BEYOND IMITATION: ROLE-AWARE MULTI-TEACHER ON-POLICY DISTILLATION FOR MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has demonstrated its effectiveness in improving multimodal reasoning with multimodal large language model (MLLM) teachers. In this paper, we investigate how a text-only LLM teacher can complement MLLM supervision in multimodal OPD. We find that directly distilling from the LLM teacher can substantially degrade student performance. Our analysis attributes this degradation to the conditioning mismatch between multimodal and text-only teachers, which can weaken image-dependent predictions and introduce conflicting token-level guidance. We treat the LLM as a reference to constrain excessive text-only alignment rather than as targets to imitate. Based on this insight, we propose **Role-Aware Heterogeneous Teacher Guidance (RouTe)**, which aligns the student with the MLLM teacher at visually dependent positions while maintaining controlled separation from the LLM teacher at reasoning-critical positions. RouTe further adapts the separation strength according to the student's relative divergence from the two teachers. Experiments across multiple multimodal reasoning benchmarks demonstrate significant and consistent performance gains.

1 INTRODUCTION

On-policy distillation (OPD) has demonstrated its effectiveness in improving the reasoning capabilities of large language models (LLMs) by leveraging a stronger LLM teacher to provide dense, token-level supervision on student-generated trajectories (Xiao et al., 2026; Lu & Lab, 2025). Building on this success, recent studies have extended OPD to multimodal large language models (MLLMs), using MLLM teachers to supervise student-generated trajectories and enhance their multimodal reasoning capabilities (Yuan et al., 2026; Liu et al., 2026; Tian et al., 2026; Li et al., 2026b; Yoon et al., 2026). Specifically, existing efforts have primarily improved supervision from MLLM teachers in two ways: **(1)** augmenting the teacher inputs to provide more fine-grained perceptual supervision (Yuan et al., 2026; Tian et al., 2026; Li et al., 2026b); and **(2)** refining the distillation objective to emphasize critical tokens within student-generated trajectories (Liu et al., 2026).

While the above methods primarily improve supervision from MLLM teachers, recent work has also explored transferring reasoning capabilities from a text-only LLM teacher to an MLLM student on text-only tasks (Bousselham et al., 2025). This naturally raises the question: *Can MLLM and LLM teachers be jointly leveraged to guide an MLLM student on multimodal reasoning tasks?* However, our preliminary experiments reveal substantial performance degradation when directly distilling from a text-only LLM teacher. The key challenge lies in their conditioning mismatch: during multimodal reasoning, the student and MLLM teacher have access to visual inputs, whereas the LLM teacher relies solely on textual context. Our analysis further shows that this mismatch can weaken image-dependent predictions and introduce conflicting token-level guidance. These observations suggest that MLLM and LLM teachers should play distinct roles in multimodal OPD.

Under this view, the MLLM teacher provides visually grounded supervision, while the text-only LLM serves as a reference for reasoning patterns that arise without visual information. Based on this, we propose **Role-Aware Heterogeneous Teacher Guidance (RouTe)**, which assigns distinct roles to heterogeneous teachers: the MLLM teacher provides visually grounded supervision, while the LLM teacher constrains the student's reliance on text-only reasoning preferences.

Specifically, RouTe consists of three components that determine where each teacher should act, how each teacher should guide, and how their guidance should be balanced.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

• **Teacher-specific token identification.** We identify visually dependent tokens by measuring the change in token likelihoods with and without visual inputs, and reasoning-critical tokens according to token-level entropy, allowing the two teachers to focus on positions most relevant to their respective roles.

• **Role-specific teacher guidance.** At visually dependent positions, we align the student with the MLLM teacher's distribution, while separating the student from the LLM teacher's text-only predictions at reasoning-critical positions.

• **Adaptive Guidance Calibration.** We dynamically adjust the strength of LLM guidance according to the relative divergence between the student and the two teachers, enabling effective utilization of both teachers without excessive reliance on the LLM's text-only prior.

In summary, RouTe enables heterogeneous teachers to complement each other by assigning them distinct roles in multimodal OPD. This role-aware design allows the student to benefit from text-only reasoning references without compromising visually grounded supervision. Extensive experiments across multiple multimodal reasoning benchmarks and various student–teacher configurations demonstrate consistent performance improvements.

2 PRELIMINARY

On-policy distillation (OPD) (Lu & Lab, 2025) transfers the knowledge of a strong teacher model to a student model using trajectories sampled from the student itself. In contrast to conventional off-policy distillation, where the student is optimized on a fixed collection of teacher-generated responses, OPD continuously queries the teacher along the states visited by the evolving student policy. We formulate OPD in the context of multimodal large language models. Let $\boldsymbol{x} = (\mathcal{I}, \mathcal{Q}) \in \mathcal{D}$ denote a multimodal input from the training dataset $\mathcal{D}$, where $\mathcal{I}$ and $\mathcal{Q}$ represent the image and the corresponding textual question, respectively. Given $(\mathcal{I}, \mathcal{Q})$, the student policy π_θ autoregressively generates a response $\boldsymbol{y} = (y_1, \ldots, y_T)$ according to

$$\pi_\theta(\boldsymbol{y} \mid \mathcal{I}, \mathcal{Q}) = \prod_{t=1}^{T} \pi_\theta(y_t \mid \mathcal{I}, \mathcal{Q}, \boldsymbol{y}_{<t}), \tag{1}$$

where $\boldsymbol{y}_{<t} = (y_1, \ldots, y_{t-1})$ is the response prefix before the t-th token. For each student-generated trajectory, the frozen teacher policy π_T provides next-token distributions conditioned on the same multimodal input and response prefixes. OPD then minimizes the reverse Kullback–Leibler (KL) divergence between the student and teacher distributions along these on-policy trajectories (Lu & Lab, 2025):

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{(\mathcal{I},\mathcal{Q})\sim\mathcal{D}} \mathbb{E}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\mathcal{I},\mathcal{Q})} \left[\frac{1}{|\boldsymbol{y}|} \sum_{t=1}^{|\boldsymbol{y}|} \mathbb{D}_{\mathrm{KL}}(\pi_\theta(\cdot \mid \mathcal{I}, \mathcal{Q}, \boldsymbol{y}_{<t}) \,\|\, \pi_T(\cdot \mid \mathcal{I}, \mathcal{Q}, \boldsymbol{y}_{<t})) \right]. \tag{2}$$

Minimizing $\mathcal{L}_{\mathrm{OPD}}$ encourages the student to approach the teacher's token-level behavior on its own multimodal generation trajectories. The sequence-length normalization ensures that responses of different lengths contribute comparably to the training objective.

3 REVISITING HETEROGENEOUS TEACHER GUIDANCE IN MLLM OPD

The complementary capabilities of multimodal and text-only LLM teachers make their combination appear attractive. However, the two teachers make predictions based on different information: the MLLM teacher has access to visual inputs, whereas the LLM teacher relies solely on textual context. Before introducing RouTe, we analyze the potential effect of direct LLM distillation on visual dependence and examine disagreement between the two teachers' token-level guidance.

3.1 THE STRUCTURAL RISK OF DIRECT LLM DISTILLATION

As illustrated in Fig. 1(a), the key difference lies in whether the teacher prediction can adapt to changes in visual evidence. Consider the textual context at step t, $c_t = (\mathcal{Q}, \boldsymbol{y}_{<t})$. For clarity,

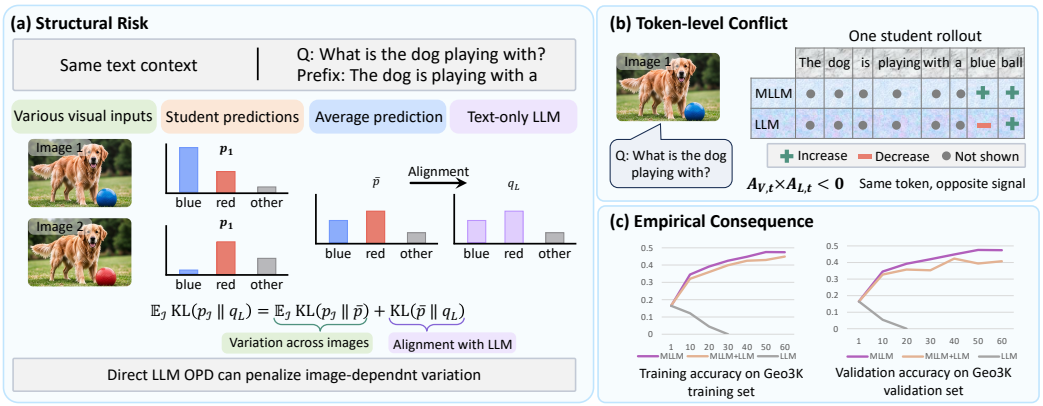

Figure 1: We examine heterogeneous teacher guidance from three aspects: (a) direct LLM distillation can suppress image-dependent variation, (b) MLLM and LLM teachers can induce conflicting token-level updates, and (c) such incompatibility is reflected in downstream performance.

we denote the next-token distributions of the student π_θ, MLLM teacher π_{ϕ_V}, and text-only LLM teacher π_{ϕ_L} as:

$$\begin{cases} p_{\theta,t}(\cdot \mid \mathcal{I}, c_t) = \pi_\theta(\cdot \mid \mathcal{I}, \mathcal{Q}, \boldsymbol{y}_{<t}), \\ q_{\phi_V,t}(\cdot \mid \mathcal{I}, c_t) = \pi_{\phi_V}(\cdot \mid \mathcal{I}, \mathcal{Q}, \boldsymbol{y}_{<t}), \\ q_{\phi_L,t}(\cdot \mid c_t) = \pi_{\phi_L}(\cdot \mid \varnothing, \mathcal{Q}, \boldsymbol{y}_{<t}), \end{cases} \tag{3}$$

respectively. The essential asymmetry is that $q_{\phi_V,t}$ can directly condition on the image, whereas the text-only LLM teacher cannot. Consequently, for a fixed textual context c_t, $q\phi_L, t$ remains unchanged as the visual input varies. Consistent with Eq. 2, the direct LLM distillation objective can be written as:

$$\mathcal{L}_{\phi_L,t} = \mathbb{E}_{\mathcal{I}|c_t}\left[\mathbb{D}_{\mathrm{KL}}(p_{\theta,t}(\cdot \mid \mathcal{I}, c_t) \,\|\, q_{\phi_L,t}(\cdot \mid c_t))\right]. \tag{4}$$

For a fixed textual context, this objective draws student predictions conditioned on different images toward the same text-only LLM distribution. To characterize the variation in student predictions induced by visual inputs, we average the student's next-token distributions over images while keeping the textual context fixed:

$$\bar{p}_{\theta,t}(\cdot \mid c_t) = \mathbb{E}_{\mathcal{I}|c_t}\left[p_{\theta,t}(\cdot \mid \mathcal{I}, c_t)\right]. \tag{5}$$

This average distribution serves as a reference for quantifying how the student's predictions vary with visual inputs. The expected KL divergence in Eq. 4 can be decomposed as:

$$\mathcal{L}_{\phi_L,t} = \mathbb{E}_{\mathcal{I}|c_t}\left[\mathbb{D}_{\mathrm{KL}}(p_{\theta,t}(\cdot \mid \mathcal{I}, c_t) \,\|\, \bar{p}_{\theta,t}(\cdot \mid c_t))\right] + \mathbb{D}_{\mathrm{KL}}(\bar{p}_{\theta,t}(\cdot \mid c_t) \,\|\, q_{\phi_L,t}(\cdot \mid c_t)). \tag{6}$$

The first term captures the variation in student predictions induced by visual inputs under the same textual context, while the second measures the discrepancy between the average student prediction and the text-only LLM distribution. Minimizing this term therefore encourages the student's predictions to become less sensitive to changes in the visual input under the same textual context.

Therefore, direct LLM distillation not only encourages alignment with the LLM, but also penalizes image-dependent variation, which can be undesirable when the next prediction still relies on visual evidence. This does not imply that a text-only teacher is always harmful: the generated prefix $\boldsymbol{y}_{<t}$ may already contain the visual information needed for the next decision. However, when additional visual evidence is still required, alignment with an image-invariant LLM distribution may weaken the student's reliance on that evidence.

3.2 Token-Level Guidance Conflict

As illustrated in Fig. 1(b), we next examine whether the two teachers produce conflicting token-level update signals when supervising the same student-generated trajectory. Let $\boldsymbol{y} \sim \pi_\theta(\cdot \mid \mathcal{I}, \mathcal{Q})$ denote a rollout from the current student policy.

For each teacher $\phi_T \in \{\phi_V, \phi_L\}$, we define the token-level distillation advantage as

$$A_{\phi_T,t} = \log q_{\phi_T,t}(y_t) - \log p_{\theta,t}(y_t), \tag{7}$$

where the conditioning variables follow Eq. 3 and are omitted for readability. To characterize the update induced by each teacher at a sampled token, we consider the per-token surrogate objective of reverse-KL OPD. Holding the sampled token and its distillation advantage fixed during differentiation, we write the surrogate objective as:

$$\widehat{\mathcal{L}}_{\phi_T,t}(\theta) = -\,\mathrm{sg}[A_{\phi_T,t}] \log p_{\theta,t}(y_t), \tag{8}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. The corresponding gradient-descent direction is

$$\boldsymbol{d}_{\phi_T,t} = -\nabla_\theta \widehat{\mathcal{L}}_{\phi_T,t}(\theta) = A_{\phi_T,t} \nabla_\theta \log p_{\theta,t}(y_t). \tag{9}$$

Since both teachers supervise the same sampled student token, they share the same score vector $\nabla_\theta \log p_{\theta,t}(y_t)$. Therefore, their local update directions differ only through the corresponding scalar distillation advantages:

$$\langle \boldsymbol{d}_{\phi_V,t}, \boldsymbol{d}_{\phi_L,t} \rangle = A_{\phi_V,t} A_{\phi_L,t} \cdot \|\nabla_\theta \log p_{\theta,t}(y_t)\|_2^2 . \tag{10}$$

For a nonzero score vector, the two teachers induce opposing local update directions whenever their distillation advantages have opposite signs, *i.e.,* $A_{\phi_V,t} A_{\phi_L,t} < 0$. Thus, opposite signs of $A_{\phi_V,t}$ and $A_{\phi_L,t}$ identify conflicting teacher guidance, where one teacher increases the likelihood of the sampled token while the other decreases it.

To quantify the phenomenon, for each rollout we compute the token-level conflict rate as:

$$R_{\text{conflict}}(\boldsymbol{y}) = \frac{1}{|\boldsymbol{y}|} \sum_{t=1}^{|\boldsymbol{y}|} \mathbb{I}[A_{\phi_V,t} A_{\phi_L,t} < 0] . \tag{11}$$

On the Geometry3K validation set, we compute this rate for each student rollout and observe that over 42% of rollouts have a token-level conflict rate exceeding 50%.

Empirical analysis. As shown in Fig. 1(c), the above incompatibility is also reflected in end-task performance. In a controlled comparison on Geometry3K, MLLM-only OPD consistently improves the student, whereas direct distillation from the text-only LLM leads to substantial performance degradation. Moreover, directly combining the two teachers as distillation targets still underperforms MLLM-only OPD, indicating that the LLM cannot be simply incorporated as an additional imitation target.

These findings motivate assigning distinct roles to the two teachers. The MLLM teacher provides visually grounded supervision, while the text-only LLM serves as a reference for reasoning patterns supported by textual context alone. Accordingly, we use the MLLM teacher for alignment at visually dependent positions and maintain controlled separation from the LLM teacher at reasoning-critical positions, which forms the basis of RouTe.

4 ROUTE: ROLE-AWARE HETEROGENEOUS TEACHER GUIDANCE

In this section, we introduce RouTe, our role-aware heterogeneous teacher guidance framework for multimodal OPD. Specifically, we first identify visually dependent tokens and apply MLLM distillation selectively at these positions to provide visually grounded supervision (Section 4.1). We then identify reasoning-critical tokens and use the text-only LLM teacher as a separation reference at the selected positions (Section 4.2). Finally, we combine the two teacher-specific objectives and dynamically calibrate the strength of LLM separation according to the student's relative divergence from the two teachers (Section 4.3). We follow the distribution notation defined in Eq. 3, and the ovewview is illustrated in Fig. 2.

4.1 VISUALLY GROUNDED MLLM GUIDANCE

Visually dependent token estimation. The contribution of visual evidence varies across response tokens. Applying MLLM teacher distillation uniformly can allocate much of the supervision to tokens that require little visual information. We therefore compare the student's predictions with and without the image, keeping the question and generated prefix unchanged. Let $\mathcal{B} = \{(\mathcal{I}^i, \mathcal{Q}^i, \boldsymbol{y}^i)\}_{i=1}^{B}$

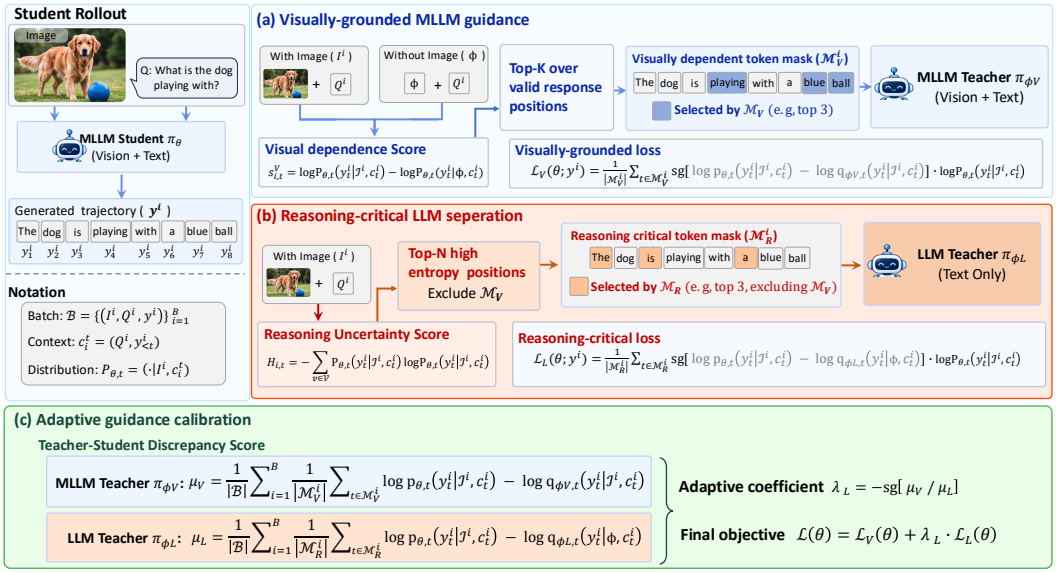

Figure 2: **Overview of RouTe.** Given student-generated rollouts, RouTe applies MLLM guidance to visually dependent tokens and LLM separation to reasoning-critical tokens, and adaptively calibrates their combination based on student–teacher discrepancies.

denote a batch of B input–rollout samples, with $c_t^i = (\mathcal{Q}^i, \boldsymbol{y}_{<t}^i)$ denoting its textual context. For the sampled token y_t^i, its visual dependence score is

$$s_{i,t}^V = \log p_{\theta,t}(y_t^i \mid \mathcal{I}^i, c_t^i) - \log p_{\theta,t}(y_t^i \mid \varnothing, c_t^i), \tag{12}$$

where $\varnothing$ denotes removal of the visual input from the student. A large $s_{i,t}^V$ indicates a strong visual influence on token y_t^i, suggesting that its prediction relies heavily on visual evidence beyond the information already available in the prefix.

Selective MLLM guidance. We select visually dependent tokens by ranking their scores across the valid response positions in the entire rollout batch:

$$\mathcal{M}_V = \mathrm{TopK}\big(\{s_{i,t}^V\}_{1\le i\le B,\, 1\le t\le |\boldsymbol{y}^i|}\big), \tag{13}$$

where TopK returns the index pairs (i,t) with the highest scores, with i identifying a rollout and t a response-token position. The index set $\mathcal{M}_V$ serves as the visual dependence token mask, and $\mathcal{M}_V^i = \{t \mid (i,t) \in \mathcal{M}_V\}$ contains the selected positions in rollout i. On these positions, we apply MLLM guidance through the sampled-token OPD surrogate:

$$\mathcal{L}_V(\theta; \boldsymbol{y}^i) = \frac{1}{|\mathcal{M}_V^i|}\sum_{t\in\mathcal{M}_V^i} \mathrm{sg}\left[\log \frac{p_{\theta,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}{q_{\phi_V,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}\right] \cdot \log p_{\theta,t}(y_t^i \mid \mathcal{I}^i, c_t^i). \tag{14}$$

Averaging over the selected positions prevents the loss scale from depending on the number of selected tokens.

4.2 Reasoning-Critical LLM Separation

Reasoning-critical token selection. Visual dependence alone does not identify every position that may affect the reasoning trajectory. The student may remain uncertain about how to continue the reasoning trajectory even with access to visual information. For token selection, we measure this uncertainty using the entropy of the student's next-token distribution:

$$H_{i,t} = -\sum_{v\in\mathcal{V}} p_{\theta,t}(v \mid \mathcal{I}^i, c_t^i) \log p_{\theta,t}(v \mid \mathcal{I}^i, c_t^i), \tag{15}$$

where $\mathcal{V}$ denotes the student's token vocabulary. High entropy indicates greater uncertainty about the next token. We use this uncertainty as a proxy for identifying positions that may affect subsequent reasoning. We first select high-entropy positions across all valid response tokens in the rollout batch, then exclude positions selected by the visual mask:

$$\mathcal{M}_R = \mathrm{TopN}\big(\{H_{i,t}\}_{1\le i\le B,\, 1\le t\le |\boldsymbol{y}^i|}\big) \setminus \mathcal{M}_V, \tag{16}$$

where TopN returns the index pairs with the highest entropies. $\mathcal{M}_R$ serves as the reasoning-critical token mask, and $\mathcal{M}_R^i = \{t \mid (i,t) \in \mathcal{M}_R\}$ denote its selected positions.

Selective LLM teacher separation. At the selected reasoning positions, the LLM serves as a reference for text-only reasoning preferences. On these positions, the sampled-token OPD surrogate for the LLM teacher is

$$\mathcal{L}_L(\theta; \boldsymbol{y}^i) = \frac{1}{|\mathcal{M}_R^i|} \sum_{t \in \mathcal{M}_R^i} \mathrm{sg}\left[\log \frac{p_{\theta,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}{q_{L,t}(y_t^i \mid c_t^i)}\right] \cdot \log p_{\theta,t}(y_t^i \mid \mathcal{I}^i, c_t^i). \tag{17}$$

We assign this term a negative coefficient in the combined objective, reversing the local LLM OPD guidance at the selected reasoning positions. The LLM thus serves as a separation reference intended to discourage reliance on text-only reasoning preferences.

4.3 Adaptive Guidance Calibration

A fixed LLM separation strength cannot account for changes in the student during training. Strong separation may be unnecessary when the student already differs substantially from the LLM, whereas a small coefficient may provide insufficient guidance when their predictions are close. We therefore calibrate the LLM separation strength according to the student's relative discrepancy from the MLLM and LLM teachers on the current rollout batch. The coefficient is computed after token selection and then used to combine the two teacher-specific objectives during student updates.

Batch-level calibration. Let θ_0 denote the student parameters before updates on the current rollout batch. For the nonempty token sets $\mathcal{M}_V$ and $\mathcal{M}_R$, we compute the corresponding batch-level discrepancies as

$$\begin{aligned} \mu_V &= \frac{1}{|\mathcal{M}_V|} \sum_{(i,t)\in\mathcal{M}_V} \log \frac{p_{\theta_0,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}{q_{V,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}, \\ \mu_L &= \frac{1}{|\mathcal{M}_R|} \sum_{(i,t)\in\mathcal{M}_R} \log \frac{p_{\theta_0,t}(y_t^i \mid \mathcal{I}^i, c_t^i)}{q_{L,t}(y_t^i \mid c_t^i)}, \end{aligned} \tag{18}$$

Each statistic is averaged over the corresponding selected tokens across the entire batch. With the MLLM teacher coefficient fixed to one, we set

$$\lambda_L = -\,\mathrm{sg}\left[\frac{\mu_V}{\mu_L}\right], \qquad \mu_L \neq 0. \tag{19}$$

The coefficient is shared by all samples in the current rollout batch, held fixed during updates on that batch, and recomputed for the next rollout batch.

Weighted teacher guidance. Let $\mathcal{L}_V(\theta)$ and $\mathcal{L}_L(\theta)$ denote the teacher-specific OPD training losses corresponding to Sections 4.1 and 4.2, respectively. We weight the LLM term by λ_L and combine it with the MLLM term:

$$\mathcal{L}_{\mathrm{RouTe}}(\theta) = \mathcal{L}_V(\theta) + \lambda_L \mathcal{L}_L(\theta). \tag{20}$$

Discussion. RouTe assigns distinct roles to heterogeneous teachers according to their input modalities and applies their guidance at different token positions. It concentrates grounded MLLM supervision on visually dependent tokens and uses the LLM as a text-only reference at reasoning-critical positions. Together, token selection and batch-level calibration determine where and how strongly each teacher guides the student, preserving their distinct roles throughout on-policy distillation.

5 Experiment

In this section, we validate the effectiveness of our RouTe strategy.

Table 1: Main results across five multimodal reasoning benchmarks. For RouTe, DeepSeek-R1-Distill-Qwen-7B is used as the text-only LLM teacher. The best and second-best results within each student setting are shown in **bold** and underline, respectively.

Setup	MathVerse	MathVision	MathVista	WeMath	HallusionBench	Avg.
Existing Multimodal Reasoning Methods						
R1-OneVision-7B (Yang et al., 2025)	46.1	22.5	63.9	62.1	65.6	52.0
OpenVLThinker-7B (Deng et al., 2025)	48.0	25.0	71.5	67.8	70.8	56.6
MM-Eureka-Qwen-7B (Meng et al., 2025)	50.5	28.3	71.5	65.5	68.3	56.8
ThinkLite-7B-VL (Wang et al., 2025)	50.2	27.6	72.7	69.2	71.0	58.1
NoisyRollout-7B (Liu et al., 2025)	53.2	28.5	72.6	69.6	72.1	59.2
Durian-7B (Li et al., 2026a)	52.8	28.8	72.3	69.2	72.9	59.2
Student: Qwen2.5-VL-3B						
Base	31.6	25.5	55.6	52.5	58.8	44.8
+ GRPO	34.1	23.8	60.9	61.9	60.8	48.3
MLLM Teacher: Qwen2.5-VL-3B-GRPO						
+ OPD	35.5	27.5	58.8	60.9	60.4	48.6
+ RouTe	37.6	**28.8**	**63.1**	62.1	61.4	50.6
MLLM Teacher: Qwen2.5-VL-7B-GRPO						
+ OPD	39.2	26.2	58.8	62.1	61.9	49.6
+ RouTe	**42.4**	**29.1**	**63.2**	**66.7**	**64.4**	**53.2**
Student: Qwen2.5-VL-7B						
Base	42.4	24.5	64.4	60.7	58.8	50.2
+ GRPO	45.3	26.6	66.6	68.7	69.1	55.3
MLLM Teacher: Qwen2.5-VL-7B-GRPO						
+ OPD	50.4	29.4	71.6	68.3	69.1	57.8
+ RouTe	**53.9**	**29.9**	**73.5**	**71.0**	**74.3**	**60.5**

5.1 Experimental Settings

Dataset. For training, we leverage the Geometry3K dataset in our experiments. Its training and validation splits contain approximately 2.1K and 0.3K examples, respectively.

Benchmarks. We evaluate our RouTe on five benchmarks: MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MathVista (Lu et al., 2024), and WeMath (Qiao et al., 2025) for mathematical reasoning, as well as HallusionBench (Guan et al., 2024) for visual perception.

Models. We conduct experiments on the Qwen2.5-VL (Bai et al., 2025) model family. Specifically, we use Qwen2.5-VL-3B-Instruct and Qwen3-VL- 7B-Instruct as student models. We use Qwen2.5-VL-3B-Instruct and Qwen3-VL-7B-Instruct after GRPO (Shao et al., 2024) training as the MLLM teacher and DeepSeek-R1-Distill-Qwen-7B as the LLM teacher.

Evaluation. We adopt the NoisyRollout (Liu et al., 2025) evaluation suite for all reported checkpoints. Responses are generated with greedy decoding using vLLM, and Gemini-3.5-Flash-001 is used as the judge model to parse the generated answers according to the suite's evaluation protocol.

Table 2: Ablation configurations. Standard MLLM-only OPD uses all valid response tokens; RouTe w/o LLM retains visual token selection.

	MathVerse	MathVision	MathVista	WeMath	HalluBench
Qwen2.5-VL-3B	31.6	25.5	55.6	52.5	58.8
Standard MLLM-only OPD	39.2	26.2	58.8	62.1	61.9
RouTe w/o LLM	40.8	25.3	60.4	63.8	60.5
RouTe, fixed LLM coefficient	42.2	26.8	62.6	64.1	61.9
RouTe	42.4	29.1	63.2	66.7	64.4

5.2 Main Results

Table 1 demonstrates the effectiveness of RouTe across different student sizes and its advantages over representative multimodal reasoning methods. For Qwen2.5-VL-3B, RouTe achieves average scores of 50.6 and 53.2 with the 3B and 7B MLLM teachers, respectively, benefiting from both same-sized and larger teachers. With the 7B MLLM teacher held fixed, increasing the student size from 3B to 7B further raises the average score from 53.2 to 60.5, a gain of 7.3 points. Compared with GRPO and standard OPD, RouTe with the 7B MLLM teacher improves the average score by 4.9 and 3.6 points for the 3B student, and by 5.2 and 2.7 points for the 7B student, respectively. These improvements over OPD hold across all five benchmarks and all three student–MLLM-teacher configurations. Moreover, the 7B RouTe student outperforms all listed existing multimodal reasoning methods on every benchmark, exceeding the highest competing average score of 59.2 by 1.3 points. Together, these results support the effectiveness of role-aware heterogeneous teacher guidance across the evaluated model scales.

5.3 Ablation Studies

We compare the variants in Table 5.3 to examine the contributions of teacher-specific token selection, LLM guidance, and adaptive calibration. All variants use the same student initialization, training data, and evaluation protocol.

Token selection. We first compare standard MLLM-only OPD, which supervises all valid response tokens, with *RouTe w/o LLM*, which retains the visual mask $\mathcal{M}_V$ but removes the LLM loss. This comparison demonstrate the effectiveness of selective visual supervision. We additionally replace the visual mask with an equal-sized set of uniformly sampled response positions in the MLLM-only setting, testing the visual dependence score against random selection at the same token budget.

Adaptive guidance calibration. We replace the batch-dependent LLM coefficient with a fixed negative value, leaving both token selection rules unchanged. This comparison assesses whether adapting the coefficient across rollout batches improves on a fixed guidance strength.

6 Related Works

In this section, we overview the relevant studies. Specifically, we first review on-policy distillation and its extensions to multimodal reasoning, then discuss related token-level optimization strategies, and finally highlight the key differences between our approach and related studies.

6.1 On-Policy Distillation

On-policy distillation (OPD) has demonstrated its effectiveness in improving the reasoning capabilities of large language models (LLMs), where a stronger LLM teacher provides dense, token-level supervision over student-generated trajectories (Lu & Lab, 2025). By providing supervision on trajectories from the evolving student policy, OPD alleviates the distribution mismatch associated with distillation from fixed teacher-generated demonstrations (Agarwal et al., 2024). Building on its success in reasoning tasks, recent studies have extended OPD to tool-integrated reasoning and multi-turn agentic tasks, exploring how effective teacher supervision can be maintained over longer and more complex interactions (Wang et al., 2026a; Zhong et al., 2026; Tan et al., 2026). Alongside this expansion in task scope, multi-teacher OPD integrates capabilities from multiple teachers into a single student model, with increasing attention to preserving and balancing these capabilities during distillation (Xiao et al., 2026; Chen et al., 2026; Gao et al., 2026).

6.2 Distillation for Multimodal LLMs

We organize related studies around three themes: enhancing visual perception, strengthening reasoning, and jointly improving perception and reasoning.

Enhancing perception. Perception-oriented methods improve visual supervision by augmenting teacher inputs or refining the distillation objectives. At the input level, evidence-based image crops (Yuan et al., 2026) and recoverable visual cues (Tian et al., 2026) make question-relevant

details more accessible to the teacher, allowing students to learn from the refined visual evidence while retaining their original inputs. At the objective level, existing methods modify distillation loss to emphasize learning from visual evidence by either comparison across various visual conditions (Liu et al., 2026) or loss decomposition to prioritize the perceptual capability (Yoon et al., 2026).

Strengthening reasoning. Reasoning-oriented methods improve the student's reasoning capability through knowledge transfer and more effective training supervision. Existing approaches transfer reasoning capabilities from text-only LLM teachers through alignment on text-only tasks (Bousselham et al., 2025), or distill predictions conditioned on intermediate visual thoughts into students that do not generate intermediate images at inference (Li et al., 2026b). Other approaches restrict distillation to high-quality student trajectories and introduce teacher-derived hints to improve supervision reliability and support exploration (Yang et al., 2026).

Jointly improving perception and reasoning. Joint approaches integrate supervision for visual perception and reasoning within a unified training framework. Existing strategies incorporate expert trajectories into RLVR to guide the use of visual evidence in reasoning (Lu et al., 2026a), or introduce separate perception and reasoning signals through black-box adversarial alignment before RLVR (Wang et al., 2026b). Another direction combines MLLM and text-only LLM teachers, providing image descriptions to the latter and weighting their predictions according to token-level uncertainty to integrate visual grounding and reasoning supervision (Yin et al., 2026).

6.3 Token-Level Optimization

Token-level optimization distinguishes the contributions of individual tokens to visual perception and reasoning, allowing training to emphasize informative positions. Existing approaches combine student predictive entropy with visual sensitivity or hidden-state similarity to reweight policy updates for perception-related and reasoning-related tokens (Lu et al., 2026b; Li et al., 2026c). Other approaches allocate token-level advantages using both local visual support and estimates of a token's influence on subsequent visually grounded reasoning (Song et al., 2026).

6.4 Differences from Prior Work

Unlike methods that transfer LLM reasoning through text-only alignment (Bousselham et al., 2025) or combine heterogeneous teacher predictions into a shared distillation target (Yin et al., 2026), RouTe assigns distinct roles to the two teachers: the MLLM teacher provides visually grounded alignment, while the text-only LLM serves as a separation reference rather than an imitation target.

RouTe also differs from token-level optimization methods that mainly use visual sensitivity or predictive uncertainty to reweight supervision (Lu et al., 2026b; Li et al., 2026c; Liu et al., 2026; Yoon et al., 2026; Aniri et al., 2026). Instead, RouTe uses these signals to determine where different teacher roles should be applied, enabling teacher-specific guidance within multimodal OPD.

7 Conclusion

In this work, we investigate how heterogeneous teachers should be used in multimodal on-policy distillation. We find that directly distilling from a text-only LLM teacher can degrade student performance due to the mismatch between multimodal and text-only conditioning, motivating its use as a text-only reference rather than an imitation target. Our analysis further shows that treating both teachers as equivalent distillation targets can weaken visual dependence and introduce conflicting token-level guidance. Based on this insight, we propose **RouTe**, which combines visually grounded supervision from an MLLM teacher with controlled separation from an LLM teacher's predictions. RouTe applies these two forms of guidance at visually dependent and reasoning-critical positions, respectively, and adapts the separation strength according to the student's relative divergence from both teachers. Experiments across multiple multimodal reasoning benchmarks and student–teacher configurations demonstrate substantial and consistent performance improvements, highlighting the importance of assigning distinct roles to heterogeneous teachers.

AI USE STATEMENT

We used generative AI tools to improve the clarity and organization of the manuscript, including language polishing and assistance with presenting related work. The authors reviewed and revised the AI-assisted text and take full responsibility for the final content, including all claims, citations, and conclusions. Generative AI tools were not used to implement the proposed method, process experimental data, or conduct experiments.

ETHICS STATEMENT

This work uses existing open-source datasets and models to study multimodal on-policy distillation. It does not involve human participant recruitment or the collection of new personal data.

REPRODUCIBILITY STATEMENT

The paper describes the RouTe training objective, teacher and student models, implementation details, and evaluation protocols to support reproducibility. Upon acceptance, we will publicly release the training and evaluation code, model checkpoints, and configuration files to facilitate reproduction of our results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: visual on-policy self-distillation with modality balance. *CoRR*, abs/2608.05131, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: reasoning transfer from llms to vision-language models via on-policy distillation. *CoRR*, abs/2510.23497, 2025.

Tianlei Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. Counteraction-aware multi-teacher on-policy distillation for general capability recovery with domain preservation. *arXiv preprint arXiv:2605.27115*, 2026.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *CoRR*, abs/2503.17352, 2025.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-mopd: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation. *arXiv preprint arXiv:2608.19098*, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Jinghan Li, Junfeng Fang, Jinda Lu, Yuan Wang, Xiaoyan Guo, Tianyu Zhang, Xiang Wang, and Xiangnan He. Enhancing multi-modal llms reasoning via difficulty-aware group normalization. *CoRR*, abs/2602.21743, 2026a.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *CoRR*, abs/2606.18974, 2026b.

Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026c.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *CoRR*, abs/2605.21924, 2026.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *CoRR*, abs/2504.13055, 2025.

Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Mingzhu Chen, Jiancan Wu, Kuien Liu, and Xiang Wang. Beyond where to look: Trajectory-guided reinforcement learning for multimodal rlvr. *arXiv preprint arXiv:2603.26126*, 2026a.

Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Guoyin Wang, Jiancan Wu, Xiang Wang, and Xiangnan He. Bridging perception and reasoning: Token reweighting for rlvr in multimodal llms. *arXiv preprint arXiv:2603.25077*, 2026b.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, abs/2503.07365, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *ACL*, 2025.

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

Xinxin Song, Siyuan Li, Tingxiong Xiao, and Jinli Suo. Anchoring what matters: A dual-level learning framework for visually-grounded multimodal reasoning. *arXiv preprint arXiv:2609.18057*, 2026.

Qitai Tan, Zefang Zong, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2606.27814*, 2026.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation. *CoRR*, abs/2606.05718, 2026.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024.

Sudong Wang, Weiquan Huang, Xiaomin Yu, Zuhao Yang, Hehai Lin, Keming Wu, Chaojun Xiao, Chen Chen, Wenxuan Wang, Beier Zhu, Yunjian Zhang, and Chengwei Qin. PRISM: pre-alignment via black-box on-policy distillation for multimodal reinforcement learning. *CoRR*, abs/2604.28123, 2026b.

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

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Fan Yang, Rui Meng, Trudi Di Qi, Ali Ezzati, and Yuxin Wen. KEPO: knowledge-enhanced preference optimization for reinforcement learning with reasoning. *CoRR*, abs/2602.00400, 2026.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *CoRR*, abs/2503.10615, 2025.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *CoRR*, abs/2607.02592, 2026.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark A. Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In *ICML*, 2026.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *CoRR*, abs/2605.18740, 2026.

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. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

8 APPENDIX

8.1 ALGORITHM OF ROUTE

Algorithm 1 Role-Aware Heterogeneous Teacher Guidance (RouTe)

Require: Student policy π_θ; MLLM teacher π_{ϕ_V}; LLM teacher π_{ϕ_L}; input batch $\{(\mathcal{I}^i, \mathcal{Q}^i)\}_{i=1}^B$; token selection budgets K and N
Ensure: Updated student policy π_θ

1. Generate on-policy trajectories

Snapshot the current student parameters: $\theta_0 \leftarrow \theta$
for $i = 1, \dots, B$ **do**
  Sample a response: $\boldsymbol{y}^i \sim \pi_{\theta_0}(\cdot \mid \mathcal{I}^i, \mathcal{Q}^i)$
  Construct textual contexts: $c_t^i \leftarrow (\mathcal{Q}^i, \boldsymbol{y}_{<t}^i)$ for $t = 1, \dots, |\boldsymbol{y}^i|$
end for
Construct the rollout batch: $\mathcal{B} \leftarrow \{(\mathcal{I}^i, \mathcal{Q}^i, \boldsymbol{y}^i)\}_{i=1}^B$
Collect valid response positions: $\mathcal{T} \leftarrow \{(i,t) : 1 \le i \le B,\ 1 \le t \le |\boldsymbol{y}^i|\}$

2. Identify teacher-specific token positions

for each $(i,t) \in \mathcal{T}$ **do**
  Evaluate the student with visual input: $p_{i,t}^I(\cdot) \leftarrow p_{\theta_0,t}(\cdot \mid \mathcal{I}^i, c_t^i)$
  Evaluate the student without visual input: $p_{i,t}^{\varnothing}(\cdot) \leftarrow p_{\theta_0,t}(\cdot \mid \varnothing, c_t^i)$
  $s_{i,t}^V \leftarrow \log p_{i,t}^I(y_t^i) - \log p_{i,t}^{\varnothing}(y_t^i)$ ▷ Eq. 12
  $H_{i,t} \leftarrow -\sum_{v \in \mathcal{V}} p_{i,t}^I(v) \log p_{i,t}^I(v)$ ▷ Eq. 15
end for
$\mathcal{M}_V \leftarrow \mathrm{TopK}(\{s_{i,t}^V\}_{(i,t)\in\mathcal{T}})$ ▷ Eq. 13
Select high-entropy candidates: $\mathcal{C}_R \leftarrow \mathrm{TopN}(\{H_{i,t}\}_{(i,t)\in\mathcal{T}})$
$\mathcal{M}_R \leftarrow \mathcal{C}_R \setminus \mathcal{M}_V$ ▷ Eq. 16
Construct per-rollout masks: $\mathcal{M}_V^i \leftarrow \{t : (i,t) \in \mathcal{M}_V\}$, $\mathcal{M}_R^i \leftarrow \{t : (i,t) \in \mathcal{M}_R\}$

3. Evaluate teachers and calibrate separation

for each $(i,t) \in \mathcal{M}_V$ **do**
  Obtain the MLLM teacher probability: $q_{i,t}^V \leftarrow q_{V,t}(y_t^i \mid \mathcal{I}^i, c_t^i)$
  Compute the sampled-token log-ratio: $d_{i,t}^V \leftarrow \log p_{i,t}^I(y_t^i) - \log q_{i,t}^V$
end for
for each $(i,t) \in \mathcal{M}_R$ **do**
  Obtain the text-only LLM teacher probability: $q_{i,t}^L \leftarrow q_{L,t}(y_t^i \mid c_t^i)$
  Compute the sampled-token log-ratio: $d_{i,t}^L \leftarrow \log p_{i,t}^I(y_t^i) - \log q_{i,t}^L$
end for
$\mu_V \leftarrow \frac{1}{|\mathcal{M}_V|} \sum_{(i,t)\in\mathcal{M}_V} d_{i,t}^V$ ▷ Eq. 18
$\mu_L \leftarrow \frac{1}{|\mathcal{M}_R|} \sum_{(i,t)\in\mathcal{M}_R} d_{i,t}^L$ ▷ Eq. 18
$\lambda_L \leftarrow -\,\mathrm{sg}[\mu_V / \mu_L]$ ▷ Eq. 19

4. Optimize the role-aware objective

Fix $\mathcal{M}_V$, $\mathcal{M}_R$, and λ_L for updates on the current rollout batch
Evaluate student log-probabilities at selected positions with gradients enabled
Compute $\mathcal{L}_V(\theta)$ on $\mathcal{M}_V$ ▷ Eq. 14
Compute $\mathcal{L}_L(\theta)$ on $\mathcal{M}_R$ ▷ Eq. 17
$\mathcal{L}_{\mathrm{RouTe}}(\theta) \leftarrow \mathcal{L}_V(\theta) + \lambda_L \mathcal{L}_L(\theta)$ ▷ Eq. 20
Update θ by minimizing $\mathcal{L}_{\mathrm{RouTe}}(\theta)$, keeping both teachers frozen
return π_θ