TEACHING THE WAY, NOT THE ANSWER: PRIVILEGED TUTORING DISTILLATION FOR MULTIMODAL POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent post-training methods, particularly Reinforcement Learning with Verifiable Rewards (RLVR), have significantly enhanced the reasoning ability of Large Vision-Language Models (LVLMs). However, the sparse nature of verifiable rewards provides little token-level supervision for failed rollouts, often leading to inefficient exploration in complex multimodal reasoning tasks. Although policy distillation can offer dense guidance, external teacher based methods introduce substantial computational overhead, while answer conditioned tuning methods potentially expose answer-level information and induce shortcut-like generation behavior. To overcome above limitations, we propose a novel Privileged Tutoring Distillation Policy Optimization (PTD-PO) method for RLVR, whose core idea is to supplement supervision with guidance on the reasoning path rather than the answer itself. Specifically, our PTD-PO first constructs Way-revealing Structured Privileged Hints by considering both spatial attention guidance and intermediate textual reasoning steps, which are further employed to guide the student model toward corrective reasoning paths through in-context-learning. particularly, the student model is still optimized under the original question-only context, while only failed rollouts are aligned with the hint-augmented reference model at the token-distribution level. To further stabilize distillation under the distribution shift between guided and unguided contexts, we introduce a Top-K Jensen-Shannon divergence with tail compensation, focusing on aligning informative tokens while reducing memory overhead. Experiments on LVLMs ranging from 2B to 8B parameters show that our PTD-PO can provide effective dense guidance without relying on answer-revealing supervision and consistently outperforms state-of-the-art counterparts by a clear margin.

1 INTRODUCTION

Recent advances in Large Vision-Language Models (LVLMs) have enabled strong performance on complex multimodal reasoning tasks, where models need to ground visual evidence, perform multi-step inference, and produce verifiable answers (Liu et al., 2023; Yue et al., 2024; Lu et al., 2023; Bai et al., 2025a; Dai et al., 2023; Li et al., 2023). Reinforcement Learning with Verifiable Rewards (RLVR) has become a promising post-training paradigm for further eliciting these above abilities, as it replaces costly human preference annotation with outcome-level feedback (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025; Team et al., 2025; Askell et al., 2021). However, verifiable rewards are usually assigned only to the final answer (Lightman et al., 2023; Chen et al., 2024), where failed rollouts provide little information on incorrect intermediate visual grounding or reasoning steps. Therefore, it makes credit assignment difficult and often cause exploration failure in large reasoning spaces (Prakash & Buvanesh, 2025; Tran et al., 2025; Ren et al., 2026).

A natural way to alleviate reward sparsity is replacement or supplement of outcome-level RLVR with dense token-level guidance (Agarwal et al., 2024; Gu et al., 2024). On-Policy Distillation (OPD) provides such guidance by aligning student rollouts with a teacher policy during training. However, it typically requires online inference from an external teacher over student-generated trajectories, incurring substantial computational overhead, lowering training efficiency, and potentially introducing tokenizer or vocabulary mismatch (Fu et al., 2026; Minixhofer et al., 2025). A more efficient

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.

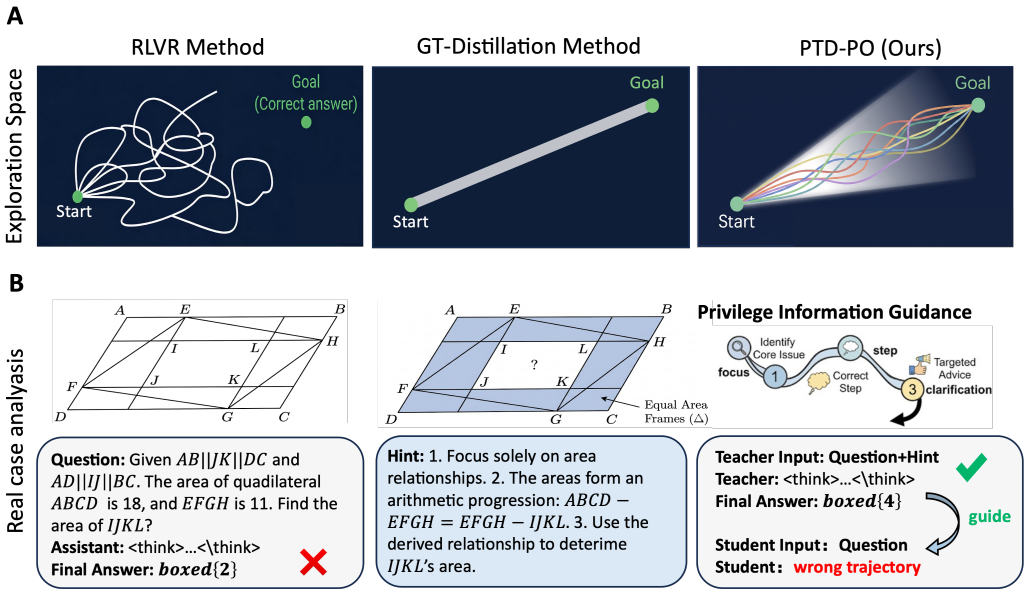


Figure 1: (a) PTD-PO complements sparse-reward RLVR with way-revealing structured privileged hints, avoiding the overly deterministic behavior of answer-revealing distillation. (b) Given a failed geometry example, the actor gives a wrong answer from the question-only context. The way-revealing hint guides the reference model to focus on area relationships and equal-area frames, and to infer the relation among $ABCD$, $EFGH$, and $IJKL$ without exposing the final area. The hint-augmented teacher then provides dense corrective supervision for the failed trajectory.

alternative is contextual self-distillation, which constructs the teacher from the student model itself under additional conditioning (Zhao et al., 2026; Hübotter et al., 2026). Generally, existing methods directly condition the self-teacher on ground-truth answers or complete solution traces to obtain strong supervision. Although advanced efforts are made, we argue that such answer-revealing contexts leads to ineffective teaching behavior, where potentially may bias the teacher toward answer-level shortcuts and sharpened token distributions, rather than guiding the student model to discover the reasoning path and so reducing diversity of reasoning trajectories. The similar phenomenon on degraded exploration was also observed in prior work (Zhang et al., 2026; Yang et al., 2026a). Therefore, a question is naturally rasied up: Can we provide the effective yet efficient dense supervisions on multimodal policy optimization without answer-revealing contexts and costly online external teachers?

To address this question, we propose Privileged Tutoring Distillation Policy Optimization (PTD-PO), inspired by an ancient Taoist proverb[1]. As illustrated in Figure 1, PTD-PO uses way-revealing hints to guide a reference model in correcting failed student trajectories, while the student remains conditioned only on the original question. Our PTD-PO is developed with two key components, i.e., Way-revealing Structured Privileged Hint (WSPH) and failure-routed Privileged Tutoring Distillation (PTD), under the learning-with-privileged-information paradigm (Vapnik & Vashist, 2009; Sharmanska et al., 2013). Specifically, WSPH construction combines where-to-look visual grounding with how-to-reason guidance, followed by refinement to remove answer-revealing or unreliable hints. Based on WSPH, PTD routes tutoring signals to failed rollouts and aligns tutor and student token distributions under asymmetric contexts, enabling targeted supervision for reasoning recovery while preserving the student's original problem context. Furthermore, we introduce a Top-K Jensen-Shannon Divergence with tail compensation in PTD to efficiently align informative token distributions while accounting for the remaining probability mass under long-context reasoning. As such, our PTD-PO can transform sparse outcome feedback into targeted reasoning-path supervision through, teaching the student how to recover from failures rather than imitate answers, thereby improving training stability and generalization. We evaluate PTD-PO on LVLMs ranging from 2B to 8B parameters across complex multimodal reasoning benchmarks(Wang et al., 2025c), clearly

[1]Give a man a fish and you feed him for a day. Teach a man to fish and you feed him for a lifetime.

demonstrating its ability to provide dense reasoning guidance while avoiding answer-revealing supervision during RLVR post-training.

In summary, our contributions are as follows:

- We disclose the issues of ineffective teaching behavior in sparse outcome rewards and answer-revealing self-distillation for multimodal policy optimization. They encourage us to propose an effective privileged tutoring distillation (PTD-PO) with core idea of multimodal policy optimization with way-revealing tutoring rather than answer-revealing self-distillation.
- To this end, our PTD-PO presents Way-revealing Structured Privileged Hint (WSPH) and failure-routed Privileged Tutoring Distillation (PTD) with Top-K Jensen-Shannon Divergence with tail compensation, which routes dense supervision to failed trajectories by conditioning the reference model on way-revealing hints of both visual evidence and text reasoning, efficiently transforming sparse outcome feedback into dense yet informative reasoning-path supervision.
- Extensive experiments show that PTD-PO consistently improves LVLM performance across model scales and a diverse range of multimodal reasoning benchmarks. Furthermore, our PTD-PO can be seamlessly integrated with different RLVR optimizers, highlighting its strong generalization ability.

2 METHODOLOGY

2.1 PRELIMINARY

Policy Optimization Policy optimization is a fundamental paradigm for multimodal model post-training, where the policy is optimized using feedback from sampled responses. Group Relative Policy Optimization (GRPO) as one of most popular methods (Shao et al., 2024) leverages group-relative rewards without requiring an explicit value function. Given a multimodal question x, GRPO samples a group of G responses $\{y_i\}_{i=1}^{G}$ with $y_i = (y_{i,1}, \ldots, y_{i,T_i})$, where T_i denotes the response length. Each response receives a verifiable reward r_i, from which the group-relative advantage is computed as $A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G}) + \epsilon}$, where ϵ is a small scalar for numerical stability. Let $\pi_{\theta_{\text{old}}}$ denote the rollout policy, for each token $y_{i,t}$ with prefix $y_{i,<t}$, GRPO defines $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$ and optimizes the clipped objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^{G}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\min(\rho_{i,t}(\theta) A_i, \text{clip}(\rho_{i,t}(\theta), 1-\epsilon_c, 1+\epsilon_c) A_i) \right) \right], \quad (1)$$

where ϵ_c is a clipping coefficient. While GRPO avoids an explicit value function, its learning signal remains primarily outcome-level. When all sampled responses fail, it provides limited guidance for correcting intermediate reasoning steps. This motivates on-policy distillation to complement answer-level feedback with dense token-level supervision.

On-Policy Distillation and Self-Distillation. On-policy distillation (Ye et al., 2026) provides dense token-level supervision by aligning the student policy π_θ with a teacher policy π_T. Given an input x and a rollout y_i sampled from the student, where $s_{i,t} = (x, y_{i,<t})$ denotes the tokens generated before step t. The standard online policy distillation objective is define as:

$$\mathcal{J}_{\text{OPD}}(\theta) = \mathbb{E}_{x,\{y_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\text{KL}} \left(\pi_T(\cdot \mid s_{i,t}) \| \pi_\theta(\cdot \mid s_{i,t}) \right) \right]. \quad (2)$$

Compared with GRPO (Eq. (1)) that only provides scalar outcome-level rewards, OPD in Eq. (2) specifies a token-level target distribution at each reasoning step and thus offers more informative optimization signals. However, when π_T is instantiated as a larger external model, on-policy distillation requires additional teacher inference and potentially suffers from tokenizer or vocabulary mismatch.

To avoid the cost and dependency of external teachers, self-distillation (Agarwal et al., 2024) instantiates the teacher policy π_T in the above policy distillation objective using the actor model itself, where SDPO (Hübotter et al., 2026) uses the exponential moving average of the actor parameters, $\text{EMA}(\pi_\theta)$, for π_T, while OPSD (Zhao et al., 2026) uses the frozen reference model (π_{ref}) for π_T.

Although this removes the need for an external teacher, the corrective supervision provided by self-distillation depends on the additional context supplied to the teacher. We analyze this dependence below.

2.2 Analysis on limitations of answer-revealing policy optimization

Analysis 1: Limited Informative Reward Groups in Answer-Only Policy Optimization. Answer-only policy optimization relies on outcome-level feedback derived from the correctness of sampled responses. To examine the informativeness of such supervision, we analyze the group-level reward composition in GRPO as a representative example. For each question x, we sample G responses and categorize each group by the number of successful responses, $m(x) = \sum_{i=1}^{G} r_i$, into all-fail ($m = 0$), informative ($0 < m < G$), and all-pass ($m = G$) groups. As shown in Figure 2, all-fail groups dominate the early training stage, while informative groups remain limited to roughly one third of all sampled groups and all-pass groups gradually become dominant. This indicates that answer-only policy optimization frequently provides homogeneous or low-information reward signals, limiting its effectiveness for correcting failed trajectories. This observation shares similar philosophy with recent studies (He et al., 2026; Yu et al., 2026a), showing that homogeneous rewards within a group lead to advantage collapse or near-zero learning signals in GRPO.

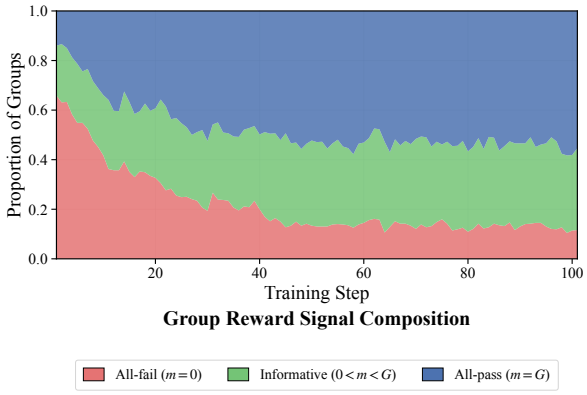


Figure 2: Group-level reward composition during GRPO training. Informative groups ($0 < m < G$) remain limited, while all-fail and all-pass groups yield homogeneous rewards.

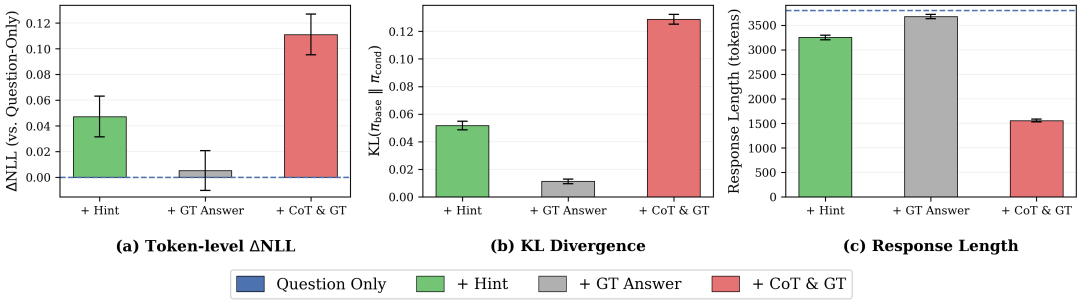


Figure 3: Context-conditioned self-teacher statistics. (a) Token-level ΔNLL, (b) full-distribution KL divergence, and (c) response length show that complete solution traces induce the largest distribution shift and shortcut-like generation behavior, while our Way-revealing Structured Privileged Hint context provide a moderate guidance signal.

Analysis 2: Answer-Revealing Contexts Induce Distribution Shift and Shortcut-Like Teacher Behavior. We further examine how answer-revealing information changes the token distribution of a frozen self-teacher. All conditions use the same self-teacher and differ only in context. As shown in Figure 3(a) and (b), providing only the final GT answer has a marginal effect on the self-teacher distribution, where both token-level ΔNLL and full-distribution KL remain close to the question-only baseline. In contrast, conditioning the self-teacher on both the complete CoT solution and GT answer induces a much larger distribution shift, with substantially higher ΔNLL on student-generated tokens and the largest KL divergence. Moreover, as shown in Figure 3(c), the CoT-and-GT condition sharply reduces response length, suggesting that the self-teacher increasingly relies on the provided solution rather than autonomous reasoning. These results highlight a concern

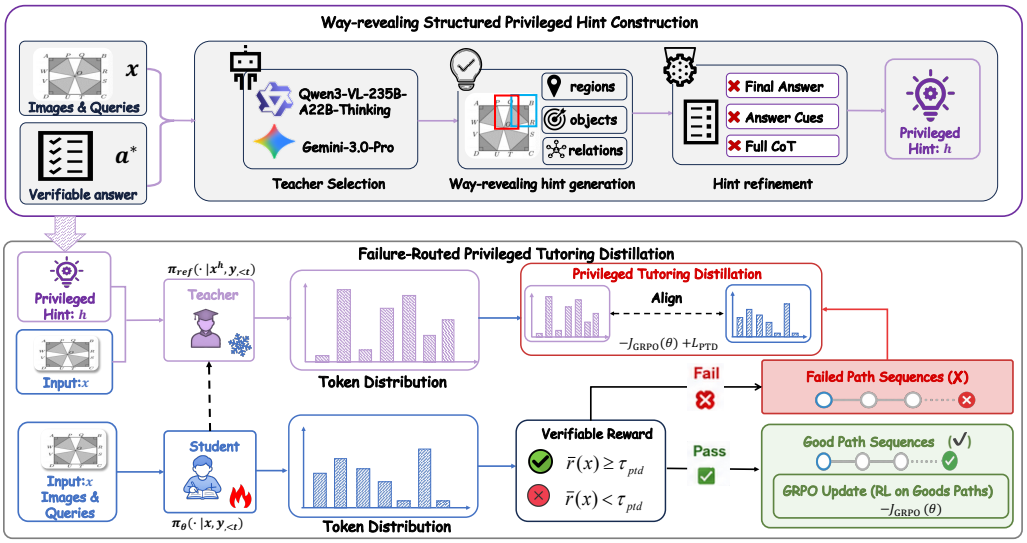


Figure 4: Overview of PTD-PO, which combines RLVR with privileged-tutoring distillation from the Way-revealing Privileged Hint conditioned reference model on failed trajectories.

with answer-revealing supervision: additional answer information can alter the self-teacher's reasoning behavior and induce shortcut-like generation. Above observations encourage us to provide an alternative effective guidance for policy optimization without exposing complete answer-level information. Therefore, we propose a Privileged Tutoring Distillation Policy Optimization (PTD-PO) method to perfrom policy optimization with way-revealing tutoring rather than answer-revealing self-distillation, whose details are described as follows.

2.3 PTD-PO: Privileged Tutoring Distillation Policy Optimization

To address the limited supervision for failed RLVR trajectorie, PTD-PO is proposed by turning them into targets for corrective tutoring while preserving exploration. As shown in Figure 4, our PTD-PO consists of two key components: Way-revealing Structured Privileged Hint (WSPH) construction and failure-routed Privileged Tutoring Distillation (PTD). Specifically, the actor samples responses under the original prompt and receives verifiable rewards to identify failed trajectories. Building on this, WSPH provides structured spatial and reasoning hints for these trajectories, while PTD routes the tutoring signals by aligning tutor–student token distributions under asymmetric contexts. The resulting signal complements the reward-driven objective to update the actor, turning failed rollouts into dense supervision for learning reasoning strategies rather than answer-conditioned shortcuts. The following sections detail how WSPH constructs non-spoiling guidance and how PTD effectively distills it into the actor.

Way-revealing Structured Privileged Hint Construction. For each training example, we construct a structured privileged hint from the multimodal question x and verifiable answer $a^\star$, where answer is only used to construct a way-revealing hint and is never exposed to the student policy. The hint provides only partial, process-level guidance, thereby inducing a milder privileged-context shift from the original question. To preserve this property, the hint is constrained to indicate a valid reasoning direction without revealing answer-identifying information.

Specifically, we use Qwen3-VL-235B-A22B-Thinking as the primary external teacher and Gemini-3.0-Pro as a fallback model to generate the hints. The construction consists of two stages: way-revealing hint generation and hint refinement. Given $(x, a^\star)$, a rule-guided system prompt instructs the teacher to identify task-relevant visual evidence, including key regions, objects, and spatial relations, and formulate a concise and informative reasoning direction based on these cues. We then perform lightweight hint refinement to filter invalid candidates, including those with answer leakage, missing or irrelevant visual evidence, or overly complete reasoning traces. The resulting validated

hints serve as the final privileged annotations for subsequent policy training. Figure 5 (a) illustrates a representative construction example, while the complete prompt, refinement rules, and additional cases are provided in Appendices C.1 and F.

Failure-Routed Privileged Tutoring Distillation. Building on WSPH, we incorporate the way-revealing hints into on-policy distillation to provide effective supervision on corrective lines of reasoning for failed trajectories. Given a multimodal question x, the rollout policy samples a group of G responses $\{y_i\}_{i=1}^{G}$ under the original question-only context, i.e.,

$$y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \qquad r_i = R(x, y_i) \in \{0, 1\}, \tag{3}$$

where $R(\cdot)$ is the verifiable reward function. We characterize the difficulty and training informativeness of the sampled group by its empirical accuracy, i.e.,

$$\bar{r}(x) = \frac{1}{G} \sum_{i=1}^{G} r_i, \tag{4}$$

where small $\bar{r}(x)$ indicates most trajectories fail to reach the correct answer, and the answer-level reward provides limited information about how their intermediate visual grounding or reasoning steps should be corrected. Therefore, instead of applying privileged tutoring distillation uniformly to all rollouts, we route PTD supervision to the incorrect responses in the sampled group. Specifically, given an activation threshold τ_{ptd}, the set of token positions selected for PTD is defined as:

$$\mathcal{F} = \{(i, t) \mid \bar{r}(x) < \tau_{\text{ptd}},\ r_i = 0,\ m_{i,t} = 1\}, \tag{5}$$

where $m_{i,t}$ is the response mask. Particularly, we use $\tau_{\text{ptd}} = 1.0$ as the default setting, indicating that PTD is applied to all failed trajectories.

For each question x, we construct a way-revealing privileged hint h and form the hint-augmented teacher context $x^h = [x; h]$. In the OPD objective of Eq. (2), we replace the teacher distribution $\pi_T(\cdot \mid s_{i,t})$ with the hint-conditioned frozen reference distribution $q^h_{i,t} = \pi_{\text{ref}}(\cdot \mid x^h, y_{i,<t})$, while retaining the question-only student distribution $p_{i,t} = \pi_\theta(\cdot \mid x, y_{i,<t})$. This asymmetric construction decouples the information used to produce dense supervision from the information available to the student policy. As such, the teacher policy can use privileged spatial and textual hints to infer a corrective reasoning direction, whereas the student must learn to reproduce the improved token distribution without directly observing the hint. To this end, our privileged tutoring distillation loss is defined over the selected failed positions, i.e.,

$$\mathcal{L}_{\text{PTD}} = \frac{\sum_{(i,t)\in\mathcal{F}} D_{\text{PTD}}\left(q^h_{i,t}, p_{i,t}\right)}{|\mathcal{F}| + \epsilon}, \tag{6}$$

where D_{PTD} measures the discrepancy between the hint-conditioned teacher distribution and the question-only student distribution. Since the two distributions are conditioned on different contexts, a directional KL divergence potentially introduce a one-sided matching bias, particularly when the teacher distribution is sharp. We therefore base D_{PTD} on the symmetric and bounded Jensen–Shannon divergence (JSD). For distributions q and p, the full JSD is defined as $D_{\text{JSD}}(q, p) = \frac{1}{2} D_{\text{KL}}(q \| m) + \frac{1}{2} D_{\text{KL}}(p \| m)$, where $m = \frac{1}{2}(q + p)$.

Finally, we perform policy optimization by combining the GRPO with our PTD regularizer:

$$\mathcal{L}_{\text{PTD-PO}} = -J_{\text{GRPO}}(\theta) + \lambda_{\text{PTD}} \mathcal{L}_{\text{PTD}}, \tag{7}$$

where λ_{PTD} controls the effect of privileged distillation.

Top-K JSD with Tail Compensation. Obviously, retaining full-vocabulary distributions at every response position is memory-intensive. To reduce this overhead while preserving the dominant probability mass, we approximate the full JSD with a compact Top-K support and explicitly compensate for the remaining tail mass. At each selected position (i, t), we take the union of the teacher and student Top-K tokens, retaining high-probability tokens favored by either distribution:

$$\mathcal{S}_{i,t} = \text{Top-K}(p_{i,t}) \cup \text{Top-K}(q^h_{i,t}). \tag{8}$$

Then, the probability mass outside this support is aggregated into a single tail bucket:

$$p^{\text{tail}}_{i,t} = 1 - \sum_{v \in \mathcal{S}_{i,t}} p_{i,t}(v), \qquad q^{h,\text{tail}}_{i,t} = 1 - \sum_{v \in \mathcal{S}_{i,t}} q^h_{i,t}(v). \tag{9}$$

Table 1: Comparison of different training methods with various LVLMs on general multimodal reasoning and vision-dependent multimodal reasoning tasks.

Method / Setting	General Multimodal Reasoning						Vision-Dependent Multimodal Reasoning					Overall
	MMK12	Geo3K	MathVerse	MathVista	We-Math	AVG	MMMU-Pro	Counting	MathVerse$_V$	LogicVista	AVG	AVG
Qwen3-VL-2B-Thinking												
SFT	38.59	41.95	55.50	49.05	59.55	48.93	20.56	53.44	51.25	36.83	40.52	45.19
OPSD	44.99	38.48	52.69	48.52	56.90	48.32	17.38	61.19	46.41	30.20	38.80	44.08
GRPO	51.30	45.36	53.49	56.14	58.36	52.93	23.31	84.50	49.91	33.31	47.76	50.63
HDPO	50.22	49.36	66.60	58.86	69.81	58.97	29.10	83.38	63.21	48.55	56.06	57.68
PAPO	49.78	44.64	60.00	56.60	64.27	55.06	24.88	88.38	54.78	43.70	52.94	54.11
PTD-PO(Ours)	**56.86**	**50.52**	**70.53**	**60.38**	**72.64**	**62.19**	**32.26**	**88.88**	**66.62**	**52.21**	**59.99**	**61.21**
Qwen3-VL-4B-Thinking												
SFT	59.59	57.38	70.85	57.07	71.24	63.23	29.91	72.38	66.39	46.50	53.80	59.03
OPSD	60.97	58.13	67.98	55.23	68.06	62.07	23.20	72.19	67.87	41.50	51.19	57.24
GRPO	64.53	62.62	75.25	68.38	78.30	69.82	40.19	91.69	73.03	58.56	65.87	68.06
HDPO	66.26	62.75	76.25	67.83	76.93	70.00	39.93	91.75	73.84	59.06	66.15	68.29
PAPO	65.11	62.69	77.73	**70.95**	80.32	71.36	40.03	92.81	74.31	58.91	66.52	69.20
PTD-PO(Ours)	**73.84**	**64.79**	**80.45**	69.81	**81.67**	**74.11**	**41.20**	**92.88**	**76.15**	**60.32**	**67.64**	**71.23**
Qwen3-VL-8B-Thinking												
SFT	55.97	59.09	67.72	57.81	70.47	62.21	34.14	76.06	66.74	46.09	55.76	59.34
OPSD	68.73	58.19	69.84	59.44	71.72	65.58	29.47	74.56	69.37	45.52	54.73	60.76
GRPO	70.75	60.17	77.19	67.97	80.17	71.25	42.83	88.56	73.98	57.38	65.69	68.78
HDPO	71.05	**64.60**	78.04	66.86	78.88	71.89	**45.87**	90.06	74.55	59.40	67.47	69.92
PAPO	71.17	62.53	77.31	66.20	79.10	71.26	42.90	88.69	74.27	59.28	66.29	69.05
PTD-PO(Ours)	**76.21**	64.52	**79.81**	**71.39**	**82.31**	**74.85**	44.15	**92.06**	**76.41**	**59.87**	**68.12**	**71.86**

As such, we have the compressed distributions:

$$\widetilde{p}_{i,t} = \left(p_{i,t}(v)_{v\in\mathcal{S}_{i,t}}, p_{i,t}^{\text{tail}}\right), \qquad \widetilde{q}_{i,t}^{h} = \left(q_{i,t}^{h}(v)_{v\in\mathcal{S}_{i,t}}, q_{i,t}^{h,\text{tail}}\right). \tag{10}$$

After above approximations, our PTD divergence in Eq. (6) can be computed as

$$D_{\text{PTD}}\left(q_{i,t}^{h}, p_{i,t}\right) \approx D_{\text{JSD}}\left(\widetilde{q}_{i,t}^{h}, \widetilde{p}_{i,t}\right). \tag{11}$$

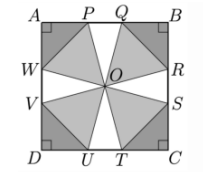
It is worth mentioning that the tail bucket preserves omitted mass without renormalizing the retained tokens, so the objective in Eq. (11) is a mass-preserving coarse-grained lower bound on full JSD whose gap contains only within-tail discrepancies (see Appendix D.3), ensuring the computational reliability. Importantly, it reduces the storage required for distillation distributions from $\mathcal{O}(BTV)$ to $\mathcal{O}(BTK)$ (with $|\mathcal{S}_{i,t}| \leq 2K$), while retaining high-probability alignment signals needed for stable privileged tutoring distillation.

3 Experiments

3.1 Experimental Setup

Training and evaluation. We use ViRL39K as the training corpus for multimodal reasoning post-training. Specifically, ViRL39K contains 38,870 verifiable vision-language question-answering examples and covers diverse reasoning scenarios, including mathematics, science, charts, diagrams, tables, documents, and spatial reasoning (Lu et al., 2023; Yue et al., 2024; Xiao et al., 2024; Zhang et al., 2024; Johnson et al., 2017; Lu et al., 2021). To support different paradigms, we construct three forms of auxiliary supervision for each example: our Way-revealing Structured Privileged Hints (WSPH), regular hints, and chain-of-thought answer trajectories with the final solution. For evaluation, we employ the PAPO (Wang et al., 2025c) benchmark suite and its unified verifiable evaluation protocol, grouping tasks into general and vision-dependent multimodal reasoning and sampling eight rollouts per question to report average accuracy for each benchmark.

Implementation Details. For the main experiments, all model sizes are trained and evaluated with a maximum response length of 4096 tokens. For RL-based methods, including GRPO, self-distillation methods and PTD-PO, we use a rollout number of 8 during training. To reduce computational cost, ablation studies are conducted under a smaller but matched setting, with rollout number 5 and maximum response length 2048 tokens. More implementation details, hyperparameters, reward rules, and training configurations are provided in Appendix B.



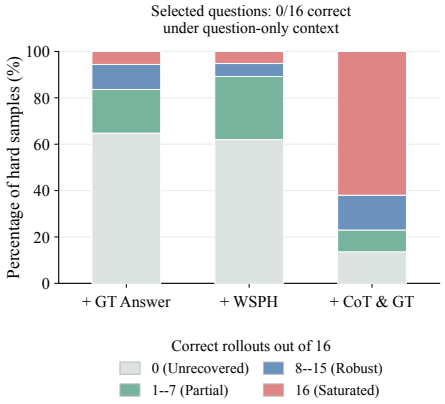
Question

Square *ABCD* is partitioned by four isosceles and four equilateral triangles. Compare the areas of $\triangle PQO$ and $\triangle BRQ$.

WSPH

- Ignore all shaded regions; focus solely on triangles *PQO* and *BRQ*.
- *BRQ* is right isosceles; express its area in terms of one leg length.
- *PQO* has a horizontal base on *AB*; find its base and vertical height from *O*.
- Use equilateral *ROQ* to relate the relevant segment lengths.
- Apply its side equality, solve the unknown, and compute both areas.

(a) WSPH example



(b) Hard-case recovery distribution

Figure 5: WSPH design and hard-case recovery on Qwen3-VL-4B-Thinking. (a) **WSPH example.** The hint identifies relevant subfigures and links their geometric constraints without revealing the answer. (b) **Hard-case recovery.** Recovery distributions across teacher contexts on questions with no successful question-only rollouts.

3.2 Main Results

Table 1 compares our PTD-PO with some state-of-the-arts spanning supervised fine-tuning, self-distillation, and multimodal RLVR by Qwen3-VL-Thinking models of different scales. PTD-PO consistently achieves the best overall performance on 2B, 4B, and 8B models, outperforming SFT, OPSD, and GRPO baseline (Zhao et al., 2026; Shao et al., 2024) on both general and vision-dependent multimodal reasoning tasks. Compared with HDPO (Ding, 2026) that directly conditions privileged self-distillation on complete CoT trajectories and ground-truth answers, our PTD-PO achieves clearly better results by transforming answer-derived privileged supervision into non-answer-revealing reasoning guidance rather than directly conditioning the tutor on the solution. PTD-PO also outperforms PAPO (Wang et al., 2025c) substantially, which augments multimodal RLVR with perception-aware visual grounding objectives. These improvements suggest that reasoning-path tutoring provides gains beyond perception-oriented policy optimization. Overall, these results establish privileged tutoring distillation as a consistently effective complement to sparse outcome supervision across model scales. Training dynamics in Appendix E.3 further confirm its advantage.

3.3 Compatibility and Broader Comparison

Furthermore, we examine whether our PTD-PO can serve as a general auxiliary module across different RLVR objectives. As shown in Table 3 of Appendix E.1, we also integrate PTD with DAPO (Yu et al., 2026a) and GSPO (Zheng et al., 2025) while retaining their original optimization objectives. The consistent improvements across these optimizers indicate that the proposed tutoring signal is not tied to a particular policy optimization rule, but can provide complementary corrective supervision for failed trajectories under different RLVR formulations. We also provide a broader comparison of PTD-PO with recent open-source multimodal reasoning methods, including VGPO (Wang et al., 2026) and CFPO (Yu et al., 2026b).

3.4 Ablation Study

Effect of Way-revealing Tutoring Context. To examine the tutoring behavior induced by WSPH, we use Qwen3-VL-4B-Thinking and sample 16 rollouts for each of 384 examples under different teacher contexts. Figure 5(b) focuses on hard questions with 0/16 correct rollouts under the question-only context. WSPH increases the fraction of these questions with 1–7 successful rollouts from 18.8% under GT-answer prompting to 27.2%. In contrast, CoT+GT yields 16/16 successes on 62.0% of these questions, compared with 5.2% under WSPH. These results distinguish partial recovery

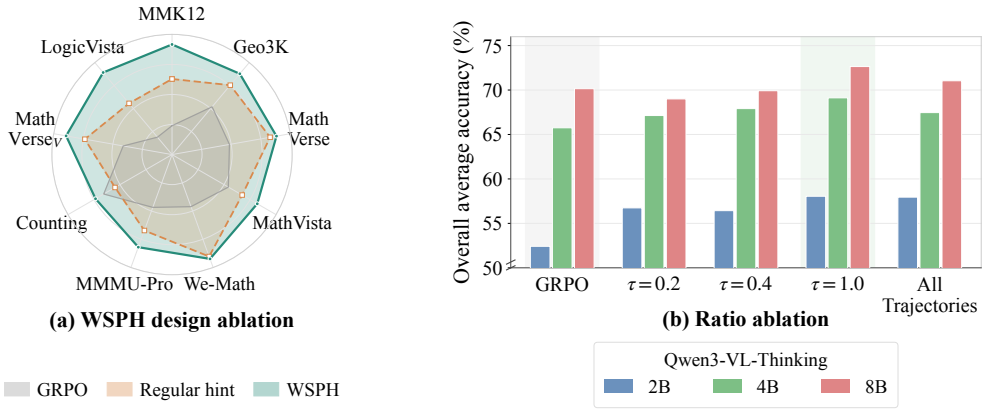


Figure 6: Ablations of PTD-PO. (a) Effect of hint design: GRPO, PTD with regular hints, and PTD with WSPH. (b) Effect of activation thresholds and trajectory selection on overall average accuracy.

under WSPH from the predominantly all-pass behavior under full solution conditioning. Together with the distribution shift and response shortening in Figure 3, they support using WSPH to provide corrective guidance with less reliance on solution-revealing contexts.

Effect of Way-revealing Structure. Figure 6(a) shows that even regular hints generally outperform GRPO, indicating that privileged reasoning direction alone provides useful supervision beyond sparse outcome rewards. WSPH consistently extends this improvement across all benchmarks, with a more pronounced advantage over regular hints on vision-dependent tasks such as LogicVista, Counting, and MathVerse$_V$. This separation indicates that the additional gain does not merely come from supplying more text: explicitly identifying relevant visual evidence and relations while filtering distractors helps the tutor correct perceptual selection and cross-modal binding errors that a regular hint leaves unresolved. The smaller separation on MathVerse and We-Math further supports this interpretation, since these tasks rely more on symbolic or textual reasoning once the problem has been parsed. Complete per-benchmark results are reported in Table 5 of Appendix E.4.

Effect of PTD Activation Threshold. Finally, we assess effect of the trajectory selection strategy on our PTD-PO, where τ_{ptd} denotes the group-level activation threshold and "All Trajectories" applies PTD to all rollouts regardless of correctness. Figure 6(b) shows that applying PTD to all failed trajectories with $\tau_{\text{ptd}} = 1.0$ achieves the highest overall average accuracy across the 2B, 4B, and 8B models. Smaller thresholds restrict PTD to more severely failed groups, reducing corrective supervision and yielding lower scores. Extending PTD to all trajectories also reduces performance, especially on larger models: correct rollouts already provide reliable reward-driven signals, and forcing them to match the WSPH-conditioned tutor may over-regularize the policy. These results support focusing PTD on failed trajectories, where answer-level rewards provide limited guidance.

4 Conclusion

In this paper, we presented PTD-PO, an RLVR framework that combines Way-revealing Structured Privileged Hints (WSPH) with failure-routed privileged tutoring distillation. WSPH provides where-to-look visual grounding and how-to-reason process guidance without exposing answer-identifying information or complete solution traces, while the student remains under the original question-only context. Failed rollouts are aligned with the WSPH-conditioned frozen reference model through a memory-efficient Top-K JSD objective with tail compensation. Experiments on LVLMs from 2B to 8B show consistent improvements over RLVR and distillation baselines; regular hints already improve upon GRPO, while WSPH yields further gains, especially on vision-dependent tasks. These results clearly demonstrate that our idea of way-revealing privileged tutoring distillation can effectively complement sparse verifiable rewards by correcting perceptual grounding and reasoning-path errors without turning training into answer imitation.

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 *The twelfth international conference on learning representations*, 2024.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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. *arXiv e-prints*, pp. arXiv–2503, 2025.

Ken Ding. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025a.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Zhihang Lin, Mingbao Lin, Yuan Xie, and R Cppo Ji. Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 37, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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. *arXiv preprint arXiv:2504.13055*, 2025.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

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. *arXiv preprint arXiv:2310.02255*, 2023.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Benjamin Minixhofer, Ivan Vulić, and Edoardo M Ponti. Cross-tokenizer distillation via approximate likelihood matching. 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Jatin Prakash and Anirudh Buvanesh. What can you do when you have zero rewards during rl? *arXiv preprint arXiv:2510.03971*, 2025.

Yanwei Ren, Haotian Zhang, Likang Xiao, Xikai Zhang, Jiaxing Huang, Jiayan Qiu, Baosheng Yu, Quan Chen, and Liu Liu. Recycling failures: Salvaging exploration in rlvr via fine-grained off-policy guidance. *arXiv preprint arXiv:2602.24110*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *Proceedings of the IEEE international conference on computer vision*, pp. 825–832, 2013.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv e-prints*, pp. arXiv–2503, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in rl training of llms. *arXiv preprint arXiv:2509.18314*, 2025.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.

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. *arXiv preprint arXiv:2504.07934*, 2025b.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*, 2026. URL https://arxiv.org/abs/2604.09349v2.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025c.

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

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11694–11702, 2026b.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual policy optimization for multimodal reasoning. *arXiv preprint arXiv:2606.23206*, 2026b. URL https://arxiv.org/abs/2606.23206v1.

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. *arXiv preprint arXiv:2605.18740*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Wenjing Zhang, Jiangze Yan, Jieyun Huang, Yi Shen, Shuming Shi, Ping Chen, Ning Wang, Zhaoxiang Liu, Kai Wang, and Shiguo Lian. Heal: Hindsight entropy-assisted learning for reasoning distillation. *arXiv preprint arXiv:2603.10359*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model, 2025. *URL https://arxiv.org/abs/2503.05132*, 9:10–16.

APPENDIX

APPENDIX CONTENTS

A Related Work **15**

B More Implementation Details **15**

C Data Construction and Quality Control **16**

C.1 WSPH Construction 16

C.2 Chain-of-Thought Answer Construction 19

D Proofs for Theoretical Analysis **19**

D.1 A Unified View of GRPO and Token-Level Distillation 19

D.2 Why Jensen-Shannon Divergence Instead of Directional KL 21

D.3 Approximation Property of Top-K JSD with Tail Compensation 23

D.4 Optimization Rationale for Applying PTD to Failed Trajectories 24

E Additional Experiments and Analysis **26**

E.1 Compatibility With Other RLVR Optimizers 26

E.2 Sensitivity Analysis on the Top-K Support Size 27

E.3 Training Dynamics and Efficiency Analysis 27

E.4 Detailed Ablation Results 28

F Case Study **29**

G Limitations and Future Work **30**

A Related Work

RLVR for Multimodal Reasoning. Reinforcement Learning with Verifiable Rewards (RLVR) uses automatically checkable outcomes rather than learned RLHF rewards (Schulman et al., 2017; Ouyang et al., 2022; Askell et al., 2021). Text-based advances include DeepSeekMath, DeepSeek-R1, Kimi k1.5, and improved policy optimizers (Shao et al., 2024; Guo et al., 2025; Team et al., 2025; Yu et al., 2026a; Zheng et al., 2025; Zhao et al., 2025). Multimodal extensions such as Vision-R1, R1-VL, MM-Eureka, OpenVLThinker, and Reason-RFT combine cold-start data, rule-based rewards, and GRPO-style optimization (Huang et al., 2025b; Zhou et al.; Meng et al., 2025; Deng et al., 2025; Tan et al., 2025). ThinkLite-VL uses MCTS-guided sample selection (Wang et al., 2025b), VL-Rethinker combines selective sample replay with forced rethinking (Wang et al., 2025a), and NoisyRollout augments images to enhance exploration (Liu et al., 2025). For visual grounding, PAPO adds an implicit perception loss (Wang et al., 2025c), and VPPO-RL reweights trajectory advantages and selects policy-update tokens using visual dependency (Huang et al., 2025a). Other methods explore contrastive perception learning and visual re-focusing (Lin et al., 2025; Yang et al., 2026b). VGPO combines visual attention compensation with token- and trajectory-level advantage reweighting (Wang et al., 2026), while CFPO regularizes predictions through counterfactual suppression of critical visual cues (Yu et al., 2026b). In contrast, PTD-PO densely corrects failed trajectories through distillation from a frozen self-teacher conditioned on WSPH, while preserving the student's question-only context.

On-Policy and Self-Distillation for Dense Supervision. Another closely related direction is on-policy distillation (OPD), which supervises student-generated trajectories with dense teacher distributions, combining the distributional relevance of on-policy learning with token-level feedback from knowledge distillation (Bai et al., 2025a; Ye et al., 2026; 2025). Compared with off-policy distillation on fixed teacher traces, OPD can better expose the teacher to student-side errors, but it still typically depends on an external stronger teacher and incurs nontrivial rollout-time supervision cost (Li et al., 2026b). Recent self-distillation methods further reduce this dependency by letting the same model serve as both student and teacher under different contexts, e.g., conditioning the teacher on ground-truth solutions, rich feedback, successful sibling rollouts, or other privileged signals to provide dense token-level correction (Zhao et al., 2026; Hübotter et al., 2026; Li et al., 2026a). Vision-OPD extends this idea to multimodal fine-grained perception by using a crop-conditioned policy as a privileged teacher to supervise a full-image policy on on-policy rollouts (Yuan et al., 2026). More relevant to our work, HDPO introduces privileged self-distillation into RL by directly using ground-truth answers to construct a self-teacher for cliff prompts where standard RL yields vanishing gradients (Ding, 2026). However, HDPO focuses on text-only mathematical reasoning and relies on answer-revealing privileged information, while Vision-OPD focuses on crop-based regional perception; in contrast, our method targets multimodal reasoning and uses WSPH to provide non-answer-revealing visual-grounding and reasoning guidance for correcting failed trajectories.

B More Implementation Details

Overall Setup. All experiments are implemented with PyTorch and conducted on Qwen3-VL-Thinking models with different parameter scales. For all training methods, we use ViRL39K as the training corpus and train each model for two epochs. The vision encoder is kept consistent with the original model configuration, and all RL-based methods are trained under the same data mixture, rollout setting, response length budget, and evaluation protocol to ensure fair comparison. For PTD-PO, WSPH is used only to condition the frozen reference model for teacher-side distribution generation and is never included in the student rollout or evaluation context.

Training Details. For the main experiments, the maximum response length is set to 4096 tokens during both training and evaluation. Ablation studies use a maximum response length of 2048 tokens and five training rollouts per prompt. SFT is optimized with standard next-token prediction on the constructed solution trajectories. For RL-based methods, each prompt is sampled under the original question-only context, and the policy generates multiple rollouts for verifiable reward optimization. The reward is composed of a format reward and an accuracy reward, with weights of 0.1 and 0.9, respectively, where the accuracy reward is computed by rule-based answer matching whenever possible. GRPO and PTD-PO share the same rollout configuration and reward design.

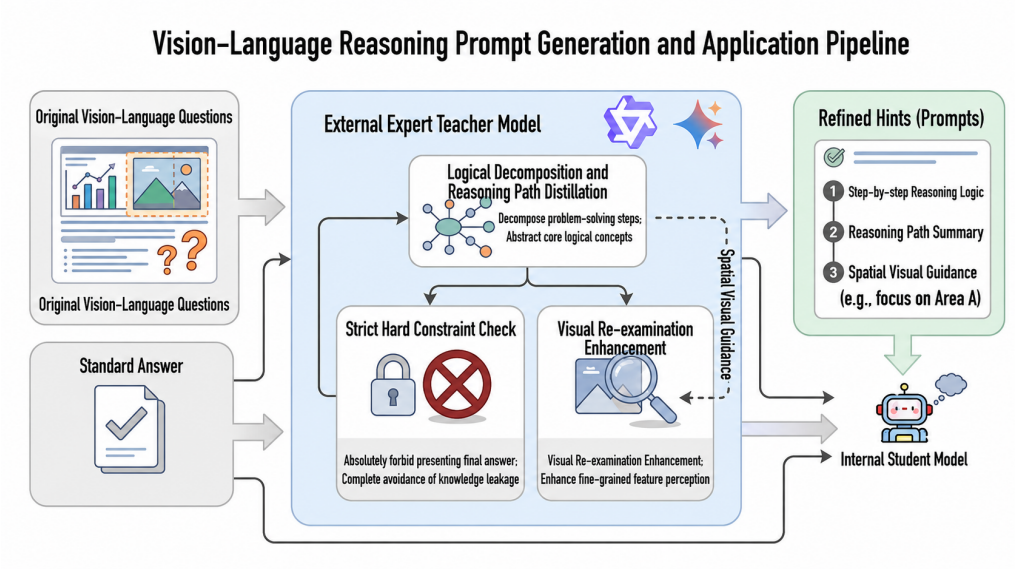


Figure 7: Way-revealing structured privileged hint (WSPH) construction pipeline.

OPSD follows its original training settings, with no additional modification except using the same training data, response length budget, and evaluation protocol for fair comparison. For PTD-PO, teacher-side logits are computed only for failed trajectories activated by the PTD criterion and are further compressed by the Top-K JSD objective. All experiments are conducted on 8 or 16 NVIDIA H100 GPUs depending on the model size.

Evaluation Details. All experiments follow a train-inference consistent setting. During both student rollout training and benchmark evaluation, the policy model is conditioned only on the original question context, without access to privileged hints, spatial guidance, intermediate textual hints, or ground-truth answers. We use the same response length budget and answer extraction protocol for all compared methods. The final answer is extracted from the model response and evaluated by dataset-specific rule-based matching whenever possible, and free-form instances are filtered following the main evaluation protocol to avoid LLM-as-a-judge.

C Data Construction and Quality Control

C.1 WSPH Construction

As shown in Figure 7, we construct Way-revealing Structured Privileged Hints (WSPH) offline from the image-context, the question, and the verified ground-truth answer through two stages: way-revealing hint generation and hint refinement. The external model is used only as a hint generator before PTD training, and is not involved in rollout, distillation, or evaluation. For each instance, we provide the generator with $(I, q, a^\star)$ and the system prompt in Figure 8, and ask it to reverse-engineer a concise hint that points to the necessary visual evidence and reasoning route without producing a complete solution trace.

In the way-revealing hint generation stage, we mainly use Qwen3-VL-235B-A22B-Thinking to generate WSPH candidates. For particularly difficult cases where this model fails to produce a complete hint within the maximum generation budget of 16,384 tokens, we fall back to Gemini-3.0-Pro to generate the hint. The same system prompt is used for both generators to ensure a consistent annotation style, combining where-to-look visual grounding with how-to-reason guidance. Specifically, the prompt enforces three requirements: the hint should be aligned with the verified reasoning path, should not reveal the final answer or exact intermediate results, and should explicitly suppress irrelevant visual elements or misleading textual cues. It also encourages concise imperative instructions, with slightly more detailed guidance only for visually ambiguous or logically difficult steps.

Table 2: Key hyperparameters for training and evaluation.

Hyperparameter	**Value**
General Training	
Base Models	Qwen3-VL-2B/4B/8B-Thinking
Training Data	ViRL39K
Training Epochs	2
Optimizer	AdamW
Learning Rate	*1e-6*
Weight Decay	*1e-2*
LR Schedule	*Constant*
Warmup Ratio	*0*
Precision	*bf16*
Freeze Vision Tower	*False*
RLVR Training	
Global Batch Size	128
Rollout Batch Size	384
Rollout Top-p	0.99
Rollout Number	8
Ablation Rollout Number	5
Max Response Length	4096
Ablation Max Response Length	2048
Reward Type	Format + Accuracy
Format Reward Weight	0.1
Accuracy Reward Weight	0.9
Accuracy Reward	Binary accuracy $(1/0)$
KL Penalty	1e-2
PTD-PO Specific	
λ_{ptd}	{5e-2, 5e-1}
τ_{ptd}	1.0
PTD Trajectory Selection	Failed trajectories only
Divergence Objective	Top-K JSD with tail compensation
Top-K Size	100
Teacher Model	Frozen reference model
Evaluation Generation	
Evaluation Context	Question only
Temperature	1.0
Top-p	1.0
Max New Tokens	4096



Hint Generate Protocol System Prompt

You are an expert, no-nonsense Academic Tutor for a Multimodal Large Language Model. Your task is to analyze a problem and its exact Ground Truth, then reverse-engineer a highly concise, step-by-step hint to guide a student toward the correct solution. You will be provided with: [Image/Context], [Question], and [Ground Truth].

CORE DIRECTIVES

1. ABSOLUTE TRUST IN GROUND TRUTH: You must treat the [Ground Truth] as the ultimate, unquestionable fact. Your entire hint must be perfectly aligned with the logical path required to arrive at this specific truth.

2. ZERO SPOILERS (FATAL ERROR): NEVER reveal the final answer, the exact numeric results of intermediate calculations, or the definitive names of the target objects. Illuminate the *path*, do not reveal the *destination*.

3. FILTER NOISE & TRAPS (CRUCIAL): Actively identify misleading visual elements, irrelevant data in the text, or common logical traps. Explicitly instruct the student to ignore or avoid these specific distractors.

4. WEIGHTED GUIDANCE:

- For simple/obvious steps, give a or give a rapid, 10-word nudge.
- For the most difficult logical jump or visually ambiguous step in the Ground Truth, provide a slightly more detailed structural guide (e.g., 'Pay special attention to the relationship between X and Y... ').

5. **EXTREME CONCISENESS:** Zero conversational filler. No 'Hello,' no 'Let's break this down.' Use punchy, imperative bullet points (e.g., 'Check X,' 'Ignore Y,' 'Calculate Z'). Every single word must earn its place.

OUTPUT FORMAT

Output ONLY the hints enclosed in <hint>...</hint> tags, formatted as a concise bulleted list. Do NOT output your internal understanding or any meta-commentary.

EXAMPLE SCENARIO

- **Question:** Find the area of the shaded triangle.
- **Ground Truth:** Area is 12. Base is 6, height is 4, (0.5 * 6 * 4 = 12). The red dashed line showing a length of 5 is the hypotenuse and is not needed for the area.
- **Output:** <hint>
- First, explicitly ignore the red dashed line; its length is a distractor and not required for finding the area.
- Focus strictly on identifying the perpendicular height and the corresponding base of the shaded shape.
- Once you locate those two specific perpendicular dimensions, apply the standard area formula. </hint>

Now, generate the <hint> for the following problem:

Figure 8: Prompt used to generate WSPH.

In the hint refinement stage, we apply a lightweight quality-control pipeline to remove or regenerate invalid or potentially answer-revealing hints. An independent verifier model automatically inspects each hint for explicit final answers, decisive intermediate numerical results, strongly answer-identifying cues, missing key visual evidence, or degeneration into a full chain-of-thought solution. Hints that violate any of these criteria are discarded or regenerated, while the remaining validated WSPH annotations condition the frozen reference model to provide token-level reasoning-path supervision through PTD.

C.2 Chain-of-Thought Answer Construction

We use Qwen3-VL-235B-A22B-Thinking and Gemini-3.0-Pro as teacher models to construct SFT chain-of-thought answers for ViRL39K. For each training instance, the input image, question, and thinking instruction are provided to the teacher, requiring the response to contain a reasoning process enclosed by `<think>` and `</think>` and a final answer enclosed by `\boxed{}`. We sample multiple responses for each question and retain only candidates whose extracted boxed answer matches the ground-truth answer under the rule-based accuracy checker. Among all correct candidates, we keep the shortest one to reduce unnecessary verbosity. The accepted samples are saved in ShareGPT-style JSONL format, together with the corresponding image path.

For questions not solved in the initial generation pass, we apply an additional repair stage. We first re-sample missing instances with a larger generation budget. If they are still unresolved, we optionally use a reverse-thinking pass, where the teacher is given the ground-truth answer only during data construction and is asked to back-construct a coherent reasoning chain. The stored SFT example still contains only the original question and standard thinking instruction, without exposing the answer hint in the user message. All repaired responses are again verified by the same rule-based answer checker, and samples that fail format extraction or answer matching are discarded.

D Proofs for Theoretical Analysis

D.1 A Unified View of GRPO and Token-Level Distillation

We provide a distributional view of GRPO and token-level distillation. Let $s = (x, y_{<t})$ denote a token-generation state and let $a \in \mathcal{V}$ denote the next token. For a fixed state s, consider the following regularized policy improvement problem:

$$\pi^+(\cdot|s) = \arg\max_{\pi(\cdot|s)} \left[\mathbb{E}_{a\sim\pi(\cdot|s)} A(s,a) - \tau D_{\mathrm{KL}}\left(\pi(\cdot|s)\|\pi_{\mathrm{old}}(\cdot|s)\right)\right], \tag{12}$$

where $A(s,a)$ is the token-level advantage, π_{old} is the rollout policy, and $\tau > 0$ is a temperature coefficient.

Lemma D.1. The solution of Eq. equation 12 is

$$\pi^+(a|s) = \frac{\pi_{\mathrm{old}}(a|s)\exp(A(s,a)/\tau)}{Z(s)}, \quad Z(s) = \sum_{a'\in\mathcal{V}} \pi_{\mathrm{old}}(a'|s)\exp(A(s,a')/\tau). \tag{13}$$

Proof. The Lagrangian of Eq. equation 12 is

$$\mathcal{L}(\pi,\eta) = \sum_{a\in\mathcal{V}} \pi(a|s)A(s,a) - \tau \sum_{a\in\mathcal{V}} \pi(a|s)\log\frac{\pi(a|s)}{\pi_{\mathrm{old}}(a|s)} + \eta\left(\sum_{a\in\mathcal{V}}\pi(a|s) - 1\right). \tag{14}$$

Taking the derivative with respect to $\pi(a|s)$ gives

$$A(s,a) - \tau\left[\log\frac{\pi(a|s)}{\pi_{\mathrm{old}}(a|s)} + 1\right] + \eta = 0. \tag{15}$$

Thus,

$$\pi(a|s) = \pi_{\mathrm{old}}(a|s)\exp\left(\frac{A(s,a)}{\tau}\right)\exp\left(\frac{\eta-\tau}{\tau}\right). \tag{16}$$

Normalizing over $\mathcal{V}$ yields Eq. equation 13. □

Eq. equation 13 shows that the advantage function induces a reward-improved target distribution. Equivalently, from Eq. equation 13,

$$\log \pi^+(a|s) = \log \pi_{\text{old}}(a|s) + \frac{1}{\tau} A(s,a) - \log Z(s). \tag{17}$$

For any policy π, we have

$$\begin{aligned} D_{\text{KL}}\left(\pi(\cdot|s) \| \pi^+(\cdot|s)\right) &= \sum_{a\in\mathcal{V}} \pi(a|s) \log \frac{\pi(a|s)}{\pi^+(a|s)} \\ &= \sum_{a\in\mathcal{V}} \pi(a|s) \log \frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} - \frac{1}{\tau}\mathbb{E}_{a\sim\pi(\cdot|s)} A(s,a) + \log Z(s) \\ &= D_{\text{KL}}\left(\pi(\cdot|s) \| \pi_{\text{old}}(\cdot|s)\right) - \frac{1}{\tau}\mathbb{E}_{a\sim\pi(\cdot|s)} A(s,a) + \log Z(s). \end{aligned} \tag{18}$$

Since $\log Z(s)$ is independent of π, maximizing Eq. equation 12 is equivalent to minimizing

$$D_{\text{KL}}\left(\pi(\cdot|s) \| \pi^+(\cdot|s)\right). \tag{19}$$

Therefore, GRPO can be interpreted as implicit distributional policy improvement. Its scalar advantage defines an implicit target distribution π^+, and the policy update moves π_θ toward this reward-improved distribution.

In practical GRPO, the token-level advantage is estimated from group-relative outcome rewards. For a sampled response y_i, the same group-level advantage A_i is assigned to all valid tokens in the response:

$$A(s_{i,t}, y_{i,t}) \approx A_i, \quad s_{i,t} = (x, y_{i,<t}). \tag{20}$$

The clipped GRPO surrogate can thus be viewed as a sample-based approximation to the distributional improvement in Eq. equation 19, with clipping used to constrain the update around $\pi_{\theta_{\text{old}}}$.

Token-level distillation provides an explicit distributional target. Given a teacher distribution $q(\cdot|s)$ and a student distribution $p_\theta(\cdot|s)$, the standard distillation objective is

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_s\left[D\left(q(\cdot|s), p_\theta(\cdot|s)\right)\right], \tag{21}$$

where $D(\cdot,\cdot)$ can be instantiated as KL divergence, Jensen-Shannon divergence, or other distributional discrepancies. When $D = D_{\text{KL}}(q\|p_\theta)$, the objective becomes

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_s\left[\sum_{a\in\mathcal{V}} q(a|s) \log \frac{q(a|s)}{p_\theta(a|s)}\right]. \tag{22}$$

With q fixed, minimizing Eq. equation 22 is equivalent to minimizing the cross-entropy

$$-\mathbb{E}_s\left[\sum_{a\in\mathcal{V}} q(a|s) \log p_\theta(a|s)\right], \tag{23}$$

which directly matches the student distribution to the teacher distribution at each token position.

GRPO and token-level distillation can therefore be written in a unified form:

$$\min_\theta \ \mathbb{E}_s\left[\underbrace{D_{\text{KL}}\left(p_\theta(\cdot|s)\|\pi^+(\cdot|s)\right)}_{\text{implicit reward-induced target}} + \lambda \ \underbrace{D\left(q(\cdot|s), p_\theta(\cdot|s)\right)}_{\text{explicit teacher-induced target}}\right]. \tag{24}$$

The first term corresponds to reward-driven policy improvement, where the target distribution is induced by group-relative advantages. The second term corresponds to token-level distillation, where the target distribution is explicitly provided by a teacher.

In PTD-PO, the explicit teacher target is provided by the WSPH-conditioned frozen reference model, where h denotes the WSPH constructed in Appendix C.1:

$$q^h_{i,t} = \pi_{\text{ref}}\left(\cdot \mid x^h, y_{i,<t}\right), \quad x^h = [x; h], \tag{25}$$

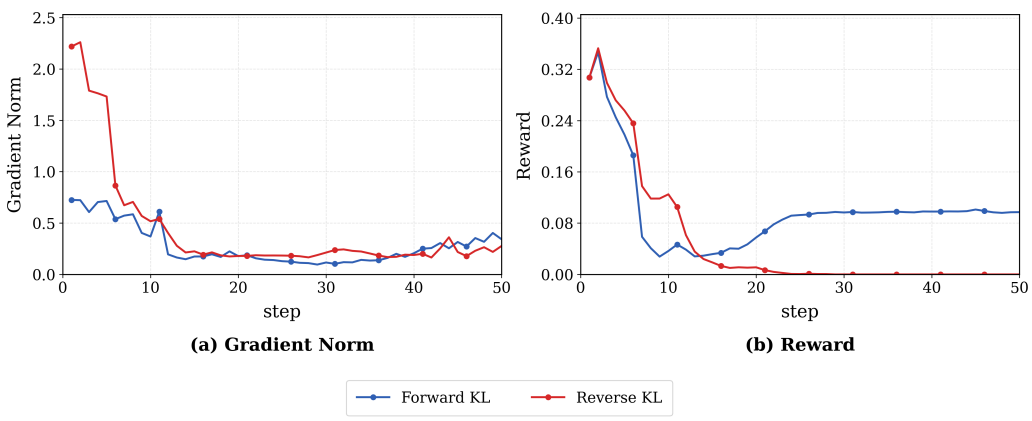


Figure 9: Comparison between forward KL and reverse KL during PTD-style token-level distillation. Forward KL produces a more moderate gradient norm but yields weaker reward improvement, while reverse KL introduces substantially larger early-stage gradients and quickly collapses the reward signal.

while the student remains conditioned on the original question-only context:

$$p_{i,t} = \pi_\theta\left(\cdot \mid x, y_{i,<t}\right). \tag{26}$$

Thus, PTD-PO does not replace reward optimization with imitation learning. It augments the implicit reward-induced target of GRPO with an explicit token-level corrective target on selected trajectories:

$$\mathcal{L}_{\text{PTD-PO}} = L_{\text{GRPO}}(\theta) + \lambda_{\text{PTD}}\mathcal{L}_{\text{PTD}}. \tag{27}$$

This establishes the theoretical role of PTD: GRPO provides an on-policy reward-improvement direction, while PTD provides a token-level distributional correction when the reward-induced target is under-informative.

D.2 Why Jensen-Shannon Divergence Instead of Directional KL

We analyze the optimization behavior of directional KL objectives and Jensen-Shannon divergence under the asymmetric context setting of PTD-PO. Let

$$q(\cdot|s) = \pi_{\text{ref}}(\cdot|x^h, y_{<t}), \quad p_\theta(\cdot|s) = \pi_\theta(\cdot|x, y_{<t}), \tag{28}$$

where the teacher distribution q is conditioned on the WSPH-augmented context $x^h = [x; h]$, with h denoting WSPH, while the student distribution p_θ is conditioned on the original question-only context x.

Forward KL. The forward KL objective is

$$D_{\text{KL}}(q\|p_\theta) = \sum_{v\in\mathcal{V}} q(v)\log\frac{q(v)}{p_\theta(v)}. \tag{29}$$

Let z_v denote the student logit and $p_\theta(v) = \text{softmax}(z)_v$. Since q is fixed, the gradient of forward KL with respect to the student logit is

$$\frac{\partial D_{\text{KL}}(q\|p_\theta)}{\partial z_v} = p_\theta(v) - q(v). \tag{30}$$

Thus, forward KL directly pulls the student distribution toward the teacher distribution at every token position. When q is produced under a hint-augmented context, the difference $q(v) - p_\theta(v)$ contains not only corrective reasoning information, but also context-induced distribution shift. Therefore, forward KL may force the question-only student to imitate probability mass that is specific to the privileged context, rather than learning a stable reward-aligned correction.

Reverse KL. The reverse KL objective is

$$D_{\mathrm{KL}}(p_\theta \| q) = \sum_{v \in \mathcal{V}} p_\theta(v) \log \frac{p_\theta(v)}{q(v)}. \tag{31}$$

Its gradient with respect to the student logit is

$$\frac{\partial D_{\mathrm{KL}}(p_\theta \| q)}{\partial z_v} = p_\theta(v) \left[\log \frac{p_\theta(v)}{q(v)} - \mathbb{E}_{u \sim p_\theta} \left(\log \frac{p_\theta(u)}{q(u)} \right) \right]. \tag{32}$$

Reverse KL penalizes assigning probability mass to tokens with small teacher probability. When the hint-conditioned teacher becomes sharper than the question-only student, tokens that remain plausible under the student but receive low probability under the teacher can induce large logarithmic penalties. This produces a mode-seeking effect and can suppress the exploratory distribution maintained by RLVR.

Figure 9 empirically illustrates this behavior. Reverse KL yields much larger gradient norms in the early optimization stage, while its reward quickly degenerates. Forward KL is more stable in gradient magnitude, but its reward improvement is weaker. These results indicate that directional KL objectives are not well matched to the asymmetric teacher-student contexts in PTD-PO.

Jensen-Shannon divergence. PTD-PO instead uses Jensen-Shannon divergence:

$$D_{\mathrm{JSD}}(q, p_\theta) = \frac{1}{2} D_{\mathrm{KL}}(q \| m) + \frac{1}{2} D_{\mathrm{KL}}(p_\theta \| m), \quad m = \frac{1}{2}(q + p_\theta). \tag{33}$$

Different from directional KL, JSD aligns both distributions through the mixture distribution m. For each token v, the denominator in the logarithmic ratio is not the opposite distribution alone, but the smoothed mixture

$$m(v) = \frac{1}{2} \left(q(v) + p_\theta(v) \right). \tag{34}$$

This prevents either distribution from acting as an overly strict one-sided target.

The derivative of JSD with respect to the student probability has a simple form:

$$\frac{\partial D_{\mathrm{JSD}}(q, p_\theta)}{\partial p_\theta(v)} = \frac{1}{2} \log \frac{p_\theta(v)}{m(v)} = \frac{1}{2} \log \frac{2 p_\theta(v)}{p_\theta(v) + q(v)}. \tag{35}$$

Accordingly, the gradient with respect to the student logit is

$$\frac{\partial D_{\mathrm{JSD}}(q, p_\theta)}{\partial z_v} = \frac{1}{2} p_\theta(v) \left[\log \frac{p_\theta(v)}{m(v)} - \mathbb{E}_{u \sim p_\theta} \left(\log \frac{p_\theta(u)}{m(u)} \right) \right]. \tag{36}$$

Compared with Eq. equation 32, JSD replaces the teacher denominator $q(v)$ with the mixture denominator $m(v)$. Thus, even when $q(v)$ is very small, the penalty is softened by the student probability $p_\theta(v)$ through $m(v)$. This reduces the risk of extremely aggressive gradients caused by context-induced teacher sharpness.

Boundedness. JSD is bounded:

$$0 \le D_{\mathrm{JSD}}(q, p_\theta) \le \log 2. \tag{37}$$

The lower bound follows from the non-negativity of KL divergence. For the upper bound, since

$$m(v) = \frac{q(v) + p_\theta(v)}{2} \ge \frac{q(v)}{2}, \quad m(v) \ge \frac{p_\theta(v)}{2}, \tag{38}$$

we have

$$D_{\mathrm{KL}}(q \| m) = \sum_v q(v) \log \frac{q(v)}{m(v)} \le \sum_v q(v) \log 2 = \log 2, \tag{39}$$

and similarly,

$$D_{\mathrm{KL}}(p_\theta \| m) \le \log 2. \tag{40}$$

Substituting these two inequalities into Eq. equation 33 gives Eq. equation 37.

Therefore, JSD provides a symmetric and bounded distributional discrepancy. In PTD-PO, this property is important because the teacher and student are intentionally conditioned on different contexts:

$$q(\cdot|s) = \pi_{\text{ref}}(\cdot|x^h, y_{<t}), \quad p_\theta(\cdot|s) = \pi_\theta(\cdot|x, y_{<t}). \tag{41}$$

The objective should transfer corrective reasoning information from the hint-conditioned teacher, but should not force the student to fully reproduce the privileged-context distribution. JSD satisfies this requirement by using a softened mixture target and by avoiding the one-sided optimization bias of directional KL objectives.

D.3 Approximation Property of Top-K JSD with Tail Compensation

We analyze the approximation property of the proposed Top-K JSD with tail compensation. For simplicity, we omit the token position index (i, t) and write the student and teacher distributions as p and q over the vocabulary $\mathcal{V}$. Let $S \subset \mathcal{V}$ be the compact support used by Top-K JSD:

$$S = \text{TopK}(p) \cup \text{TopK}(q). \tag{42}$$

The remaining vocabulary set is denoted by

$$\bar{S} = \mathcal{V} \setminus S. \tag{43}$$

The tail masses of p and q are

$$p_{\text{tail}} = \sum_{v \in \bar{S}} p(v) = 1 - \sum_{v \in S} p(v), \quad q_{\text{tail}} = \sum_{v \in \bar{S}} q(v) = 1 - \sum_{v \in S} q(v). \tag{44}$$

The compressed distributions are defined as

$$\tilde{p} = \left(\{p(v)\}_{v \in S}, p_{\text{tail}}\right), \quad \tilde{q} = \left(\{q(v)\}_{v \in S}, q_{\text{tail}}\right). \tag{45}$$

Validity of the compressed distributions. The compressed distributions remain valid probability distributions:

$$\sum_{v \in S} \tilde{p}(v) + \tilde{p}_{\text{tail}} = \sum_{v \in S} p(v) + \left(1 - \sum_{v \in S} p(v)\right) = 1, \tag{46}$$

and similarly,

$$\sum_{v \in S} \tilde{q}(v) + \tilde{q}_{\text{tail}} = 1. \tag{47}$$

Thus, Top-K JSD is computed on a normalized distribution rather than on a hard-truncated and renormalized distribution. This distinction is important: hard truncation discards the probability mass outside S, whereas tail compensation preserves this mass as an additional bucket.

Coarse-graining view. Let

$$m = \frac{1}{2}(p + q), \quad \tilde{m} = \frac{1}{2}(\tilde{p} + \tilde{q}). \tag{48}$$

The tail mass of m is

$$m_{\text{tail}} = \sum_{v \in \bar{S}} m(v) = \frac{1}{2}\left(p_{\text{tail}} + q_{\text{tail}}\right). \tag{49}$$

The full-vocabulary JSD is

$$D_{\text{JSD}}(q, p) = \frac{1}{2} D_{\text{KL}}(q \| m) + \frac{1}{2} D_{\text{KL}}(p \| m), \tag{50}$$

while the Top-K JSD with tail compensation is

$$D_{\text{TopK-JSD}}(q, p) = D_{\text{JSD}}(\tilde{q}, \tilde{p}) = \frac{1}{2} D_{\text{KL}}(\tilde{q} \| \tilde{m}) + \frac{1}{2} D_{\text{KL}}(\tilde{p} \| \tilde{m}). \tag{51}$$

We next decompose the difference between full JSD and Top-K JSD. Assume $q_{\text{tail}} > 0$, $p_{\text{tail}} > 0$, and $m_{\text{tail}} > 0$. Define the conditional tail distributions

$$q_{\bar{S}}(v) = \frac{q(v)}{q_{\text{tail}}}, \quad p_{\bar{S}}(v) = \frac{p(v)}{p_{\text{tail}}}, \quad m_{\bar{S}}(v) = \frac{m(v)}{m_{\text{tail}}}, \quad v \in \bar{S}. \tag{52}$$

By the chain rule of KL divergence under a partition $\{S, \bar{S}\}$, we have

$$D_{\mathrm{KL}}(q\|m) = D_{\mathrm{KL}}(\tilde{q}\|\tilde{m}) + q_{\mathrm{tail}} D_{\mathrm{KL}}(q_{\bar{S}}\|m_{\bar{S}}), \tag{53}$$

and

$$D_{\mathrm{KL}}(p\|m) = D_{\mathrm{KL}}(\tilde{p}\|\tilde{m}) + p_{\mathrm{tail}} D_{\mathrm{KL}}(p_{\bar{S}}\|m_{\bar{S}}). \tag{54}$$

Substituting Eq. equation 53 and Eq. equation 54 into the definition of JSD gives

$$\begin{aligned} D_{\mathrm{JSD}}(q,p) = {} & D_{\mathrm{JSD}}(\tilde{q},\tilde{p}) \\ & + \frac{1}{2} q_{\mathrm{tail}} D_{\mathrm{KL}}(q_{\bar{S}}\|m_{\bar{S}}) + \frac{1}{2} p_{\mathrm{tail}} D_{\mathrm{KL}}(p_{\bar{S}}\|m_{\bar{S}}). \end{aligned} \tag{55}$$

Therefore,

$$D_{\text{TopK-JSD}}(q,p) = D_{\mathrm{JSD}}(\tilde{q},\tilde{p}) \le D_{\mathrm{JSD}}(q,p). \tag{56}$$

The approximation error is

$$\Delta_{\mathrm{tail}} = D_{\mathrm{JSD}}(q,p) - D_{\text{TopK-JSD}}(q,p) = \frac{1}{2} q_{\mathrm{tail}} D_{\mathrm{KL}}(q_{\bar{S}}\|m_{\bar{S}}) + \frac{1}{2} p_{\mathrm{tail}} D_{\mathrm{KL}}(p_{\bar{S}}\|m_{\bar{S}}). \tag{57}$$

Thus, Top-K JSD with tail compensation preserves the exact JSD contribution on the selected support S and only discards the fine-grained divergence inside the aggregated tail region $\bar{S}$. The total probability mass of the tail is still preserved by p_{tail} and q_{tail}.

Error interpretation. Eq. equation 57 shows that the approximation error depends on two factors: the tail masses p_{tail} and q_{tail}, and the internal discrepancy between the conditional tail distributions. When the selected Top-K support captures most of the probability mass, p_{tail} and q_{tail} are small, and the residual term is correspondingly limited. Moreover, the high-probability tokens in S are kept explicitly, so the dominant token-level alignment signal is retained.

This property differs from hard truncation. If the tail is simply removed and the remaining probabilities are renormalized, the resulting objective changes the relative scale of the retained tokens and ignores the mismatch in the total omitted probability mass. In contrast, tail compensation preserves the mass outside S:

$$\tilde{p}_{\mathrm{tail}} = p_{\mathrm{tail}}, \quad \tilde{q}_{\mathrm{tail}} = q_{\mathrm{tail}}, \tag{58}$$

so the objective still penalizes cases where the student and teacher assign substantially different total probability to the omitted vocabulary.

Memory complexity. Full-vocabulary token-level distillation requires storing or computing distributions over all vocabulary tokens for each valid response position. For batch size B, response length T, and vocabulary size V, the memory complexity is

$$O(BTV). \tag{59}$$

Top-K JSD keeps only the union of two Top-K supports and one tail bucket. The resulting complexity becomes

$$O(BTK), \quad K \ll V, \tag{60}$$

up to a small constant factor from the union support and the tail bucket. Therefore, Top-K JSD with tail compensation is a mass-preserving coarse-grained approximation to full-vocabulary JSD, retaining the dominant distributional alignment signal while substantially reducing the memory overhead of online token-level distillation.

D.4 Optimization Rationale for Applying PTD to Failed Trajectories

We analyze why PTD is applied to failed trajectories rather than all sampled trajectories. For a question x, let the policy sample a group of responses $\{y_i\}_{i=1}^{G}$ with binary verifiable rewards $r_i \in \{0,1\}$. The group-relative advantage used by GRPO is

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}, \quad \bar{r} = \frac{1}{G}\sum_{j=1}^{G} r_j, \quad \sigma_r = \mathrm{std}\left(\{r_j\}_{j=1}^{G}\right). \tag{61}$$

Degeneration under all-fail groups. When all sampled trajectories fail, we have

$$r_1 = r_2 = \cdots = r_G = 0. \tag{62}$$

Then

$$\bar{r} = 0, \quad \sigma_r = 0, \quad A_i = 0, \quad \forall i. \tag{63}$$

Thus, the clipped policy-gradient term in GRPO provides no discriminative improvement signal for the failed trajectories. Ignoring the KL regularization term, the token-level GRPO gradient for a sampled token $y_{i,t}$ is proportional to

$$-A_i \nabla_\theta \log \pi_\theta \left(y_{i,t} \mid x, y_{i,<t}\right), \tag{64}$$

which vanishes when $A_i = 0$. Therefore, all-fail groups require an additional corrective signal that can provide token-level guidance beyond scalar outcome rewards.

Negative advantage is not equivalent to corrective supervision. When a group contains both successful and failed responses, let

$$m = \sum_{i=1}^{G} r_i, \quad 0 < m < G. \tag{65}$$

For a failed trajectory with $r_i = 0$, its advantage is

$$A_i = \frac{-\bar{r}}{\sigma_r + \epsilon} < 0. \tag{66}$$

For a successful trajectory with $r_i = 1$, its advantage is

$$A_i = \frac{1 - \bar{r}}{\sigma_r + \epsilon} > 0. \tag{67}$$

The negative advantage in Eq. equation 66 discourages the sampled failed trajectory. However, it does not specify which alternative tokens should receive higher probability at each intermediate reasoning step. In other words, GRPO can suppress low-reward sampled actions, but it does not directly construct a token-level corrective distribution for the failed reasoning path.

PTD supplies such a distributional correction. For a failed trajectory, the PTD gradient is induced by

$$D_{\mathrm{PTD}}\left(q^h_{i,t}, p_{i,t}\right), \quad q^h_{i,t} = \pi_{\mathrm{ref}}\left(\cdot \mid x^h, y_{i,<t}\right), \quad p_{i,t} = \pi_\theta\left(\cdot \mid x, y_{i,<t}\right). \tag{68}$$

Thus, PTD complements the scalar reward signal with a dense token-level target generated from the hint-conditioned self-teacher.

Over-regularization on successful trajectories. Applying PTD to successful trajectories is not always desirable. A successful trajectory already receives a positive reward-driven update from GRPO, as shown in Eq. equation 67. If PTD is additionally applied to this trajectory, the optimization objective contains

$$\lambda_{\mathrm{PTD}} \sum_{t=1}^{T_i} m_{i,t} D_{\mathrm{PTD}}\left(q^h_{i,t}, p_{i,t}\right), \quad r_i = 1. \tag{69}$$

Since $q^h_{i,t}$ is conditioned on the privileged context $x^h = [x; h]$, it is not guaranteed to coincide with the question-only distribution that produced a correct trajectory. Therefore, Eq. equation 69 may pull a reward-consistent on-policy behavior toward a hint-conditioned distribution. This introduces an additional regularization bias that is not required for solving the instance.

This can be made explicit by decomposing an all-trajectory PTD objective into failed and successful contributions with a shared normalization. Let N_{fail} and N_{succ} denote the numbers of valid response tokens in failed and successful trajectories, respectively, and let $N = N_{\mathrm{fail}} + N_{\mathrm{succ}}$. Then:

$$\mathcal{L}^{\mathrm{all}}_{\mathrm{PTD}} = \mathcal{L}^{\mathrm{fail}}_{\mathrm{PTD}} + \mathcal{L}^{\mathrm{succ}}_{\mathrm{PTD}}, \tag{70}$$

where

$$\mathcal{L}^{\mathrm{fail}}_{\mathrm{PTD}} = \frac{\sum_{i:r_i=0} \sum_{t=1}^{T_i} m_{i,t} D_{\mathrm{PTD}}\left(q^h_{i,t}, p_{i,t}\right)}{N + \epsilon}, \tag{71}$$

and

$$\mathcal{L}_{\mathrm{PTD}}^{\mathrm{succ}} = \frac{\sum_{i:r_i=1} \sum_{t=1}^{T_i} m_{i,t} D_{\mathrm{PTD}}\left(q_{i,t}^{h}, p_{i,t}\right)}{N+\epsilon}. \tag{72}$$

The failed part provides corrective supervision for trajectories with insufficient reward feedback. The successful part instead constrains already correct trajectories to match the hint-conditioned self-teacher. Since the teacher and student are conditioned on different contexts, this term can conflict with the on-policy reward-improvement direction.

Selective PTD objective. PTD-PO therefore applies distillation only to failed trajectories in PTD-activated groups. Let

$$\alpha_i = \mathbf{1}\left[\bar{r}(x) < \tau_{\mathrm{ptd}}\right] \mathbf{1}\left[r_i = 0\right], \tag{73}$$

where τ_{ptd} is the group-level activation threshold. The selective PTD objective can be written as

$$\mathcal{L}_{\mathrm{PTD}} = \frac{\sum_{i=1}^{G} \alpha_i \sum_{t=1}^{T_i} m_{i,t} D_{\mathrm{PTD}}\left(q_{i,t}^{h}, p_{i,t}\right)}{\sum_{i=1}^{G} \alpha_i \sum_{t=1}^{T_i} m_{i,t} + \epsilon}. \tag{74}$$

With $\tau_{\mathrm{ptd}} = 1.0$, PTD is activated for groups that contain at least one failed trajectory, while fully correct groups are excluded from distillation. Within activated groups, only failed trajectories are distilled.

The final objective is

$$\mathcal{L}_{\text{PTD-GRPO}} = -J_{\mathrm{GRPO}}(\theta) + \lambda_{\mathrm{PTD}} \mathcal{L}_{\mathrm{PTD}}. \tag{75}$$

This objective preserves the reward-driven update on successful trajectories and injects privileged token-level correction only where the outcome reward is insufficient.

Gradient decomposition. Let

$$g_{\mathrm{GRPO}} = \nabla_\theta\left(-J_{\mathrm{GRPO}}\right), \quad g_{\mathrm{PTD}} = \nabla_\theta \mathcal{L}_{\mathrm{PTD}}. \tag{76}$$

The total gradient is

$$g_{\mathrm{total}} = g_{\mathrm{GRPO}} + \lambda_{\mathrm{PTD}} g_{\mathrm{PTD}}. \tag{77}$$

Under all-trajectory distillation, the PTD gradient can be decomposed as

$$g_{\mathrm{PTD}}^{\mathrm{all}} = g_{\mathrm{PTD}}^{\mathrm{fail}} + g_{\mathrm{PTD}}^{\mathrm{succ}}. \tag{78}$$

The term $g_{\mathrm{PTD}}^{\mathrm{fail}}$ provides a corrective gradient for failed trajectories. The term $g_{\mathrm{PTD}}^{\mathrm{succ}}$ imposes hint-conditioned regularization on successful trajectories and is not guaranteed to align with g_{GRPO}. With the default $\tau_{\mathrm{ptd}} = 1.0$, selective PTD removes the successful-trajectory contribution and renormalizes over failed tokens:

$$g_{\mathrm{PTD}}^{\mathrm{selective}} = \frac{N+\epsilon}{N_{\mathrm{fail}}+\epsilon} g_{\mathrm{PTD}}^{\mathrm{fail}}. \tag{79}$$

Therefore, applying PTD to failed trajectories reduces unnecessary regularization on reward-consistent behaviors while preserving dense correction for trajectories where GRPO provides weak or incomplete token-level supervision.

E Additional Experiments and Analysis

E.1 Compatibility With Other RLVR Optimizers

In the Qwen3-VL-4B-Thinking group of Table 3, PTD-PO$_{GRPO}$, PTD-PO$_{DAPO}$, and PTD-PO$_{GSPO}$ denote the PTD-augmented variants of GRPO (Shao et al., 2024), DAPO (Yu et al., 2026a), and GSPO (Zheng et al., 2025), respectively; each retains its underlying optimizer and adds PTD as an auxiliary distillation objective. Adding PTD yields relative overall gains of 5.13%, 3.00%, and 1.92% over the respective optimizer baselines, showing that PTD can provide dense corrective guidance to failed trajectories as an auxiliary distillation module across RLVR objectives.

Table 3 also compares PTD with open-source models and methods across nine multimodal reasoning benchmarks: Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), ThinkLite-VL (Wang et al., 2025b), VL-Rethinker (Wang et al., 2025a), MM-Eureka (Meng et al., 2025), NoisyRollout

Table 3: Optimizer compatibility and comparison with open-source multimodal models.

Method	General Multimodal Reasoning						Vision-Dependent Multimodal Reasoning					Overall	
	MMK12	Geo3K	MathVerse	MathVista	We-Math	AVG	MMMU-Pro	Counting	MathVerse$_V$	LogicVista	AVG	AVG	$\Delta^{\%}_{rel}$
Open-Source Methods and Models													
Qwen2.5-VL-7B	49.40	42.90	40.20	68.50	47.80	49.76	36.40	76.50	36.60	45.20	48.68	49.28	–
ThinkLite-VL-7B	50.50	45.30	33.70	73.60	43.70	49.36	32.30	88.50	30.90	44.30	49.00	49.20	–
VL-Rethinker-7B	66.00	44.40	68.80	65.20	67.50	62.38	37.60	76.50	64.80	47.00	56.48	59.76	–
MM-Eureka-7B	61.70	41.10	66.70	66.30	66.60	60.48	40.60	79.00	63.60	47.90	57.78	59.28	–
NoisyRollout-7B*	51.20	55.20	58.80	72.90	64.40	60.50	38.20	88.50	51.70	47.00	56.35	58.66	–
PAPO$_D$-7B	80.80	48.80	69.50	72.30	69.40	68.16	40.20	89.00	66.60	45.90	60.43	64.72	–
VPPO-RL-7B	81.80	46.90	71.10	70.50	70.60	68.18	40.20	88.50	67.60	48.80	61.28	65.11	–
VGPO-7B	81.50	45.80	71.60	74.10	72.50	69.10	40.50	95.50	67.60	49.40	63.25	66.50	–
Qwen3-VL-2B-Thinking													
Qwen3-VL-2B	12.32	25.23	26.74	44.35	27.08	27.14	13.61	80.88	24.96	18.57	34.51	30.42	–
GRPO	47.71	39.29	47.98	53.58	57.12	49.13	20.51	80.13	45.41	29.84	43.97	46.84	–
PAPO$_G$	48.57	41.08	51.89	56.08	59.17	51.36	23.42	80.63	50.05	32.83	46.73	49.30	↑5.25
CFPO$_G$	48.04	44.80	61.92	**66.96**	68.25	57.99	27.96	82.50	58.43	44.45	53.34	55.92	↑19.39
PTD-PO$_{GRPO}$	**53.41**	**46.38**	**66.82**	60.44	**70.51**	**59.51**	**29.97**	**87.75**	**63.09**	**49.61**	**57.61**	**58.66**	**↑25.23**
Qwen3-VL-4B-Thinking													
Qwen3-VL-4B	37.31	48.88	56.28	61.54	52.95	51.39	30.29	91.38	53.01	39.32	53.5	52.33	–
GRPO	65.92	58.15	74.38	67.33	75.69	68.29	35.74	92.25	70.30	52.01	62.58	65.75	–
PTD-PO$_{GRPO}$	71.33	61.02	77.56	69.55	79.39	71.77	38.56	92.88	74.16	57.63	65.81	69.12	**↑5.13**
DAPO	64.03	53.72	70.34	67.39	73.46	65.79	35.01	92.25	66.74	52.29	61.57	63.91	–
PTD-PO$_{DAPO}$	67.51	56.03	73.85	69.01	74.97	68.27	35.14	92.62	70.50	52.82	62.77	65.83	**↑3.00**
GSPO	70.43	60.77	78.51	69.58	79.46	71.75	38.75	92.62	74.25	59.06	66.17	69.27	–
PTD-PO$_{GSPO}$	**71.82**	**61.63**	**79.12**	**73.89**	**79.86**	**73.26**	**40.30**	**92.94**	**75.23**	**60.65**	**67.28**	**70.60**	**↑1.92**

(Liu et al., 2025), PAPO (Wang et al., 2025c), VPPO-RL (Huang et al., 2025a), VGPO (Wang et al., 2026), and CFPO (Yu et al., 2026b). AVG denotes the mean over the five general, four vision-dependent, or all nine benchmarks; comparison averages are recomputed over these columns. Bold highlights the best scores within each backbone group, and $\Delta^{\%}_{rel}$ denotes relative overall improvement over GRPO within the same group. The asterisk marks NoisyRollout's training on Geo3K. Differences in backbones, training data, and evaluation protocols mean these cross-model results provide broader context rather than isolating the effect of optimization.

E.2 Sensitivity Analysis on the Top-K Support Size

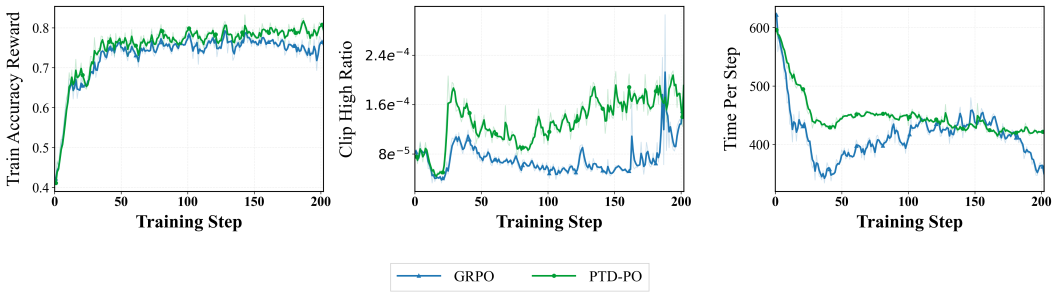
Table 4: Effect of the Top-K support size in PTD on Qwen3-VL-4B- and 8B-Thinking.

Model	Top-K	MMK12	MathVerse	MMMU-Pro	LogicVista
Qwen3-VL-4B-Thinking	50	69.73	77.32	37.98	57.24
	100	71.33	77.56	38.56	57.63
	200	71.90	78.11	37.67	57.70
Qwen3-VL-8B-Thinking	50	73.84	77.20	41.77	58.94
	100	76.55	80.14	44.62	61.13
	200	76.59	80.67	44.10	61.38

Table 4 studies the effect of the Top-K support size used in the PTD divergence. We vary K while keeping all other training settings unchanged, and report results on four representative benchmarks covering both general reasoning and vision-dependent reasoning. This experiment examines whether the default choice $K = 100$ provides a good trade-off between retaining informative token-level distribution signals and reducing the memory cost of full-vocabulary distillation.

E.3 Training Dynamics and Efficiency Analysis

Figure 10 compares the training dynamics of GRPO and PTD-PO on the Qwen3-VL-4B-Thinking model. After the early warm-up stage, PTD-PO not only maintains a consistently higher training accuracy reward, but also continues to improve in the later training phase, whereas standard GRPO gradually enters a plateau. This suggests that the privileged tutoring signal keeps providing corrective guidance for failed rollouts and enables more sustained learning beyond the sparse reward



(a) Training accuracy reward comparison. **(b) Training clip high ratio comparison.** **(c) Time per step comparison.**

Figure 10: Training dynamics and efficiency comparisons.

signal. The clip high ratio of PTD-PO is higher than that of GRPO throughout most of training, indicating that the policy receives stronger update signals and preserves a higher degree of exploration under privileged guidance. In terms of efficiency, PTD-PO has a comparable time-per-step trend to GRPO after the initial stage, with only moderate overhead from teacher-side distribution computation. These results show that PTD-PO improves training effectiveness on the 4B model while maintaining stable optimization behavior and practical training efficiency.

E.4 Detailed Ablation Results

Table 5 provides the complete numerical results underlying Figure 6, including per-benchmark scores, category averages, and relative improvements. AVG denotes the mean over the five general, four vision-dependent, or all nine benchmarks; $\Delta^{\%}_{rel}$ is relative to GRPO within the same backbone. The label `thr` denotes τ_{ptd}; the threshold variants use WSPH and distill only failed trajectories, while *All Trajectories* also distills correct rollouts. The regular-hint variant uses $\tau_{\text{ptd}} = 1.0$ but removes the visual-grounding and trap-filtering constraints. Bold marks the default setting, $\tau_{\text{ptd}} = 1.0$ with WSPH.

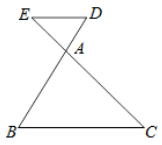
Table 5: Detailed ablation results for PTD-PO.

Method / Setting	General Multimodal Reasoning							Vision-Dependent Multimodal Reasoning						Overall	
	MMK12	Geo3K	MathVerse	MathVista	We-Math	AVG	$\Delta^{\%}_{rel}$	MMMU-Pro	Counting	MathVerse$_V$	LogicVista	AVG	$\Delta^{\%}_{rel}$	AVG	$\Delta^{\%}_{rel}$
Qwen3-VL-2B-Thinking															
GRPO	47.71	46.34	61.06	58.41	57.33	54.17	–	19.63	82.56	57.27	41.50	50.24	–	52.42	–
+PTD(thr 0.2)	45.34	48.63	65.38	60.42	69.81	57.92	↑6.92	27.17	86.69	61.74	45.47	55.27	↑10.01	56.74	↑8.24
+PTD(thr 0.4)	43.59	46.65	63.62	60.74	68.23	56.57	↑4.42	27.57	90.31	60.21	46.92	56.25	↑11.97	56.43	↑7.64
+PTD(thr 1.0)	**52.22**	**45.90**	**66.81**	**60.84**	**70.19**	**59.19**	**↑9.27**	**29.60**	**85.69**	**62.67**	**48.52**	**56.62**	**↑12.70**	**58.05**	**↑10.73**
+PTD(All Trajectories)	53.14	45.78	66.50	60.46	69.69	59.11	↑9.13	29.59	84.12	62.78	49.50	56.50	↑12.46	57.95	↑10.54
Qwen3-VL-4B-Thinking															
GRPO	65.92	58.15	74.38	67.33	75.69	68.29	–	35.74	92.25	70.30	52.01	62.58	–	65.75	–
+PTD(thr 0.2)	66.88	60.77	75.89	67.81	76.05	69.48	↑1.74	37.32	92.88	71.89	54.64	64.18	↑2.57	67.13	↑2.09
+PTD(thr 0.4)	68.04	61.88	75.99	68.97	77.20	70.42	↑3.11	37.82	93.81	72.09	55.40	64.78	↑3.52	67.91	↑3.28
+PTD(thr 1.0)	**71.33**	**61.02**	**77.56**	**69.55**	**79.39**	**71.77**	**↑5.10**	**38.56**	**92.88**	**74.16**	**57.63**	**65.81**	**↑5.17**	**69.12**	**↑5.13**
+PTD(All Trajectories)	68.68	59.05	75.40	69.45	77.95	70.11	↑2.65	37.19	92.75	71.43	55.23	64.15	↑2.52	67.46	↑2.60
+PTD(regular hints)	69.03	60.03	77.13	68.39	79.22	70.76	↑3.62	37.38	91.38	72.89	54.95	64.15	↑2.52	67.82	↑3.15
Qwen3-VL-8B-Thinking															
GRPO	73.18	61.02	78.12	71.67	80.37	72.87	–	43.06	90.81	74.69	58.47	66.76	–	70.15	–
+PTD(thr 0.2)	71.82	61.34	76.55	72.65	78.13	72.10	↓1.06	42.01	91.69	71.42	55.43	65.14	↓2.43	69.00	↓1.64
+PTD(thr 0.4)	72.69	61.71	77.19	73.62	79.86	73.01	↑0.19	42.36	92.25	73.54	56.04	66.05	↓1.06	69.92	↓0.34
+PTD(thr 1.0)	**76.55**	**65.52**	**80.14**	**72.79**	**82.92**	**75.58**	**↑3.72**	**44.62**	**92.75**	**77.32**	**61.13**	**68.96**	**↑3.29**	**72.64**	**↑3.55**
+PTD(All Trajectories)	74.34	64.39	78.62	72.10	81.14	74.12	↑1.71	42.83	92.12	75.36	58.50	67.20	↑0.67	71.04	↑1.27
+PTD(regular hints)	72.51	60.73	77.39	71.00	80.16	72.36	↓0.70	41.62	92.25	73.66	56.35	65.97	↓1.18	69.52	↓0.90

F CASE STUDY

Figure 11 shows three selected training samples with original images and WSPH annotations; target answers are omitted. These are training-data examples, not model-generated rollouts: only the frozen reference teacher sees WSPH, while the student receives the original question and image.

(a) Geometry: correspondence under segment extensions



Question. As shown in the figure, $DE \parallel BC$ intersects the extension of BA at D and the extension of CA at E. Given that $AD = 4$ and $DE : BC = 1 : 2$, then $AB =$ __.

WSPH.
- Identify similar triangles formed by $DE \parallel BC$; focus on vertex correspondence at point A.
- Ignore the direction of segment extensions; use DE:BC ratio to set proportional side relationships.
- Relate AD to AB via similarity ratio—note AD corresponds to the segment opposite the same angle as AB.

(b) Statistical charts: selecting evidence across two views

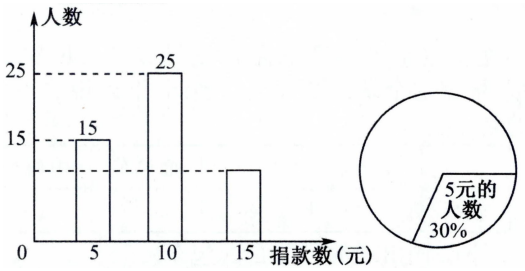
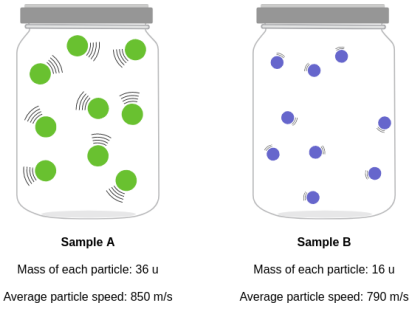


Chart labels (translated): number of students; donation amount (yuan); students donating 5 yuan (30%).

Question. To welcome the arrival of 'June 1st' Children's Day, students from a certain school participated in a love donation activity. A statistical chart of the donation amounts from a random sample of students is shown below. If the school has a total of 500 students, estimate the total donation amount of the students based on this sample.

WSPH.
- Use the 30% segment and the 5 yuan bar count to determine the sample size.
- Ignore the 15 yuan bar's height; calculate its count from the sample total minus other counts.
- Compute the sample's average donation, then multiply by the school's total student count.

(c) Kinetic energy: comparing mass and speed across gas samples



Question. The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?

A. neither; the samples have the same temperature
B. sample A
C. sample B

WSPH.
- Ignore the number of particles and container details; they are irrelevant for temperature comparison.
- Calculate the product of each particle's mass and the square of its average speed for both samples.
- Focus on comparing these products directly; the higher value indicates the higher temperature.

Figure 11: Original training examples with WSPH. The hints provide visual grounding and reasoning guidance, while final answers are omitted.

G LIMITATIONS AND FUTURE WORK

Limitations. Although PTD-PO improves failed-trajectory learning through dense privileged guidance, it does not fully eliminate reasoning failures in highly compositional multimodal problems. In addition, the relative gains of PTD-PO become less pronounced on larger models compared with smaller ones. One possible reason is the mismatch between model capability and training-data difficulty: stronger models produce fewer failed trajectories during RLVR training, which reduces the number of cases where PTD can be activated to provide corrective supervision. Since PTD-PO is mainly designed to improve learning from failed rollouts, its effective scope naturally depends on the availability of sufficiently challenging examples that expose such failures. Therefore, PTD-PO should be viewed as an effective way to improve corrective supervision for failed rollouts, rather than a complete solution to all visual grounding and symbolic reasoning errors. Constructing harder and more capability-matched multimodal RLVR training data may further enlarge the benefit of privileged-information distillation, especially for stronger models.

Future Work. An important future direction is to construct harder and more capability-matched RLVR training data for stronger LVLMs. Since PTD-PO mainly provides corrective supervision for failed trajectories, larger models may benefit less when the current training data is not sufficiently challenging and produces too few informative failures. Designing more difficult multimodal reasoning tasks, especially those requiring fine-grained visual grounding, symbolic verification, and long-chain reasoning, could expose more failure cases and further enlarge the effective scope of privileged-information distillation. Beyond static multimodal reasoning, another promising direction is to extend PTD-style training to multimodal agent settings. In agentic tasks, models need to interact with environments, use tools, update observations, and make long-horizon decisions. Future work may construct non-goal-revealing privileged hints from intermediate observations, tool feedback, or expert action traces, so as to guide failed agent trajectories without directly revealing the final goal state. Such an extension could help improve exploration, grounding, and recovery behaviors in multimodal agents.