2024), MathVision (Ke Wang et al., 2024), and MathVerse (Renrui Zhang et al., 2024) for four multimodal LLMs (LLaVA-Next-8B (Li et al., 2024a), Idefics-8B (Laurençon et al., 2024), Qwen2-VL-7B (Wang et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025)) before and after merging with their corresponding text-only reasoning experts. Scores are shown with relative improvements ($rel.$) over the base model.

Benchmark	LLaVA-Next-LLaMA3-8B			Idefics-8B			Qwen2-VL-7B			Qwen2.5-VL-7B		
	Base	+Dart-Uniform	$rel.$	Base	+MetaMath	$rel.$	Base	+Qwen2-Math	$rel.$	Base	+DeepSeek-R1	$rel.$
MathVista	37.4	38.2	**+0.8**	51.8	53.2	**+1.4**	61.2	60.2	**-1.0**	67.9	65.8	**-2.1**
MathVision	13.8	15.8	**+2.0**	17.1	11.8	**-5.3**	21.1	21.7	**+0.6**	25.0	22.7	**-2.3**
MathVerse	16.0	17.4	**+1.4**	11.0	12.4	**+1.4**	26.9	26.7	**-0.2**	41.4	33.2	**-8.2**

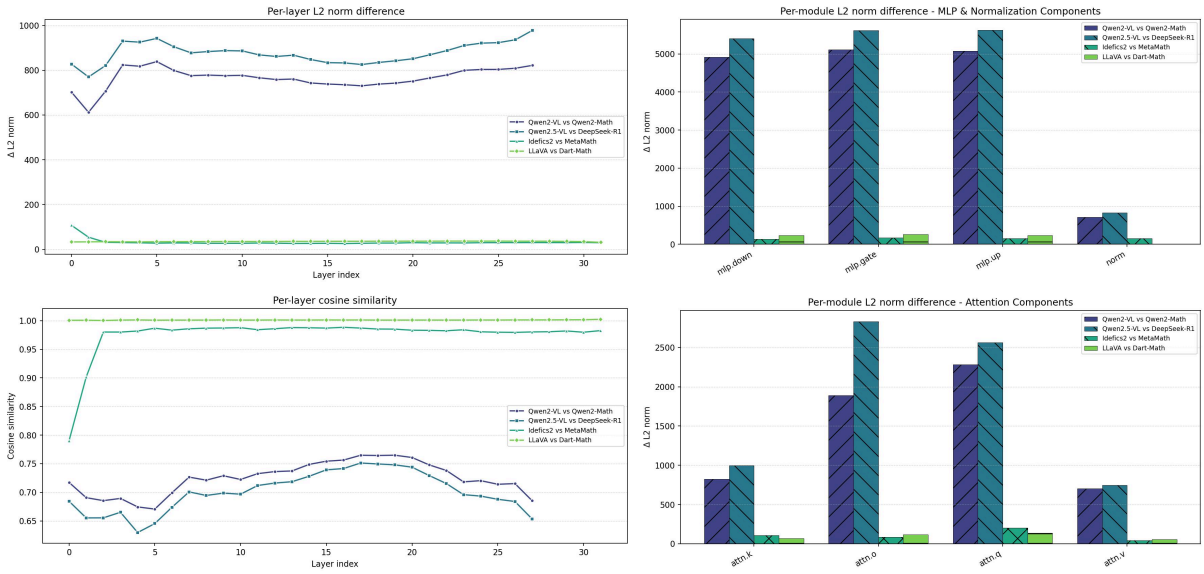

Figure 2: **Layer/Module-wise analysis of model merging pairs.** We compare LLaVA-Next-8B vs. Dart-Uniform, Idefics-8B vs. MetaMath, Qwen2-VL-7B vs. Qwen2-Math-7B, and Qwen2.5-VL-7B vs. DeepSeek-R1-Qwen-7B. *Top Left*: per-layer $\mathcal{L}_2$ norm differences (**magnitude**). *Bottom Left*: per-layer cosine similarity (**direction**). *Top Right*: average $\mathcal{L}_2$ norm differences for FFN layers and normalization layers. *Bottom Right*: average $\mathcal{L}_2$ norm differences for attention projections (Q/K/V/O).

space, while Qwen variants are substantially more dispersed. Moreover, the parameter magnitudes of multimodal Qwen models diverge sharply from their reasoning counterparts, which likely explains the failure of naive merging in this family. These results suggest that model merging is not universally a "free lunch", its success depends strongly on how post-training reshapes the underlying parameter space.

3.3 Directional Reasoning Injection for Fine-Tuning MLLMs

We reformulate the task as mapping a reasoning expert ϕ_{reason} and a multimodal LLM ϕ_{VL} into a reasoning-capable multimodal model:

$$(\phi_{\text{VL}}, \phi_{\text{reason}}) \;\mapsto\; \phi_{\text{VL}\oplus\text{reason}}.$$

As demonstrated in Sec. 3.2, typical merging methods like BR2V (Chen et al., 2025a) merge parameters (task vectors) relative to the base model:

$$\begin{aligned}\phi_{\text{VL}\oplus\text{reason}} = \phi_{\text{base}} &+ \beta\big(\phi_{\text{VL}} - \phi_{\text{base}}\big) \\ &+ (1-\beta)\big(\phi_{\text{reason}} - \phi_{\text{base}}\big).\end{aligned} \tag{1}$$

However, this approach often fails in practice. Large discrepancies between ϕ_{VL} and ϕ_{reason} make performance highly sensitive to β: even small distributional mismatches can yield large shifts in weights. Learning an optimal β is expensive because it requires storing all candidate models in GPU memory. Moreover, when the two models diverge heavily in magnitude, naive interpolation can cause unstable updates or gradient explosions. These drawbacks suggest that parameter-space merging is neither stable nor efficient for large-scale MLLMs.

From parameter merging to directional injection. Instead of interpolating parameters, we propose to inject reasoning knowledge into the *opti-*

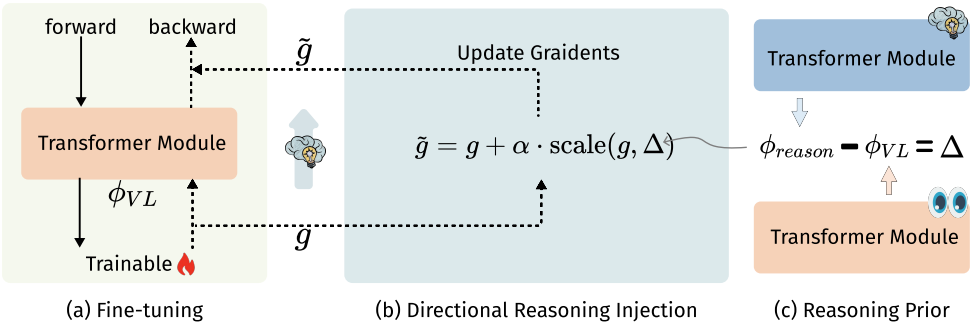

Figure 3: **Overview of Directional Reasoning Injection (DRIFT).** (a) Standard fine-tuning of a multimodal LLM ϕ_{VL}, where gradients g are applied directly to update trainable modules. (b) DRIFT modifies gradients by injecting a reasoning prior: $\tilde{g} = g + \alpha \cdot \text{scale}(g, \Delta)$, where Δ encodes the reasoning direction and scale$(\cdot)$ adjusts how Δ interacts with g. (c) The reasoning prior Δ is constructed as the parameter difference between a text-only reasoning model ϕ_{reason} and the multimodal variant ϕ_{VL}. Our method enables reasoning knowledge to be transferred without destabilizing parameter-space merging.

mization trajectory. Our key insight is that the gap between variants encodes domain-specific knowledge (e.g., reasoning). Rather than directly applying this gap in weight space, which may distort multimodal alignment, we leverage it as a *directional prior* that guides gradient updates.

We define the difference between a reasoning model and a multimodal variant:

$$\Delta = \phi_{\text{reason}} - \phi_{\text{VL}}, \tag{2}$$

restricted to reasoning-relevant modules (MLP projections, attention projection layers, and normalization layers). This Δ serves as the *reasoning direction*. During multimodal supervised fine-tuning (SFT) with limited multimodal CoT data, we leave model weights intact and instead bias gradients towards the reasoning direction. For a parameter w with gradient g, we compute the guided gradient:

$$\tilde{g} = g + \alpha \cdot \text{scale}(g, \Delta), \tag{3}$$

where α controls prior strength and scale$(\cdot)$ adjusts how Δ interacts with g. We explore three variants:

- **Absolute:** $\tilde{g} = g + \alpha\Delta$, directly pulling weights toward the reasoning prior.
- **Grad-Norm:** $\tilde{g} = g + \alpha\|g\|\frac{\Delta}{\|\Delta\|}$, aligning updates with the direction of Δ while preserving the gradient magnitude of g.
- **Grad-Norm w/ Adaptive α:** $\tilde{g} = g + \alpha'\|g\|\frac{\Delta}{\|\Delta\|}$, where $\alpha' = \alpha \cdot \frac{1+\cos(g,\Delta)}{2}$, adapting strength based on gradient-delta alignment.

Discussion. The proposed *Directional Reasoning Injection* (DRIFT) offers two main benefits. First, it preserves the standard multimodal SFT pipeline: training remains on multimodal data, but optimization is nudged toward reasoning directions, enabling gradual knowledge transfer without destabilizing pre-merge operations or requiring large-scale multimodal CoT supervision. Second, it is lightweight: the reasoning prior Δ is computed once, stored on the CPU, and only transferred to the GPU when needed for gradient updates. DRIFT introduces no additional parameters and modifies only the backward pass, making it both memory-efficient and easily scalable to large MLLMs.

4 Experiments

4.1 Dataset Collection

Reasoning transfer in DRIFT requires only a small amount of high-quality multimodal reasoning data. Prior work (ThinkLite (Wang et al., 2025)) shows that challenging questions are more effective than large volumes of easy ones. Building on this insight, we start from the ThinkLiteVL-11K dataset, which contains 11K high-quality image–question pairs. However, this dataset provides only answers without accompanying reasoning chains. To address this, we use ThinkLite models to distill chain-of-thought (CoT) annotations and filter out examples with incorrect answers or invalid outputs. The retained reasoning traces are enclosed within `<think></think>` tags to clearly separate the CoT from the final answer. After filtering, we obtain a curated set of 4K high-quality multimodal reasoning examples, which serve as the foundation for our proposed DRIFT.

Table 2: **Evaluation results on multimodal reasoning benchmarks.** We compare our gradient-based merging approach with standard parameter-space merging baselines. Results are reported on *MathVista*, *MathVision*, *MathVerse*, *WeMath* (strict/loose), and *LogicVista*. Best results are in **bold**. Note: Improvements are reported relative to Baseline.

Model	MathVista	MathVision	MathVerse	WeMath		LogicVista	Avg.
				strict	loose		
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	67.9	25.0	41.4	34.3	52.8	46.7	44.7
Parameter merging with DeepSeekR1-Qwen-Distill-7B							
Task Arithmetic (Ilharco et al.)	65.8$_{-2.1}$	22.7$_{-2.3}$	33.2$_{-8.2}$	30.1$_{-4.2}$	51.2$_{-1.6}$	42.0$_{-4.7}$	40.8$_{-3.9}$
Layer Swap (Bandarkar et al.)	63.6$_{-4.3}$	22.9$_{-2.1}$	37.9$_{-3.5}$	32.1$_{-2.2}$	50.1$_{-2.7}$	35.1$_{-11.6}$	40.3$_{-4.4}$
TIES (Yadav et al., 2023)	63.6$_{-4.3}$	23.1$_{-1.9}$	39.5$_{-1.9}$	33.4$_{-0.9}$	51.7$_{-1.1}$	42.1$_{-4.6}$	42.2$_{-2.5}$
DARE-TIES (Yu et al., 2024)	66.3$_{-1.6}$	23.6$_{-1.4}$	38.3$_{-3.1}$	33.7$_{-0.6}$	52.6$_{-0.2}$	42.0$_{-4.7}$	42.8$_{-1.9}$
DARE-Linear (Yu et al., 2024)	66.0$_{-1.9}$	22.3$_{-2.7}$	35.5$_{-5.9}$	30.8$_{-3.5}$	51.2$_{-1.6}$	42.5$_{-4.2}$	41.4$_{-3.3}$
Reasoning Injection from DeepSeekR1-Qwen-Distill-7B							
DRIFT (Ours)	**69.9**$_{+2.0}$	**26.6**$_{+1.6}$	**43.9**$_{+2.5}$	**38.5**$_{+4.2}$	**60.2**$_{+7.4}$	**47.2**$_{+0.5}$	**47.7**$_{+3.0}$

4.2 Experimental Setting

In particular, to construct a strong multimodal reasoning model, we select `DeepSeek-R1-Qwen-Distill-7B` (DeepSeek-AI, 2025) as the text-only reasoning expert and `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025) as the multimodal backbone. The `DeepSeek-R1` family is designed to elicit explicit reasoning traces, while `Qwen2.5-VL` provides strong visual grounding and perception. Investigating whether combining these complementary capabilities yields a more powerful multimodal reasoning model is our central question.

We implement our method on top of the `LLaMAFactory` codebase (Zheng et al., 2024), ensuring reproducibility and compatibility with existing fine-tuning workflows. Training follows the standard supervised fine-tuning pipeline, with DRIFT integrated as a lightweight plugin. The reasoning direction Δ is precomputed once and cached on the CPU, then transferred to the GPU only when needed for gradient updates. During backpropagation, we register additional gradient hooks that inject Δ into online gradients, enabling reasoning-aware optimization with negligible overhead. We train the model for three epochs with a learning rate of 1×10^{-6}.

For evaluation, we focus on multimodal reasoning benchmarks, particularly those involving mathematical reasoning: MathVista (Pan Lu et al., 2024) testmini subset, MathVision (Ke Wang et al., 2024), MathVerse (Renrui Zhang et al., 2024) vision-only subset, WeMath (Runqi Qiao et al., 2024), and LogicVista (Xiao et al., 2024). These datasets contain not only general visual question answering tasks but also problems that explicitly require reasoning, making them suitable testbeds for our approach. We adopt `VLMEvalKit` (Duan et al., 2024) for standardized evaluation and to minimize randomness, following the official protocols of each benchmark.

4.3 Comparison with Parameter Merging-based Methods

As discussed in Sec. 3.2, parameter-space merging has emerged as a popular approach for injecting reasoning into multimodal models. However, its effectiveness is far from guaranteed: naive merging often yields no gain, particularly when the underlying models diverge significantly in parameter space. We compare against several representative merging approaches, including Task Arithmetic (Ilharco et al.), Layer Swap (Bandarkar et al.), TIES (Yadav et al., 2023), and DARE (Yu et al., 2024). These methods operate by directly manipulating model weights via vector addition or interpolation, layer replacement, or sparsity/importance masking, to combine complementary skills without full retraining. We follow the hyperparameter selection practice of Chen et al. (2025a) for fair comparison.

As shown in Tab. 2, we merge the strong reasoning model DeepSeek-R1-Qwen-Distill-7B (DeepSeek-AI, 2025) into Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Surprisingly, none of the merging methods improve performance; in fact, several degrade it. We hypothesize that this failure stems from the large distributional discrepancy between the reasoning model and the multimodal variant, consistent with our earlier analysis in Sec. 3.2. This finding underscores the fragility of parameter-level merging and motivates the need for a more robust alternative.

Our Gradient-based Alternative. In contrast,

Table 3: **Evaluation results on visual reasoning benchmarks.** We report performance on MathVista, MathVision, MathVerse, WeMath (strict), and LogicVista across *open-source models*, and *reasoning fine-tuning methods*. † indicates results reproduced by ourselves, while others are reported by Open Vision Reasoner (Wei et al., 2025). Our DRIFT results are bold, and improvements relative to our SFT baseline are reported.

Model	MathVista	MathVision	MathVerse	WeMath	LogicVista
Open-source Models					
LLaVA-OneVision-7B (Li et al., 2024c)	62.6	17.6	17.6	17.7	32.0
InternLM-XComposer2.5 (Zhang et al., 2024a)	64.0	17.8	16.2	14.1	34.7
InternVL3-8B (Zhu et al., 2025)	70.5	28.6	33.9	37.5	43.6
InternVL2.5-8B (Chen et al., 2024a)	64.5	17.0	22.8	23.5	36.0
InternVL2-8B (Chen et al., 2024b)	58.3	20.0	20.4	20.2	33.6
QvQ-72B-Preview (Team, 2024)	70.3	34.9	48.2	39.0	58.2
Kimi-VL-16B (Team et al., 2025)	66.0	21.8	34.1	32.3	42.7
Qwen2-VL-7B (Wang et al., 2024b)	61.6	19.2	25.4	22.3	33.3
Qwen2.5-VL-7B (Bai et al., 2025)	67.9†	25.0†	41.4†	34.3†	46.7†
Reasoning Fine-tuning Methods					
R1-Onevision-7B (Yang et al., 2025)	64.1	29.9	40.0	–	61.8
OpenVLThinker-7B (Deng et al., 2025)	65.3	23.0	38.1	35.2	44.5
R1-VL-7B (Zhang et al., 2025)	63.5	24.7	40.0	–	–
X-REASONER (Liu et al., 2025a)	69.0	29.6	–	–	–
Ours (SFT)	68.7	25.1	42.0	33.3	45.6
DRIFT (Ours)	**69.9**$_{+1.2}$	**26.6**$_{+1.6}$	**43.9**$_{+1.9}$	**38.5**$_{+5.2}$	**47.2**$_{+1.6}$

DRIFT sidesteps the instability of direct parameter interpolation by explicitly encoding reasoning directions during supervised fine-tuning. The multimodal model begins with full vision–language capability inherited from the base, and fine-tuning data naturally couples perception and reasoning. DRIFT leverages this setting by nudging gradients slightly toward the reasoning direction, reinforcing reasoning signals without disrupting multimodal alignment. This design yields consistent improvements across benchmarks, surpassing both the baseline and parameter-merging methods (e.g., $+4.5$ points on MathVista compared to Task Arithmetic). These results highlight that DRIFT provides an effective mechanism for transferring reasoning ability, offering robustness where parameter-level merging is brittle.

4.4 Comparison with Training-based Methods

A prominent line of work aims to endow multimodal LLMs with reasoning ability through additional training, typically requiring either large-scale multimodal CoT supervision or specialized fine-tuning strategies such as reinforcement learning. Representative examples include R1-OneVision (Yang et al., 2025), OpenVLThinker (Deng et al., 2025), and X-Reasoner (Liu et al., 2025a), all of which demand curated multimodal reasoning datasets and substantial training budgets. As shown in Tab. 3, these approaches achieve competitive performance, but only at the cost of generating or collecting large-scale CoT traces (see Fig. 1 for performance and dataset size comparison).

In contrast, our method avoids such heavy supervision. By introducing *Directional Reasoning Injection*, we leverage a lightweight reasoning prior distilled from a text-only expert and inject it into multimodal training via gradient guidance. This design preserves the simplicity of standard SFT pipelines while enabling efficient reasoning transfer. Empirically, DRIFT achieves consistent gains over the SFT baseline on all the benchmarks. Although training-heavy methods such as X-Reasoner or R1-OneVision sometimes achieve higher absolute scores, DRIFT reaches competitive performance with orders of magnitude less reasoning-specific data and training time. The efficiency benefits of DRIFT are: existing reasoning-focused methods require **days of training** with SFT or RL, while DRIFT requires only SFT-style training and completes in **roughly two hours**.

Overall, these results, together with the efficiency analysis, validate our central claim: reasoning transfer can be achieved not only through resource-intensive multimodal fine-tuning, but also via lightweight gradient-space priors that exploit the gap between text-only reasoning experts and multimodal models.

4.5 Analysis of DRIFT

Is Reasoning Prior Useful? Tab. 3 shows that while supervised fine-tuning (SFT) is a strong baseline, incorporating the reasoning prior via DRIFT

Table 4: **Comparison of scaling strategies in DRIFT.** We report performance on *MathVista*, *MathVerse*, and *LogicVista*. Scores are shown with relative improvements ($rel.$) over the SFT baseline. Merging candidates include attention projection layers (ATTN), Feedforward layers (MLP), input normalization and output normalization layers (Norm), and the output language model projection head (LM Head).

	Scaling Strategy	Merge Candidates	MathVista		MathVerse		LogicVista	
			Score	$rel.$	Score	$rel.$	Score	$rel.$
SFT	–	–	68.7	–	42.0	–	45.6	–
DRIFT	Absolute		65.7	-3.0	39.5	-2.5	25.9	-19.7
	Grad-Norm	{ATTN, MLP}	68.8	+0.1	43.9	+1.9	46.1	+0.5
	Grad-Norm w/ Adaptive α		69.9	+1.2	43.9	+1.9	47.2	+1.6
	Grad-Norm	{ATTN}	68.8	+0.1	44.4	+2.4	49.4	+3.8
		{MLP}	68.5	-0.5	42.6	+0.6	46.3	+0.7
		{ATTN, MLP, Norm}	68.6	-0.1	43.0	+1.0	46.8	+1.2
		{ATTN, MLP, Norm, LM Head}	68.6	-0.1	42.7	+0.7	48.5	+2.9
	Grad-Norm w/ Adaptive α	{ATTN}	68.8	+0.1	43.2	+1.2	48.3	+2.7
		{MLP}	69.0	+0.3	43.3	+1.3	46.3	+0.7
		{ATTN, MLP, Norm}	69.1	+0.4	42.8	+0.8	48.3	+2.7
		{ATTN, MLP, Norm, LM Head}	68.8	+0.1	43.3	+1.3	48.3	+2.7

consistently yields further gains. For instance, DRIFT achieves $+1.9$ points on *MathVerse* and $+5.2$ on *WeMath*, compared to the SFT baseline. These gains suggest that the reasoning prior extracted from text-only experts is indeed useful in guiding multimodal training, providing complementary reasoning signals beyond what the multimodal instruction data alone can supply. Importantly, the improvements are achieved without relying on costly multimodal CoT annotations.

On the Role of Merging Candidates. To understand which components benefit most from reasoning injection, we vary the set of modules to which DRIFT is applied (see Tab. 4). We start from the attention layers, and find that applying DRIFT only to attention layers achieves the strongest performance on *LogicVista*, with additional improvements on *MathVerse*. In contrast, restricting to feedforward layers yield modest or inconsistent gains, and including normalization layers often leads to diminished performance. Extending to the LM head provides mixed results – limited impact on *MathVerse* but noticeable gains on *LogicVista*. These findings suggest that attention modules are the most sensitive to reasoning priors, while over-extending to normalization layers can inject noise rather than useful signals.

On the Role of Merging Strategies. Different strategies for incorporating the reasoning prior lead to distinct behaviors. The *Absolute* update rule degrades performance across all benchmarks, likely because it pulls parameters too aggressively toward the reasoning model, disrupting multimodal alignment. In contrast, gradient-based scaling strategies (*Grad-Norm* and *Grad-Norm w/ Adaptive* α) yield stable improvements. Notably, *Grad-Norm w/ Adaptive* α achieves the highest MathVista score (69.9, $+1.2$), showing that adapting the prior based on the gradient–delta relation provides a balanced integration. This highlights that subtle guidance, rather than direct overwriting, is the key to successfully transferring reasoning capabilities.

Overall, these analyses reinforce our central claim: reasoning priors are beneficial, but their utility depends strongly on *where* they are applied (attention layers vs. others) and *how* they are integrated (gradient guidance vs. absolute interpolation). DRIFT's design provides a stable mechanism for exploiting these priors.

5 Conclusion

We study how to transfer reasoning from text-only LLMs to multimodal LLMs without large-scale multimodal CoT supervision. We show that parameter-space merging is brittle when models are misaligned, and propose *Directional Reasoning Injection for Fine-Tuning*, a gradient-based method that injects reasoning priors from expert models during fine-tuning. DRIFT consistently outperforms standard SFT and remains competitive with training-intensive approaches, demonstrating that lightweight gradient-space priors offer an efficient and scalable solution for reasoning transfer.

6 Limitations

While DRIFT demonstrates that lightweight gradient-space priors can effectively transfer reasoning from text-only experts to multimodal models, several limitations remain. First, our method relies on the availability of strong text-only reasoning experts, which constrains applicability in domains where such experts are weak or unavailable. Second, although DRIFT avoids destabilizing multimodal alignment in our experiments, its reliance on precomputed reasoning directions may introduce biases or diminish performance when tasks require perception-heavy reasoning. Third, we primarily evaluate on mathematical and logical reasoning benchmarks; further validation on diverse multimodal tasks such as commonsense reasoning, scientific understanding, or open-domain visual question answering is needed to assess generality. Finally, while DRIFT reduces training costs compared to reinforcement learning or large-scale multimodal CoT supervision, it still adds overhead relative to standard SFT and does not yet guarantee interpretability of the injected reasoning signals.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. In *Forty-second International Conference on Machine Learning*.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025a. Bring reason to vision: Understanding perception and reasoning through model merging. In *Forty-second International Conference on Machine Learning*.

Yang Chen, Yufan Shen, Wenxuan Huang, Sheng Zhou, Qunshu Lin, Xinyu Cai, Zhi Yu, Jiajun Bu, Botian Shi, and Yu Qiao. 2025b. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback. *arXiv preprint arXiv:2507.20766*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, and 1 others. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

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.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. 2024a. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024b. Mindmerger: Efficiently boosting llm reasoning in non-english languages. *Advances in Neural Information Processing Systems*, 37:34161–34187.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with MATH-Vision dataset. arXiv preprint arXiv:2402.14804.

Séamus Lankford, Haithem Afli, and Andy Way. 2023. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. *Information*, 14(12):638.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

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

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

Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024d. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 365–374.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, and 1 others. 2025a. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*.

Ruyang Liu, Chen Li, Yixiao Ge, Thomas H Li, Ying Shan, and Ge Li. 2024. Bt-adapter: Video conversation is feasible without video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667.

Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. 2025b. Othink-mr1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. 2024. Lisa: Layer-wise importance sampling for memory-efficient large language model fine-tuning. *Advances in Neural Information Processing Systems*, 37:57018–57049.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Leonardo Ranaldi and Andre Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. *arXiv preprint arXiv:2405.00402*.

Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. 2025. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. In *Proceedings of the AAAI Symposium Series*, volume 5, pages 384–388.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? arXiv preprint arXiv:2403.14624.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, and 1 others. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

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

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, and 1 others. 2025. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2409.02813.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

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 *Forty-first International Conference on Machine Learning*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large

language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, and 1 others. 2024a. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024b. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

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

A Complete Related Works

A.1 Multimodal Reasoning in Large Language Models

Following the success of chain-of-thought prompting in enabling large language models (LLMs) to solve complex problems step by step, researchers have increasingly explored whether similar reasoning capabilities exist in multimodal large language models (MLLMs). Among the many domains for evaluation, mathematical reasoning has emerged as one of the most prominent. Lu et al. (2023) introduced MathVista, a visual mathematics benchmark designed to assess the problem-solving abilities of MLLMs on math tasks that require visual understanding. Similarly, Xiao et al. (2024) proposed LogicVista, which evaluates integrated logical reasoning skills over visual concepts. Additional benchmarks, including MathVision (Wang et al., 2024a), MathVerse (Renrui Zhang et al., 2024), and WeMath (Qiao et al., 2024), extend this line of research by covering diverse mathematical problem types and difficulty levels, with a strong emphasis on the vision modality.

Many methods have been proposed to enhance the reasoning ability of MLLMs. Ratzlaff et al. (2025); Li et al. (2024d); Ranaldi and Freitas (2024) explore instruction tuning to teach MLLMs to reason over visual concepts. Similarly, Subramaniam et al. (2025); Huang et al. (2024b); Dong et al. (2025) adopt supervised fine-tuning (SFT) to further improve MLLM performance. More recent works (Wan et al., 2025; Liu et al., 2025b; Chen et al., 2025b) demonstrate that reinforcement learning (RL) approaches can effectively enhance the reasoning capabilities of MLLMs while maintaining strong generalization across diverse tasks. Among these methods, both SFT and RL have shown remarkable potential. SFT is generally lightweight and efficient, but its effectiveness depends heavily on the availability of high-quality, diverse multimodal datasets. RL methods, on the other hand, are less constrained by dataset diversity and can yield robust improvements, though they are more computationally expensive and require substantial resources for training.

A.2 Efficient Fine-Tuning of LLMs

Given the high memory and computational cost of full-parameter fine-tuning, numerous studies have proposed methods to reduce these costs and improve training efficiency. These approaches can generally be divided into parameter-efficient and data-efficient fine-tuning methods.

Parameter-Efficient Fine-Tuning. Hu et al. (2022) introduced LoRA, which reduces trainable parameters by injecting and training a low-rank decomposition within the model's weight matrices. Subsequent works have refined LoRA with various enhancements, including QLoRA (Dettmers et al., 2023), LoRA+ (Hayou et al., 2024), and LiSA (Pan et al., 2024). Another line of work focuses on adapter-based methods, where small trainable modules are inserted into the model while keeping the base parameters frozen. Examples include AdaptMLLM (Lankford et al., 2023), LLaMA-Adapter (Zhang et al., 2024b; Gao et al., 2023), and Bt-Adapter (Liu et al., 2024).

Data-Efficient Fine-Tuning. Another research direction seeks to improve fine-tuning efficiency by carefully curating or compressing the training data. For instance, Lin et al. (2024) propose pruning and selecting representative samples to maximize data utility. He et al. (2024) leverage external MLLMs to select high-quality multimodal data for training. Additionally, methods such as those proposed by Shang et al. (2024) and Cai et al. (2024) reduce the number of visual tokens used for training, thereby

Method	SFT	RL	Est. time
OpenVLThinker-7B (Deng et al., 2025)	✓	✗	>1 day
R1-OneVision-7B (Yang et al., 2025)	✓	✗	>1 day
X-REASONER (Liu et al., 2025a)	✓	✓	>2 days
Ours (DRIFT)	✓	✗	$\approx$2 hrs

Table 5: **Training schemes and estimated wall-clock cost.** Existing methods require at least one day of training, while DRIFT completes in about two hours under comparable hardware.

accelerating both fine-tuning and inference.

Model Merging. An even more efficient alternative, model merging repurposes fine-tuned models by directly combining parameters through simple arithmetic (Ilharco et al.; Yadav et al., 2023; Yu et al., 2024), requiring no additional training or inference cost. Although well studied in vision models (Huang et al., 2024a; Gargiulo et al., 2025), its use in MLLMs remains limited. Recent work, such as BR2V (Chen et al., 2025a), demonstrates the potential of merging for transferring reasoning into multimodal models. Nonetheless, large parameter discrepancies and cross-modal transfer of reasoning remain open challenges. Our work addresses these by injecting reasoning priors from LLMs into MLLMs via gradient space merging.

B Parameter-Space Merging Method Setup

We experiment with several parameter-space merging strategies, where models are combined without additional training by directly manipulating their parameters. The hyper-parameters in Tab. 6 correspond to: (i) λ coefficients that control the interpolation ratio between two models (Ilharco et al.); (ii) α scaling factors used in data-aware reweighting (e.g., in DARE (Yu et al., 2024)); and (iii) for layer swapping (Bandarkar et al.), the number of layers replaced. All hyperparameter choices are sourced from BR2V (Chen et al.).

C Training Time Comparison

Training efficiency is a critical factor when scaling reasoning-capable MLLMs. Most existing approaches rely on either large-scale supervised fine-tuning (SFT) with multimodal CoT data or reinforcement learning (RL) on specialized reasoning benchmarks. Both settings typically require multiple days of training on high-end GPU clusters, limiting their practicality for rapid iteration or deployment.

Method	Hyper-parameters
Baseline	-
Task Arithmetic	$(\lambda = 0.9,\ 0.1)$
TIES	$(\lambda = 1.6,\ \alpha = 0.2)$
Dare-TIES	$(\lambda = 1.6,\ \alpha = 0.2)$
Dare-Linear	$(\lambda = 1.6,\ \alpha = 0.2)$
Layer Swap	$(\lambda = 0.9,\ 0.1,\ k = 5)$

Table 6: Hyper-parameter setup for different parameter-space merging methods. (λ, α, k) denote interpolation ratios, scaling factors, and number of swapped layers, respectively.

As summarized in Tab. 5, representative methods such as OpenVLThinker, R1-OneVision, and X-REASONER all involve either full SFT or RL and require more than one day of training. In contrast, our method, DRIFT, requires only SFT-style training with gradient guidance and completes within roughly two hours under comparable hardware. This dramatic reduction in cost is achieved because DRIFT (i) avoids a huge amount of multimodal CoT data collection, (ii) adds only lightweight gradient-time operations with a precomputed prior, and (iii) leaves the forward pass unchanged.

In practice, this efficiency means DRIFT can be integrated into existing SFT pipelines with negligible additional overhead, making it far more scalable for both research and production settings.

C.1 Dataset Collection Details

We leverage the ThinkLite (Wang et al., 2025) model to distill multimodal reasoning data on the ThinkLite-VL-Hard-11K dataset. The prompt used to elicit reasoning traces is illustrated in Figure 4.

After generating candidate responses, we apply a multi-step filtering process to ensure data quality. First, we verify whether the final answer is enclosed in `\boxed{}` and matches the ground-truth solution. Second, we check the correctness of the reasoning format enclosed by `<think>` and `</think>`. Fi-

System Prompt: You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.

User Prompt: <question>

Figure 4: Example prompt used to distill reasoning traces from the ThinkLite model.

nally, we retain the highest-quality subset, resulting in 4K verified samples from the original 11K examples.

C.2 Training Loss of Gradient Merging Strategies

We compare training loss curves of different gradient merging strategies against the SFT baseline on the same dataset. As shown in Fig. 5, the *Absolute* strategy introduces instability, leading to large spikes in the early stages. *Grad-Norm* reduces this effect but still shows noticeable fluctuations. In contrast, *Grad-Norm with Adaptive* α closely follows the stable SFT baseline while yielding improved convergence.

- **Absolute:** $\tilde{g} = g + \alpha\Delta$, directly pulling weights toward the reasoning prior.
- **Grad-Norm:** $\tilde{g} = g + \alpha\|g\|\frac{\Delta}{\|\Delta\|}$, aligning updates with the direction of Δ while preserving the gradient magnitude of g.
- **Grad-Norm w/ Adaptive α:** $\tilde{g} = g + \alpha'\|g\|\frac{\Delta}{\|\Delta\|}$, where $\alpha' = \alpha \cdot \frac{1+\cos(g,\Delta)}{2}$ adapts the strength based on gradient–delta alignment.

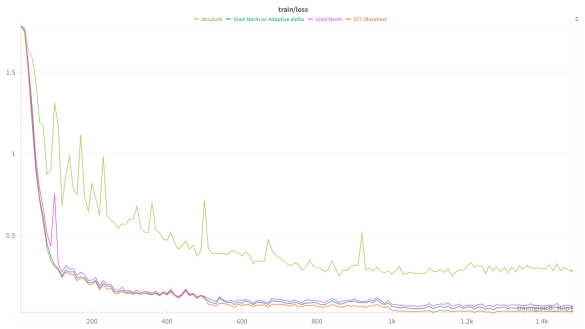

Figure 5: Training loss curves for different gradient merging strategies compared with the SFT baseline. Adaptive Grad-Norm achieves stable optimization while improving performance over standard SFT.

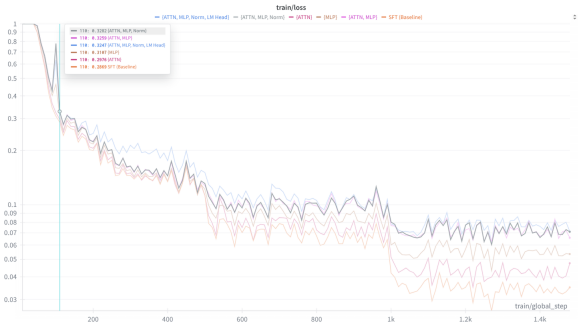

Figure 6: Training loss curves for gradient merging candidates compared with the SFT baseline. The {ATTN} strategy avoids training spikes, while other candidates show instability before convergence.

D Training Loss of Gradient Merging Candidates

We compare the training loss curves of different gradient merging candidates against the SFT baseline on the same dataset. As shown in Fig. 6, merging on {ATTN} yields the most stable curve without spikes, while all other variants exhibit noticeable fluctuations in the early training stage. For clarity, we also plot the loss in log scale and zoom in around the spike region to highlight differences across methods:

- {ATTN}
- {MLP}
- {ATTN + MLP}
- {ATTN + MLP + Norm}
- {ATTN + MLP + Norm + LM Head}

D.1 Ablation on the Types of Priors

In the main paper, we define the *directional prior* as the parameter difference between a text-only reasoning model and its multimodal variant:

$$\Delta = \phi_{\text{reason}} - \phi_{\text{VL}}. \tag{4}$$

This Δ captures the *reasoning direction* and serves as the guidance signal for gradient modulation during fine-tuning.

To further understand the impact of prior choice, we examine alternative constructions of Δ. Specifically, for `DeepSeekR1-Qwen-Distill-7B`, we test two variants: (i) a **direct prior** computed as the difference between the reasoning model

Table 7: **Ablation on types of reasoning priors.** We compare different constructions of the reasoning prior Δ under the DRIFT framework on multimodal reasoning benchmarks. Results are reported on *MathVista*, *MathVision*, *MathVerse*, *WeMath* (strict/loose), and *LogicVista*. Best results are in **bold**, and improvements relative to the SFT baseline are shown as subscripts in green (+) or red (–).

Reasoning Prior Δ	MathVista	MathVision	MathVerse	WeMath		LogicVista	Avg.
				strict	loose		
$\phi_{\text{DeepSeekR1}} - \phi_{\text{Qwen-VL}}$	**69.9**$_{+2.3}$	**26.6**$_{+1.5}$	**43.9**$_{+0.8}$	**38.5**$_{+2.5}$	**60.2**$_{+1.8}$	**47.2**$_{+3.4}$	**47.7**$_{+2.1}$
$\phi_{\text{DeepSeekR1}} - \phi_{\text{Qwen-Math}}$	67.6	25.1	43.1	36.0	58.4	43.8	45.6

and its multimodal variant, and (ii) a **hierarchical prior** derived between the reasoning expert (`DeepSeekR1-Qwen-Distill-7B`) and the base model used during its fine-tuning (`Qwen-Math-7B`).

The results in Tab. 7 show that the direct prior generally achieves higher and more consistent improvements across benchmarks, suggesting that the reasoning direction from a distilled expert (`DeepSeekR1-Qwen-Distill-7B`) more effectively aligns with multimodal training gradients. In contrast, the hierarchical prior (`DeepSeekR1` vs. `Qwen-Math`) provides smaller gains, possibly because it encodes earlier-stage reasoning behaviors less compatible with multimodal features. These findings highlight that the choice of reasoning prior directly affects transfer quality, and the most effective priors could be those closely matched to the target model's reasoning granularity.

E Future Work

Building on our findings, several directions remain open for exploration. First, extending DRIFT beyond mathematical and logical reasoning to domains such as scientific understanding, embodied perception, and real-world decision-making would test its generality. Second, developing adaptive strategies that dynamically select or combine reasoning priors, rather than relying on a fixed direction, could improve robustness when transferring across diverse tasks. Third, integrating DRIFT with reinforcement learning or preference optimization may further enhance reasoning without sacrificing multimodal grounding. Finally, improving interpretability of injected reasoning signals, through visualization or attribution, would provide stronger insights into how reasoning knowledge is transferred, fostering trust and transparency in multimodal systems.

F Broader Impact

This work highlights a lightweight path for transferring reasoning abilities from text-only experts to multimodal models, offering efficiency benefits and reduced reliance on costly multimodal supervision. By lowering the resource barrier, DRIFT may help democratize access to multimodal reasoning systems in academic and industrial settings. However, transferring reasoning across domains also raises important considerations. First, biases embedded in text-only experts may propagate into multimodal models, amplifying inaccuracies or cultural biases in downstream tasks. Second, more capable multimodal reasoning systems may be misused in sensitive domains such as surveillance, misinformation generation, or automated decision-making, where reliability and transparency are critical. Third, although DRIFT reduces compute costs, it still benefits institutions with access to pretrained reasoning experts, potentially reinforcing existing inequities in model development.