IS BETTER TEACHER SUPERVISION ENOUGH? UNLOCKING STUDENT-SIDE LEARNING IN MULTIMODAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

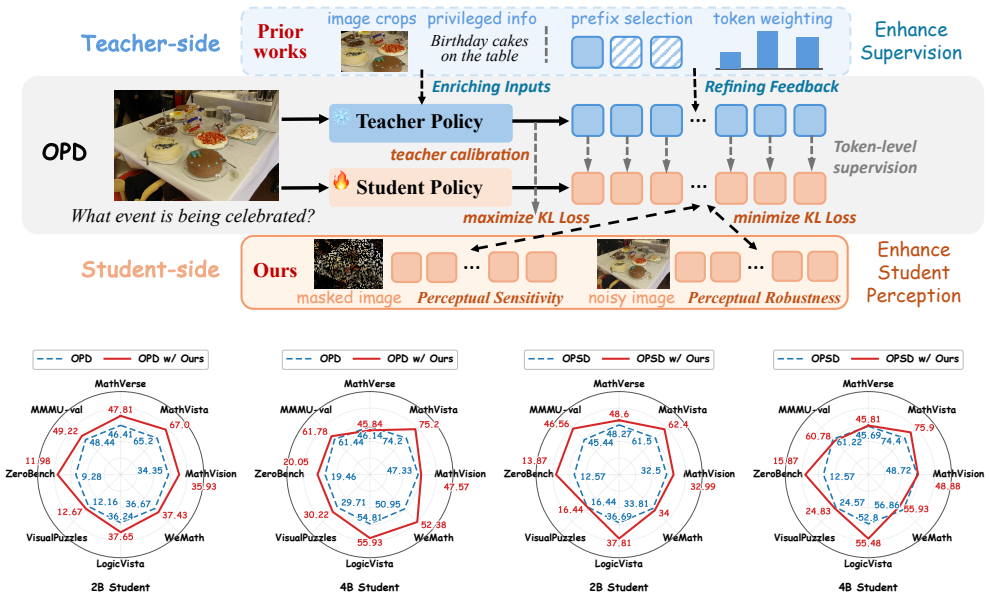


Figure 1: **Beyond enhancing teacher-side supervision:** Previous methods improve the teacher-side signal, while S-OPD adds two student-side objectives to enhance perception. This improves performance across model scales (2B and 4B) and distillation paradigms (OPD and OPSD).

ABSTRACT

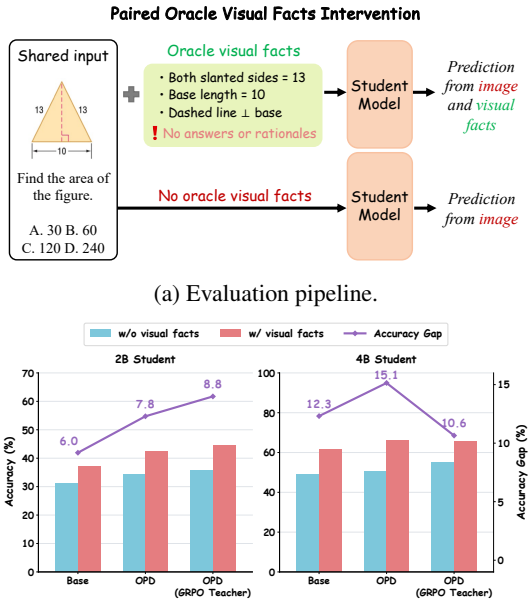
On-policy distillation (OPD) improves reasoning by providing token-level supervision from a teacher on a student's own trajectories. Existing methods primarily focus on enhancing this teacher-side guidance (e.g., by enriching teacher inputs and refining teacher feedback), yet we find that limited student perception is another critical bottleneck in multimodal OPD. By providing oracle visual facts, the performance of OPD-trained students can still be substantially improved for both weak and strong teachers. To address this bottleneck, we propose S-OPD, a simple multimodal on-policy distillation framework that explicitly strengthens student perceptual learning through two objectives. Specifically, *Teacher-calibrated Policy Contrast* separates student policies under original and masked images with teacher-based token-level gating, strengthening the student's reliance on visual evidence during reasoning. *Policy Agreement* aligns student policies under original and noise-perturbed images, further improving perceptual robustness to visual noise. Notably, our method can be seamlessly plugged into existing OPD frameworks, requiring no additional data annotations, model parameters or inference operations. Extensive experiments on eight benchmarks across student scales and distillation paradigms demonstrate consistent performance improvements, with gains of up to 4.25 points on LogicVista. When combined with existing teacher-side supervision methods, our method can yield further gains. Code is available at `https://anonymous.4open.science/r/S-OPD-0322/`.

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.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Bai et al., 2025; Zhu et al., 2025) have demonstrated strong capabilities in solving tasks that require models to integrate visual information with multi-step reasoning, like visual question answering (Liu et al., 2024; Ma et al., 2026), geometry problem solving (Lu et al., 2021; 2024) and chart-oriented reasoning (Masry et al., 2022; Chen et al., 2025). This motivates efforts to equip compact models for these capabilities with lower deployment costs. On-policy distillation (OPD) (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025) is appealing as it trains a student on its own trajectories using feedback from a strong teacher, helping the student learn better behavior at its own rollout distribution during training. Compared to outcome-based reinforcement learning (Shao et al., 2024; Guo et al., 2025), OPD provides dense token-level supervision by scoring student-generated prefixes with the teacher model, allowing intermediate decisions to receive direct supervision.

Recent efforts to improve OPD primarily focus on two aspects: enriching teacher inputs and refining teacher feedback. The first provides teachers with additional privileged information, such as reference rationales (Zhao et al., 2026) and in multimodal settings, image crops (Yuan et al., 2026) or recoverable visual cues (Tian et al., 2026b). The second utilizes differences of teacher predictions conditioned on various visual inputs to reweight the token-level supervision (Liu et al., 2026; Liang et al., 2026) or reconstruct teacher targets (Zhang et al., 2026). Despite these advances, genuine multimodal reasoning requires models to integrate visual perception, a critical component that provides visual evidence, with multi-step textual reasoning (Zhang et al., 2024; Tong et al., 2024; Wang et al., 2026). Prior approaches seek to improve the teacher-side guidance that the student receives. Yet this guidance is conveyed through text-token supervision. Students with limited visual perception may struggle to ground the guidance in the visual inputs, limiting the benefits of distillation. This motivates us to ask: *Is student-side visual perception capability another critical bottleneck for multimodal OPD?*



(a) Evaluation pipeline.

(b) Accuracy and accuracy gaps for 2B and 4B students.

Figure 2: **Oracle visual facts evaluation on Geometry3K.**

To investigate this question, we evaluate base and OPD-trained students, including Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct (Bai et al., 2025), on the same Geometry3K problems with and without oracle visual facts (Figure 2a). These facts describe detailed diagram content without providing answers or reasoning steps, reducing the demand on visual perception (see Section 3.1 for more details). The resulting accuracy gap indicates how much students benefit when the required visual information is explicitly available in textual form. A larger gap suggests that students still struggle to extract task-relevant visual evidence on their own despite training. As shown in Figure 2b, we find that substantial gaps persist after OPD, even with stronger teachers (Li et al., 2026a) that are trained with GRPO (Shao et al., 2024), suggesting that teacher supervision alone does not sufficiently address students' perceptual limitations.

The persistent gap raises a further question: *How does student visual perception influence reasoning performance after OPD on different benchmarks?* To examine this, we systematically analyze four benchmarks covering mathematical, logical and general domains (see Section 3.1 for more details). Specifically, we estimate the KL divergence between two student policies under original and masked images to measure the visual sensitivity of the student. We then divide samples into four quartiles according to this measure and compare model performance across groups. As shown in Figure 3, samples with higher KL divergence–i.e. those exhibiting greater sensitivity to visual input–generally achieve better performance across both student scales. This observation demonstrates a strong pos-

itive correlation between the student visual perception and reasoning performance, highlighting the necessity of explicitly strengthening student perceptual learning within OPD.

Building upon our earlier observations, *how can we exploit OPD's token-level learning structure to explicitly strengthen student perception?* We propose **S-OPD**, a simple multimodal OPD framework that explicitly strengthens student perceptual learning with two KL objectives. Our key idea is to shape student perceptual sensitivity according to whether an input change alters visual evidence: *visual evidence removal should induce a corresponding policy change*, while *perturbations preserving task-relevant content should maintain consistent predictions.* To enhance perceptual sensitivity to evidence removal, Teacher-calibrated Policy Contrast (TPC) maximizes the KL divergence between student policies under original and masked images. To focus this objective on *visually dependent* tokens (Huang et al., 2026), we use differences in the teacher's predictions under the same two images to gate which tokens KL maximization is applied to. Ablations of different gating strategies in Section 4.2 demonstrate the importance of this teacher-based selection. However, encouraging policy divergence under evidence removal alone does not constrain the student's response to irrelevant visual variations. We therefore introduce Policy Agreement (PA) as a complementary regularizer that minimizes the KL divergence between student policies under original and noise-perturbed images.

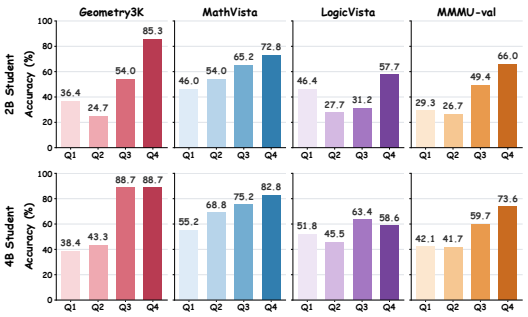


Figure 3: **Performance across quartiles of estimated KL divergence** between student policies under original and masked images (Q1–Q4, lowest to highest). A clear trend is observed: Higher-KL groups generally achieve higher accuracy.

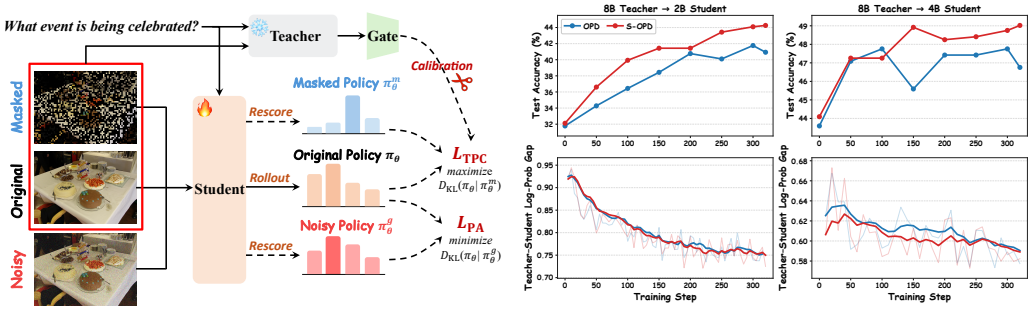
As shown in Figure 1, despite its simplicity, S-OPD achieves consistent performance gains on various reasoning tasks across student scales and distillation paradigms. In particular, methods that enhance teacher-side supervision achieve further performance gains when equipped with S-OPD, highlighting the additional benefits of our student perceptual learning.

Our contributions are:

- We investigate an overlooked student-side visual perception bottleneck in multimodal on-policy distillation, and reveal a positive association between students' perceptual sensitivity and reasoning performance across diverse tasks and models.
- We propose S-OPD, a simple multimodal OPD framework that enhances student perceptual learning with two KL objectives targeting perceptual sensitivity and stability.
- Extensive experiments demonstrate consistent performance improvements across multiple student scales, distillation paradigms and benchmarks, with gains reaching 4.25 points on LogicVista. Methods that enhance teacher-side supervision achieve further gains when augmented with S-OPD.

2 Related Work

On-Policy Distillation. Knowledge distillation (Hinton et al., 2015) trains a student to match a larger teacher's predictive distribution, while sequence-level distillation (Kim & Rush, 2016) uses teacher-generated sequences as training targets. However, training on fixed or teacher-generated sequences creates a mismatch with inference (Agarwal et al., 2024), where students condition on their own generated prefixes. On-policy distillation (Gu et al., 2024; Lu & Lab, 2025) addresses this mismatch by having the teacher provide dense token-level supervision on trajectories sampled by the student. Recent work improves this supervision along two dimensions: enriching teacher inputs (Zhao et al., 2026; Tian et al., 2026a) to improve teacher advantage and refining teacher feedback (Jin et al., 2026; Jang et al., 2026; Xu et al., 2026) to improve the reliability of token-level guidance. These efforts primarily enhance teacher-side supervision, while S-OPD focuses on student-side learning to help students better exploit the teacher guidance they receive.



(a) Details of two learning objectives in S-OPD. (b) Training dynamics of 2B and 4B student models.

Figure 4: (a) Details of S-OPD with original, masked, and noisy images. The teacher compares original and masked images to gate student policy contrast, while policy agreement aligns the student's original and noisy policies. (b) S-OPD achieves higher performance and reduces the teacher-student log-probability gap during training on Geometry3K.

Multimodal On-Policy Distillation. Recent work improves multimodal supervision from three perspectives: enriching its sources through text-to-vision reasoning transfer (Bousselham et al., 2026) or heterogeneous teacher capabilities (Yin et al., 2026); constructing informative teacher–student asymmetry through privileged teacher crops (Yuan et al., 2026), recoverable visual cues (Tian et al., 2026b), or augmented student views (Li et al., 2026b); and refining distillation signals through visual interventions (Sun et al., 2026) and modality-aware constraints (Aniri et al., 2026). Within signal refinement, VA-OPD (Liu et al., 2026) uses teacher visual advantage to allocate supervision across rollouts and token groups and VAD (Zhang et al., 2026) reconstructs visually attributable distillation targets. Beyond these advances, S-OPD directly shapes student visual perceptual sensitivity through two cross-view KL objectives, explicitly promoting sensitivity to evidence removal and stability under content-preserving perturbations.

Perceptual Learning for Multimodal Reasoning. Prior work has explored multiple training and inference strategies for improving visual perception. Visual contrast uses prediction differences to strengthen visual grounding in decoding (Leng et al., 2024) and preference optimization (Xie et al., 2024), while consistency learning promotes the robustness of neural networks in stability training (Zheng et al., 2016) and semi-supervised learning (Tarvainen & Valpola, 2017). In multimodal reinforcement learning, Video-R1 (Feng et al., 2025) contrasts temporally ordered and shuffled video inputs to encourage temporal reasoning, and PAPO (Wang et al., 2026) encourages separation between original- and masked-image policies to strengthen visual perception. Token-selective methods (Huang et al., 2026) focus optimization on tokens with higher visual dependency.

Unlike prior applications of visual contrast and consistency, S-OPD addresses the student perceptual bottleneck that limits the effectiveness of teacher supervision in OPD. Notably, S-OPD uses the changes in teacher-assigned token probabilities to select tokens for policy contrast. Teacher feedback thus provides both distillation targets and token-level calibration for student perceptual learning.

3 METHOD

We propose **S-OPD**, a simple framework that explicitly shapes student visual perception sensitivity through two KL objectives. Teacher-Calibrated Policy Contrast (TPC) maximizes the KL divergence between student policies under original and masked images at tokens selected using the teacher's cross-image responses, *strengthening perceptual sensitivity to evidence removal*. Policy Agreement (PA) minimizes the KL divergence between student policies under original and mildly perturbed images, *improving perceptual stability under content-preserving perturbations.* Figure 1 illustrates the overall framework and Figure 4 shows the details and training dynamics of our method.

3.1 PROBLEM FORMULATION

Let $\mathcal{D}$ denote the training distribution of multimodal inputs $x = (q, I)$, where q is a question and I is an image. Let π_θ denote the student policy and π_T the frozen teacher policy. For each input,

the student samples a response $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$, where T is the response length. We define $c_t = (q, I, y_{<t})$ as the context at step t, under which both policies are evaluated using the same student-generated prefix.

Following prior OPD studies (Gu et al., 2024; Lu & Lab, 2025), we train the student to minimize the reverse KL divergence to the teacher:

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

In practice, we estimate the KL divergence using the sampled-token k_1 estimator, $\log \pi_\theta(y_t \mid c_t) - \log \pi_T(y_t \mid c_t)$, evaluated on student-sampled tokens.

Bottleneck analysis of multimodal on-policy distillation. We first investigate whether student visual perception remains a bottleneck for multimodal OPD. We train Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct (Bai et al., 2025) on ViRL39K for 1 epoch with standard OPD (Lu & Lab, 2025). We use Qwen3-VL-8B-Instruct (Bai et al., 2025) and its variant trained with GRPO on ViRL39K (Wang et al., 2025) as teachers, following the use of post-trained teachers to strengthen supervision (Li et al., 2026a). Base and OPD-trained students are evaluated on the Geometry3K test split with and without oracle visual facts. These facts are constructed by verbalizing official `diagram_logic_form` annotations with fixed templates, without accessing answers or rationales. We insert the facts after the answer choices as a `Diagram observations` block, keeping all other inputs and decoding settings fixed. Construction and audit details can be found in Appendix A.3. As shown in Figure 2b, oracle facts substantially improve accuracy even after OPD with the stronger GRPO teacher, indicating persistent student limitations in visual perception, i.e. the ability to recover visual evidence.

Visual perception sensitivity and reasoning performance. We next examine how student visual perception sensitivity relates to reasoning performance on Geometry3K (Lu et al., 2021), MathVista (Lu et al., 2024), LogicVista (Xiao et al., 2024), and MMMU (Yue et al., 2024), covering mathematical, logical and general reasoning. For each student response generated from the original image, we measure the average token-level KL divergence between student predictive distributions under original and randomly masked images, keeping the question and response prefixes fixed:

$$s(x, y) = \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_\theta(\cdot \mid c_t^m)), \tag{2}$$

where c_t^m replaces the original image in c_t with its masked version and higher s indicates greater sensitivity to evidence removal. Within each benchmark and student scale, we divide examples into quartiles by their estimated scores. Figure 3 shows that higher-scoring groups generally achieve higher accuracy across both student scales.

Together, these findings motivate explicitly shaping how students respond to visual evidence during OPD. S-OPD augments teacher–student supervision with student-side objectives that encourage sensitivity to evidence removal and consistency under mild perturbations. To define these objectives, we introduce masked and noisy image views:

$$I^m = \mathcal{M}(I; \rho), \qquad I^g = \mathrm{clip}(I + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{3}$$

where $\mathcal{M}$ randomly masks a proportion ρ of image patches, σ is the Gaussian noise standard deviation, $\mathbf{I}$ is the identity matrix, and clip restricts pixels to their valid range. Masking reduces available visual evidence, while mild noise is intended to preserve task content. The corresponding contexts, $c_t^m = (q, I^m, y_{<t})$ and $c_t^g = (q, I^g, y_{<t})$, share the original response, enabling comparisons at the same position. We provide the visualization of masked and noisy images in Appendix A.1.

3.2 Teacher-Calibrated Policy Contrast

To enhance student visual perception, TPC encourages the student to use task-relevant visual evidence. Specifically, it increases the discrepancy between the student's policies under original and masked images at tokens where the teacher indicates that the removed evidence is relevant.

To focus this contrast on tokens affected by masking, we select positions where the teacher assigns greater probability under the original image:

$$\Delta_t^T = \log \frac{\pi_T(y_t \mid c_t)}{\pi_T(y_t \mid c_t^m)}, \qquad g_t = \mathbb{I}\left[\Delta_t^T > \tau\right], \tag{4}$$

where τ specifies the required log-probability drop and is set to 0 by default. Despite using a zero margin, the teacher gate remains selective in practice, activating on only 46.9% of response tokens; detailed per-rollout statistics are provided in Appendix B.3. This gate uses the teacher's response to masking as a token-level signal of visual dependence (Huang et al., 2026). At the selected positions, TPC maximizes the KL divergence between the student's original and masked policies, weighted by the teacher's log-probability drop:

$$\mathcal{L}_{\mathrm{TPC}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|q,I)}\left[\frac{1}{Z}\sum_{t=1}^{T} g_t\, \Delta_t^T\, D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_\theta(\cdot \mid c_t^m)\right)\right], \tag{5}$$

where $Z = \max(1, \sum_{t=1}^{T} g_t)$ normalizes by the number of selected tokens. In practice, we use the sampled-token k_3 estimator, $r_t^m - \log r_t^m - 1$, where $r_t^m = \pi_\theta(y_t \mid c_t^m)/\pi_\theta(y_t \mid c_t)$, and stop gradients through the masked-image policy. Thus, the teacher determines where contrast is applied, while the student's cross-image discrepancy provides the contrastive learning signal.

3.3 Policy Agreement

To make the student's use of visual evidence robust to local image variations, PA encourages consistent predictions under mild Gaussian perturbations. These perturbations are intended to preserve task-relevant content, so the student's predictions should remain stable despite changes in individual pixel values. Accordingly, PA minimizes the KL divergence between the student's original and noisy policies at the same response prefixes:

$$\mathcal{L}_{\mathrm{PA}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|q,I)}\left[\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_\theta(\cdot \mid c_t^g)\right)\right]. \tag{6}$$

We implement the KL divergence using the sampled-token k_3 estimator, $r_t^g - \log r_t^g - 1$, where $r_t^g = \pi_\theta(y_t \mid c_t^g)/\pi_\theta(y_t \mid c_t)$, and stop gradients through the noisy-image policy. The noisy-image policy thus provides a fixed target for each update, encouraging the student to maintain consistent predictions under perturbations that preserve task-relevant visual evidence.

3.4 Overall Objective

S-OPD combines teacher–student distillation with explicit objectives for visual perception:

$$\mathcal{L}_{\text{S-OPD}}(\theta) = \mathcal{L}_{\mathrm{OPD}}(\theta) + \lambda_{\mathrm{TPC}}\mathcal{L}_{\mathrm{TPC}}(\theta) + \lambda_{\mathrm{PA}}\mathcal{L}_{\mathrm{PA}}(\theta), \tag{7}$$

where λ_{TPC} and λ_{PA} control the strengths of policy contrast and agreement. Alongside token-level teacher guidance, these objectives train the student to respond to relevant evidence removal and remain stable under mild visual perturbations.

For each input, S-OPD samples only a single rollout under the original image. Notably, S-OPD can be seamlessly plugged into existing OPD frameworks, requiring no additional model parameters or inference operations. Its additional computation is confined to training, with a modest and practical overhead (see Appendix B.6).

4 Experiments

Our experiments test whether strengthening student visual perception enables more effective use of teacher supervision in multimodal OPD. We address four questions. **Q1:** Does S-OPD improve multimodal OPD across student scales and teacher strengths? **Q2:** Are these gains sustained across various tasks beyond the training domain? **Q3:** How do TPC and PA contribute to these improvements, and does teacher calibration enhance the effectiveness of student policy contrast? **Q4:** What additional computational cost does S-OPD introduce?

Data. For the main experiments, we train on ViRL39K (Wang et al., 2025). We evaluate the models across eight benchmarks covering three capabilities: mathematical reasoning on MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), and WeMath (Qiao et al., 2024); logical reasoning on LogicVista (Xiao et al., 2024) and VisualPuzzles (Song et al., 2025); and general visual reasoning on ZeroBench-sub (Roberts et al., 2025) and MMMU-val (Yue et al., 2024). Geometry3K (Lu et al., 2021) is reserved for ablation studies due to its compact size and focused geometry domain.

Models and baselines. We use Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct as student models, with Qwen3-VL-8B-Instruct serving as the teacher (Bai et al., 2025). For each student size, we conduct on-policy distillation with either the original 8B teacher or an enhanced teacher obtained by training the 8B model with GRPO (Shao et al., 2024) on ViRL39K for one epoch. We compare our method against four baselines in the main experiments: the corresponding base model, GRPO, vanilla OPD (Lu & Lab, 2025), and VA-OPD (Liu et al., 2026). We also include OPSD (Zhao et al., 2026) as a baseline in Section 4.2. Details are in Appendix A.

4.1 Main Results

Table 1 reports results across all benchmarks for both 2B and 4B students. We compare S-OPD against the corresponding distillation baseline under the same teacher configuration.

S-OPD benefits different student scales and teacher strengths. S-OPD improves average performance across all four student–teacher configurations. With the 8B teacher, the gains are larger for the 2B student than for the 4B student (+1.37 vs. +0.61 points), suggesting greater room for improvement through student-side learning at the smaller scale. Strengthening the teacher with GRPO raises the OPD baseline, yet S-OPD provides further gains of 0.92 and 0.82 points respectively. The gain decreases for the 2B student but increases for the 4B student, indicating that the benefit of student-side optimization depends on both student capacity and teacher strength.

Gains span reasoning domains but vary across tasks. S-OPD improves the average score within mathematical, logical, and general reasoning in every student–teacher configuration. The distribution of gains nevertheless differs across scales and teachers. With the 8B teacher, the 2B student improves on all eight benchmarks, including MathVista (+1.80) and ZeroBench (+2.70). With the GRPO-trained teacher, the most pronounced gain shifts to logical reasoning for the 4B student, which improves by 4.25 points on LogicVista, while its general reasoning performance changes little overall. These results suggest that student-side perceptual learning benefits diverse reasoning tasks, with the largest gains depending on the student–teacher configuration.

Student-side learning complements enhanced teacher supervision. Adding our objectives to VA-OPD improves average performance by 1.03 points for the 2B student and 0.72 points for the 4B student, with notable improvements including MathVision (+2.70) for the 2B student and MMMU-val (+1.67) for the 4B student. These gains show that improving teacher supervision leaves additional opportunities on the student side: S-OPD remains effective when combined with a method that already enhances the supervision students receive.

4.2 Ablation Studies

Component analysis. Table 2 isolates the contributions of TPC and PA across both 2B and 4B student scales, and shows that TPC and PA contribute differently to in-domain learning and out-of-domain generalization. TPC provides stronger in-domain gains, improving Geometry3K accuracy by 1.63 and 2.26 points for the 2B and 4B students, respectively, compared with 1.00 and 1.27 points from PA. PA is particularly effective for transfer to general reasoning, achieving the highest scores at both scales, and also improves logical reasoning for the 2B student. These gains suggest that encouraging consistency under visual noise helps students generalize beyond the training domain. Combining both modules improves over OPD on Geometry3K and all three other categories at both scales, yielding average gains of 1.60 and 1.04 points. The results support their complementary roles: TPC strengthens learning from visual evidence within the training domain, while PA promotes perceptual stability that supports cross-domain transfer.

Table 1: Main results on eight benchmarks. We compare the distillation performance against the corresponding baseline. For OPD and S-OPD, we use either Qwen3-VL-8B-Instruct or its GRPO-trained variant as the teacher; the subscript G denotes the latter teacher setting. Avg. is computed over the eight benchmarks.

	Mathematical Reasoning				Logical Reasoning		General Reasoning		Overall
Method	**MathVerse**	**MathVista**	**MathVision**	**WeMath**	**LogicVista**	**VisualPuzzles**	**ZeroBench**	**MMMU**	**Avg.**
Qwen3-VL-2B-Instruct									
Base model	45.51	61.2	30.29	32.29	34.52	16.18	13.17	45.78	34.87
GRPO	48.43	62.3	31.62	36.57	42.73	23.16	14.67	49.67	38.64
OPD	46.41	65.2	34.35	36.67	36.24	12.16	9.28	48.44	36.09
S-OPD	47.81 (+1.40)	67.0 (+1.8)	35.93 (+1.58)	37.43 (+0.76)	37.65 (+1.41)	12.67 (+0.51)	11.98 (+2.70)	49.22 (+0.78)	37.46 (+1.37)
OPD_G	50.68	67.6	35.95	39.24	39.49	14.89	12.57	47.33	38.47
S-OPD$_G$	52.76 (+2.08)	68.3 (+0.7)	36.03 (+0.08)	40.29 (+1.05)	39.25 (-0.24)	15.62 (+0.73)	13.78 (+1.21)	49.11 (+1.78)	39.39 (+0.92)
VA-OPD	45.48	63.7	31.77	33.14	36.91	10.79	11.38	47.56	35.09
w/ S-OPD	47.51 (+2.03)	64.3 (+0.6)	34.47 (+2.70)	34.19 (+1.05)	38.03 (+1.12)	11.56 (+0.77)	10.78 (-0.60)	48.11 (+0.55)	36.12 (+1.03)
Qwen3-VL-4B-Instruct									
Base model	42.28	73.8	49.31	55.43	54.81	25.43	15.57	61.56	47.27
GRPO	39.11	73.5	48.39	52.57	59.73	41.35	20.36	62.78	49.72
OPD	46.14	74.2	47.33	50.95	54.81	29.71	19.46	61.44	48.01
S-OPD	45.84 (-0.30)	75.2 (+1.0)	47.57 (+0.24)	52.38 (+1.43)	55.93 (+1.12)	30.22 (+0.51)	20.05 (+0.59)	61.78 (+0.34)	48.62 (+0.61)
OPD_G	46.60	76.6	48.55	55.71	54.14	33.99	20.36	64.33	50.04
S-OPD$_G$	46.81 (+0.21)	77.0 (+0.4)	48.22 (-0.33)	57.43 (+1.72)	58.39 (+4.25)	34.25 (+0.26)	20.05 (-0.31)	64.76 (+0.43)	50.86 (+0.82)
VA-OPD	44.01	72.6	46.45	51.05	52.80	30.39	14.67	60.11	46.51
w/ S-OPD	44.49 (+0.48)	73.5 (+0.9)	46.81 (+0.36)	51.52 (+0.47)	53.47 (+0.67)	30.99 (+0.60)	15.27 (+0.60)	61.78 (+1.67)	47.23 (+0.72)

Table 2: Component analysis with 2B and 4B students on Geometry3K. Avg. denotes the arithmetic average over all benchmarks including the test split of Geometry3K.

	2B Student					4B Student				
Method	**Geometry3K**	**Mathematical**	**Logical**	**General**	**Avg.**	**Geometry3K**	**Mathematical**	**Logical**	**General**	**Avg.**
OPD	40.93	45.98	25.83	30.40	37.48	46.76	54.38	42.89	36.76	47.06
w/ TPC	42.56	47.10	26.05	29.77	38.07	**49.02**	54.94	42.88	37.99	47.83
w/ PA	41.93	46.78	27.91	**31.27**	38.60	48.03	54.42	41.46	**39.06**	47.41
S-OPD	**44.26**	**47.39**	**28.02**	30.92	**39.08**	**49.02**	**55.03**	**43.33**	38.57	**48.10**

Effectiveness of teacher-calibrated gating. We examine whether the teacher provides useful guidance on where to apply policy contrast (PC). As shown in Table 3, teacher-calibrated gating outperforms student-based gating and no gating by 1.15 and 1.02 average points, respectively. More importantly, comparing these results with the PA-only variant in Table 2 reveals that TPC is not inherently beneficial: applying PC to all tokens reduces the average score from 38.60 to 38.06. This comparison shows that teacher calibration is critical for preventing indiscriminate contrastive supervision from harming the student policy. Student-based gating offers no improvement over applying TPC to all tokens, suggesting that the student's own visual sensitivity is insufficient to identify where policy contrast is beneficial. These results support *using the teacher already available in OPD to guide the student's perceptual learning* beyond providing distillation targets. To test whether the gains simply arise from selecting fewer tokens, we randomly select the same number of tokens as the teacher gate for each response and reassign the original teacher-derived weights to these randomly selected tokens. This matched random baseline falls short of teacher-calibrated gating by 0.69 average points, indicating that the teacher identifies informative tokens for TPC.

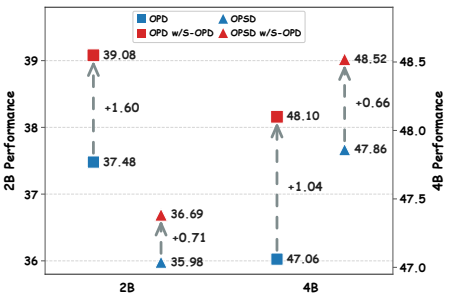


Figure 5: S-OPD gains across student scales and distillation paradigms.

Compatibility across distillation paradigms and student scales. Figure 5 evaluates S-OPD across different student scales and teacher advantage sources trained on Geometry3K. Standard OPD obtains supervision from a larger teacher, whereas OPSD (Zhao et al., 2026) uses privileged information to strengthen supervision in a self-teacher setting. S-OPD improves both paradigms: under standard OPD, it increases the average performance of the 2B and 4B students by 1.60 and

Table 3: Comparison of gating strategies for TPC with the remaining S-OPD configuration fixed. OPD is included as a reference. Random gating selects the same number of tokens as teacher-calibrated gating for each response. Bold values indicate the best results.

Gating strategy	Token selection	Geometry3K	Mathematical	Logical	General	Avg.
OPD (reference)	–	40.93	45.98	25.83	30.40	37.48
No gating	All	42.56	46.37	26.55	30.72	38.06
Student-based	Adaptive	41.43	46.80	25.44	30.92	37.93
Random	Count-matched	40.61	**47.47**	26.57	**30.95**	38.39
Teacher-calibrated (Ours)	Adaptive	**44.26**	47.39	**28.02**	30.92	**39.08**

1.04 points, respectively; under OPSD, the corresponding gains are 0.71 and 0.66 points. Larger gains under standard OPD suggest that student-side optimization offers more room for improvement in this setting, while the positive gains under OPSD show that it remains effective when teacher advantage comes from additional information rather than greater model capacity. Together, these results demonstrate the applicability of S-OPD across different student sizes and forms of teacher supervision, with the extent of its benefit depending on the distillation setting.

Impact of the coefficients of TPC and PA. We examine whether S-OPD requires precise coefficient tuning by varying λ_{TPC} and λ_{PA} for students trained on Geometry3K. As shown in Table 4, average performance varies by only 0.74 points for TPC and 0.50 points for PA. Although individual domains favor different coefficients, the overall performance remains stable, allowing us to obtain comparable results across a range of objective weights. We obtain the highest average score with $\lambda_{\text{TPC}} = 0.005$ and $\lambda_{\text{PA}} = 0.02$ on Geometry3K. We set both coefficients to 0.02 for the main experiments on ViRL39K and the corresponding ablations are in Appendix B.4.

Table 4: Ablation of the TPC and PA coefficients.

Value	Geometry3K	Mathematical	Logical	General	Avg.
		λ_{TPC}			
0.005	44.26	47.39	**28.02**	30.92	**39.08**
0.01	**44.92**	**47.73**	25.83	**30.93**	38.82
0.02	43.05	47.00	26.64	30.36	38.34
		λ_{PA}			
0.01	**44.92**	46.85	26.55	30.91	38.58
0.02	44.26	**47.39**	**28.02**	30.92	**39.08**
0.04	44.09	47.12	26.13	**31.38**	38.62

Impact of the masking ratio and Gaussian noise level. We vary the masking ratio and Gaussian noise level to identify effective perturbation strengths for TPC and PA. In Table 5, increasing ρ from 0.2 to 0.6 raises the average score from 38.49 to 39.08, but further increasing it to 0.8 lowers the score to 38.74. Similarly, Gaussian noise performs best at $\sigma = 0.2$, with weaker and stronger noise yielding lower average scores. We thus obtain the best overall results with substantial masking for TPC and moderate noise for PA. The preferred strength varies by domain, suggesting that different tasks may place distinct demands on visual perception. We choose $\rho = 0.6$ and $\sigma = 0.2$ for both Geometry3K ablations and ViRL39K main experiments.

Table 5: Ablation of the masking ratio and Gaussian noise level.

Value	Geometry3K	Mathematical	Logical	General	Avg.
		Masking ratio ρ			
0.2	43.43	47.11	24.94	**32.34**	38.49
0.4	42.43	47.18	26.65	31.20	38.54
0.6	44.26	**47.39**	**28.02**	30.92	**39.08**
0.8	**44.43**	47.34	26.02	31.42	38.74
		Gaussian noise level σ			
0.1	42.93	46.85	26.55	30.91	38.36
0.2	**44.26**	**47.39**	**28.02**	30.92	**39.08**
0.3	43.05	47.12	26.13	**31.38**	38.51

5 Conclusion

In this paper, we identify student visual perception as a persistent bottleneck in multimodal on-policy distillation, even when teacher supervision is strengthened. Motivated by this, we propose S-OPD to explicitly improve student perceptual sensitivity and robustness within the token-level distillation process. Teacher-calibrated Policy Contrast encourages sensitivity to visual evidence removal at teacher-selected tokens, while Policy Agreement promotes consistent predictions under visual noise. Experiments on eight benchmarks demonstrate consistent average improvements across student scales and teacher strengths, with further gains when integrated into existing methods that enhance teacher supervision. These benefits require no additional annotations, model parameters, or inference-time operations, with only moderate training overhead. Our findings establish student-side learning as a complementary direction for advancing multimodal OPD.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and the retrieval and discovery of related literature. We reviewed all AI-assisted revisions to ensure that they accurately reflected the intended meaning and verified the suggested references against their original sources. We take full responsibility for the final content of this paper, including all text, claims, citations, and other research artifacts.

ETHICS STATEMENT

This study uses publicly available models, datasets, and benchmarks and involves no human-subject experiments or collection of new personal data. Teacher supervision and model-based evaluation may inherit errors and biases from the underlying models. These outputs are used within a controlled research setting and are not treated as human-verified ground truth. The work focuses on benchmark evaluation rather than deployment in high-stakes settings. Improvements in multimodal reasoning nevertheless carry the potential misuse risks associated with general-purpose vision-language models.

REPRODUCIBILITY STATEMENT

Our experiments build on publicly available datasets, pretrained models, and evaluation tools. Our code is available at: `https://anonymous.4open.science/r/S-OPD-0322/`. The ViRL39K and Geometry3K datasets used in our experiments are provided in `https://figshare.com/s/58a217f8663ed09a458b`. Section 3 describes the S-OPD framework and its training objectives. Section 4 specifies the datasets, model configurations, and baselines used in our experiments. Appendix A.1 reports the training framework, hardware, and hyperparameters, while Appendix A.2 details the evaluation protocols, decoding settings, prompts, and scoring procedures. Additional component analyses and computational cost measurements are provided in Appendix B.

REFERENCES

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

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. OPD-V: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning transfer from llms to vision-language models via on-policy distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Lei Chen, Xuanle Zhao, Zhixiong Zeng, Jing Huang, Yufeng Zhong, and Lin Ma. Chart-r1: Chain-of-thought supervision and reinforcement for advanced chart reasoner. *arXiv preprint arXiv:2507.15509*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, 2026.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2026a.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026b.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

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

Pan Lu, 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)*, 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. In *International Conference on Learning Representations*, 2024.

Qinghe Ma, Zhen Zhao, Yiming Wu, Jian Zhang, Lei Bai, and Yinghuan Shi. Are tools always beneficial? learning to invoke tools adaptively for dual-mode multimodal llm reasoning. *arXiv preprint arXiv:2605.19852*, 2026.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics*, 2022.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. ZeroBench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-Zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kanghui Tian, Siyuan Liu, Tianxiang Jiang, Shuai Dong, Yizhuo Li, Tian Ding, Yuan Guo, Songze Li, Haowen Hou, Congcong Wang, and Yi Wang. What should a self-teacher see? privileged context design for on-policy self-distillation. *arXiv preprint arXiv:2609.25623*, 2026a.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2025.

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

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

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

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024.

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.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

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

A Implementation Details

A.1 Training Details

All the experiments are implemented on top of the `verl` framework (Sheng et al., 2025) and run on NVIDIA A100 80GB GPUs. We train the student model with PyTorch Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) and use vLLM (Kwon et al., 2023) for on-policy rollout generation and frozen-teacher forward passes. Table 6 summarizes the key training configurations for the main ViRL39K experiments and the Geometry3K ablations.

Table 6: Key training configurations for the ViRL39K experiments and the Geometry3K ablations.

Setting	ViRL39K Distillation	ViRL39K GRPO	Geometry3K Distillation
Training data	ViRL39K	ViRL39K	Geometry3K train
Max. prompt length	4,096	4,096	1,024
Max. response length	4,096	4096	2,048
Prompt batch size	192	192	128
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Rollouts per prompt (n)	1	5	1
Training epochs	1	1	20
λ_{TPC}	0.02	–	0.005
λ_{PA}	0.02	–	0.02

Details of masking. We construct the masked image in RGB pixel space before applying the model-specific image processor. Specifically, we partition the original-resolution image into non-overlapping 16×16 pixel blocks, including partial blocks along the image boundaries, and independently sample a mask indicator $m_{ij} \sim \text{Bernoulli}(0.6)$ for each block. If $m_{ij} = 1$, all pixels in that block are replaced with black pixels, i.e., RGB $(0, 0, 0)$. Thus, 0.6 denotes the expected masking ratio rather than an exact per-image proportion. We modify only the pixel values and preserve the image dimensions; no image patches or visual tokens are removed. Consequently, the original and masked images have the same visual-token sequence length and two-dimensional positional structure after preprocessing. Figure 6 visualizes images under different masking ratios.

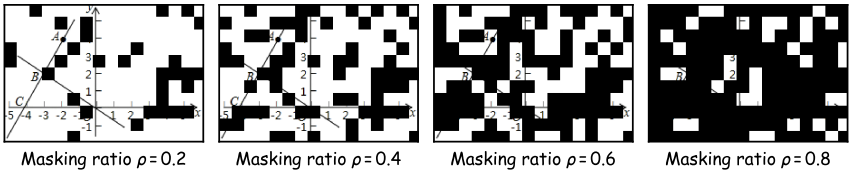


Figure 6: Masked images used for TPC at different masking ratios ρ. Randomly selected image patches are replaced with black pixels, removing increasing amounts of visual evidence as ρ increases.

Details of Gaussian perturbation. We construct the noisy image by adding Gaussian noise before applying the model-specific image processor. Given an unsigned 8-bit RGB image v, we first rescale its pixel values from $[0, 255]$ to $[0, 1]$ and independently sample $\epsilon_{hwc} \sim \mathcal{N}(0, \sigma^2)$ for every spatial location and color channel, with $\sigma = 0.2$. The perturbed image is computed as

$$v_{\text{n}} = \text{round}\Big[255 \cdot \text{clip}\Big(\frac{v}{255} + \epsilon,\, 0,\, 1\Big)\Big], \tag{8}$$

and is converted back to an unsigned 8-bit RGB image before standard visual preprocessing. Therefore, the reported noise scale is defined with respect to the $[0, 1]$ pixel range; equivalently, $\sigma = 0.2$ corresponds to approximately 51 intensity levels on the $[0, 255]$ scale. Figure 7 visualizes images under different Gaussian perturbation noise.

A.2 Evaluation Details

All the checkpoints are evaluated on `MathVerse_MINI`, `MathVista_MINI`, `MathVision`, `WeMath`, `LogicVista`, `VisualPuzzles`, `ZeroBench_sub`, `MMMU_DEV_VAL` with greedy

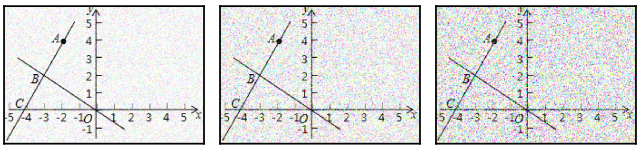


Figure 7: Noisy images used for PA at different Gaussian noise levels σ. Independent zero-mean Gaussian noise is added to each RGB channel in the $[0, 1]$ pixel range, followed by clipping. Larger σ produces stronger perturbations while retaining the overall image structure.

Table 7: Generation user prompts used for each evaluation dataset. Braced terms denote fields from each VLMEvalKit example, and `[Options]` denotes the formatted list of answer choices.

Dataset	User prompt template
`MathVerse_MINI`	`{question}`
`MathVista_MINI`	`{question}`
`MathVision`	`{question}`
`WeMath`	`Question: {question}` `Options: [Options]`
`LogicVista`	`{question}`

`VisualPuzzles`	`{question}`
	`Options: [Options]`
	`Solve the multiple-choice question and then answer with the option letter from the given choices. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes), where LETTER is one of the options. Think step by step before answering.`

`ZEROBench_sub`	`{question}` `Let's think step by step and give the final answer in curly braces, like this: {final answer}.`

`MMMU_DEV_VAL`	`Question: {question}`
	`Options: [Options]`
	`Please select the correct answer from the options above.`

decoding using the default VLMEvalKit (Duan et al., 2024) dataset definitions and a common Qwen3-VL wrapper in the main experiments. We use the shared system prompt `You are a helpful assistant.` and dataset-specific user prompts listed in Table 7.

Rule-based benchmarks use their official exact-match or multiple-choice scorers. Benchmarks requiring answer extraction or semantic grading use a locally served Qwen3-30B-A3B. MathVerse is averaged over its Vision Intensive, Vision Only, Text Dominant, Text Lite, and Vision Dominant subsets; the text-only subset is excluded. We report MathVista on its 1,000-example mini split, WeMath using Core (Strict), and MMMU on the validation split.

A.3 Oracle Visual-Fact Intervention Details

Oracle-fact construction. We use oracle visual facts as a test-time intervention to make the geometric information encoded in each diagram explicit. For every Geometry3K example, we deterministically convert `diagram_logic_form` from the official `logic_form.json` annotations into natural-language statements describing geometric entities, incidence relations, lengths, angles, parallelism, perpendicularity, equality, congruence, and other geometric properties. The renderer has no access to the answer, solution, program, theorem sequence, or text-based logic forms. It also rejects `Find` and `UseTheorem` expressions, including nested occurrences. We compare two paired inputs that share the same image, question, answer choices, system prompt, and output instruction. The oracle condition differs only by the insertion of the rendered statements under `Diagram observations:` in the same user message. Thus, the intervention augments the diagram information available to the model without changing its parameters or removing the original image.

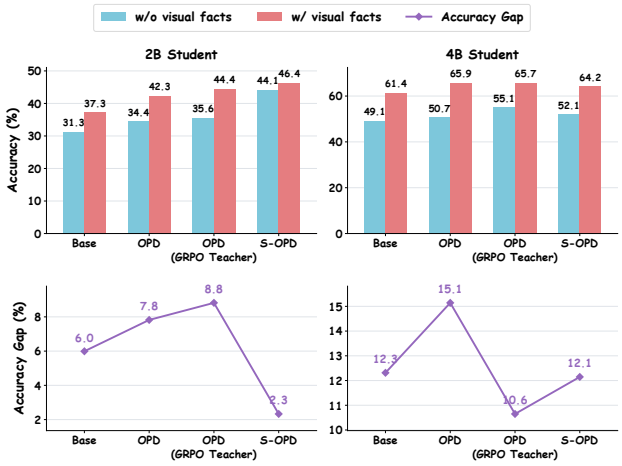


Figure 8: **Oracle visual-fact intervention on Geometry3K.** Top panels report accuracy with the original image alone (*w/o visual facts*) and with additional oracle visual facts derived from diagram annotations (*w/ visual facts*); bottom panels show the corresponding accuracy gap in percentage points. Compared with standard OPD, S-OPD improves accuracy without visual facts and reduces the accuracy gap from 7.8 to 2.3 points for the 2B student and from 15.1 to 12.1 points for the 4B student.

Evaluation protocol. We evaluate every checkpoint on the same official 601-example test split of Geometry3K. Inference uses greedy decoding (temperature 0) and a maximum generation length of 4096 tokens. We report answer-content accuracy after mapping predicted option letters to their corresponding choice values and normalizing common LaTeX and numeric variants. The *Accuracy Gap* is the paired difference between accuracy with oracle visual facts and image-only accuracy.

Results. We revisit the oracle visual-fact intervention from our motivational study to examine whether S-OPD alleviates the student-side perceptual bottleneck that motivated our method. As shown in Figure 8, compared with standard OPD, S-OPD improves image-only accuracy and reduces the Accuracy Gap at both model scales. For the 2B student, image-only accuracy increases from 34.4% to 44.1%, while accuracy with oracle visual facts improves from 42.3% to 46.4%, narrowing the gap from 7.8 to 2.3 percentage points. S-OPD thus achieves the highest accuracy in both conditions and the smallest gap among the compared methods at this scale. This result directly complements our motivational finding: whereas stronger teacher-side supervision alone leaves a substantial gap, explicitly strengthening student perception helps the student recover more task-relevant evidence from the diagram itself. For the 4B student, S-OPD improves image-only accuracy from 50.7% to 52.1% and reduces the gap from 15.1 to 12.1 points. Here, the evidence is more qualified, as oracle-fact accuracy also decreases from 65.9% to 64.2%, and OPD with the GRPO teacher retains the smallest gap. Overall, these results, particularly the strong improvement for the 2B student, support our central motivation that student-side perceptual learning is an important complement to teacher-side supervision in multimodal OPD.

Audit. Our data audit finds no unsupported expressions, parse failures, missing images, missing answers, or use of forbidden annotation sources in the rendered facts.

B Ablation Details

B.1 Visual Perception Beyond Reasoning Benchmarks

To further examine whether S-OPD improves student visual perception, we evaluate the distilled students on three additional benchmarks, V*Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026) and HallusionBench (Guan et al., 2024). Specifically, V*Bench and ZoomBench assess fine-grained visual understanding, while HallusionBench explicitly diagnoses failures arising from visual

Table 8: Comparison of OPD and S-OPD across different student model sizes and training datasets. All models are evaluated on three out-of-domain visual benchmarks.

Benchmark	Trained on Geometry3K				Trained on ViRL39K			
	2B		4B		2B		4B	
	OPD	S-OPD	OPD	S-OPD	OPD	S-OPD	OPD	S-OPD
V*Bench	73.30	**74.35**	**74.87**	74.35	74.87	**75.92**	62.67	**63.30**
ZoomBench	41.42	**42.37**	43.55	**44.26**	**41.30**	40.95	43.67	**44.14**
HallusionBench	55.73	**56.26**	64.77	**64.98**	52.58	**53.00**	**74.35**	**74.35**

Table 9: Detailed component analysis results on Geometry3K. The Geometry3K test split is used for in-domain performance evaluation. Bold and underlined numbers indicate the best and second-best results within each student size, respectively.

	In-domain	Mathematical Reasoning				Logical Reasoning		General Reasoning		Overall
Method	Geo3K-Test	MVerse	MVista	MVision	WeMath	LogicVista	VisPuzzles	ZeroBench	MMMU-val	Avg.
Qwen3-VL-2B-Instruct										
Vanilla OPD	40.93	49.28	63.3	34.96	36.38	39.83	11.82	11.68	**49.11**	37.48
w/ TPC	42.56	49.84	65.2	35.07	**38.29**	40.71	11.39	11.98	47.56	38.07
w/ PA	41.93	50.69	**65.3**	34.36	36.76	**43.14**	12.67	**14.97**	47.56	38.60
S-OPD	**44.26**	**50.76**	**65.3**	**35.43**	38.05	42.51	**13.53**	13.17	48.67	**39.08**
Qwen3-VL-4B-Instruct										
Vanilla OPD	46.76	44.24	74.1	46.68	52.48	55.81	29.97	14.07	59.44	47.06
w/ TPC	**49.02**	45.14	**74.4**	47.07	**53.14**	55.70	**30.05**	15.87	60.11	47.83
w/ PA	48.03	43.83	74.1	**47.63**	52.10	55.26	27.65	15.57	**62.56**	47.41
S-OPD	**49.02**	**45.38**	**74.4**	47.20	**53.14**	**56.60**	**30.05**	**16.47**	60.67	**48.10**

illusions and language hallucinations. As shown in Table 8, S-OPD improves or matches OPD in 10 out of the 12 model–data configurations.

The improvements are consistent across different student capacities and training distributions. On V*Bench, S-OPD improves the 2B student by 1.05 points when trained on either Geometry3K or ViRL39K, and improves the 4B student trained on ViRL39K by 0.63 points. On ZoomBench, it yields gains of 0.95 and 0.71 points for the 2B and 4B students trained on Geometry3K, respectively, while improving the ViRL39K-trained 4B student by 0.47 points. Importantly, S-OPD also consistently improves or preserves performance on HallusionBench across all four settings, with gains of up to 0.53 points. Although a few configurations exhibit small fluctuations, the overall trend across model sizes, training datasets, and diagnostic tasks indicates that the benefits of S-OPD extend beyond the in-domain reasoning benchmarks. In particular, the gains on fine-grained perception and hallucination-oriented evaluations provide additional evidence that S-OPD encourages students to ground their predictions more effectively in visual evidence.

B.2 Main Component Analysis

Table 9 examines TPC and PA across both student scales. TPC improves in-domain accuracy and all four mathematical reasoning benchmarks for both students, consistent with its role in linking predictions to supporting visual evidence through teacher-calibrated policy contrast. PA yields more varied gains, with notable improvements on LogicVista and ZeroBench for the 2B student and MMMU for the 4B student. This pattern suggests that agreement under mild visual perturbations can support generalization beyond the geometric training domain, although its benefits depend on the task and student scale.

Combining both objectives achieves the highest overall average at both scales, improving over vanilla OPD by 1.60 and 1.04 points for the 2B and 4B students, respectively. Gains extend to both in-domain accuracy and out-of-domain average performance. While the combination does not preserve every component's best individual score, its stronger aggregate results support the complementary roles of sensitivity to evidence removal and stability under mild visual variation in student learning.

Table 10: Ablation of the TPC and PA coefficients on ViRL39K.

Value	Mathematical	Logical	General	Avg.
λ_{TPC}				
0.005	**47.07**	25.09	29.28	37.13
0.01	46.20	25.05	28.71	36.54
0.02	47.04	**25.16**	**30.60**	**37.46**
λ_{PA}				
0.01	46.41	24.37	28.57	36.44
0.02	47.04	**25.16**	**30.60**	**37.46**
0.04	**47.15**	24.84	30.44	37.39

B.3 Empirical Sparsity of the Teacher Gate

We examine whether the zero-margin criterion ($\tau = 0$) makes the teacher gate overly dense. In a representative S2B–T8B training run with 60% random patch masking, we analyze 3,840 logged rollouts containing 8,649,621 valid response tokens. The gate retains 4,054,346 tokens, corresponding to a token-weighted activation rate of 46.87%. When each rollout is weighted equally, the mean and median activation rates are 48.29% and 48.39%, respectively, with an interquartile range of 43.50–53.48% and a 5th–95th percentile range of 32.36–65.64%. These results show that masking does not induce a uniformly positive teacher likelihood drop: even with $\tau = 0$, more than half of the response tokens are filtered out on average.

B.4 Impact of the Coefficients of TPC and PA

We extend the coefficient ablations on Geometry3K in the main text to ViRL39K. As shown in Table 10, average performance varies by 0.92 points across the tested TPC coefficients and 1.02 points across the PA coefficients, with all settings outperforming the OPD baseline of 36.09. Both coefficients achieve the highest average score at 0.02, while increasing λ_{PA} to 0.04 yields nearly identical performance (37.39 vs. 37.46). Together with the Geometry3K results, these findings show that S-OPD remains effective across a range of coefficient choices on both training datasets. We set $\lambda_{\text{TPC}} = \lambda_{\text{PA}} = 0.02$ by default in the main experiments on ViRL39K, as this configuration achieves the highest average score.

B.5 Token-level Log-probability Analysis

Full-rollout log-probability analysis. Figure 9 compares the teacher–student token-level log-probability gap over complete rollouts on ViRL39K and Geometry3K. S-OPD exhibits similar training dynamics to OPD and reaches comparable final gaps at both student scales, with smaller gaps during parts of training for the 4B student, particularly on Geometry3K. Since this metric averages over initial planning, intermediate reasoning, and final-answer generation, it can obscure changes within individual stages. Together with the accuracy gains and the reductions in middle-segment gaps analyzed below, these results suggest that S-OPD preserves overall teacher alignment while improving learning within the reasoning process, consistent with its goal of strengthening the student's use of visual evidence.

Segment-level log-probability analysis. We divide each rollout into four equal-length segments, Q1–Q4, to examine alignment throughout the response. A typical solution roughly progresses from establishing an approach, through inspecting visual evidence and deriving intermediate results, to producing the final answer, although these stages need not align exactly with segment boundaries. Under this interpretation, improvements in the middle segments are particularly relevant to learning how to connect visual evidence with reasoning. S-OPD maintains a smaller Q2 gap through much of training for the 4B student on Geometry3K and reduces the Q3 gap during later training for the 2B student on ViRL39K. Both 4B settings also show smaller Q3 gaps during portions of training. This pattern suggests that S-OPD benefits intermediate solution construction rather than concentrating its effect on final-answer prediction. Together with the comparable full-rollout gaps and improved accuracy, these observations are consistent with the intended mechanism: visual contrast and alignment help students use teacher supervision more effectively within the reasoning process.

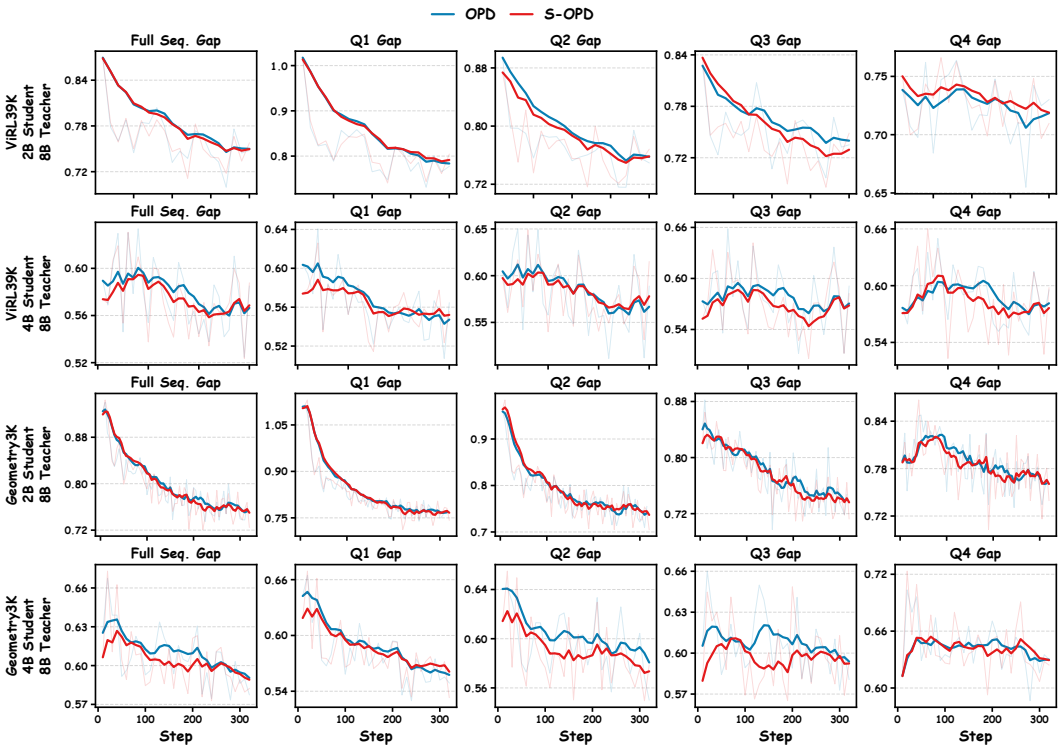


Figure 9: Token-level log-probability analysis of 2B and 4B students on ViRL39K and Geometry3K during training. We report the teacher–student token-level log-probability gap under different teacher–student configurations.

Table 11: Computational overhead of S-OPD. Both methods use four NVIDIA A100 80GB GPUs, two for the student and two for the teacher. Training time excludes initialization, checkpoint saving, and validation. Overhead is relative to OPD in time per step.

	ViRL39K, 2B		**Geometry3K, 4B**	
Metric	**OPD**	**S-OPD**	**OPD**	**S-OPD**
Tokens/response	2235.9	2233.2	802.4	800.3
Time/step (s)	189.44	243.56	92.26	107.06
Gen. (ms/token)	0.233	0.252	0.315	0.333
Overhead	–	+28.6%	–	+16.0%

B.6 Computational Overhead Analysis

S-OPD introduces additional training computation through teacher evaluation of masked images for token gating and student evaluation of masked and noisy images for TPC and PA. All branches reuse the same student-generated response, requiring no additional autoregressive rollouts, trainable parameters, or annotations. Table 11 quantifies the resulting overhead: average time per step increases from 189.44 to 243.56 seconds for the 2B student on ViRL39K and from 92.26 to 107.06 seconds for the 4B student on Geometry3K, corresponding to moderate increases of 28.6% and 16.0%, respectively. Average response lengths remain nearly unchanged in both settings. These auxiliary computations are used only during training; inference retains the original student architecture and decoding procedure, with no additional computational overhead.

C Qualitative Case Studies

C.1 Sample-level Rollout Analysis

To complement the aggregate benchmark results, we examine three sample-level cases from MathVerse, LogicVista, and MMMU-val. These examples are selected using two criteria: the OPD baseline produces an incorrect answer, whereas S-OPD produces the correct answer; and solving the problem requires information that is available only from the image. The cases cover complementary forms of visual reasoning, including reading geometric dimensions, propagating motion through spatial relations, and recognizing fine-grained visual patterns. Rather than serving as additional quantitative evidence, these examples illustrate how the two models differ in their use of visual evidence along the reasoning process.

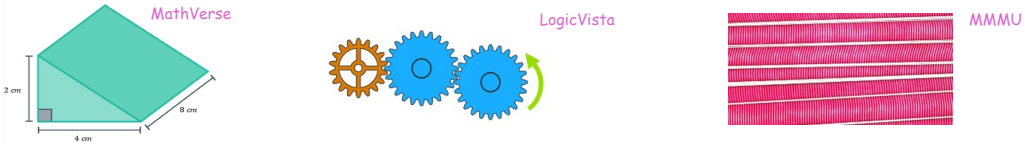


Figure 10: Qualitative examples from MathVerse, LogicVista, and MMMU-val where 2B S-OPD correctly grounds its reasoning in visual evidence, while 2B OPD produces incorrect answers.

C.1.1 MathVerse: Triangular Prism Volume Calculation

OPD baseline (incorrect, answer is 8 cm^3)

Official extracted answer: 8.
Although the response identified 2 cm, 4 cm, and 8 cm, it repeatedly confused the sides of the triangle with the lengths of the prism, and the final output was truncated. The original response was 14k characters long.

S-OPD (Ours) (correct, answer is 32 cm^3)

Looking at the image:

- The triangle has a base of 4 cm and a height of 2 cm.
- The side of the prism that is perpendicular to the base is 8 cm.

$$\text{Base Area} = \frac{1}{2} \times 4 \times 2 = 4 \text{ cm}^2$$

$$V = 4 \times 8 = 32 \text{ cm}^3$$

So, the volume of the triangular prism is 32 cm^3.

Analysis. This case tests whether the model can assign the three visible measurements to the correct geometric quantities. The OPD baseline successfully reads the values 2, 4, and 8 from the image, indicating that its failure is not caused by completely missing the relevant visual content. Instead, it repeatedly confuses the dimensions of the triangular base with the length of the prism. This role-assignment error leads to an incorrect extracted answer of $8\ \text{cm}^3$ and a long, self-correcting response that ultimately terminates without a consistent derivation.

S-OPD uses the same visual measurements more coherently. It identifies 4 cm and 2 cm as the base and height of the right-triangular face, respectively, and interprets 8 cm as the length of the prism. It therefore computes the triangular area as $\frac{1}{2} \times 4 \times 2 = 4\ \text{cm}^2$ and the volume as $4 \times 8 = 32\ \text{cm}^3$. The contrast between the responses suggests that, for this example, **S-OPD improves the integration and semantic assignment of perceived quantities rather than merely extracting more numbers from the image**.

C.1.2 LogicVista: Gear Rotation Direction Reasoning

OPD baseline (incorrect, answer is A)

The rightmost blue gear is rotating counterclockwise, so the middle blue gear must rotate clockwise, and the orange gear must rotate clockwise as well.
Answer: (A) Clockwise

S-OPD (Ours) (correct, answer is B)

- The rightmost blue gear is rotating counterclockwise, as indicated by the green arrow.
- Since it is meshed with the middle blue gear, the middle blue gear must rotate clockwise.
- The middle blue gear is also meshed with the orange gear. Therefore, the orange gear must rotate in the opposite direction to the middle blue gear.

Therefore, the correct answer is (B) Counterclockwise.

Analysis. As shown in Figure 10, this example requires the direction of motion to be propagated through two successive gear contacts. The OPD baseline correctly recognizes the green arrow and infers that the rightmost gear rotates counterclockwise. It also correctly determines that the middle gear must rotate clockwise. However, it fails to apply the direction reversal at the second contact and consequently predicts that the orange gear also rotates clockwise. The error therefore occurs during relational composition rather than initial visual perception.

S-OPD explicitly follows the two visible gear contacts. Starting from the counterclockwise rotation of the rightmost gear, it reverses the direction once for the middle gear and a second time for the orange gear, yielding the correct answer, counterclockwise. This case illustrates that **effective visual grounding must be maintained across multiple reasoning steps**: correctly reading the arrow alone is insufficient unless the model consistently applies the spatial interaction represented by each gear contact.

C.1.3 MMMU-val: Muscle Tissue Image Identification

OPD baseline (incorrect, answer is A)

The image depicts a series of pink, cylindrical structures with a distinct striped or ribbed pattern. These are characteristic of cardiac muscle.
Therefore, the correct answer is: A. Cardiac muscle

S-OPD (Ours) (correct, answer is C)

The image depicts a series of pink, cylindrical structures with a distinct striped or ridged pattern. These are characteristic of muscle tissue, specifically skeletal muscle fibers.
Cardiac muscle is typically arranged in a more complex, branching pattern. Skeletal muscle is composed of long, cylindrical, striated fibers arranged in parallel bundles. The image shows these fibers in a neat, parallel arrangement.
Therefore, the correct answer is C. Skeletal muscle.

Analysis. Both models detect the prominent striated appearance of the tissue, but they interpret the arrangement of the fibers differently. The OPD baseline associates striation with cardiac muscle and selects option A, without sufficiently accounting for the long, unbranched, and highly parallel organization visible in the image. Its prediction is therefore based on a relevant but non-discriminative visual attribute.

S-OPD considers both the shared attribute and the features that distinguish the candidate tissue types. In particular, it contrasts the branching organization expected of cardiac muscle with the long, cylindrical, parallel fibers characteristic of skeletal muscle. **This fine-grained comparison leads to the correct selection of option C**. Unlike the preceding examples, which emphasize quantitative and spatial reasoning, this case demonstrates the use of multiple visual attributes for category discrimination.

C.2 Token-level Probability Visualization and Analysis

To examine how teacher gating modulates visually induced policy differences at the token level, we visualize three representative examples covering mathematical reasoning, abstract visual logic, and scientific diagram understanding. For each example, the student first generates a fixed response y from the original image I and the question q. We then score the same response and prefix under the original image I and a patch-masked image $\widetilde{I}$ using both the student and teacher models. This controlled comparison isolates how removing visual evidence changes the probability assigned to each generated token.

Figure 11 visualizes the three cases. The signed student policy contrast is defined as

$$\Delta_t^S = \log p_S(y_t \mid q, I, y_{<t}) - \log p_S(y_t \mid q, \widetilde{I}, y_{<t}), \tag{9}$$

where orange indicates that the original image increases the probability of token y_t, blue indicates the opposite direction, and color intensity represents the absolute magnitude. The teacher gate strength is

$$s_t = \left[\log p_T(y_t \mid q, I, y_{<t}) - \log p_T(y_t \mid q, \widetilde{I}, y_{<t})\right]_+, \tag{10}$$

with darker purple denoting stronger teacher support from the original visual evidence. Finally, the gated policy contrast combines the teacher gate with the non-negative student contrast:

$$C_t^{\text{gated}} = s_t \left(\exp(d_t) - d_t - 1\right), \qquad d_t = \log p_S(y_t \mid q, \widetilde{I}, y_{<t}) - \log p_S(y_t \mid q, I, y_{<t}). \tag{11}$$

Darker yellow therefore marks tokens for which the student policy changes substantially and the teacher simultaneously provides positive visual support.

The three cases reveal consistent but task-dependent behavior. In the MathVerse example, the student contrast responds broadly to visually grounded terms such as "unit," "sec," and the coordinates of $P(x, y)$, whereas the teacher gate places stronger emphasis on mathematical entities such as θ, $\sec\theta$, P, and the predicted answer. Their combination concentrates the gated contrast on the quantities required to infer $\sec\theta$ from the unit-circle diagram. In LogicVista, the student is sensitive to several local shape descriptions, while the teacher more selectively emphasizes tokens describing the structural pattern, including intersections and shape categories such as "cross", "triangle", and "square". The resulting gated contrast highlights the tokens used to distinguish the two sets. In the MMMU-val example, the contrast is concentrated around branch identities, node relations, and the answer token corresponding to the most recent common ancestor, reflecting the relational structure of the phylogenetic tree.

Across all three task types, the teacher gate and student policy contrast exhibit clearly different token-level emphasis. The teacher signal is therefore not a simple copy or uniform rescaling of the student's visual sensitivity. Instead, the gated policy contrast preserves student discrepancies only when they are supported by the teacher's response to the visual evidence, concentrating the learning signal on visually grounded and decision-relevant reasoning steps. The consistent behavior across distinct visual-reasoning domains supports the intended role of our teacher-gating mechanism as a selective token-level routing strategy.

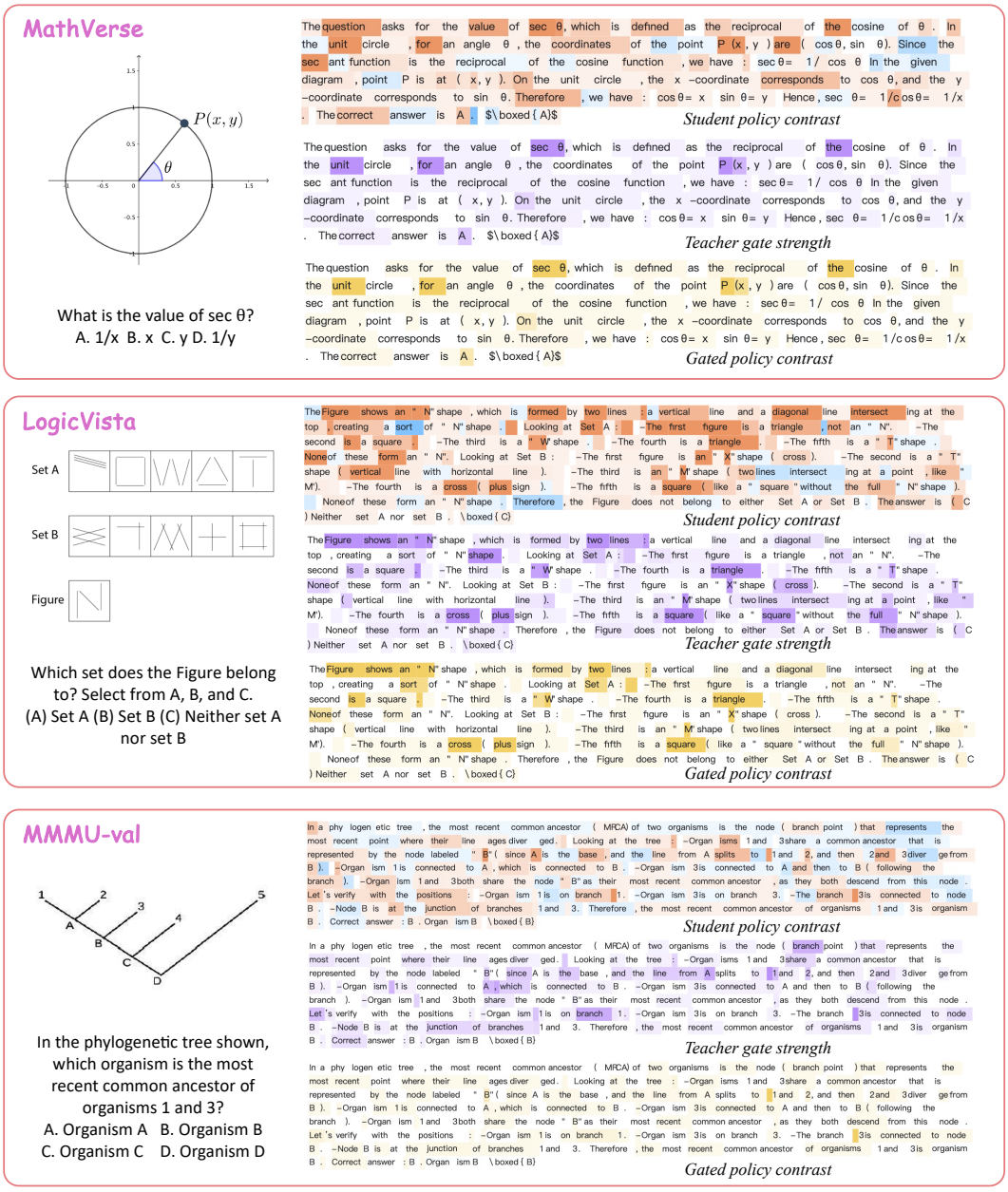


Figure 11: **Qualitative visualization of token-level policy contrast and teacher gating across three types of visual reasoning**: mathematical geometry (MathVerse), abstract visual logic (LogicVista), and scientific diagram understanding (MMMU-val). In the *student policy contrast*, orange and blue indicate positive and negative changes in token log-probability between the original and masked visual inputs, respectively, while darker colors denote larger absolute changes. For the *teacher gate strength*, darker purple indicates a larger positive teacher log-probability gap between the original and masked images. For the *gated policy contrast*, darker yellow indicates a larger teacher-modulated policy discrepancy. The teacher and student exhibit distinct token-level emphasis across all three tasks, and the gated policy contrast selectively concentrates on key visual elements and decision-relevant reasoning steps, demonstrating the intended effect of our teacher-gating mechanism.