ICoM: Interleaved CoT with Adaptive Visual Focusing and Layer-specific Merging for Advanced Mathematical Reasoning

Anonymous ACL submission

Abstract

Vision Large Language Models (VLLMs) have achieved remarkable progress in multimodal reasoning. However, they often generate text-only reasoning steps based on internal priors, making it difficult to dynamically focus on critical visual regions. Multimodal interleaved Chain-of-Thought (CoT) paradigm, built on visual modules can incorporate visual inputs, but they typically require additional tools and multi-step interactions. To address these issues, we propose a coupled framework, ICoM, that integrates interleaved CoT driven adaptive visual focusing with layer-specific merging. ICoM employs a Q-Former to adaptively retrieve the most relevant regions from the original image via interleaved tokens, and are inserted before each textual reasoning step to enable visual focus. We train Qwen2-VL-2B with a three-stage SFT+RL pipeline on the open-source MINT-CoT dataset. To enhance reasoning in a cost effective way, we linearly merge only layers 19–27 of the post trained Qwen2-VL-2B language component with the corresponding parameters of Qwen2-Math-1.5B-Instruct. Experiments show that ICoM-2B is competitive with state-of-the-art VLLMs (e.g., LLaVA-Reasoner-8B and Mulberry-7B) across six benchmarks. Notably, ICoM-2B outperforms GPT-4o-0513 by 2.13% on MathVista and 0.22% on MMStar. Code will be released once the paper is accepted.

1 Introduction

With the evolution of vision large language models (VLLMs) (Bai et al., 2025a; Hurst et al., 2024; Anthropic, 2024) and reinforcement learning (RL) with verifiable rewards (DeepSeek-AI, 2025; Shao et al., 2024a), VLLMs have made significant progress in jointly processing images and text inputs to perform complex tasks. However, prevailing reasoning paradigms rely solely on the VLLM's internal priors and express intermediate reasoning steps through textual tokens. This leads to inherent limitations in visually dense scenarios, where the model struggles to autonomously identify key visual regions (Su et al., 2025a) (e.g., tiny objects and subtle spatial relationships), impeding effective interaction with information-rich images. Therefore, bridging the gap between perception and reasoning, together with strengthening region-focused attention, has emerged as a critical problem.

Some studies (Peng et al., 2023; You et al., 2024) use region bounding boxes or masks during training to enable local region understanding. These methods, howerver, depend on additional annotations and predefined regions, limiting their ability to dynamically select relevant visual information during inference. In contrast, multimodal chain-of-thought (MCoT) (Zhang et al., 2024b) conducts intermediate reasoning by combining text-only chains with structured visual evidence (e.g., knowledge graphs (Mondal et al., 2024)). More recently, interleaved paradigms (Gao et al., 2025) incorporate additional visual content during reasoning through cropping tools or specialized sketching models. While effective in general settings (Li et al., 2025b; Zheng et al., 2025), they typically require external tools and multi-step interactions, which can result in insufficient modeling of mathematical visual cues such as symbols, axis scales, and geometric relations. MINT-CoT (Chen et al., 2025e) further improves mathematical reasoning by interleaving visual tokens before each reasoning step, yet its pointwise filtering makes it difficult to capture step-specific context and multi-view information. Cross-modal parameter merging (Chen et al., 2025d) has emerged as an efficient training-free solution that transfers the strong reasoning capability of LLMs into VLLMs, offering a principled way to strengthen the coupling between perception and reasoning. Its potential for fine-grained mathematical reasoning remains underexplored.

To address these challenges, we propose ICoM, a coupled framework that combines adaptive vi-

sual focusing driven by interleaved visual chain-of-thought (CoT) with layer-specific merging, as illustrated in Figure 1. Specifically, we introduce the Q-Former (Li et al., 2023) that uses the interleaved token as a query to retrieve the most relevant visual regions from the input image and incorporates them before each textual reasoning step to enable visual focusing. We build on two baseline models, Qwen2-VL-2B-Instruct and Qwen2.5-VL-3B-Instruct, and train ICoM in three stages using a combination of supervised fine-tuning (SFT) and RL: text-only CoT SFT, interleaved CoT SFT, and interleaved CoT RL. This training pipeline consistently improves mathematical reasoning, but the model's abstract mathematical priors remain constrained by the scale and distribution of the training data. To address this, after the three-stage training of Qwen2-VL-2B, we linearly merge its LM parameters with the corresponding layers of Qwen2-Math-1.5B-Instruct in the late layers (19–27). Notably, our ICoM-2B is competitive with state-of-the-art VLLMs (e.g., LLaVA-Reasoner-8B and Mulberry-7B) across six benchmarks. For example, ICoM-2B outperforms GPT-4o-0513 and R1-VL-7B by 2.13% and 2.43% on the mathematical benchmark MathVista, and by 0.22% and 4.92% on the comprehensive benchmark MMStar.

2 Related Work

Reinforcement Learning for VLLMs. RL is critical for adapting VLLMs to complex tasks beyond their pretraining distribution, and an increasing number of studies are applying it to multimodal reasoning settings. These methods typically follow a multi-stage pipeline, first performing SFT on costly distilled data and then applying RL to further enhance reasoning. For example, VL-Rethink (Wang et al., 2025) explores a more direct RL strategy to promote slow thinking in VLLMs and introduces selective sample replay (SSR) to mitigate the vanishing advantage problem in Group Relative Policy Optimization (GRPO) (Shao et al., 2024b). Perception-R1 (Yu et al., 2025) directly encodes image patches, effectively integrating test-time augmentation with RL fine-tuning. STAR-R1 (Li et al., 2025c) further demonstrates RL's effectiveness for spatial and concrete reasoning.

CoT with Images. The CoT with Images (Interleaved CoT) paradigm has emerged as a promising direction for enhancing multimodal reasoning capabilities. Textual CoT (Wei et al., 2022), however, often falters when tasks require information beyond textual descriptions (Hao et al., 2025; Jiang et al., 2025). In contrast, this paradigm allows models to invoke visual operations (e.g., local zoom-in) or leverage external visual modules (e.g., cropping tools and specialized sketching models (Hu et al., 2024; Zhou et al., 2024)), enabling progressive region exploration and narrowing the solution space . MVoT (Li et al., 2025a) introduces an interleaved action representation for maze solving. OpenThinkIMG (Su et al., 2025b) proposes an end-to-end framework for learning to use visual tools. Despite their advances in multimodal reasoning, these approaches often depend on external tools and multi-step interactions, rendering the reasoning process indirect and brittle.

3 Methodology

To address the limitations of multimodal interleaved CoT paradigms and strengthen models' internal reasoning capability, we propose ICoM in Section 3. We then introduce Q-Former interleaved selection of relevant visual tokens (Section 3.2) and the three stage training scheme (Section 3.3). Finally, we perform layer-specific merging across modalities in the late layers (19–27) in Section 3.4.

3.1 Multimodal Chain-of-Thought (MCoT)

A typical VLLM consist of a vision tower, a language model, and a projector that bridges these two parts. The vision tower processes images, enabling the model to perceive visual content, while the language model serves as the reasoning engine, processing knowledge and generating responses. Then, the VLLM takes the image I and the text $T = (\text{Instruction}, \text{Question})$ as inputs and produces a final answer:

$$\text{answer} = \text{VLLM}(I, T). \quad (1)$$

Compared with the direct prediction in Equation 1, multimodal MCoT further elicits the VLLMs to generate a sequence of intermediate textual reasoning steps $\{r^1, r^2, \cdots, r^m\}$ before the final answer:

$$\{r^j\}_{j=1}^{m}, \text{answer} = \text{VLLM}(I, T). \quad (2)$$

These intermediate reasoning steps are generated from global image features, so the model struggles to focus on key local regions during reasoning.

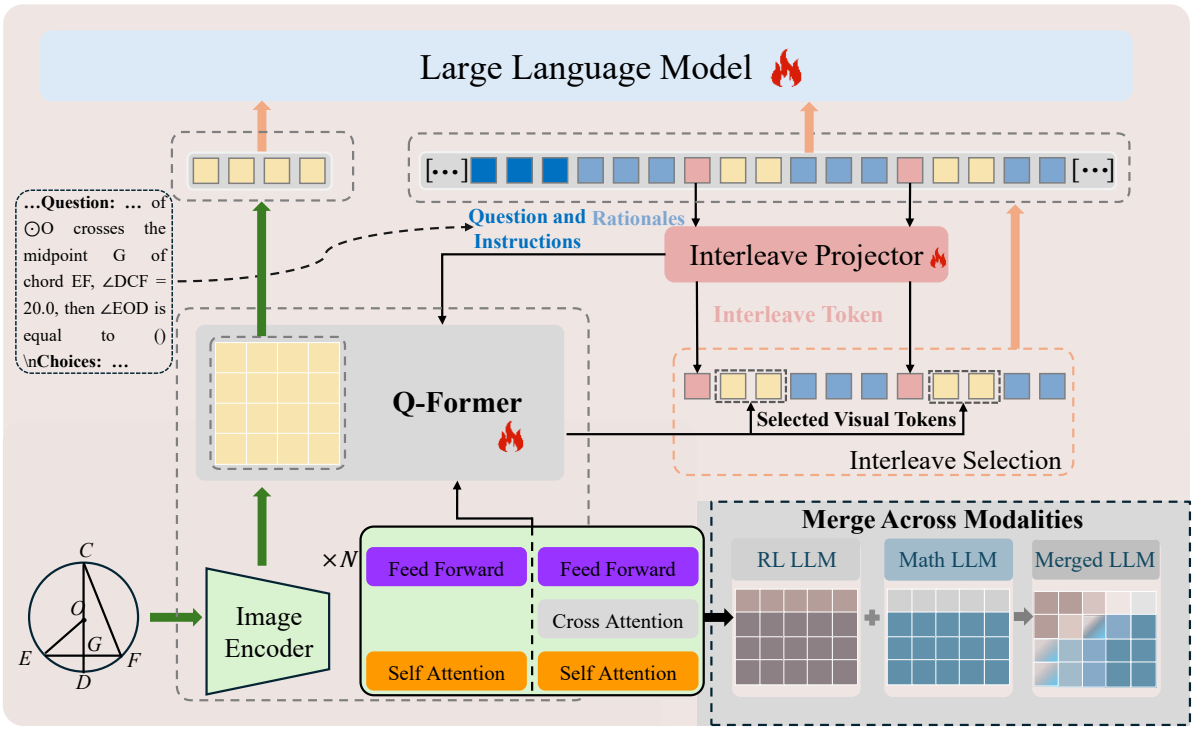


Figure 1: Architecture of ICoM, including Q-Former based selection of relevant visual tokens before each reasoning step and layer-specific (e.g., early, middle, and late layers) merging across modalities .

3.2 Interleaved Visual Chain-of-Thought

3.2.1 Interleave Token

We employ an interleave token, which is used to select visual tokens that are relevant to the mathematical concepts involved in that step (e.g., "radius OC" and "segment OD"). Following established practice (Chen et al., 2025e), we utilize this special token to adaptively focus on key visual regions and reduce interference from irrelevant visual content (e.g., "auxiliary lines" and "annotation text") so as to facilitate the reasoning process.

When an interleave token is output at step j, its hidden state h^j_{inter} is passed through a projector P:

$$q^j_{\text{inter}} = P(h^j_{\text{inter}}). \tag{3}$$

3.2.2 Visual Focusing with Q-Former

Considering that multi-head attention naturally supports parallel retrieval across multiple views of information such as scales, text, and geometric relations, it can cover heterogeneous evidence and local–global relations more effectively than a single cosine similarity measure. As shown in Figure 1, we leverage the Q-Former, using its multi-head attention and stacked layers to achieve iterative focusing from coarse to fine and to couple "where to look" with "how to solve" in an end-to-end manner.

Specifically, the vision encoder E extracts visual features from the input image I as $V = \mathrm{E}(I) = \{v_k\}_{k=1}^K$, where $v_k \in \mathbb{R}^d$ represents the k-th visual token. At the j-th reasoning step, Q-Former uses q^j_{inter} as the query to reweight the visual tokens:

$$\mathbf{w}^j = \mathrm{QFormer}(q^j_{\text{inter}}, V) \in \mathbb{R}^K, \tag{4}$$

where $\mathbf{w}^j = (w^j_1, w^j_2, \cdots, w^j_K)$ is a selection distribution over the K visual tokens, satisfying $w^j_k \geq 0$ and $\sum_{k=1}^K w^j_k = 1$.

The score of each token w^j_k is compared against a predefined threshold δ, and visual tokens with scores above this threshold are dynamically selected:

$$\mathbf{v}^j = \{v^j_k | w^j_k > \delta\}, \tag{5}$$

where $k \in \{1, 2, \cdots, K\}$. The selected tokens are interleaved into the reasoning process at step j. Formally, we interleave visual content with text-based reasoning steps during inference to produce the final answer:

$$\{(\mathbf{v}^j, r^j)\}_{j=1}^m, \ \text{answer} = \mathrm{VLLM}(I, T). \tag{6}$$

Through this selection mechanism, important visual regions are interleaved into the model before each textual step, improving visual grounding.

3.3 Training Strategy

Inspired by the MINT-CoT training set, we adopt a three-stage training scheme for ICoM.

Stage 1: Text-only CoT SFT. To enable the VLLM to learn a general reasoning pattern, we train the base model on the text-only CoT reasoning data from MINT-CoT.

Stage 2: Interleaved CoT SFT. Building on Stage 1, we train ICoM to dynamically select visual tokens using the interleaved token and to adapt to reasoning with interleaved visual content. To optimize textual reasoning and visual alignment, the model is fine-tuned using a language modeling loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t\in\mathcal{T}} \log P_\theta(y_t \mid y_{<t}, I, T)\,, \tag{7}$$

where $\mathcal{T} \in \{1, 2, \cdots, T\}$ ranges over the positions of the text tokens, and the output sequence is $Y = \{y_1, y_2, \cdots, y_T\}$.

We do not supervise the cross-entropy loss for predicting the interleave token. Instead, we manually concatenate it at each step, and during inference, we concatenate the interleave token whenever the "### Step" marker is generated. To explicitly supervise the selection behavior in Equation 5, we impose a binary cross-entropy (BCE) loss on the selection decisions over all visual tokens at each reasoning step, together with a validity mask, stepwise adaptive positive-class reweighting, and loss clipping to mitigate class imbalance and stabilize optimization. Additional theoretical details of stepwise visual token selection loss are provided in Appendix A. The BCE loss is:

$$\mathcal{L}_{\mathrm{BCE}} = \frac{\sum_j \sum_k m_{j,k}\, \tilde{\ell}_{j,k}}{\sum_j \sum_k m_{j,k} + \varepsilon}, \tag{8}$$

Finally, based on Equation 7 and Equation 8, we have the training objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{BCE}}. \tag{9}$$

This combined loss guides ICoM to jointly optimize text generation and visual token selection, while performing interleaved reasoning.

Stage 3: Interleaved CoT RL. To improve the consistency of long-chain reasoning and, under the guidance of the task objective, adaptively explore more effective visual tokens, we incorporate GRPO into training. For each group of G reasoning trajectories $\{Y_j\}_{j=1}^{G}$, we assign a binary reward $r_j \in \{0, 1\}$ based on answer correctness. Within each group, the advantage of the j-th trajectory is $\hat{A}_j \triangleq \frac{r_j - \mathrm{mean}(\mathbf{r})}{\mathrm{std}(\mathbf{r})}$, where $\mathbf{r}$ denotes the set of rewards in the group. The policy loss over the generated tokens is formulated as:

$$\mathcal{L}_{\mathrm{GRPO}} = -\,\mathbb{E}_{\{Y_j\}_{j=1}^{G}}\Bigg[\frac{1}{G}\sum_{j=1}^{G}\Bigg(\frac{P_\theta(Y_j)}{P_{\theta_{\mathrm{old}}}(Y_j)}\,\hat{A}_j - \beta\, D_{\mathrm{KL}}\big(P_\theta \,\|\, P_{\mathrm{ref}}\big)\Bigg)\Bigg]. \tag{10}$$

3.4 Layer-specific Merging Across Modalities

Previous studies (Yin et al., 2025; Shi et al., 2025; Chen et al., 2025a) have shown that the early layers of VLLMs primarily capture visual perception abilities and world knowledge, whereas mathematical CoT reasoning mainly emerges in the middle and later layers. Although cross-modal language model (LM) parameter merging (Chen et al., 2025d) can improve mathematical reasoning, it also introduces text-only optimized parameter shifts into early layers, potentially undermining multimodal alignment and visual perception. To balance high-level reasoning enhancement with the stability of early-layer multimodal representations, we propose a layer-specific task vector merging strategy that injects external reasoning capability into the reasoning subspace of the language head.

Let Θ denote the language parameters of models that share the same backbone architecture (e.g., identical hidden size, number of layers, and number of heads), based on which we define task vectors in a unified LM parameter space.

Taking the Qwen2 family as an example, Θ_{base} is the set of language parameters of the LLM Qwen2-Instruct, and Θ_{vl} is that of the VLLM Qwen2-VL-Instruct. We set $\tau_{\mathrm{perc}}^{\mathrm{LM}} \triangleq \Theta_{\mathrm{VL}} - \Theta_{\mathrm{base}}$, which represents the perception task vector induced by transitioning from the base LLM to the visually aligned VLLM. We apply the three-stage training scheme described in Section 3.3 to Qwen2-VL-Instruct, yielding three sets of LM parameters: Θ_{text}, Θ_{inter} and Θ_{rl}. Its training trajectory can be represented as:

$$\Theta_{\mathrm{base}} \xrightarrow{\tau_{\mathrm{perc}}^{\mathrm{LM}}} \Theta_{\mathrm{VL}} \xrightarrow{\Delta_{\mathrm{text}}} \Theta_{\mathrm{text}} \xrightarrow{\Delta_{\mathrm{inter}}} \Theta_{\mathrm{inter}} \xrightarrow{\Delta_{\mathrm{rl}}} \Theta_{\mathrm{rl}}$$

where Δ_{text}, Δ_{inter}, and Δ_{rl} are the parameter increments at each stage relative to the preceding

stage. The task vectors with respect to the base LLM Θ_{base} are given by:

$$\begin{aligned}
\tau_{\text{text}}^{\text{LM}} &\triangleq \Theta_{\text{text}} - \Theta_{\text{base}} = \tau_{\text{perc}}^{\text{LM}} + \Delta_{\text{text}}, \\
\tau_{\text{inter}}^{\text{LM}} &\triangleq \Theta_{\text{inter}} - \Theta_{\text{base}} = \tau_{\text{perc}}^{\text{LM}} + \Delta_{\text{text}} + \Delta_{\text{inter}}, \\
\tau_{\text{rl}}^{\text{LM}} &\triangleq \Theta_{\text{rl}} - \Theta_{\text{base}} = \tau_{\text{perc}}^{\text{LM}} + \Delta_{\text{text}} + \Delta_{\text{inter}} + \Delta_{\text{rl}}.
\end{aligned} \tag{11}$$

Furthermore, Θ_{math} is the set of language parameters of the math-specialized LLM Qwen2-Math-Instruct, built on the same backbone. The mathematical task vector is:

$$\tau_{\text{math}}^{\text{LM}} \triangleq \Theta_{\text{math}} - \Theta_{\text{base}}. \tag{12}$$

$\tau_{\text{math}}^{\text{LM}}$ is from large scale text-only mathematical corpora and primarily encode linguistic mathematical knowledge (e.g., symbolic reasoning patterns). It can be interpreted as a parameter offset that injects mathematical reasoning priors into the base LLM, thereby compensating for the limitations of multimodal training on purely symbolic difficult problems and long tail mathematical concepts. $\tau_{\text{rl}}^{\text{LM}}$ denotes the composite task vector accumulated over the perception alignment, interleaved CoT, and RL reasoning stages.

To inject external mathematical reasoning priors into the high layer reasoning subspace of LM, we linearly combine $\tau_{\text{math}}^{\text{LM}}$ and $\tau_{\text{rl}}^{\text{LM}}$ in a unified task vector space. A hyperparameter $\lambda \in [0, 1]$ controls the strength of the injected mathematical task vector, resulting in the fused LM parameters:

$$\Theta_{\text{merge}}^{\text{LM}} = (\lambda \tau_{\text{rl}}^{\text{LM}} + (1-\lambda)\tau_{\text{math}}^{\text{LM}} + \Theta_{\text{base}})[\text{layer1} : \text{layer2}], \tag{13}$$

In particular, by reusing the visual components from Stage 3 (e.g., the vision encoder, connector, and Q-Former), we preserve the learned visual alignment and interleaving strategy. Equation 13 is equivalent to linearly interpolating the LM parameters over the same layer range:

$$\Theta_{\text{merge}}^{\text{LM}} = (\lambda \Theta_{\text{rl}} + (1-\lambda)\Theta_{\text{math}})[\text{layer1} : \text{layer2}], \tag{14}$$

Here, $[\text{layer1} : \text{layer2}]$ indicates that parameters are merged for layers layer1 to layer2, while all other language parameters remain identical to Θ_{rl}.

4 Experiments

4.1 Setup

Implementionb Details. In this paper, we leverage the open-source LLaMA-Factory (Zheng et al., 2024) and R1-V (Chen et al., 2025c) training frameworks, build on two baseline models, Qwen2-VL-2B-Instruct and Qwen2.5-VL-3B-Instruct, and train our ICoM model in three stages using a combination of SFT and RL. The threshold δ is set to 0.7 to filter attention scores, and the Q-Former depth is set to $N = 2$. During training, all model parameters except the vision encoder are updated, including the MLP connector, the LLM, the interleave projector, and the Q-Former. To enable isomorphic parameter merging within the ICoM framework, we adopt Qwen2-1.5B-Instruct as the base LLM and Qwen2-Math-1.5B-Instruct as the mathematical reasoning model. Once Qwen2-VL-2B-Instruct completes the three-stage training, we assign a weight of 0.9 to its textual component and 0.1 to the reasoning task vector, i.e., $\lambda = 0.9$. All experiments are conducted on eight NVIDIA L40 GPUs. For Qwen2.5-VL-3B-Instruct, however, the merging operation cannot be performed due to the absence of a corresponding base LLM and math-specialized model with the same configuration.

Training Datasets. To enhance the mathematical reasoning capability of ICoM, we employ a 54K dataset constructed from MINT-CoT, where the reasoning steps are annotated with corresponding grid indices. Each sample consists of a math problem and an image as input, together with both a text-only CoT and an interleaved visual CoT as output. This dataset supports fine-grained visual reasoning, overcomes the limitations of bounding box based region selection, and provides the data foundation for our three-stage training scheme.

Evaluation protocols. To evaluate the effectiveness of ICoM, we conduct experiments on six multimodal benchmarks, grouped into two categories. The mathematical reasoning benchmarks include MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024a), MathVision (Wang et al., 2024a), and MMMath (Sun et al., 2024), while the comprehensive benchmarks consist of MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024). Additionally, we compare ICoM against three types of baselines: closed-source, open-source general and open-source reasoning VLLMs. All evaluations are conducted using the vlmevalkit framework (Duan et al., 2024) for consistency and reproducibility. For most benchmarks, we follow the framework's original evaluation pipeline. For tasks where answer extraction and correctness could not be determined by exact matching, we adopt GPT-4o-mini or GPT-4-0125 as an LLM-as-a-Judge.

Model	Mathematical Benchmarks				Comprehensive Benchmarks		Overall
	MathVista	MathVerse	MathVision	MMMath	MMStar	MMMU	Avg_{All}
Closed-Source Models							
GPT-4o-0513 (Hurst et al., 2024)	63.8	50.2	30.4	31.8	64.7	69.1	51.67
Claude 3.5 Sonnet-620 (Anthropic, 2024)	67.7	–	–	–	65.1	68.3	–
GPT-5-Mini (OpenAI, 2025)	59.6	36.5	46.6	–	61.3	67.9	–
Open-Source General Models (2B–8B)							
Qwen2-VL-7B (Wang et al., 2024b)	58.2	–	16.3	–	60.7	54.1	–
Qwen2.5-VL-7B (Bai et al., 2025b)	68.2	49.2	25.1	–	63.9	58.6	–
InternVL3-2B (Zhu et al., 2025)	57.0	25.3	21.7	–	60.7	48.6	–
InternVL3-8B (Zhu et al., 2025)	71.6	39.8	29.3	–	68.2	62.7	–
Open-Source Reasoning Models (7B–8B)							
R1-VL-7B (Zhang et al., 2025a)	63.5	40.0	24.7	–	60.0	–	–
Vision-R1-7B (Huang et al., 2025)	73.5	52.4	–	40.2	–	–	–
OpenVLThinker-7B (Deng et al., 2025)	72.3	50.3	25.9	–	–	42.9	–
VLAA-Thinker-7B (Chen et al., 2025b)	59.6	–	19.8	–	–	–	–
LLaVA-Reasoner-8B (Zhang et al., 2025b)	50.6	–	–	–	54.0	40.0	–
Mulberry-2B (Yao et al., 2024)	51.7	–	–	13.9	51.3	42	–
Mulberry-7B (Yao et al., 2024)	63.1	–	–	23.7	61.3	55	–
MINT-CoT-2B (Chen et al., 2025e)	61.9†	30.91†	24.18†	13.83†	62.4†	54.29†	41.25†
Qwen2.5-VL-3B (Baseline)	62.3	47.6	21.2	13.73†	55.9	53.1	42.31
+ Three-stage training	74.11	49.53	33.06	25.34	66.2	61.6	51.64
Δ over the Baseline Model	+11.81	+1.93	+11.86	+11.61	+10.3	+8.5	+9.33
Qwen2-VL-2B (Baseline)	43.0	15.95†	12.4	2.8†	48.0	41.1	27.21
+ Three-stage training	65.1	33.78	25.49	15.35	63.83	55.7	43.21
Δ over the Baseline Model	+22.1	+17.83	+13.09	+12.55	+15.83	+14.6	+16
+ Merge (ICoM-2B)	65.93	38.19	30.5	22.34	64.92	59.56	46.91
Δ over the Baseline Model	+22.93	+22.24	+18.1	+19.54	+16.92	+18.46	+19.7

Table 1: Comparison of our three-stage trained models and the merged model (ICoM-2B) against closed-source, open-source general, and open-source reasoning VLLMs across mathematical and comprehensive benchmark suites (higher is better). † scores are obtained from our own evaluations using a unified evaluation protocol.

4.2 Main Results

Three-Stage ICoM Training Boosts CoT Reasoning. We first conduct experiments on the baseline models Qwen2-VL-2B and Qwen2.5-VL-3B in three stages, combining SFT and RL. Figure 3 shows the training loss curves of Qwen2-VL-2B during text-only CoT SFT and interleaved CoT SFT. A detailed analysis of these training losses is provided in the Appendix B. As shown in Table 1, we observe that three-stage training brings clear performance improvements over the two baselines, i.e., +16% and +9.33% on averaged over six benchmarks, confirming the effectiveness of this training strategy in enhancing ICoM's multimodal reasoning capability. Under the same data and training settings, our Q-Former based interleaved selection outperforms the cosine based static selection in MINT-CoT across all benchmarks, yielding an overall average improvement of 1.96% and indicating that adaptive multi-view visual focusing is better suited for multi-step CoT reasoning.

Furthermore, Qwen2-VL-2B trained in three stages improves over its baseline on the mathematical benchmarks MathVista, MathVerse, and MMMath by 22.1%, 17.83% and 12.55%, respectively. Although the training data are predominantly mathematical, the model achieves gains of 15.83% and 14.6% on the general benchmarks MMStar and MMMU, validating that this strategy not only substantially strengthens mathematical CoT reasoning but also enhances general visual understanding and cross task generalization, without sacrificing visual perception or alignment. Compared with state-of-the-art models, including both open- and closed-source models, Qwen2.5-VL-3B trained in three stages surpasses most open-source models on multiple mathematical benchmarks, demonstrating its superior CoT reasoning capability.

ICoM-2B is Competitive with State-of-the-Art VLLMs. After completing three-stage training on Qwen2-VL-2B, we further apply linear merging between its language component and the math-

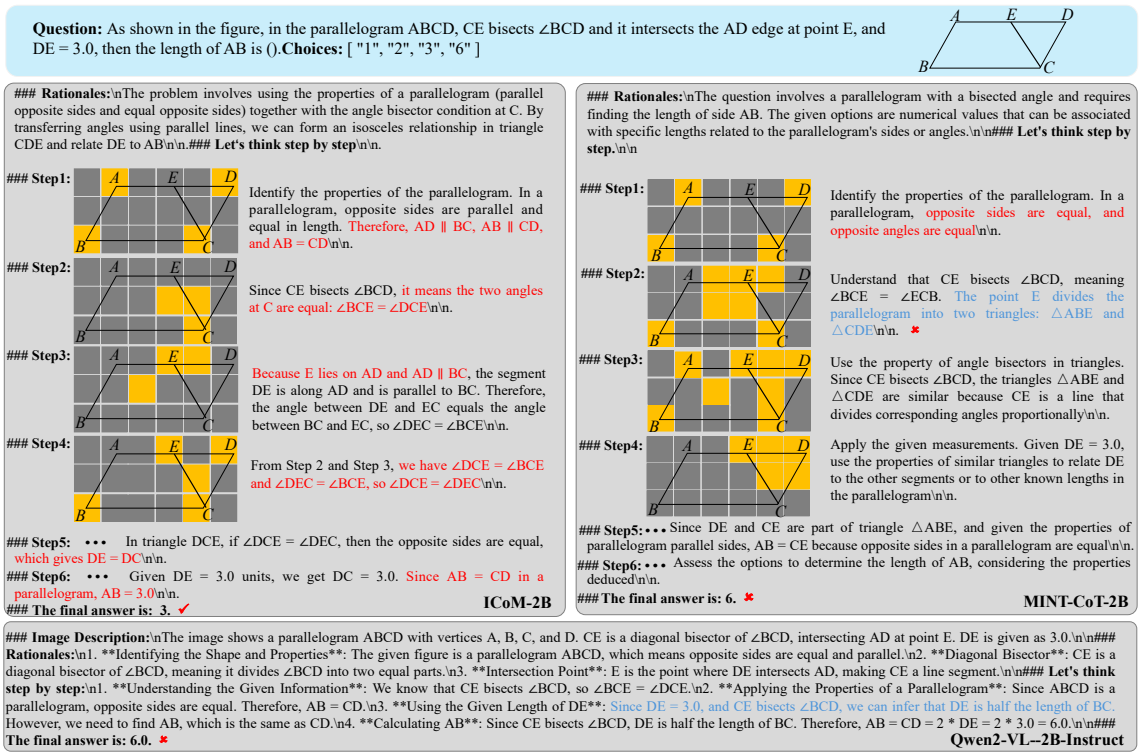


Figure 2: Qualitative results of Qwen2-VL-2B-Instruct, MINT-CoT-2B and ICoM-2B. ICoM-2B exhibits stronger CoT reasoning with more accurate visual-region focusing.

specialized LLM Qwen2-Math-1.5B-Instruct at layers 19–27, yielding ICoM-2B. Detailed ablations on merging using the interleaved CoT RL stage trained model are provided in the Appendix C. We benchmark ICoM-2B against representative state-of-the-art models (e.g., closed-source, open-source general and open-source reasoning models). As shown in Table 1, ICoM-2B outperforms GPT-4o-0513, Qwen2-VL-7B, and R1-VL-7B by 2.13%, 7.73% and 2.43% on the mathematical benchmark MathVista, and by 0.22%, 4.22% and 4.92% on the comprehensive benchmark MMStar. Moreover, ICoM-2B overall surpasses LLaVA-Reasoner-8B, Mulberry-7B. These results indicate that, despite its smaller parameter size, ICoM-2B remains competitive with larger state-of-the-art models. For example, Figure 2 shows that ICoM-2B consistently focuses on key regions at each reasoning step (e.g., point C and segment CE). In contrast, MINT-CoT-2B exhibits more diffuse attention and fails to capture key geometric cues, resulting in an incorrect answer.

4.3 Ablation Study

Effect of different training stages. We conduct an ablation study using Qwen2-VL-2B as the baseline to further investigate the contribution of each stage within the ICoM framework. As shown in Table 2, model performance increases monotonically across the three stages. The text-only CoT SFT stage substantially improves average performance by 11.94%. From Table 4, we observe consistent improvements across all five MathVerse categories. Gains are most pronounced in the text-dominant categories (TL and TD), indicating that text-only CoT supervision helps the model internalize general reasoning structures and improves its core reasoning ability. After introducing interleaved CoT SFT, the average performance across benchmarks improves by 2.33%, with larger gains in VO, VD, and VI, confirming the effectiveness of interleaved visual tokens for fine-grained visual reasoning. Finally, interleaved CoT RL further increases the overall average to 43.21% and yields an additional 0.2%–2% gain across all MathVerse categories, resulting in more robust mathematical reasoning.

Effect of merging at different layer ranges. As a core component of our proposed ICoM, we investigate layer-specific merging of the LM. After completing the three-stage training on Qwen2-VL-2B, we linearly merge its LM with a math-

Method	Mathematical Benchmarks				Comprehensive Benchmarks		Overall
	MathVista	MathVerse	MathVision	MMMath	MMStar	MMMU	Avg_{All}
Qwen2-VL-2B (Baseline)	43.0	15.95†	12.4	2.8†	48.0	41.1	27.21
+Text-only CoT SFT	60.10	30.83	21.54	11.62	60	50.80	39.15
+Interleaved CoT SFT	63.6	32.88	23.03	13.74	62.29	53.33	41.48
+Interleaved CoT RL	**65.1**	**33.78**	**25.49**	**15.35**	**63.83**	**55.7**	**43.21**

Table 2: Ablation study of the three progressive training stages on different benchmarks.

Method	Mathematical Benchmarks				Comprehensive Benchmarks		Overall
	MathVista	MathVerse	MathVision	MMMath	MMStar	MMMU	Avg_{All}
+Three-stage training	65.1	33.78	25.49	15.35	63.83	55.7	43.21
Layer0-Layer8	64.65	35.73	26.88	20.2	63.8	56.02	44.55
Layer9-Layer18	**66.55**	36.59	28.2	19.25	64.13	58.16	45.48
Layer19-Layer27	65.93	**38.19**	**30.5**	**22.34**	**64.92**	**59.56**	**46.91**
Layer0-Layer27	66	35.7	26.06	16.3	64.04	56.56	44.11

Table 3: Ablation of layer wise merging (i.e., early 0–8, middle 9–18, and late 19–27) applied to the LM component of a three stage trained Qwen2-VL-2B (Θ_{rl}) with a math specialized LLM.

Method	VO	VD	TL	TD	VI	ALL
Qwen2-VL-2B	13.58†	13.30†	16.14†	20.6†	16.11†	15.95†
+Text-only SFT	25.71	30.41	33.67	32.64	31.74	30.83
+Interleaved SFT	28.72	31.74	**35.09**	35.37	33.50	32.88
+Interleaved RL	**28.75**	**33.38**	34.62	**36.38**	**35.75**	**33.78**

Table 4: Quantitative results on MathVerse (Overall, Text-Dominant (TD), Text-Lite (TL), Vision-Integrated (vI), Vision-Dominant (VD), and Vision-Only (VO) categories) for the three progressive training stages.

λ	MathVista	MMStar	MathVision	MMMU
0.8	50.9	46.87	20.39	49.78
0.85	58.7	63.27	24.38	54.8
0.9	65.93	64.92	30.5	59.56

Table 5: Ablation of the layer-wise task vector merging hyperparameter λ.

specialized LLM at different layer ranges, including early layers (0–8), middle layers (9–18), and late layers (19–27). Table 3 shows that, from a layer-wise perspective, merging the middle-to-late layers overall outperforms early-layer merging. Specifically, while early-layer merging increases the overall average from 43.21% to 44.55%, it leads to slight drops on MathVista and MMStar. Merging the middle layers further improves the average to 45.48%. The best performance is achieved by merging the late layers (19–27), with an overall average of 46.91%, delivering the strongest results across benchmarks. For example, MathVerse improves from 35.73% to 38.19%, indicating that injecting mathematical priors primarily into later layers is beneficial for strengthening symbolic reasoning and multi-step CoT capability. In contrast, merging all layers (0–27) underperforms late layer merging. These results suggest that injecting external mathematical knowledge primarily into late reasoning layers, while keeping early layer multimodal representations stable, is a more effective layer wise merging strategy.

Effect of the merging hyperparameter. To analyze the effect of the hyperparameter λ in layer-specific merging, we conduct an ablation study across multiple benchmarks. We vary the parameter in 0.05 increments over $\{0.80, 0.85, 0.90\}$. As shown in Table 5, performance improves consistently with increasing λ, and peaks at $\lambda = 0.9$ for ICoM-2B. In contrast, $\lambda = 0.8$ leads to a significant degradation across all benchmarks, suggesting that overly strong injection of the math task shift can overwrite the multimodal representations learned during three stage training.

5 Conclusion

This paper presents ICoM, a coupled framework that integrates interleaved visual CoT driven adaptive visual focusing with layer-specific merging to enhance mathematical reasoning. It addresses two key challenges of grounding critical visual regions and mitigating limited internal reasoning capability. ICoM achieves competitive performance and generalization compared with state-of-the-art VLLMs. Future work will explore the reasoning mechanisms of VLLMs.

Limitations

Although ICoM improves visual focusing and mathematical reasoning, it has several limitations. The gains from our three-stage SFT+RL pipeline may depend on the scale and distribution of the training data and the reward design, potentially reducing robustness under domain shifts. Layer-specific merging also typically requires architecture-compatible counterparts (e.g., matched layer configurations and parameterization), which can limit applicability when aligned math-specialized variants are unavailable and may introduce additional hyperparameter tuning and inference overhead. In future work, we will further explore more data efficient and architecture agnostic alternatives.

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

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

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

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025a. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6688–6706.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025b. SFT or rl? an early investigation into training r1-like reasoning large vision-language models. *CoRR*, abs/2504.11468.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025c. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. Accessed: 2025-02-02.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025d. Bring reason to vision: Understanding perception and reasoning through model merging. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. 2025e. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *CoRR*, abs/2506.05331.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/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*.

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.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19520–19529. Computer Vision Foundation / IEEE.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? EMMA: an enhanced multimodal reasoning benchmark. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

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. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Alex Clark, Adam Ostrow, Anindya Welihinda, Alex Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulic, and Furu Wei. 2025a. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. 2025b. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9098–9108. Computer Vision Foundation / IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. 2025c. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18798–18806. AAAI Press.

OpenAI. 2025. GPT-5 mini model. OpenAI API documentation. Model identifier: `gpt-5-mini`. Snapshot: `gpt-5-mini-2025-08-07`.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024a. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Cheng Shi, Yizhou Yu, and Sibei Yang. 2025. Vision function layer in multimodal llms. *CoRR*, abs/2509.24791.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025a. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *CoRR*, abs/2505.15966.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 2025b. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *CoRR*, abs/2505.08617.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juan-Zi Li. 2024. MM-MATH: advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1358–1375. Association for Computational Linguistics.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *CoRR*, abs/2504.08837.

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, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective monte carlo tree search. *CoRR*, abs/2412.18319.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9382–9391. Computer Vision Foundation / IEEE.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *CoRR*, abs/2504.07954.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Jingyi Zhang, Jiaxing 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. *CoRR*, abs/2503.12937.

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. 2024a. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pages 169–186. Springer.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025b. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1631–1662. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 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)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *CoRR*, abs/2505.14362.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *CoRR*, abs/2405.13872.

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, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479.

A Details of the Interleaved CoT SFT Loss

We define the selection probability of the k-th visual token at step j as $p_{j,k} = \sigma(z_{j,k})$. The element-wise loss is defined as:

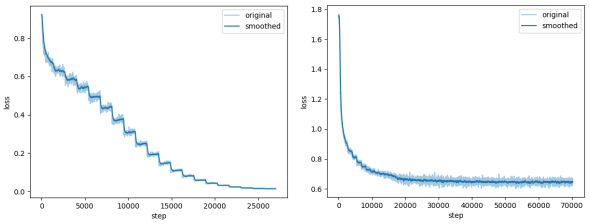
$$\ell_{j,k} = -\Big[\alpha_j \, y_{j,k} \, \log p_{j,k} + \big(1 - y_{j,k}\big) \log\big(1 - p_{j,k}\big)\Big],$$

here, $y_{j,k} \in \{0, 1\}$ is the supervision label, where $y_{j,k} = 1$ indicates that the k-th visual token is selected at step j. To mitigate the severe class imbalance, where positive instances (selected tokens) are far fewer than negative ones, we introduce a positive-class weighting factor:

$$\alpha_j = \frac{N_j}{P_j + \varepsilon},$$

Method	Mathematical Benchmarks				Comprehensive Benchmarks		Overall
	MathVista	MathVerse	MathVision	MMMath	MMStar	MMMU	Avg_{All}
Merge $(\Theta_{\text{text}}, \Theta_{\text{base}})$	61.3	32.63	22.22	12	61.27	52	40.24
Merge $(\Theta_{\text{inter}}, \Theta_{\text{base}})$	64.24	33.21	23.57	14.1	62.73	54.33	42.03
Merge $(\Theta_{\text{rl}}, \Theta_{\text{base}})$	**66**	**35.7**	**26.06**	**16.3**	**64.04**	**56.56**	**44.11**

Table 6: Ablation across benchmarks of merging Qwen2-Math-1.5B-Instruct into models obtained after each progressive training stage.



(a) (b)

Figure 3: Training loss curves for (a) Text-only CoT SFT and (b) Interleaved CoT SFT.

where $m_{j,k} \in \{0, 1\}$ is a validity mask that restricts the loss computation to positions corresponding to actual visual tokens, Here, $P_j = \sum_k m_{j,k} y_{j,k}$ and $N_j = \sum_k m_{j,k}(1 - y_{j,k})$ denote the numbers of positive and negative instances at valid positions, respectively, and $z_{j,k} = log(w_{j,k} P_j + \epsilon)$. To enhance training stability, we clip the element-wise loss $\tilde{\ell}_{j,k} = \text{clip}(\ell_{j,k}, 0, 10)$, which suppresses gradient outliers caused by extreme mispredictions or large values of α_j.

B Analysis of Training Loss

As illustrated in Figure 3(a), during text-only CoT SFT, the training loss of Qwen2-VL-2B decreases steadily with training steps and converges to near zero after approximately 2.6×10^4 steps, indicating that the model can learn the language patterns of text-only CoT. In Figure 3(b), during interleaved CoT SFT, the loss drops rapidly at early steps and then plateaus, converging at ~ 0.62. Compared with Stage 1, this stage uses the Q-Former with interleaved tokens to dynamically attend to key visual regions, better supporting vision–text interleaved reasoning and improving both textual reasoning and visual alignment.

C Ablation of Merging at Different Training Stages

To evaluate the effectiveness of merging the language component of Qwen2-VL-2B with the math-specialized model Qwen2-Math-1.5B-Instruct at different training stages, we apply the same merging strategy to the models obtained after Stages 1–3 (i.e., Θ_{text}, Θ_{inter} and Θ_{rl}) and compare them across multiple benchmarks. As shown in Table 6, the gains from merging consistently increase as training progresses, and the Stage 3 model achieves the best results, improving the average performance from 40.24% to 44.11%. The trend is more pronounced on mathematical benchmarks (e.g., MathVista, MathVision, and MMMath), which improve steadily across stages. Compared with Table 2, we observe that the Stage 3 model exhibits more substantial gains from merging beyond those achieved by training alone. This suggests that sufficient interleaved reasoning training and RL yield more robust reasoning representations and multimodal alignment, enabling more effective absorption of external mathematical priors with reduced negative transfer. In particular, we adopt the Stage 3 model for our subsequent layer-specific merging.