# Reasoning-Visual Critical Token Fine-Tuning for Multimodal Reasoning

**Anonymous submission**


## Abstract

Supervised fine-tuning (SFT) on multimodal chain-of-thought (CoT) data is widely used to improve reasoning in multimodal large language models (MLLMs). However, standard SFT applies a uniform loss to all response tokens, ignoring the unequal roles of different tokens in reasoning. In multimodal scenarios, useful supervision positions should reflect not only reasoning state transitions but also sensitivity to the visual input. In this paper, we propose RVCFT, a reasoning-visual critical token fine-tuning method for multimodal reasoning. RVCFT identifies key supervision positions using two complementary signals: it estimates reasoning relevance via conditional log-likelihood differences between a reasoning-specialized text model and a general instruction model, and measures reference-based visual sensitivity through absolute token log-likelihood changes under real versus reference images. The two signals are fused to construct a binary critical mask. During fine-tuning, the complete response remains in the teacher-forced context, while direct cross-entropy supervision is applied only at the selected positions. With a target response-token retention ratio of $50\%$, RVCFT achieves the highest average score among standard SFT and the considered token-selection baselines across multiple multimodal reasoning benchmarks, supporting the effectiveness of reasoning-visual critical supervision.


## Introduction

Supervised fine-tuning (SFT) on multimodal chain-of-thought (CoT) data has enhanced the reasoning capabilities of multimodal large language models (MLLMs) on complex visual tasks (Wei et al. 2022; Zhang et al. 2023; Chen et al. 2024b; Bai et al. 2025a). However, standard SFT assigns uniform cross-entropy loss to all response tokens, ignoring their unequal roles in reasoning. In practice, different tokens serve distinct functions, including visual grounding, intermediate deduction, and final-answer formation, and therefore contribute differently to logical correctness. Long reasoning chains also contain substantial semantic redundancy, which can dilute informative supervision and interfere with coherent reasoning (Choi et al. 2025; Yuan et al. 2025). Moreover, prior work suggests that a relatively small subset of decision points can substantially influence the direction of reasoning and state transitions (Lin et al. 2024; Wang et al. 2026a). These observations motivate token-level selective supervision that concentrates direct loss on key state-transition positions to improve multimodal fine-tuning.

However, applying existing token-selection strategies to multimodal reasoning presents two major challenges: it is hard to pinpoint reasoning bifurcations without explicit logical priors, and language-only criteria do not capture whether token predictions respond to visual input. Common token-level heuristics rely on signals such as predictive entropy (Wang et al. 2026a) or attention-based importance measures (Vaswani et al. 2017). However, token-level uncertainty or attention alone may not reliably distinguish complex reasoning transitions from ordinary linguistic variation. Moreover, multimodal CoT reasoning requires critical positions not only to carry logical deductions but also to respond to relevant visual evidence (Lu et al. 2024; Chen et al. 2024a). Unconstrained extensions of the reasoning chain can encourage over-reliance on language priors while underusing visual input, potentially producing "spurious reasoning trajectories" and visual hallucinations inconsistent with the image (Tu et al. 2026; Bi et al. 2026; Fang et al. 2026). Neither language-only reasoning signals nor visual-sensitivity signals alone are sufficient to characterize positions that are both reasoning-relevant and responsive to visual content. These limitations motivate jointly modeling token-level reasoning relevance and reference-based visual sensitivity for multimodal selective fine-tuning.

To this end, we propose RVCFT (Reasoning-Visual Critical Token Fine-Tuning), a dual-signal selective fine-tuning method that combines reasoning gain with reference-based visual sensitivity. As illustrated in Figure 1, RVCFT consists of an offline critical-mask construction stage and a mask-guided fine-tuning stage. During mask construction, RVCFT estimates token-level reasoning gain using the aligned conditional log-likelihood difference between a reasoning-specialized model and a general instruction model, providing a proxy for reasoning relevance. In parallel, it measures visual sensitivity through the magnitude of token log-likelihood changes under the real image and a fixed solid-white reference image. The two scores are normalized and fused to construct a constrained top-ratio mask with final-answer protection. During fine-tuning, the complete re-

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

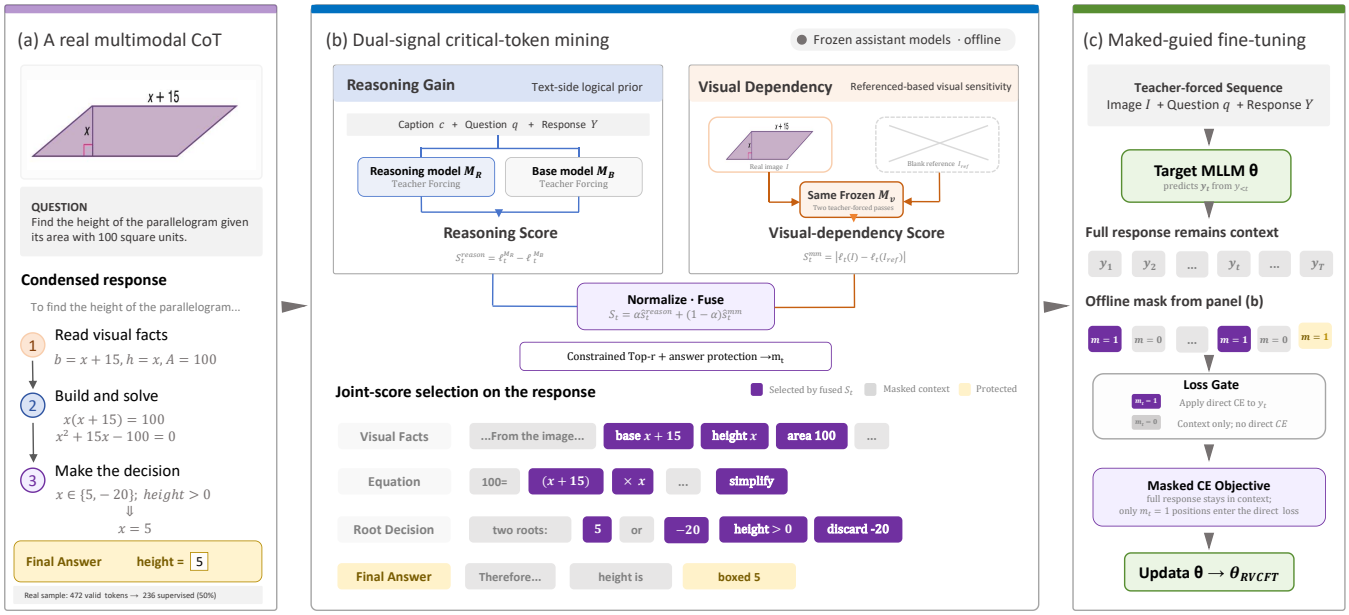


Figure 1: Overview of Reasoning-Visual Critical Token Fine-Tuning (RVCFT). (a) An illustrative multimodal CoT response contains visual facts, intermediate deductions, decision steps, and a protected final answer. (b) Frozen scoring models produce aligned reasoning-gain and reference-based visual-sensitivity scores. The normalized scores are fused to construct a constrained top-ratio mask with final-answer protection. (c) During fine-tuning, the complete response remains in the teacher-forced context, while direct cross-entropy supervision is applied only at the selected positions.

sponse remains in the teacher-forced context, while direct cross-entropy supervision is applied only at the selected positions. Experiments on MathVista (Lu et al. 2024), MathVision (Wang et al. 2024), MathVerse (Zhang et al. 2024), and WeMath (Qiao et al. 2025) show that, with a target response-token retention ratio of $50\%$, RVCFT achieves the highest overall average score among the compared methods, outperforming standard SFT and the considered token-selection baselines in average performance.

Our main contributions are summarized as follows:

- We formulate a selective-supervision approach for multimodal CoT fine-tuning that retains the complete response as context while applying direct token-level loss only at selected response positions.
- We propose RVCFT, which constructs critical-token masks by combining caption-conditioned reasoning gain with reference-based visual sensitivity.
- Experiments across multiple multimodal reasoning benchmarks show that RVCFT achieves the highest overall average performance among standard SFT and the considered token-selection baselines under the reported settings, while remaining effective on multidisciplinary reasoning tasks.

## Related Work

### Multimodal Reasoning

In recent years, MLLMs have made significant progress on vision-language understanding and interaction tasks through vision-language alignment and visual instruction tuning (Chen et al. 2024b; Bai et al. 2025a). However, multimodal reasoning requires the model not only to perceive image content, but also to complete verifiable multi-step deduction on the basis of cross-modal alignment (Wei et al. 2022; Zhang et al. 2023). Recent works reveal that extended thinking chains in MLLMs do not unconditionally translate to performance gains; instead, unconstrained CoT generation often causes models to over-rely on textual language priors while ignoring visual inputs, exacerbating visual hallucination (Tu et al. 2026; Bi et al. 2026; Fang et al. 2026). To address this, Lookback (Bi et al. 2026) guides models to refocus on visual inputs via adaptive lookback prompts during inference, while C3PO (Fang et al. 2026) compresses CoT traces and applies contrastive preference optimization to suppress non-grounded hallucination. Unlike these inference-time decoding prompts or contrastive preference alignment techniques, our RVCFT focuses on the allocation of SFT supervision signals, selectively fine-tuning response tokens that possess both high reasoning value and strong reference-based visual sensitivity.

### Token-Level Supervision

Standard SFT applies uniform loss to all output tokens. However, recent studies demonstrate that effective supervision concentrates on a small subset of critical tokens while ignoring redundant ones (Lin et al. 2024; Choi et al. 2025; Yuan et al. 2025). In textual LLMs, PGTG (Yang et al. 2023) and TDPO (Zeng et al. 2024) conduct token-granularity preference optimization. Wang et al. (2026a) reveal that a few high-entropy tokens dominate reasoning improvements. Fur-

thermore, Choi et al. (2025) demonstrate that pruning redundant tokens via attention patterns removes reasoning distraction, while CTS (Yuan et al. 2025) demonstrates that training models on conditionally compressed CoTs achieves superior efficiency without performance degradation. For MLLMs, ReGATE (Li, Kulkarni, and Fazli 2026) filters out redundant tokens by leveraging sample difficulty estimation. Grounded in these insights, RVCFT advances token-level selective supervision into multimodal reasoning by jointly quantifying caption-conditioned reasoning gain and reference-based visual sensitivity.

# Methodology

## Problem Definition

In this study, we focus on selective fine-tuning for MLLMs. The core idea is to assign a binary supervision mask to response tokens before fine-tuning, such that direct cross-entropy supervision is applied only to selected positions. This modifies standard SFT by concentrating direct supervision on critical reasoning positions.

Formally, we are given a multimodal instruction-tuning dataset $\mathcal{D} = \{(I_i, q_i, Y_i)\}_{i=1}^{N}$, where $I_i$ and $q_i$ denote the input image and textual question, respectively. For a target response sequence $Y = (y_1, \ldots, y_T)$ of length $T = |Y|$, we formulate critical token fine-tuning using a binary selection mask $m_t \in \{0, 1\}$ for each token position $t$. The token-masked training objective is defined as:

$$\mathcal{L}(\theta) = -\frac{\sum_{(I,q,Y)\in\mathcal{D}} \sum_{t=1}^{|Y|} m_t \log P_\theta(y_t \mid I, q, y_{<t})}{\sum_{(I,q,Y)\in\mathcal{D}} \sum_{t=1}^{|Y|} m_t}, \tag{1}$$

where $m_t = 1$ indicates that direct token-level supervision is applied at position $t$, and $m_t = 0$ masks out its loss gradient. Standard SFT corresponds to the special case where $m_t = 1$ for all $t \in \{1, \ldots, T\}$.

## Motivation

Multimodal CoT responses contain different types of tokens, including visual facts, reasoning steps, linguistic connectors, and final answers. However, these tokens do not share equal training value. Wang et al. (2026a) discover that in CoT reasoning and reinforcement learning for LLMs, a few high-entropy tokens often correspond to critical reasoning bifurcations and dominate optimization gains, indicating that effective supervision should focus more on key reasoning tokens. Nevertheless, relying solely on language-side reasoning signals remains insufficient in multimodal reasoning. M3CoT (Chen et al. 2024a) points out that multimodal CoT requires models to leverage both textual and visual information for step-by-step reasoning. Meanwhile, recent studies on MLLM hallucination show that models still over-rely on language priors and generate content inconsistent with images (Tu et al. 2026). Therefore, we argue that critical tokens in multimodal reasoning must be both reasoning-informative and visually sensitive, and we estimate their criticality by fusing reasoning gain with reference-based visual sensitivity.

## Reasoning-Visual Critical Token Identification

Given the target response sequence $Y = (y_1, y_2, \ldots, y_T)$, we compute a reasoning-visual critical score for each token. Because different assistant models ($M_R, M_B, M_V$) and the target model $\theta$ may employ distinct tokenizers, all log-probabilities from the assistant models are mapped back to the target sequence $Y$ via character-level span offset matching. If a target token overlaps with multiple source tokens, the source token scores are combined using a weighted average based on the overlap length. For notational simplicity, all token-level log-probabilities below refer to scores after alignment to the corresponding target-token positions.

**Reasoning Gain.** Reasoning gain estimates the reasoning value of a token. We use the likelihood gap between a reasoning-specialized model $M_R$ and a general instruction model $M_B$ as a proxy for token-level reasoning relevance. Intuitively, tokens that are more preferred by $M_R$ than by $M_B$ are more likely to correspond to reasoning-intensive steps rather than generic response continuation. Since both $M_R$ and $M_B$ are text-only models, we convert the image $I$ into a text caption $c$ in advance as an alternative context. Under the same textual context (consisting of $c$, question $q$, and the response prefix $y_{<t}$), we apply teacher-forcing on the target response using $M_R$ and $M_B$. The aligned log-probability score of the $t$-th token under model $M$ is defined as:

$$\ell_t^M = \log P_M(y_t \mid c, q, y_{<t}), \quad M \in \{M_R, M_B\}. \tag{2}$$

The reasoning gain score $s_t^{\text{reason}}$ is expressed as:

$$s_t^{\text{reason}} = \ell_t^{M_R} - \ell_t^{M_B}. \tag{3}$$

A larger $s_t^{\text{reason}}$ indicates that the token is more strongly preferred by the reasoning model and possesses stronger reasoning relevance.

**Visual Sensitivity.** Relying solely on reasoning gain cannot reveal whether a token prediction changes in response to the image. To capture this property, we use a frozen multimodal scoring model $M_V$ to compare the teacher-forced log-probabilities under the real image $I$ and a reference image $I_{\text{ref}}$:

$$\ell_t(I_{\text{in}}) = \log P_{M_V}(y_t \mid I_{\text{in}}, q, y_{<t}), \quad I_{\text{in}} \in \{I, I_{\text{ref}}\}. \tag{4}$$

Here, $I_{\text{ref}}$ is a reference RGB image of matching spatial dimensions with all pixel values uniformly set to 255 (i.e., a solid white image), providing a fixed reference for measuring how token likelihoods change when the original visual content is removed. We define the reference-based visual-sensitivity score $s_t^{\text{mm}}$ as:

$$s_t^{\text{mm}} = \left|\ell_t(I) - \ell_t(I_{\text{ref}})\right|. \tag{5}$$

We use the magnitude of the likelihood change rather than its positive part because the one-sided score maps all negative likelihood changes to zero and can therefore produce a sparse visual score distribution, discarding tokens that are visually responsive in the negative direction. Thus, $s_t^{\text{mm}}$ measures the sensitivity of a token prediction to the visual input, irrespective of the direction of the change.

**Score Fusion.** Since the scales of the two scores are different, we apply min-max normalization within each individual sample:

$$\hat{s}_t^k = \frac{s_t^k - \min_{1\le j\le T} s_j^k}{\max_{1\le j\le T} s_j^k - \min_{1\le j\le T} s_j^k + \epsilon}, \quad k \in \{\text{reason}, \text{mm}\}. \tag{6}$$

Here, $\epsilon > 0$ is a small constant used for numerical stability. The final joint critical score $S_t$ is defined as:

$$S_t = \alpha \hat{s}_t^{\text{reason}} + (1-\alpha)\hat{s}_t^{\text{mm}}, \tag{7}$$

where $\alpha \in [0,1]$ is the balancing coefficient. This score characterizes both the reasoning value and the visual sensitivity of the token.

## Critical Mask Construction

Given the joint scores $S = (S_1, \ldots, S_T)$, RVCFT models the critical token selection as a constrained top-ratio selection. We first define a protection set $\mathcal{P}$ to preserve the final answer region. If final-answer boundary markers such as `</think>`, `\boxed`, or `<answer>` appear in the response, we locate the final-answer region using the answer extractor and include all tokens from the detected answer boundary to the end of the response in $\mathcal{P}$.

The remaining tokens with finite joint scores form the candidate set:

$$\mathcal{C} = \{t \in \{1,\ldots,T\} \mid t \notin \mathcal{P},\ S_t \text{ is finite}\}. \tag{8}$$

Given a critical ratio $r \in (0,1]$, the target number of preserved tokens is $K = \lceil r \cdot T \rceil$. Since the tokens in the protection set $\mathcal{P}$ are mandatory, we determine the number of additional candidate tokens $K_c$ to select from $\mathcal{C}$ via:

$$K_c = \min\left(|\mathcal{C}|, \max(0, K - |\mathcal{P}|)\right). \tag{9}$$

Because answer protection takes precedence over the target retention ratio, the actual number of selected tokens may exceed $K$ when $|\mathcal{P}| > K$. Finally, the critical token set $\mathcal{M}$ and its corresponding binary mask $m_t$ are constructed as:

$$\mathcal{M} = \mathcal{P} \cup \text{Top}_{K_c}(\mathcal{C}; S), \quad m_t = \mathbb{I}[t \in \mathcal{M}]. \tag{10}$$

## Reasoning-Visual Critical Fine-Tuning Objective

During the fine-tuning phase, RVCFT imposes the supervised loss only at the positions where the critical mask is activated:

$$\mathcal{L}_{\text{RVCFT}}(\theta) = -\frac{\sum_{(I,q,Y)\in\mathcal{D}} \sum_{t=1}^{|Y|} m_t \log P_\theta(y_t \mid I, q, y_{<t})}{\sum_{(I,q,Y)\in\mathcal{D}} \sum_{t=1}^{|Y|} m_t}. \tag{11}$$

Let $p_{t,v}$ denote the predicted probability of the model for vocabulary token $v$ at position $t$, and let $g_t$ denote the gold token. The gradient of this objective with respect to the logit $z_{t,v}$ satisfies:

$$\frac{\partial \mathcal{L}_{\text{RVCFT}}}{\partial z_{t,v}} \propto m_t \left(p_{t,v} - \mathbb{I}[v = g_t]\right). \tag{12}$$

Therefore, when $m_t = 0$, no direct token-level supervision loss is imposed at this position. When $m_t = 1$, the position contributes the standard token-level cross-entropy term, with the overall loss normalized over the selected positions. Consequently, RVCFT concentrates direct supervision signals on the critical decision tokens jointly determined by reasoning gain and reference-based visual sensitivity.

Algorithm 1: RVCFT Mask Construction and Fine-Tuning

**Input:** Dataset $\mathcal{D}$, initial target model $P_\theta$, frozen scorers $M_R, M_B, M_V$, and captioner $G$
**Parameters:** Fusion weight $\alpha$, retention ratio $r$, and constant $\epsilon$
**Output:** Fine-tuned target model $P_\theta$

$\mathcal{D}_m \leftarrow \varnothing$
**for all** $(I, q, Y) \in \mathcal{D}$ **do**
  $T \leftarrow |Y|, \quad c \leftarrow G(I)$
  $I_{\text{ref}} \leftarrow \text{SolidWhiteLike}(I)$
  $\ell^{M_R}, \ell^{M_B} \leftarrow \text{ReasoningScore}(c, q, Y)$
  $\ell^{I}, \ell^{\text{ref}} \leftarrow \text{VisualScore}(I, I_{\text{ref}}, q, Y)$
  $s^{\text{reason}} \leftarrow \ell^{M_R} - \ell^{M_B}$
  $s^{\text{mm}} \leftarrow |\ell^{I} - \ell^{\text{ref}}|$
  **for all** $k \in \{\text{reason}, \text{mm}\}$ **do**
    $\hat{s}^k \leftarrow \text{Norm}_\epsilon(s^k)$
  **end for**
  $S \leftarrow \alpha \hat{s}^{\text{reason}} + (1-\alpha)\hat{s}^{\text{mm}}$
  $\mathcal{P} \leftarrow \text{AnswerRegion}(Y)$
  $\mathcal{C} \leftarrow \{t \in \{1,\ldots,T\} \setminus \mathcal{P} : S_t \text{ is finite}\}$
  $K \leftarrow \lceil rT \rceil$
  $K_c \leftarrow \min(|\mathcal{C}|, \max(0, K - |\mathcal{P}|))$
  $\mathcal{M} \leftarrow \mathcal{P} \cup \text{Top}_{K_c}(\mathcal{C}; S)$
  $m \leftarrow (\mathbb{I}[t \in \mathcal{M}])_{t=1}^{T}$
  $\mathcal{D}_m \leftarrow \mathcal{D}_m \cup \{(I, q, Y, m)\}$
**end for**
Fine-tune $\theta$ on $\mathcal{D}_m$ by minimizing $\mathcal{L}_{\text{RVCFT}}$
**return** $P_\theta$

The complete two-stage procedure of RVCFT is summarized in Algorithm 1. In the algorithm, ReasoningScore computes the teacher-forced scores of $M_R$ and $M_B$ under the caption context, VisualScore computes the scores of $M_V$ under the real and reference images, and both operations include character-level token alignment. $\text{Norm}_\epsilon$ denotes the sample-wise min-max normalization defined above, while AnswerRegion returns the protected answer positions or $\varnothing$ when no valid boundary is detected. Critical-token scoring and mask construction are performed once offline, and the resulting masks are reused during fine-tuning.

# Experiments

## Setup

**Training Datasets and Models.** We conduct experiments on two multimodal reasoning training corpora. The primary training corpus is DRIFT (Huang et al. 2026). DRIFT is constructed by ThinkLite (Wang et al. 2026b) through CoT distillation from ThinkLiteVL-11K and low-quality sample filtering and contains 4K multimodal reasoning samples. To examine robustness with an alternative training corpus, we further construct a training corpus from the officially released MMMU (Yue et al. 2024) test split and its annotations. We

| Method | Math Vista | Math Vision | Math Verse | WeMath | Avg. |
|---|---|---|---|---|---|
| Org. | 69.40 | 26.57 | 39.59 | 35.90 | 42.87 |
| SFT | 69.80 | 27.30 | 42.76 | 36.86 | 44.18 |
| Attention | 69.60 | **29.61** | 43.78 | 36.19 | 44.80 |
| Entropy | 70.80 | 25.99 | **44.28** | 37.43 | 44.63 |
| RVCFT | **72.30** | 27.96 | 43.40 | **37.90** | **45.39** |

Table 1: Main results on multimodal reasoning benchmarks with Qwen2.5-VL-7B-Instruct. The average score is computed over the four benchmarks, and the best result in each column is highlighted in bold. "Org." denotes the model before task-specific fine-tuning.

repurpose this split solely as a source pool for supervised-data construction because it provides a larger candidate set; we do not use it as an evaluation set. Specifically, we use Qwen3-VL-30B-A3B-Instruct (Bai et al. 2025a) to generate chain-of-thought responses, retain only samples whose final answers match the released annotations, and normalize the final answers into the `\boxed{}` format, resulting in 3488 training examples. Evaluation is performed only on the official MMMU validation split, which has no sample overlap with the constructed training corpus. The target model is Qwen2.5-VL-7B-Instruct (Bai et al. 2025b). For critical token selection, we employ DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI 2025) and Qwen2.5-7B-Instruct (Qwen et al. 2025) as the reasoning and base models, respectively. Both the multimodal scoring model $M_V$ and the captioner $G$ use separate frozen copies of the original Qwen2.5-VL-7B-Instruct checkpoint. The captioner converts each image into a textual description once offline, whereas $M_V$ scores the target response under the real and reference images.

**Baselines.** We compare RVCFT with standard SFT and two token-selection baselines:

- **SFT**: applies uniform cross-entropy supervision to all response tokens.
- **Entropy** (Wang et al. 2026a): scores each response token by its predictive entropy $H_t = -\sum_v p(v \mid I, q, y_{<t}) \log p(v \mid I, q, y_{<t})$.
- **Attention** (Vaswani et al. 2017): averages attention weights across all layers and heads, and scores each response token by its mean incoming attention from itself and subsequent response positions.

For both token-selection baselines, scores are computed offline using the frozen initial target model, and the top-$\lceil rT \rceil$ response tokens are selected within each sample. All token-selection methods use the same target retention ratio $r$.

**Training Settings.** All fine-tuning experiments are completed using LLaMA-Factory (Zheng et al. 2024) on four NVIDIA A800 GPUs, each with 80 GB of memory. Unless otherwise specified, all experiments use full-parameter fine-tuning for 3 epochs with a learning rate of $1 \times 10^{-6}$ and a global batch size of 8. We construct a fixed validation set of 100 examples from the portion of the MathVista test split that does not overlap with testmini. This validation set is used exclusively for hyperparameter selection. Based on the validation accuracy, we set the fusion coefficient to $\alpha = 0.5$ and the target retention ratio to $r = 0.5$ for all reported experiments.

| Method | Math Vista | Math Vision | Math Verse | WeMath | Avg. |
|---|---|---|---|---|---|
| w/o VS | 70.20 | 25.98 | 40.22 | 32.95 | 42.34 |
| w/o RG | 69.90 | 27.63 | 41.75 | 36.67 | 43.99 |
| w/o AP | 71.80 | 25.32 | **45.55** | 35.62 | 44.57 |
| Random | 69.70 | 26.97 | 43.52 | 36.29 | 44.12 |
| RVCFT | **72.30** | **27.96** | 43.40 | **37.90** | **45.39** |

Table 2: Ablation study of different components in RVCFT. "VS", "RG", and "AP" denote Visual Sensitivity, Reasoning Gain, and Answer Protection, respectively. All variants use the same target response-token retention ratio.

**Evaluation Settings.** For the main evaluation, we use four multimodal mathematical reasoning benchmarks: MathVista (Lu et al. 2024), MathVision (Wang et al. 2024), MathVerse (Zhang et al. 2024), and WeMath (Qiao et al. 2025). We report results on the `testmini` subsets of MathVista and MathVision, the Vision-Only subset of MathVerse, and the Strict Score metric of WeMath. To further evaluate robustness to an alternative training corpus, we also report results on the official MMMU validation split. All evaluations are conducted using VLMEvalKit (Duan et al. 2024) with its benchmark-specific default configurations and greedy decoding (`do_sample=False`).

## Main Results

Table 1 presents the main experimental results of Qwen2.5-VL-7B-Instruct across four multimodal reasoning benchmarks. All token-selection methods use the same target response-token retention ratio of $50\%$. For RVCFT, we set the fusion coefficient to $\alpha = 0.5$. Overall, RVCFT achieves the highest average score among the considered methods, reaching 45.39 and outperforming the original model and standard SFT by 2.52 and 1.21 percentage points, respectively. These results suggest that, under our experimental setting, uniformly supervising all response tokens may be less effective than concentrating direct supervision on selected critical positions. Compared with the token-selection baselines, RVCFT outperforms Attention and Entropy by 0.59 and 0.76 percentage points in average performance, respectively, while achieving the best results on MathVista and WeMath. Although Attention performs best on MathVision and Entropy achieves the highest score on MathVerse, RVCFT obtains the highest overall average. This result supports the effectiveness of jointly modeling reasoning relevance and reference-based visual sensitivity for critical-token selection.

## Ablation Study

To evaluate the contribution of each core component in RVCFT, we conduct ablation experiments. All variants use

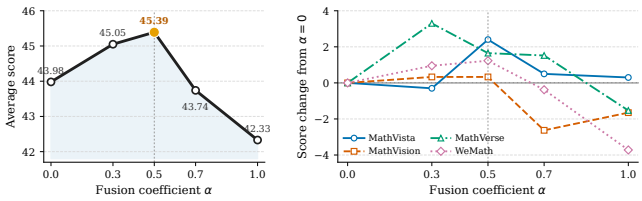


Figure 2: Effect of the fusion coefficient $\alpha$ in RVCFT. Left: Average score across the four benchmarks under different values of $\alpha$. Right: Score changes relative to $\alpha = 0$ on individual benchmarks.

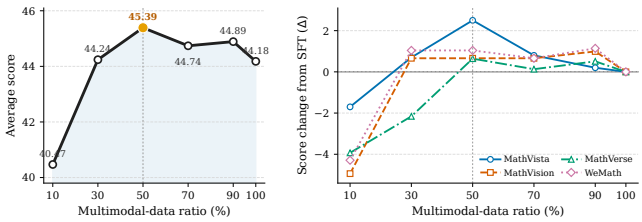


Figure 3: Effect of the target token retention ratio $r$ in RVCFT. Left: Average score across the four benchmarks under different values of $r$. Right: Performance on individual benchmarks for $r = 50\%$ and standard SFT ($r = 100\%$).

the same target retention ratio of 50% as in the main experiments. The results are shown in Table 2. RVCFT achieves the highest average score of 45.39. Removing Visual Sensitivity reduces the average score to 42.34, indicating that reasoning gain alone does not recover the performance of the complete dual-signal method. Removing Reasoning Gain decreases the average score to 43.99, indicating that visual sensitivity alone is also insufficient. When Answer Protection is removed, the average score drops to 44.57, supporting the importance of preserving direct supervision on answer-bearing tokens. Random Selection obtains an average score of 44.12, which is 1.27 percentage points lower than the full model. This result suggests that the gains of RVCFT cannot be attributed solely to reducing the number of supervised tokens and that the critical-token selection criterion contributes to its performance.

## Analyses

### Effect of Fusion Coefficient

We further analyze the impact of the fusion coefficient $\alpha$ on the performance of RVCFT. Under a fixed token retention ratio of $50\%$, varying $\alpha$ changes the relative contributions of reasoning gain and reference-based visual sensitivity to the joint critical score. As shown in Figure 2 (left), the average score first increases and then decreases as $\alpha$ increases, reaching its highest value of 45.39 at $\alpha = 0.5$. When using only visual sensitivity ($\alpha = 0$) or only reasoning gain ($\alpha = 1$), the average score decreases to 43.99 and 42.34, respectively. Figure 2 (right) further shows that $\alpha = 0.5$ outperforms both single-signal configurations across all four benchmarks. These results support the complementary contributions of reasoning gain and reference-based visual sensitivity to downstream performance.

### Effect of Token Retention Ratio

To analyze the impact of the target token-retention ratio $r$, we fix the fusion coefficient at $\alpha = 0.5$ and vary $r$ from $10\%$ to $100\%$, where $r = 100\%$ retains all response-token loss terms and is equivalent to standard SFT. As shown in Figure 3, the average performance changes non-monotonically with $r$ and reaches its highest value at $r = 50\%$. Specifically, as $r$ increases from $10\%$ to $50\%$, the average score improves from 40.47 to 45.39, suggesting that overly sparse direct supervision may exclude useful training signals. At $r = 50\%$,

| | Reference | | Token Selection | | |
|---|---|---|---|---|---|
| **Split** | **Org.** | **SFT** | **Attn.** | **Entropy** | RVCFT |
| MMMU-val | 51.78 | 52.56 | 49.33 | 51.78 | **53.56** |

Table 3: Results on the MMMU validation split using the constructed MMMU-derived fine-tuning corpus. "Org." denotes the model before task-specific fine-tuning. The best result is highlighted in bold.

RVCFT outperforms standard SFT (44.18) by 1.21 percentage points and reaches 72.30 and 43.40 on MathVista and MathVerse, respectively. It also improves over standard SFT across all four benchmarks, indicating consistent gains under this setting. At larger retention ratios, $r = 70\%$ and $r = 90\%$, the average scores decrease from the peak to 44.74 and 44.89, respectively. These results suggest that a very low retention ratio may omit useful direct supervision, whereas a high retention ratio may reintroduce low-information or redundant loss terms. Among the evaluated ratios, $r = 50\%$ provides the best empirical trade-off between filtering redundant supervision and retaining useful training signals.

### Robustness to an Alternative Training Corpus

To evaluate the robustness of RVCFT to changes in the fine-tuning corpus, we retrain the model on our constructed MMMU-derived training set and evaluate it on the disjoint MMMU validation split. This setting examines whether the token-selection strategies remain effective when the fine-tuning corpus changes from DRIFT to multidisciplinary MMMU data. As shown in Table 3, standard SFT improves the performance of the original model from 51.78 to 52.56. On top of this, RVCFT achieves the best result of 53.56, outperforming the original model and standard SFT by 1.78 and 1.00 percentage points, respectively. Notably, selection strategies based on Attention (49.33) and Entropy (51.78) fail to surpass standard SFT, with the former even performing worse than the untuned original model. This indicates that heuristic metrics relying solely on attention weights or prediction entropy easily suffer from distribution shifts. In contrast, by jointly modeling reasoning gain and visual dependency, RVCFT can more reliably filter out tokens with high supervision value across different training corpora.

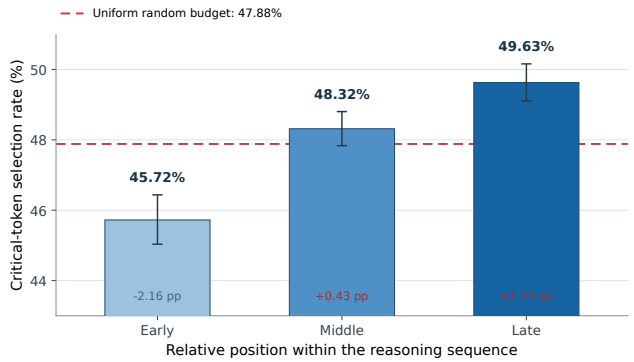


Figure 4: Positional distribution of critical tokens across Early, Middle, and Late reasoning regions. The dashed line denotes the uniform random selection baseline (47.88%). RVCFT preferentially allocates supervision budget toward later reasoning stages.

## Position Distribution of Critical Tokens.

To explore whether the token-selection mechanism of RVCFT exhibits a positional preference, we analyze the relative-position distribution of critical tokens within the reasoning chain. Specifically, for each sample, we divide the reasoning tokens before the final answer evenly into three intervals according to their relative positions: Early, Middle, and Late. Final-answer tokens belong to the predefined protection set and do not participate in adaptive filtering, so they are excluded from this analysis. The dashed line in Figure 4 represents a uniform random baseline that selects, for each sample, the same number of pre-answer reasoning tokens as RVCFT, yielding an expected selection rate of 47.88%.

As shown in Figure 4, the selection rates of RVCFT in the Early, Middle, and Late intervals are 45.72%, 48.32%, and 49.63%, respectively, showing an increasing trend across the reasoning process. Compared with the random baseline, the selection rate is 2.16 percentage points lower in the Early interval and 0.44 and 1.75 percentage points higher in the Middle and Late intervals, respectively. The selection-rate difference between the Late and Early intervals is 3.91 percentage points, with a sample-level bootstrap 95% confidence interval of $[2.74, 5.04]$. These results indicate that RVCFT does not select supervision tokens uniformly by position and tends to allocate more direct supervision to later stages of the reasoning chain.

## Case Study.

To qualitatively examine the selection behavior of RVCFT, Figure 5 presents two examples involving visual counting and geometric reasoning. In the counting example (a), the selected tokens, highlighted by black-outlined boxes, cover multiple stages of the response, including group identification ("group"), local counting ("10"), formula construction $(10 + 10 + 10 + 10)$, and the final answer ("40"). In the geometric reasoning example (b), selected tokens occur around geometric properties such as "opposite sides are equal," equation construction $(6y - 8 = 4y + 6)$, and simplification steps

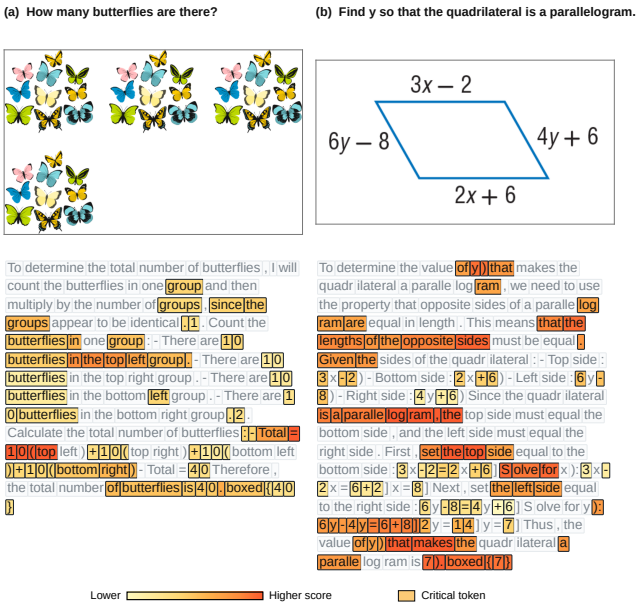


Figure 5: Qualitative visualization of critical token selection on (a) visual counting and (b) geometric reasoning. Background color intensity reflects the normalized token joint score, while black-outlined boxes highlight the selected critical tokens for supervision.

$(2y = 14$ and $y = 7)$. These examples illustrate that RVCFT selects critical supervision positions across multiple stages of the reasoning trajectory, including visual grounding, intermediate deduction, and symbolic computation, while preserving supervision on the final answer.

## Conclusion

In this paper, we propose RVCFT, a critical token fine-tuning method for multimodal reasoning. The method addresses a limitation of standard SFT, which applies uniform supervision to all response tokens despite their unequal contributions to reasoning. By jointly modeling caption-conditioned reasoning gain and reference-based visual sensitivity, RVCFT selects critical positions that reflect both reasoning-model preference and responsiveness to visual input. During fine-tuning, RVCFT retains the complete response as teacher-forced context while applying direct supervision only at the selected positions, thereby concentrating the optimization signal on critical tokens. Experiments across multiple multimodal reasoning benchmarks show that, with a target response-token retention ratio of 50%, RVCFT achieves the highest average performance among the compared methods and outperforms standard SFT on average. Additional experiments on MMMU further indicate that RVCFT remains effective beyond mathematics-focused reasoning benchmarks. Ablation studies and additional analyses support the complementary contributions of reasoning gain, reference-based visual sensitivity, and answer protection, suggesting that token-level selective supervision is a promising approach to multimodal reasoning fine-tuning.

## References


Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025b. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bi, J.; Bellos, F.; Guo, J.; Li, Y.; Huang, C.; Tang, Y.; Song, L.; Liang, S.; Zhang, Z.; Corso, J. J.; et al. 2026. When to think and when to look: Uncertainty-guided lookback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5104–5113.

Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024a. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8199–8221.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Choi, D.; Lee, J.; Tack, J.; Song, W.; Dingliwal, S.; Jayanthi, S. M.; Ganesh, B.; Shin, J.; Galstyan, A.; and Bodapati, S. B. 2025. Think Clearly: Improving Reasoning via Redundant Token Pruning. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 21437–21451. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Fang, H.; Li, J.; Kong, J.; Zhuang, T.; Gao, K.; Chen, B.; and Xia, S.-T. 2026. Seeing through the chain: Mitigate hallucination in multimodal reasoning models via cot compression and contrastive preference optimization. *arXiv preprint arXiv:2602.03380*.

Huang, C.; Zhang, Z.; Liu, J.; Sun, X.; Wu, J.; Yu, X.; Wang, Z.; Xu, C.; Barsoum, E.; and Liu, Z. 2026. DRIFT: Transferring Reasoning Priors for Efficient MLLM Fine-Tuning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 20294–20309.

Li, C.; Kulkarni, Y.; and Fazli, P. 2026. ReGATE: Learning Faster and Better with Fewer Tokens in MLLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 46430–46446.

Lin, Z.; Gou, Z.; Gong, Y.; Liu, X.; Shen, Y.; Xu, R.; Lin, C.; Yang, Y.; Jiao, J.; Duan, N.; et al. 2024. Not all tokens are what you need for pretraining. *Advances in Neural Information Processing Systems*, 37: 29029–29063.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, 23439–23554.

Qiao, R.; Tan, Q.; Dong, G.; MinhuiWu, M.; Sun, C.; Song, X.; Wang, J.; Gongque, Z.; Lei, S.; Zhang, Y.; et al. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 20023–20070.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Tu, C.; Ye, P.; Zhou, D.; Chen, T.; and Ouyang, W. 2026. Mitigating Low-Quality Reasoning in MLLMs: Self-Driven Refined Multimodal CoT with Selective Thinking and Stepwise Visual Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 9576–9584.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.-H.; Yang, J.; Zhang, Z.; et al. 2026a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38: 115452–115486.

Wang, X.; Yang, Z.; Feng, C.; Lu, H.; Li, L.; Lin, C.-C.; Lin, K.; Huang, F.; and Wang, L. 2026b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *Advances in Neural Information Processing Systems*, 38: 118818–118850.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, S.; Zhang, S.; Xia, C.; Feng, Y.; Xiong, C.; and Zhou, M. 2023. Preference-grounded token-level guidance for language model fine-tuning. *Advances in Neural Information Processing Systems*, 36: 24466–24496.

Yuan, H.; Yu, B.; Li, H.; Yang, S.; Wang, C. D.; Yu, Z.; Xu, X.; Qi, W.; and Chen, K. 2025. Not all tokens are what you need in thinking. *arXiv preprint arXiv:2505.17827*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Zeng, Y.; Liu, G.; Ma, W.; Yang, N.; Zhang, H.; and Wang, J. 2024. Token-level Direct Preference Optimization. In *International Conference on Machine Learning*, 58348–58365. PMLR.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Qiao, Y.; et al. 2024. MATH-VERSE: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *European Conference on Computer Vision*, 169–186.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; and Luo, Z. 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)*, 400–410.