RETHINKING PRIVILEGED INFORMATION EFFECTIVENESS IN MULTIMODAL ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) can provide the teacher with privileged information (PI) unavailable to the student, potentially improving teacher supervision during training. However, it remains unclear whether PI that provides a larger direct benefit to the teacher also leads to larger gains that persist in the PI-free student. We study this question in multimodal OPSD by explicitly separating two quantities: *PI access gain*, the target-task performance gain obtained when a fixed model is given direct access to PI, and *PI effectiveness*, the target-task performance gain retained by the PI-free student after OPSD relative to a matched no-PI baseline. In controlled fine-grained perception experiments, these quantities are not consistently aligned across PI conditions: ANSWER yields a larger PI access gain than HINT, whereas HINT achieves higher PI effectiveness. Student-side analyses further reveal condition-specific differences in the use of task-relevant visual evidence, with HINT showing the clearest target-region localization and selective sensitivity to target-region corruption. Extending the analysis across tasks further shows substantial variation in PI effectiveness across target tasks and concrete PI conditions. Overall, our results show that greater PI access gain does not necessarily translate into greater PI effectiveness, and that PI effectiveness is task- and condition-dependent and not intrinsic to a PI form. Motivated by this task dependence, a simple task-level PI routing strategy improves aggregate held-out performance over uniform PI baselines.

1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024) trains a student using teacher supervision on trajectories generated by the student itself. On-policy self-distillation (OPSD) (Zhao et al., 2026) adapts this paradigm to self-distillation, with teacher supervision derived from the same underlying model. It has recently gained attention as a promising post-training approach. In OPSD, potentially more effective teacher supervision is often enabled through asymmetric conditioning, where the teacher is conditioned on privileged information (PI) (Shrestha & Tessier, 2026; Zhang et al., 2026a; Zhu et al., 2026b) unavailable to the student, which can take diverse forms. Given the diversity of PI forms, a central question is: *Does PI that provides a larger direct performance gain when available to the teacher necessarily yield larger gains for the PI-free student?*

We study this problem in multimodal OPSD, where privileged information can be provided through both textual and visual modalities. We first measure the target-task performance gain obtained by a fixed base model when given direct access to PI relative to the same model without PI, which we refer to as *PI access gain*. We then measure the target-task performance gain retained by the resulting PI-free student after OPSD relative to a matched no-PI OPSD baseline, which we refer to as *PI effectiveness*. For PI effectiveness, we compare against a matched no-PI OPSD baseline to control for gains from the OPSD procedure itself, so that the metric more directly reflects the incremental benefit of a given PI condition. We formalize both quantities in §2.

We first examine the relationship between PI access gain and PI effectiveness in a controlled fine-grained perception (FGP) setting (§3). We hold the task, model, student initialization, and OPSD procedure fixed while varying the PI condition across four forms: ANSWER, TRAJECTORY, EX-

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.

PERIENCE, and HINT. This comparison reveals that PI access gain and PI effectiveness are not consistently aligned across PI conditions. Most notably, ANSWER achieves a larger PI access gain than HINT (+41.8 vs. +27.3 percentage points), whereas HINT yields greater PI effectiveness than ANSWER (+19.87 vs. +10.74 percentage points). This contrast shows that greater PI access gain does not necessarily translate into greater PI effectiveness. External evaluations further show that the effects of PI training do not extend uniformly across capabilities: HINT performs strongest overall among the PI conditions, yet its gains on perception and real-world understanding do not extend to multimodal generalization. We then return to the controlled FGP setting to characterize how this mismatch between PI access gain and PI effectiveness manifests in the resulting PI-free students.

We characterize this mismatch from two complementary student-side perspectives: **(1) Visual-overlap analysis** (§4.1) computes residual attention rollout over visual patches and measures how strongly the resulting attribution is concentrated within the task-relevant target region. Across PI conditions, HINT exhibits the clearest shift toward this region, with stronger concentration both at generation onset and before the final answer. **(2) Matched occlusion analysis** (§4.2) uses matched image interventions to test whether visual evidence within the target region is behaviorally consequential for prediction. Target-region corruption is more damaging than equally sized background corruption across models, with HINT showing the clearest target-selective pattern: its performance is largely preserved under background corruption, but drops sharply when the task-relevant region is corrupted. Together, these analyses provide a student-side characterization of the differences in PI effectiveness observed in FGP, revealing condition-specific differences in target-region localization and selective sensitivity to task-relevant visual evidence. Given the strong effectiveness of HINT in FGP, we next ask whether it remains similarly effective when the target task changes.

We therefore extend our analysis beyond fine-grained perception to knowledge-intensive perception (KIP), multi-hop spatial reasoning (MSR), and geometric reasoning (GeoR), which introduce different task demands and instantiate PI conditions accordingly. KIP offers a controlled cross-task contrast: it retains the same localized visual HINT as FGP, but shifts the target requirement from fine-grained localization to linking visual evidence with external knowledge, under which the PI-effectiveness ordering changes and ANSWER surpasses HINT. This reversal shows that the effectiveness of a given PI form can change with the target task. MSR further shows that task-specific adaptation of PI is not sufficient to guarantee higher effectiveness: although HINT is instantiated specifically for spatial reasoning, its PI effectiveness remains lower than that of ANSWER. GeoR provides a stronger boundary case: after broadening the PI space to ANSWER, ANALYSIS, TRAJECTORY, EXPERIENCE, and HINT, only TRAJECTORY yields a modest positive in-domain PI effect, and this gain does not extend consistently to broader mathematical-reasoning and multimodal evaluations. Taken together, these results show that PI effectiveness is not an intrinsic property of a PI form, but depends on the target task and its task-specific PI condition. This task dependence motivates a simple task-level PI routing strategy, which selects the PI condition for each task using validation performance and improves aggregate held-out performance over uniform PI baselines.

Our contributions are threefold. First, we explicitly distinguish *PI access gain* from *PI effectiveness*, and show that greater PI access gain does not necessarily imply greater PI effectiveness. Second, we provide a student-side characterization of PI effectiveness in fine-grained perception, revealing condition-specific differences in the use of task-relevant visual evidence. Third, we show that PI effectiveness is task- and condition-dependent rather than intrinsic to a PI form, and demonstrate the practical value of this finding through task-level PI routing.

2 PRELIMINARIES

2.1 ON-POLICY DISTILLATION

OPD trains a student using teacher supervision evaluated on trajectories generated by the student itself (Agarwal et al., 2024). Given an input x, the student samples a response $y \sim \pi_\theta(\cdot \mid x)$, yielding a sequence of visited prefixes $s_t = (x, y_{<t})$. At each prefix, the teacher provides a next-token distribution on the same state, and the student is optimized to align with this distribution:

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

Here, D denotes a token-level divergence between the student and teacher distributions. In contrast to offline distillation on a fixed set of responses, OPD evaluates teacher supervision on the evolving distribution of prefixes visited by the student policy.

2.2 Multimodal OPSD with Privileged Information

OPSD retains the student-generated rollout of OPD, but constructs teacher supervision from the same underlying model under an asymmetric training-time condition (Zhao et al., 2026; Yuan et al., 2026). For a multimodal input $x = (I, q)$, where I is an image and q a query, the student conditions only on x, while the teacher additionally observes privileged information (PI) z that is unavailable at inference time. At a student-generated prefix $y_{<t}$, the corresponding teacher and student distributions are $\pi_T(\cdot \mid x, z, y_{<t})$ and $\pi_\theta(\cdot \mid x, y_{<t})$, respectively.

2.3 Characterizing Privileged Information

We use *PI form* to denote the semantic type of privileged information, such as Answer, Hint, Experience, Trajectory, or Analysis. For a target task τ, each form is instantiated as a PI condition $c \in \mathcal{C}_\tau$, with z_c denoting the privileged signal provided under that condition. For each PI condition, we evaluate two task-level quantities: *PI access gain* and *PI effectiveness*. We additionally use distributional diagnostics to characterize how closely the student tracks the privileged teacher during OPSD (see Appendix D.2).

PI access gain. We define the PI access gain of a PI condition c on task τ as the target-task performance gain obtained when a fixed base model is given direct access to the corresponding privileged signal z_c:

$$\mathcal{G}_{\text{access}}(c, \tau) = \text{Perf}_\tau(f_{\text{base}}(x, z_c)) - \text{Perf}_\tau(f_{\text{base}}(x)) \,.$$

This quantity measures the target-task benefit of direct PI access relative to the same model without PI.

PI effectiveness. PI effectiveness quantifies the task-level gain of the PI-free student after training with privileged information, relative to a matched OPSD baseline trained without PI (see Appendix A.4). Let S_c denote the PI-free student obtained after OPSD under privileged condition c, and let $S_{\text{No-PI}}$ denote the matched student trained with the same OPSD configuration but without privileged information. On task τ, we define

$$\mathcal{E}_{\text{PI}}(c, \tau) = \text{Perf}_\tau(S_c) - \text{Perf}_\tau(S_{\text{No-PI}}).$$

Positive values indicate that training with privileged condition c provides an additional benefit over matched no-PI OPSD on the target task.

3 A Controlled Study of Privileged Information

Question. *Does greater PI access gain lead to greater PI effectiveness?*

We investigate this question in a controlled fine-grained perception setting, holding the task, model, student initialization, and OPSD procedure fixed while varying only the privileged condition. This setup allows us to separately measure PI access gain on the teacher side and PI effectiveness on the student side. We first compare these two quantities across PI conditions (§3.2), and then examine whether the differences persist beyond the target task across broader benchmarks (§3.3).

3.1 Controlled Fine-Grained Perception Setting

Setting and PI conditions. Our controlled setting is fine-grained visual perception (FGP), where answering a question requires identifying localized evidence in the input image. We use Qwen3.5-4B (Qwen Team, 2026) as the base model and vary the privileged condition across four PI forms. Answer provides the ground-truth answer, Trajectory provides sampled model trajectories, Experience summarizes reusable information derived from trajectories, and Hint provides a localized visual crop of the question-relevant region. Full PI specifications are provided in Appendix A.2. All conditions use the same student initialization and OPSD configuration, differing

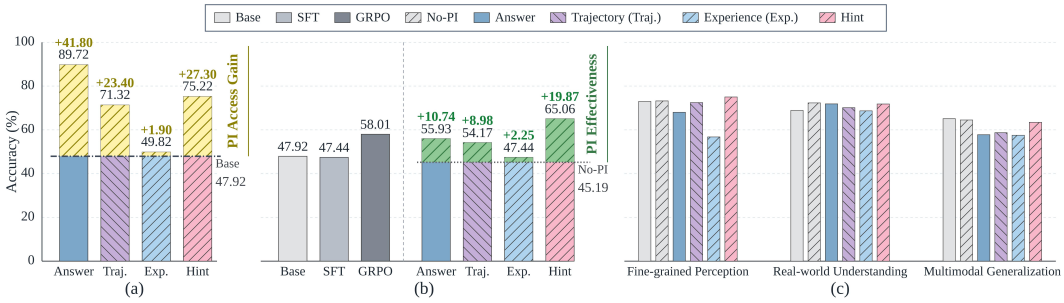


Figure 1: **Controlled comparison of privileged information in fine-grained perception. (a)** Accuracy of the fixed base model under each PI condition, with gains over the unconditioned Base annotated as **PI access gain**. **(b)** Post-training accuracy of the resulting PI-free students and reference baselines, with gains over the matched no-PI OPSD baseline annotated as **PI effectiveness**. **(c)** Average post-training accuracy across external benchmarks grouped into fine-grained perception*, real-world understanding, and multimodal generalization. Fine-grained perception* denotes benchmarks distinct from those used in the controlled FGP setting.

only in the privileged signal provided to the teacher. Full training details are provided in Appendix A.3.

Evaluation. We evaluate the controlled FGP setting in terms of the two quantities defined in §2: PI access gain and PI effectiveness. To examine whether the resulting differences extend beyond the target task, we additionally evaluate the PI-free students on external benchmarks covering fine-grained perception, real-world understanding, and multimodal generalization, with dataset composition and evaluation details provided in Appendix A.5.

3.2 PI Access Gain vs. PI Effectiveness

Figure 1(a–b) shows that PI access gain and PI effectiveness are not consistently aligned across PI conditions. On the teacher side, PI access gain follows the ordering **Answer (+41.8)** > **Hint (+27.3)** > Trajectory (+23.4) > Experience (+1.9) percentage points. After OPSD, however, PI effectiveness follows a different ordering: **Hint (+19.87)** > **Answer (+10.74)** > Trajectory (+8.98) > Experience (+2.25) percentage points. Most notably, Answer achieves the largest PI access gain (+41.8 percentage points), whereas Hint yields the largest PI effectiveness (+19.87 percentage points) This ordering reversal shows that greater PI access gain does not necessarily translate into greater PI effectiveness. We next examine how these differences manifest beyond the target FGP evaluation.

3.3 Beyond the Target Task: Capability Profiles

Figure 1(c) summarizes the resulting PI-free students across three groups of external benchmarks: fine-grained perception, real-world understanding, and multimodal generalization. On out-of-distribution (OOD) fine-grained perception benchmarks, the PI conditions exhibit a clear spread: Hint achieves the highest average accuracy (75.02%), Trajectory remains close to the Base model (72.46% vs. 72.95%), while Answer and Experience drop to 68.02% and 56.77%, respectively. On real-world understanding, the differences are smaller: Hint and Answer achieve similar average accuracy (71.79% and 71.85%), followed by Trajectory at 70.11%, while Experience remains close to the Base model (68.66% vs. 68.88%). On multimodal generalization, all PI conditions underperform the Base model (65.10%), with Hint preserving the strongest performance at 63.42%, while Trajectory, Answer, and Experience fall to 58.71%, 57.78%, and 57.52%, respectively. Taken together, these results show that the effects of PI training vary across capabilities: Hint performs strongest overall among the PI conditions, yet its gains on perception and real-world understanding do not extend to multimodal generalization.

Takeaway. ***Greater PI access gain does not necessarily translate into greater PI effectiveness.*** *The effects of PI training also vary across capabilities:* Hint *performs strongest overall among*

the PI conditions, yet its gains on perception and real-world understanding do not extend to multimodal generalization.

4 STUDENT-SIDE CHARACTERIZATION OF PI EFFECTIVENESS

Question. *How do PI-free students differ across PI conditions that vary in PI effectiveness?*

We return to the controlled FGP setting to compare the PI-free students across PI conditions with varying PI effectiveness. All analyses are conducted on the held-out FGP evaluation set. We characterize these student-side differences from two complementary perspectives: visual-overlap analysis (§4.1) examines where visual attribution is concentrated, while matched occlusion interventions (§4.2) test whether evidence in those regions is behaviorally consequential for prediction.

4.1 VISUAL OVERLAP: WHERE IS VISUAL ATTRIBUTION CONCENTRATED?

Measurement. Fine-grained perception provides a ground-truth bounding box identifying the task-relevant visual region. We compute residual attention rollout over visual patches and measure how strongly the resulting attribution is concentrated within this target region. We formalize this concentration using a task-relevant visual-overlap score. Let $\mathcal{V}$ denote the set of visual patches and $\mathcal{B} \subseteq \mathcal{V}$ the patches covered by the ground-truth target box. For a generation position t, we normalize the rollout attribution over $\mathcal{V}$ as

$$\tilde{r}_j = \frac{R_{t,j}}{\sum_{k \in \mathcal{V}} R_{t,k}}, \qquad j \in \mathcal{V}.$$

We then compute the attribution mass assigned to the target region and its spatial proportion:

$$A_{\mathcal{B}} = \sum_{j \in \mathcal{B}} \tilde{r}_j, \qquad \rho_{\mathcal{B}} = \frac{|\mathcal{B}|}{|\mathcal{V}|}.$$

The visual-overlap score is defined as

$$V_{\text{overlap}} = A_{\mathcal{B}} - \rho_{\mathcal{B}}.$$

Thus, $V_{\text{overlap}} = 0$ corresponds to uniform spatial allocation, whereas positive values indicate that visual attribution is concentrated within the target region beyond what is expected from its spatial extent alone. We evaluate V_{overlap} at two generation stages: at generation start and immediately before the final answer. This comparison tests whether task-relevant localization is already present at generation onset or emerges during the response.

Results. Figure 2 shows substantial variation in V_{overlap} across PI conditions at both generation stages. Among the PI conditions, HINT exhibits the clearest shift toward the task-relevant region, with substantially higher V_{overlap} at both generation stages. An instance-level visualization of this shift is provided in Appendix C (Figure 7). This pattern appears for both correct and incorrect trajectories, suggesting a condition-level shift in visual attribution. Within each condition, however, the correct and incorrect distributions remain strongly overlapping, with AUROCs close to chance for most models. ANSWER is the limited exception, showing modest separation immediately before the final answer (AUROC 0.557, 95% CI $[0.512, 0.602]$). These results show that visual overlap captures where attribution is concentrated, but not whether evidence in those regions is behaviorally consequential for prediction.

4.2 MATCHED OCCLUSION: DOES TASK-RELEVANT VISUAL EVIDENCE MATTER?

Measurement. We next test whether visual evidence within the task-relevant target region is behaviorally consequential using matched image interventions. For each example, we evaluate four image conditions: the original image, target-box blur over the ground-truth region, matched background blur over an equally sized non-overlapping region, and full-image blur (Figure 3, right). Comparing target-box blur with matched background blur controls for the amount of corruption and better reflects sensitivity to task-relevant visual evidence, while full-image blur provides a reference under global visual degradation.

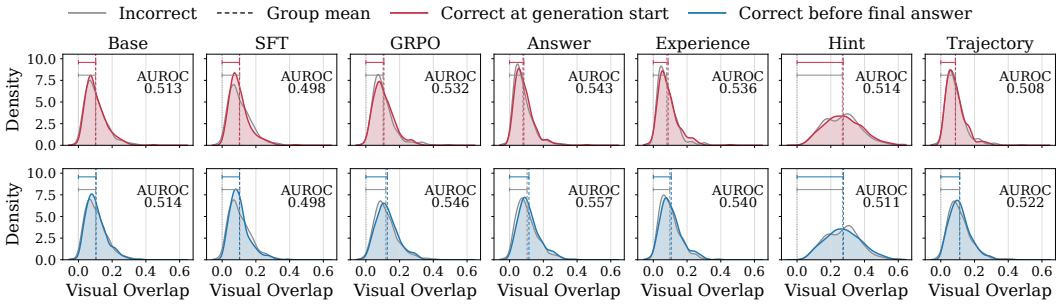


Figure 2: **Task-relevant visual overlap across PI conditions.** The top row probes visual attribution at generation start, and the bottom row immediately before the final answer. Colored and gray curves denote correct and incorrect trajectories, respectively. Dashed vertical lines mark the mean V_{overlap} for each group. AUROC measures the separation between correct and incorrect trajectories within each model.

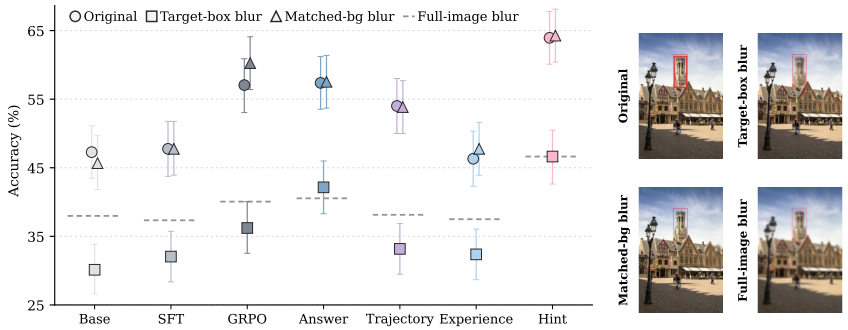


Figure 3: **Matched occlusion analysis of task-relevant visual evidence.** Left: accuracy under the original image, target-box blur, matched background blur, and full-image blur. Error bars show 95% bootstrap confidence intervals. Right: a representative example of the four intervention conditions.

Results. Across models, target-box blur is consistently more damaging than matched background blur, indicating that the annotated target region contains visual evidence that is behaviorally relevant to prediction (Figure 3, left). Among the PI conditions, HINT exhibits a particularly clear target-selective pattern: its accuracy is essentially unchanged under matched background blur ($63.94\% \rightarrow 64.26\%$) but drops sharply to 46.63% under target-box blur. Notably, the target-box-blur accuracy nearly matches the full-image-blur reference, showing that HINT's performance gain largely disappears when task-relevant evidence is removed.

Together with the visual-overlap analysis, these interventions reveal condition-specific differences in how the PI-free students use task-relevant visual evidence, with HINT showing the clearest combination of target-region attribution and selective sensitivity. These patterns provide a student-side characterization of PI effectiveness in FGP: PI conditions with different effectiveness exhibit different degrees of target-region localization and selective sensitivity to task-relevant visual evidence.

Takeaway. ***PI conditions with varying effectiveness exhibit different degrees of target-region localization and selective sensitivity to task-relevant visual evidence.*** *Among them,* HINT *shows the clearest combination of these two student-side signatures.*

5 PI EFFECTIVENESS ACROSS TASKS AND CONDITIONS

Question. *Is PI effectiveness intrinsic to a PI form, or does it vary across tasks and conditions?*

The controlled FGP study (§3) shows that PI effectiveness can vary substantially across PI conditions within a fixed task. We next broaden this analysis across tasks and task-specific PI conditions. We

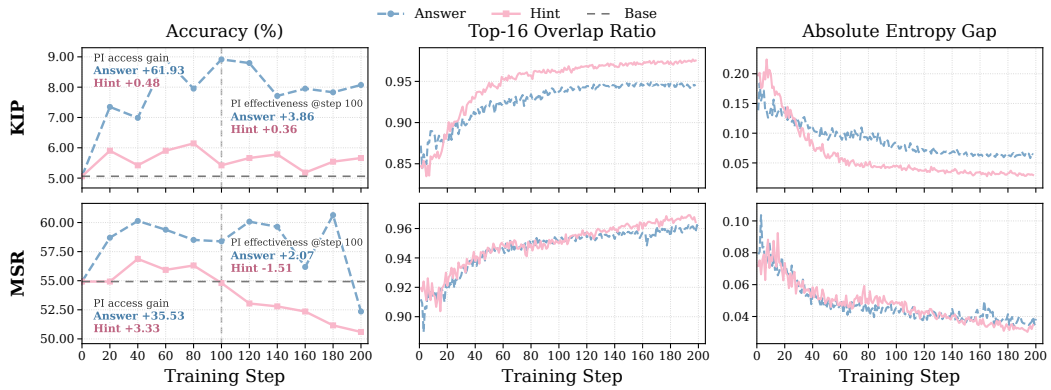


Figure 4: **Task performance and online teacher-student alignment in KIP and MSR.** The top and bottom rows correspond to KIP and MSR, respectively. Left: PI-free student evaluation accuracy, with PI access gain and PI effectiveness(@step100) annotated. Middle: Top-16 overlap ratio. Right: absolute entropy gap. Higher overlap and a smaller entropy gap indicate closer agreement.

first examine cross-task variation in KIP and MSR (§5.1), then broaden the PI space in GeoR (§5.2), and finally translate this task dependence into a simple task-level PI routing strategy (§5.3).

5.1 Cross-Task Variation in PI Effectiveness

Knowledge-Intensive Perception. KIP provides a controlled cross-task contrast with FGP: it uses the same crop-based HINT, while the task shifts from fine-grained localization to linking visual evidence with external or commonsense knowledge. With HINT instantiated identically across the two tasks, the PI effectiveness ordering reverses: at step 100, PI effectiveness is +3.86 percentage points for ANSWER and +0.36 percentage points for HINT (Figure 4, top left), opposite to the pattern observed in FGP. ANSWER also provides a larger PI access gain than HINT (+61.93 vs. +0.48 percentage points). As complementary diagnostics, HINT exhibits higher Top-16 overlap and a smaller absolute entropy gap than ANSWER during late training, despite its lower PI effectiveness (Figure 4, top middle–right). Thus, KIP shows that PI effectiveness is not intrinsic to a PI form: even the same crop-based HINT can exhibit different effectiveness across target tasks.

Multi-Hop Spatial Reasoning. MSR provides a complementary contrast: here, HINT is adapted specifically to the spatial-reasoning task rather than reusing the crop-based construction from FGP and KIP. Despite this task-specific adaptation, ANSWER achieves a larger PI access gain than HINT (+35.53 vs. +3.33 percentage points). At step 100, Figure 4 (bottom left) shows higher PI effectiveness for ANSWER than HINT (+2.07 vs. -1.51 percentage points). As in KIP, HINT shows slightly higher Top-16 overlap and a smaller absolute entropy gap than ANSWER during late training, despite its lower PI effectiveness (Figure 4, bottom middle–right). Thus, MSR shows that adapting a PI form into a task-specific condition is not sufficient to ensure higher PI effectiveness.

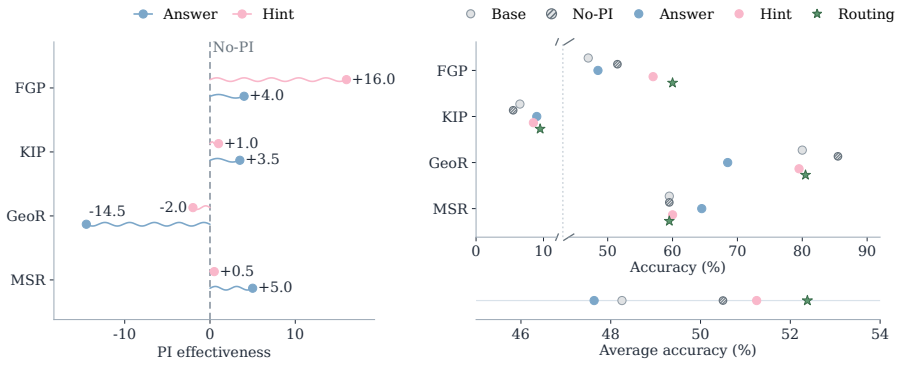
Taken together, KIP and MSR show that PI effectiveness depends on the target task and its concrete PI condition, rather than on PI form alone.

5.2 Geometric Reasoning as a Boundary Case

GeoR provides a stronger boundary case for PI effectiveness, asking whether any plausible PI condition yields a consistent benefit over matched No-PI OPSD. We evaluate five PI conditions, ANSWER, ANALYSIS, TRAJECTORY, EXPERIENCE, and HINT, under the same training and evaluation protocol. The Backbone is included as an untrained reference, while PI effectiveness is measured relative to matched No-PI OPSD. Relative to matched No-PI OPSD, TRAJECTORY yields a modest positive PI effectiveness (+2.19 percentage points), recovering much of the in-domain degradation while remaining close to the original Backbone (82.85% vs. 83.58%). EXPERIENCE remains unchanged, while HINT, ANALYSIS, and ANSWER show negative PI effectiveness (Table 1). This limited in-domain gain does not extend beyond GeoR: across all external mathematical-reasoning and

Table 1: Accuracy (%) of Qwen3.5-4B and step-100 GeoR-trained OPSD checkpoints. GeoR denotes the in-domain evaluation set. Colored superscripts indicate percentage-point changes relative to the matched No-PI OPSD reference. Bold indicates the best result among PI-conditioned variants.

Model	In-domain	Mathematical Reasoning				Multimodal Generalization			
	GeoR	MathVista	MathVision	MathVerse	WeMath (S/L)	LogicVista	MMMU	MMMU-Pro	MMStar
Reference models									
Backbone	83.58	81.90	54.31	70.76	51.90 / 75.71	63.39	66.78	55.78	72.73
No-PI OPSD	80.66	81.90	55.46	69.75	52.00 / 76.38	66.74	66.33	55.78	73.00
PI-conditioned OPSD									
Answer	54.98$^{-25.68}$	80.40$^{-1.50}$	43.82$^{-11.64}$	60.96$^{-8.79}$	47.24$^{-4.76}$ / **74.10**$^{-2.28}$	60.04$^{-6.70}$	58.67$^{-7.66}$	48.90$^{-6.88}$	**69.13**$^{-3.87}$
Analysis	71.17$^{-9.49}$	79.10$^{-2.80}$	36.64$^{-18.82}$	61.24$^{-8.51}$	36.67$^{-15.33}$ / 60.86$^{-15.52}$	56.47$^{-10.27}$	54.33$^{-12.00}$	47.98$^{-7.80}$	67.33$^{-5.67}$
Trajectory	**82.85**$^{+2.19}$	77.00$^{-4.90}$	**44.54**$^{-10.92}$	**63.83**$^{-5.92}$	41.52$^{-10.48}$ / 66.10$^{-10.28}$	60.27$^{-6.47}$	62.33$^{-4.00}$	52.20$^{-3.58}$	68.20$^{-4.80}$
Experience	80.66$^{+0.00}$	**81.00**$^{-0.90}$	41.94$^{-13.52}$	61.45$^{-8.30}$	**47.90**$^{-4.10}$ / 72.86$^{-3.52}$	60.27$^{-6.47}$	52.44$^{-13.89}$	47.86$^{-7.92}$	68.27$^{-4.73}$
Hint	80.29$^{-0.37}$	77.20$^{-4.70}$	42.04$^{-13.42}$	58.53$^{-11.22}$	43.05$^{-8.95}$ / 66.29$^{-10.09}$	**61.38**$^{-5.36}$	**64.11**$^{-2.22}$	**53.35**$^{-2.43}$	67.87$^{-5.13}$



(a) PI effectiveness on validation set (b) Held-out test accuracy

Figure 5: **Task-level PI routing in mixed-task OPSD. (a)** Validation PI effectiveness of uniform ANSWER and HINT conditioning relative to matched uniform No-PI OPSD. Routing selects the condition with the highest validation PI effectiveness for each task. **(b)** Held-out test accuracy of the uniform baselines and the task-level routing checkpoint. The lower track reports the unweighted mean across the four tasks.

multimodal-generalization benchmarks, every PI-conditioned checkpoint underperforms matched No-PI OPSD. Thus, GeoR shows that a modest in-domain PI gain does not necessarily extend to broader evaluations.

Taken together with KIP and MSR, these results further show that PI effectiveness is task- and condition-dependent rather than intrinsic to a PI form.

5.3 A PRACTICAL CONSEQUENCE: TASK-LEVEL PI ROUTING

The task-dependent PI effectiveness observed above raises a practical question for heterogeneous mixed-task training: should a single PI condition be applied uniformly across tasks, or should privileged conditioning be selected separately for each task? We study this question in a mixed-task OPSD setting over FGP, KIP, MSR, and GeoR, comparing uniform PI assignment with a simple task-level routing strategy. We first train uniform No-PI, ANSWER, and HINT checkpoints on the same mixed-task data, and then select the condition with the highest validation PI effectiveness for each task. Figure 5(a) summarizes the validation PI effectiveness used for routing, selecting HINT for FGP, ANSWER for KIP, ANSWER for MSR, and No-PI for GeoR. Using these fixed task-level conditioning assignments, we train a single routing checkpoint on the same mixed-task data and evaluate it on held-out test sets disjoint from the validation sets used for routing selection. Figure 5(b) shows that task-level routing achieves the highest average held-out accuracy at 52.38%, compared with 51.25% for uniform HINT, 50.50% for uniform No-PI, and 47.63% for uniform ANSWER. The task-level effects are heterogeneous: routing improves over the corresponding uniform condition on FGP and KIP, but remains below uniform No-PI on GeoR and uniform ANSWER on MSR. These results show that task-level routing improves aggregate mixed-task performance by adapting privileged conditioning based on task-specific validation PI effectiveness.

Takeaway. ***PI effectiveness is not an intrinsic property of a PI form, but depends on the target task and its task-specific PI condition.*** *A simple task-level routing strategy based on validation PI effectiveness improves aggregate held-out performance over uniform PI assignment in mixed-task training.*

6 Related Work

Multimodal On-Policy Self-Distillation. OPD trains a student on its own rollout distribution while obtaining token-level supervision from a teacher at student-visited states (Agarwal et al., 2024). OPSD specializes this paradigm by deriving teacher supervision from the same or a closely coupled model, with the teacher advantage typically introduced through asymmetric training-time conditioning (Zhao et al., 2026). In multimodal OPSD, this asymmetry is commonly instantiated through textual or visual privileged information provided to the teacher during training but unavailable to the student at inference. Vision-OPD, for example, instantiates visual privilege through localized image crops, conditioning the teacher on cropped evidence while training the student on the original image (Yuan et al., 2026). Subsequent work broadens this design space through localized or zoomed visual privilege (Zhang et al., 2026b; Cai et al., 2026), resolution-based privilege (Zhu et al., 2026a), clean–corrupted or asymmetrically augmented views (Wang et al., 2026a; Li et al., 2026b), and contrastive visual conditioning (Liang et al., 2026; Aniri et al., 2026). Beyond perception-oriented settings, multimodal OPSD has also been explored for spatial planning and visual reasoning using symbolic states, intermediate visual evidence, or structured multimodal reasoning signals (Luo et al., 2026; Li et al., 2026a; Wang et al., 2026b). Collectively, these works establish a broad design space for multimodal OPSD, where teacher advantage can arise from diverse forms of training-time information asymmetry.

Privileged Information in Multimodal OPSD. Within this design space, an important question is how the form and construction of privileged information affect what is retained by the PI-free student. ViCuR introduces the notion of recoverable privilege, emphasizing that privileged supervision is more useful when the evidence supporting it remains accessible from the student's original input (Tian et al., 2026). S^2VOPD studies a related form of asymmetric conditioning without privileged annotations, using asymmetric visual augmentation and showing that overly strong perturbations that remove task-relevant evidence produce less useful supervision (Li et al., 2026b). For multimodal reasoning, ViGOS identifies a complementary failure mode in which privileged reasoning supervision can bypass perceptual grounding, motivating the separation of perceptual and reasoning supervision (Wang et al., 2026b). Related work has also explored alternatives to direct answer privilege, using spatial or intermediate reasoning guidance to reduce shortcut learning (Xiang et al., 2026). Broader analyses of privileged self-distillation further show that performance gains do not necessarily imply faithful internalization of the privileged content itself (Shrestha & Tessier, 2026; Zhu et al., 2026b). Most closely related, recent work examines what privileged references contribute beyond self-distillation itself, comparing multiple reasoning references against matched reference-free distillation (Zhang et al., 2026a). It finds that a substantial portion of the observed gain can arise without privileged references, while the additional benefit of reference conditioning varies across models and rollout settings. These findings motivate separating the gains of self-distillation itself from the incremental contribution of privileged conditioning. Existing analyses, however, largely focus on a specific PI construction, failure mode, or reasoning setting, leaving the effects of heterogeneous PI across multimodal tasks less systematically understood.

7 Conclusion

We study privileged information in multimodal OPSD by distinguishing *PI access gain*, the direct performance benefit of PI access, from *PI effectiveness*, the gain retained by the PI-free student after OPSD. In controlled FGP, Answer has the highest PI access gain, while Hint has the highest PI effectiveness. Student-side analyses further reveal condition-specific differences in the use of task-relevant visual evidence. Across KIP, MSR, and GeoR, the same PI form can vary substantially in effectiveness across tasks. A simple task-level routing strategy based on validation PI effectiveness improves aggregate held-out performance in heterogeneous mixed-task training. Overall, our results show that greater PI access gain does not necessarily translate into greater PI effectiveness, and that PI effectiveness depends on the target task and its task-specific PI condition.

AI USE STATEMENT

During the preparation of this paper, we used large language models to assist with grammar checking, wording refinement, and readability improvements. We also used AI assistance to inspect and improve the presentation of selected tables and figures. In the experiments, a multimodal LLM was used as an offline experience extractor to construct the EXPERIENCE PI from stored trajectories and success or failure labels, as described in Appendix A.2. All AI-assisted revisions and AI-generated intermediate materials used in the experiments were reviewed by the authors. The authors verified that these materials did not change the intended meaning, technical content, reported results, or conclusions. The external benchmark evaluation used an LLM-based judge for grading model responses. This judge was used only for evaluation and was not used for training, data construction, or modification of the reported results. All experimental data, numerical results, tables, and figures were checked by the authors.

ETHICS STATEMENT

This work studies privileged-information conditioning in multimodal on-policy self-distillation. We do not conduct human-subject experiments, collect new personal data, or deploy models in real-world decision-making settings. Our experiments use publicly available benchmark datasets covering visual perception, knowledge-intensive perception, spatial reasoning, and geometric reasoning. The use of teacher-only privileged information introduces an information asymmetry between training and inference. We therefore evaluate both *PI access gain* and *PI effectiveness* to distinguish the benefit of direct PI access from the performance retained by the resulting PI-free student. All benchmark usage and model evaluations were conducted for research purposes, and the reported results were reviewed by the authors.

REPRODUCIBILITY STATEMENT

The main text defines the OPSD framework, PI access gain, PI effectiveness, and the evaluation protocols used in our experiments. The appendix provides the corresponding implementation and data details. Appendix A.1 describes the task datasets, source data, and retained training and evaluation counts. Appendix A.2 specifies the PI forms, modalities, and teacher-side delivery formats, while Appendix A.2 gives task-level construction examples. Appendix A.3 describes the teacher–student construction, distillation objective, optimization, batching, rollout configuration, and compute settings. Appendix A.4 explains the matched No-PI control used to measure PI effectiveness. Appendix A.5 summarizes the evaluation conventions and external benchmark coverage. Additional student-side analyses, training diagnostics, Top-K robustness analyses, and task-level routing diagnostics are provided in Appendix C, Appendix D.2, Appendix D.3, and Appendix D.4, respectively. Together, these details specify the data construction, training procedure, evaluation settings, and analysis protocols needed to reproduce the reported results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Muhammad Awais Ahmad, Tauqir Ahmed, Muhammad Aslam, Amjad Rehman, Faten S. Alamri, Saeed Ali Bahaj, and Tanzila Saba. Mathvision: An accessible intelligent agent for visually impaired people to understand mathematical equations. *IEEE Access*, 13:6155–6165, 2025. doi: 10.1109/ACCESS.2024.3514079. URL https://doi.org/10.1109/ACCESS.2024.3514079.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: visual on-policy self-distillation with modality balance. *CoRR*, abs/2608.05131, 2026. doi: 10.48550/ARXIV.2608.05131. URL https://doi.org/10.48550/arXiv.2608.05131.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, Haoli Bai, and Xiaohui Li. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *CoRR*, abs/2606.08719, 2026. doi: 10.48550/ARXIV.2606.08719. URL https://doi.org/10.48550/arXiv.2606.08719.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9586–9596, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? dissecting on-policy self-distillation. *CoRR*, abs/2608.09228, 2026. doi: 10.48550/ARXIV.2608.09228. URL https://doi.org/10.48550/arXiv.2608.09228.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *CoRR*, abs/2606.18974, 2026a. doi: 10.48550/ARXIV.2606.18974. URL https://doi.org/10.48550/arXiv.2606.18974.

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. *CoRR*, abs/2608.14144, 2026b. doi: 10.48550/ARXIV.2608.14144. URL https://doi.org/10.48550/arXiv.2608.14144.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *CoRR*, abs/2607.21556, 2026. doi: 10.48550/ARXIV.2607.21556. URL https://doi.org/10.48550/arXiv.2607.21556.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1): 145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Haocheng Luo, Jiahui Liu, Ruicheng Zhang, Zhizhou Zhong, Jiaqi Huang, Zunnan Xu, Quan Shi, Jun Zhou, and Xiu Li. Learning visual spatial planning from symbolic state via modality-gap-aware self-distillation. *CoRR*, abs/2606.06076, 2026. doi: 10.48550/ARXIV.2606.06076. URL https://doi.org/10.48550/arXiv.2606.06076.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao

Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 20023–20070. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.983. URL https://doi.org/10.18653/v1/2025.acl-long.983.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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. *arXiv preprint arXiv: 2409.19256*, 2024.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation. *CoRR*, abs/2608.18271, 2026. doi: 10.48550/ARXIV.2608.18271. URL https://doi.org/10.48550/arXiv.2608.18271.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation. *CoRR*, abs/2606.05718, 2026. doi: 10.48550/ARXIV.2606.05718. URL https://doi.org/10.48550/arXiv.2606.05718.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

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, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9568–9578. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.00914. URL https://doi.org/10.1109/CVPR52733.2024.00914.

Kohei Uehara and Tatsuya Harada. K-VQG: knowledge-aware visual question generation for common-sense acquisition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 4390–4398. IEEE, 2023. doi: 10.1109/WACV56688.2023.00438. URL https://doi.org/10.1109/WACV56688.2023.00438.

Shuai Wang, Daoan Zhang, Zhe Tang, Hao Cheng, and Jiaheng Wei. Self-boosting vision-language models with noisy student on-policy self-distillation. *CoRR*, abs/2607.23125, 2026a. doi: 10.48550/ARXIV.2607.23125. URL https://doi.org/10.48550/arXiv.2607.23125.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *CoRR*, abs/2606.19120, 2026b. doi: 10.48550/ARXIV.2606.19120. URL https://doi.org/10.48550/arXiv.2606.19120.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 7907–7915. AAAI Press, 2025. doi: 10.1609/AAAI.V39I8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *CoRR*, abs/2602.11858, 2026. doi: 10.48550/ARXIV.2602.11858. URL https://doi.org/10.48550/arXiv.2602.11858.

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, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13084–13094. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01243. URL https://doi.org/10.1109/CVPR52733.2024.01243.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

Shizhe Xiang, Ke An, Wenlong Yu, Yue Liu, Jian Luan, Pei Fu, and Qilong Wang. Teaching the way, not the answer: Privileged tutoring distillation for multimodal policy optimization. *CoRR*, abs/2606.07000, 2026. doi: 10.48550/ARXIV.2606.07000. URL https://doi.org/10.48550/arXiv.2606.07000.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal LLM logical reasoning benchmark in visual contexts. *CoRR*, abs/2407.04973, 2024. doi: 10.48550/ARXIV.2407.04973. URL https://doi.org/10.48550/arXiv.2407.04973.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *CoRR*, abs/2605.18740, 2026. doi: 10.48550/ARXIV.2605.18740. URL https://doi.org/10.48550/arXiv.2605.18740.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.736. URL https://doi.org/10.18653/v1/2025.acl-long.736.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186. Springer, 2024. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

XiuYu Zhang, Wei Chow, Junfeng Fang, Zhenkai Liang, and Tat-Seng Chua. What does privileged information add to on-policy self-distillation? *arXiv preprint arXiv:2609.20612*, 2026a.

Yan Zhang, Daiqing Wu, Huawen Shen, Can Ma, and Yu Zhou. Learn where to click from yourself: On-policy self-distillation for GUI grounding. *CoRR*, abs/2605.00642, 2026b. doi: 10.48550/ARXIV.2605.00642. URL https://doi.org/10.48550/arXiv.2605.00642.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

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. *CoRR*, abs/2601.18734, 2026. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: resolution-privileged on-policy self-distillation for multimodal large language models. *CoRR*, abs/2607.24447, 2026a. doi: 10.48550/ARXIV.2607.24447. URL https://doi.org/10.48550/arXiv.2607.24447.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *CoRR*, abs/2605.11182, 2026b. doi: 10.48550/ARXIV.2605.11182. URL https://doi.org/10.48550/arXiv.2605.11182.

Table 2: Dataset sources and final retained training and evaluation sample counts.

Task	Data source	Train samples	Eval. samples
FGP	Vision-OPD (Yuan et al., 2026)	5,617	624
KIP	K-VQG tail subset (Uehara & Harada, 2023)	7,466	830
MSR	GQA two-hop spatial-relation subset (Hudson & Manning, 2019)	11,058	1,593
GeoR	Qualified Math-VR subset (Duan et al., 2026)	2,426	274

Table 3: Privileged-information forms and task-specific conditions.

PI form	Modality	Task	PI condition
Answer	Textual	FGP / KIP / MSR / GeoR	Ground-truth answer.
Analysis	Textual	GeoR	Reference analysis and solution.
Trajectory	Textual	FGP / GeoR	Offline rollout trajectories generated by the base model.
Experience	Textual	FGP / GeoR	Multimodal LLM-generated experience summaries from trajectories.
Hint	Visual	FGP	A crop centered on the question-relevant evidence region.
		KIP	Target-object crop from the specific bounding box.
		MSR	Original image annotated with boxes around objects in the queried spatial-relation chain.
		GeoR	A task-specific diagram with auxiliary constructions.

A Experimental Setup Details

A.1 Task and Dataset Details

Table 2 summarizes the data sources and final retained training and held-out evaluation counts for FGP, KIP, MSR, and GeoR after construction and filtering. All underlying data sources are publicly available datasets. Accuracy is reported on the corresponding evaluation set.

A.2 Privileged Information Design

PI forms, modalities, and conditions. Table 3 lists the PI forms, modalities and task-specific conditions used in our experiments.

PI delivery format. Student rollouts always use the original prompt and image without PI. For textual PI, preprocessing copies the prompt to a teacher-only `teacher_prompt` field and appends the rendered signal to its last user message. Answer supplies the ground-truth answer, while Analysis, Trajectory, and Experience supply their corresponding per-example fields. Each teacher prompt begin with a condition-specific introduction, followed by the PI text and an instruction for the teacher to solve the problem using its own approach. The student rollout and policy update retain the unmodified prompt. For the visual PI form Hint, the teacher replaces each image input in the original prompt with its corresponding Hint image. The student continues to use the original image throughout rollout, optimization and evaluation.

PI examples and construction details. Figure 6 and Table 4 illustrate one example from each task. For construction, the Answer condition uses the dataset ground-truth answer, while Analysis uses the reference analysis or solution provided by the source dataset. **Trajectory** is built offline by collecting one rollout from the base model for each source training example. **Experience** is produced offline from the original problem and image, stored trajectory, and its success/failure label. A multimodal LLM returns two short, answer-free fields: successful rollouts yield *Success Mechanism* and *Reusable Heuristic*; unsuccessful rollouts yield *Failure Mechanism* and *Corrective Heuristic*. In abbreviated form, the synthesis prompt is: *given the problem, image, trajectory, and sample status, identify the specific reasoning or visual-check mechanism, then state one executable rule for a similar problem; do not provide a worked solution or an answer option.* The visual **Hint**

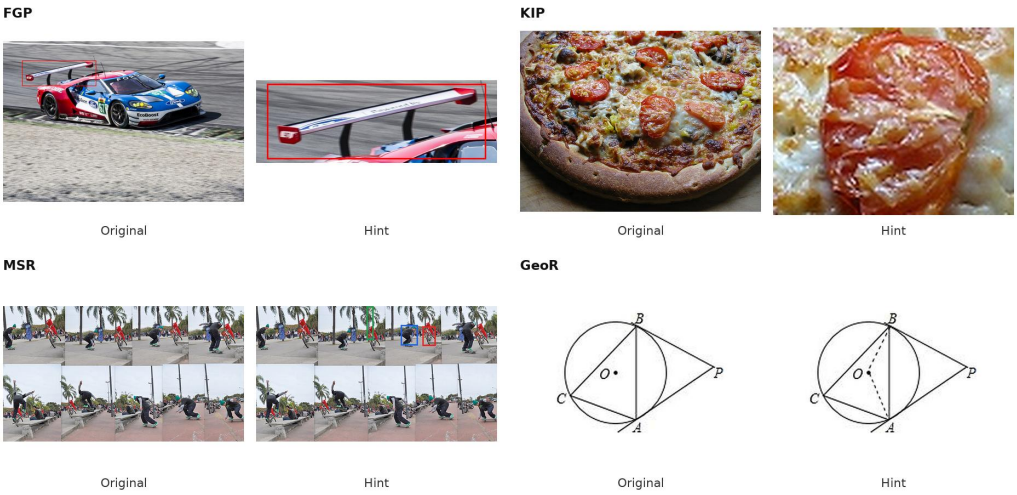


Figure 6: Visual PI examples. In each pair, the left image is the original problem image and the right image is the corresponding visual PI.

Table 4: Textual PI examples.

PI form	Teacher-side content
FGP	
	Question: What text is written on the upper surface of the car's rear wing? (A) Ford GT; (B) EcoBoost; (C) Powered by; (D) Castrol EDGE.
Answer	The ground-truth answer is: C. Powered by.
Trajectory	`[Example response] C. Powered by [Ground truth] C.`
Experience	*Success Mechanism:* The response focused exclusively on the text within the red bounding box. *Reusable Heuristic:* Isolate the target region and cross-reference its text with the answer options.
KIP	
	Question: What is the red fruit that is on top of the pizza called?
Answer	The ground-truth answer is: fruit.
MSR	
	Question: Are there any palm trees to the left of the person who is to the left of the bike?
Answer	The ground-truth answer is: yes.
GeoR	
	Question: PA and PB are tangents to circle $\odot O$ at A and B. For a point C on the circle, if $\angle P = 50^\circ$, what is $\angle ACB$? (A) 40°; (B) 50°; (C) 65°; (D) 130°.
Answer	The ground-truth answer is: C. 65°.
Analysis	Connect OA and OB. Since PA and PB are tangents, $\angle OAP = \angle OBP = 90^\circ$, hence $\angle AOB = 130^\circ$. The inscribed-angle theorem gives $\angle ACB = \frac{1}{2}\angle AOB = 65^\circ$; therefore the answer is C.
Trajectory (excerpt)	`[Example response]` PA and PB are tangents, so the central angle is 130°. *[...]* The inscribed angle is half the intercepted arc. *[...]* `[Final Answer]` C. `[Ground truth] C.`
Experience	*Success Mechanism:* Link the tangent-angle relation to the central angle, then use the placement of C to identify the intercepted arc. *Reusable Heuristic:* For two tangents, compute the minor central angle as $180^\circ - \angle P$, identify the arc opposite the inscribed-angle vertex, then halve it.

is constructed in a task-specific manner. For FGP, the source dataset provides the corresponding Hint crop. For KIP, we crop the question-specific target-object box from the original image using the bbox coordinate field in the source data. For MSR, we draw colored boxes on the original image around the two objects in the strict two-hop spatial-relation chain, using their GQA scene-graph boxes. For GeoR, we filter the source data to retain geometry problems solvable in two steps and use intermediate-step diagrams with clearly drawn auxiliary lines as HINT images.

A.3 Training Details

Training framework. We implement all OPSD experiments in verl (Sheng et al., 2024), starting from Qwen3.5-4B. PI-conditioned runs use an always-available teacher and optimize the OPSD loss. Unless stated otherwise, all PI conditions share the same optimization and rollout settings.

Teacher–student construction. The student policy is π_θ and the teacher policy $\pi_{\bar{\theta}}$ is its exponential-moving-average (EMA) copy. After each student update, the teacher parameters are updated as

$$\bar{\theta} \leftarrow \beta\,\bar{\theta} + (1-\beta)\,\theta,$$

where $\beta = 0.95$ (equivalently, the student weight is 0.05). The student receives only the original input, while the teacher additionally receives the PI for the current training condition.

Distillation objective. We optimize the student on prefixes sampled from its current policy by matching the corresponding teacher distribution with Jensen–Shannon divergence (JSD). For teacher and student next-token distributions p_T and p_S, we use

$$\mathrm{JSD}_\alpha(p_T, p_S) = \alpha\, D_{\mathrm{KL}}(p_T \| m) + (1-\alpha)\, D_{\mathrm{KL}}(p_S \| m),$$

where

$$m = \alpha p_T + (1-\alpha) p_S.$$

All experiments use $\alpha = 0.5$, yielding symmetric JSD (Lin, 1991). In the implementation, we approximate the full-vocabulary objective on a 17-way support consisting of the student's top-16 token indices and one additional tail bucket containing the remaining vocabulary mass from each distribution. The resulting 17-way distributions retain their original probability masses rather than renormalizing the top-16 entries alone. We compute the symmetric JSD on this support. Wider top-k values are used only for diagnostics. The per-token divergence is evaluated only at positions selected by the generated-response attention mask. Prompt tokens and right-padding therefore receive zero weight, and no additional EOS or special-token filtering is applied. We compute the `token-mean` over all valid response tokens in the global batch. Each token is weighted by the clipped ratio

$$\min\left\{ \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\mathrm{old}}(y_t \mid x, y_{<t})}, 2.0 \right\}$$

and by the configured token-level rollout importance weight, which is also capped at 2.0.

Optimization and batching. We optimize the actor with AdamW, using a learning rate of 2×10^{-6}, weight decay 0.01, and Adam coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning-rate scheduler remains constant after a 10-step linear warmup, and the gradient norm is clipped at 1.0. Each update collects one rollout per prompt for a batch of 48 prompts and uses an optimization mini-batch of 48 prompts. Dynamic token packing limits each actor micro-batch to 9,216 tokens per GPU, so the numbers of micro-batches and gradient-accumulation steps vary with the realized prompt and response lengths. The training stream preserves its stored order without dataset shuffling. Matched No-PI and PI-conditioned runs use the same batching configuration.

Rollout configuration. Student trajectories are generated on-policy with vLLM, using one rollout per prompt. The maximum prompt length is 8192 tokens. The maximum response length is 1024 tokens for FGP, KIP, and MSR, and 4096 tokens for GeoR to accommodate longer geometric reasoning traces. Rollout generation uses a tensor-parallel size of 1 and a GPU memory utilization target of 0.35. We sample with temperature 1.0, top-$p = 1.0$, and top-$k = -1$.

Compute and reproducibility. All experiments use four NVIDIA A100 80GB GPUs, bfloat16 precision, and the fixed random seed 42.

A.4 Why a Matched No-PI OPSD Control Is Necessary

An improvement over the initial backbone does not, by itself, establish that PI made an incremental contribution. Let S_c denote the student trained under PI condition c, $S_{\text{No-PI}}$ the matched No-PI student, and f_{base} the initial backbone. The gain over the backbone decomposes as

$$\mathrm{Perf}_\tau(S_c) - \mathrm{Perf}_\tau(f_{\text{base}}) = \underbrace{\mathrm{Perf}_\tau(S_{\text{No-PI}}) - \mathrm{Perf}_\tau(f_{\text{base}})}_{\text{gain from No-PI OPSD}} + \underbrace{\mathrm{Perf}_\tau(S_c) - \mathrm{Perf}_\tau(S_{\text{No-PI}})}_{\text{incremental effect of PI}}.$$

Table 5: Accuracy (%) of Qwen3.5-4B and step-100 FGP-trained OPSD checkpoints. FGP denotes the in-domain evaluation set. Fine-grained perception* denotes benchmarks distinct from those used in the controlled FGP setting.

Model	In-domain	Fine-grained Perception*					Real-world Understanding			Multimodal Generalization		
	FGP	VStar	ZoomBench	HR-Bench-4K	HR-Bench-8K	MMVP	MME-R-Lite	RealWorldQA	CV-Bench	MMMU	MMMU-Pro	MMStar
Reference models												
Backbone	47.92	80.63	48.05	81.50	78.88	75.67	45.70	74.12	86.81	66.78	55.78	72.73
No-PI OPSD	45.19	83.77	49.23	80.25	76.50	76.33	54.61	75.82	86.47	66.33	55.03	72.13
PI-conditioned OPSD												
Answer	55.93	74.87	54.91	70.25	69.75	70.33	53.57	75.29	86.69	57.89	46.59	68.87
Trajectory	54.17	83.77	49.70	75.12	73.38	80.33	52.32	73.20	84.80	61.44	49.02	65.67
Experience	47.44	62.30	45.80	52.25	48.50	75.00	46.27	73.86	85.86	56.78	46.59	69.20
Hint	65.06	90.05	57.63	77.38	72.38	77.67	55.13	74.38	85.87	66.33	54.34	69.60

The second term is the PI effectiveness defined in §2. A direct comparison of S_c with the backbone would conflate PI effectiveness with the gain from OPSD itself.

No-PI is not an inert condition: the EMA teacher still supplies dense targets at student-visited prefixes and can differ from the current student even without teacher-side PI. OPSD can therefore change the student's future rollouts without transferring target-specific information. Reference-free OPSD accounts for a substantial share of the gain over the base model in a related reasoning setting (Zhang et al., 2026a). Other analyses find that the paired correct reference is not necessary and that a context from another example can remain competitive with standard OPSD (Shrestha & Tessier, 2026; Ichihara et al., 2026).

Each matched No-PI run uses the same initialization, task data, student rollouts, EMA-teacher construction, optimizer, and training budget as its PI-conditioned counterpart; it removes only teacher-side PI. The difference between the two students therefore isolates the incremental effect of the privileged condition.

A.5 Evaluation Conventions

PI access gain compares the same frozen base model on the same evaluation set with and without task-specific PI. It is distinct from EMA-teacher measurements. External-benchmark tables evaluate the post-training student on the corresponding benchmark. For FGP-trained students, the external benchmarks cover out-of-distribution fine-grained perception (VStar (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench-4K, HR-Bench-8K (Wang et al., 2025), and MMVP (Tong et al., 2024b)), real-world understanding (MME-RealWorld-Lite (Zhang et al., 2025) and RealWorldQA (xAI, 2024)), and multimodal generalization (CV-Bench (Tong et al., 2024a), MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), and MMStar (Chen et al., 2024)). GeoR-trained students are additionally evaluated on mathematical reasoning (MathVista (Lu et al., 2024), MathVision (Ahmad et al., 2025), MathVerse (Zhang et al., 2024), and WeMath (Qiao et al., 2025)) and multimodal generalization (LogicVista (Xiao et al., 2024), MMMU, MMMU-Pro, and MMStar). PI effectiveness compares each PI-conditioned student with the matched No-PI checkpoint at the same training step.

B Additional Experimental Results

FGP and GeoR. Tables 5, 6, and 7 report additional results for models trained on FGP and GeoR. The tables include performance on the corresponding in-domain evaluation sets and on external benchmarks.

KIP and MSR. Table 8 reports results for models trained on KIP and MSR. It includes in-domain performance at training steps 100 and 200, with PI directly available to the teacher and for the resulting PI-free students.

C Student-Side Analysis Details

This section provides implementation details for the visual-overlap and matched-occlusion analyses introduced in Section 4.

Table 6: Accuracy (%) of Qwen3.5-4B and step-200 FGP-trained OPSD checkpoints. FGP denotes the in-domain evaluation set. Fine-grained perception* denotes benchmarks distinct from those used in the controlled FGP setting.

Model	In-domain	Fine-grained Perception*					Real-world Understanding			Multimodal Generalization		
	FGP	VStar	ZoomBench	HR-Bench-4K	HR-Bench-8K	MMVP	MME-R-Lite	RealWorldQA	CV-Bench	MMMU	MMMU-Pro	MMStar
Reference models												
Backbone	47.92	80.63	48.05	81.50	78.88	75.67	45.70	74.12	86.81	66.78	55.78	72.73
No-PI OPSD	46.15	80.10	47.81	81.50	76.12	76.33	56.70	75.42	87.07	64.00	55.43	71.87
PI-conditioned OPSD												
Answer	61.06	78.53	53.49	67.75	66.50	70.67	52.27	72.55	86.61	57.44	46.18	66.60
Trajectory	56.57	84.82	49.59	76.75	70.25	72.33	54.30	73.07	84.21	55.56	45.03	63.53
Experience	46.63	55.50	43.08	49.50	46.12	78.00	48.31	69.67	83.51	53.56	43.99	64.67
Hint	63.94	86.91	58.82	75.00	73.25	77.00	53.00	72.81	82.94	64.33	51.91	67.80

Table 7: Accuracy (%) of Qwen3.5-4B and step-200 GeoR-trained OPSD checkpoints. GeoR denotes the in-domain evaluation set.

Model	In-domain	Mathematical Reasoning				Multimodal Generalization			
	GeoR	MathVista	MathVision	MathVerse	WeMath (Strict/Loose)	LogicVista	MMMU	MMMU-Pro	MMStar
Reference models									
Backbone	83.58	81.90	54.31	70.76	51.90 / 75.71	63.39	66.78	55.78	72.73
PI-conditioned OPSD									
Answer	58.76	79.80	46.35	65.48	47.43 / 70.86	53.79	51.22	45.32	65.53
Analysis	61.90	78.70	32.84	58.30	32.10 / 54.57	52.71	54.84	45.70	66.47
Trajectory	80.66	77.80	39.08	60.89	40.00 / 61.14	56.47	59.89	49.42	67.80
Experience	75.91	78.50	46.09	63.20	39.81 / 64.00	53.12	54.56	45.95	66.33
Hint	79.56	73.30	37.83	47.59	30.76 / 54.57	58.71	62.67	50.98	66.33

Instance-level illustration. Figure 7 compares residual-attention rollout maps for the initial model and the trained variants on an FGP example. The quantitative visual-overlap analysis measures attribution to the target region across the evaluation set.

Visual overlap. For visual overlap, we replay each model's free-generation response on its original image and prompt at the corresponding checkpoint. The target box is converted to a binary visual-token mask after multimodal preprocessing, using the processor's image grid and spatial merge size. We evaluate two query-token positions. *Before generation* is the last prompt token, which predicts the first output token. *Before the final answer* is the token immediately preceding the final answer in the generated response. For this position, we append the preceding response prefix under teacher forcing. For each query position, we extract the attention matrices from eight layers of the 32-layer Qwen3.5-4B decoder, using layer indices $\{3, 7, 11, 15, 19, 23, 27, 31\}$. We average the attention weights across heads, add an identity residual, normalize each row, and multiply the matrices in model order to obtain the residual attention rollout from the query token to the visual tokens. We then normalize the resulting attribution to sum to one and compute the target-box mass and the area-corrected visual-overlap score V_{overlap}. Then we use the correctness label to partition trajectories. For each model, PI condition, and probe position, we use V_{overlap} as a continuous score and the correctness label as a binary outcome to compute the AUROC.

Matched occlusion intervention. The analysis uses the step-100 checkpoints. All interventions apply Gaussian blur with radius 12. Target-box blur replaces the region inside the target box with the corresponding region from a fully blurred copy. Matched-background blur applies the same operation to a non-overlapping rectangle with the target box's width and height. Full-image blur applies the operation to the entire image.

D Additional Training Diagnostics

These diagnostics characterize teacher–student agreement during OPSD.

Table 8: In-domain accuracy (%) on KIP and MSR under different privileged-information conditions. Teacher performance with PI and the corresponding PI-free student evaluations are reported.

Task	Condition	Teacher with PI	Backbone	PI-free step-100	PI-free step-200
KIP	Backbone	–	5.06	–	–
KIP	No-PI OPSD	–	–	5.06	4.94
KIP	Answer	66.99	–	8.92	8.07
KIP	Hint	5.54	–	5.42	5.66
MSR	Backbone	–	54.93	–	–
MSR	No-PI OPSD	–	–	56.31	54.99
MSR	Answer	90.46	–	58.38	52.35
MSR	Hint	58.25	–	54.80	50.60



Figure 7: Instance-level residual-attention rollout in fine-grained perception. The input panel marks the target region in red; the remaining panels show rollout attribution for the initial and trained variants.

D.1 Diagnostic Definitions

Top-k overlap. For teacher and student next-token distributions p_T^s and p_S^s at a student-visited state s, we compute

$$\text{Overlap@}\, k = \frac{|\text{Top}_k(p_T^s) \cap \text{Top}_k(p_S^s)|}{k}.$$

Higher overlap indicates greater agreement between the teacher's and student's high-probability token sets.

Jensen–Shannon divergence. We compute the Jensen–Shannon divergence between the teacher and student next-token distributions,

$$\text{JSD}(p_T^s, p_S^s) = \frac{1}{2} D_{\text{KL}}(p_T^s \| m^s) + \frac{1}{2} D_{\text{KL}}(p_S^s \| m^s), \qquad m^s = \frac{p_T^s + p_S^s}{2}.$$

Lower JSD indicates closer agreement between the full predictive distributions.

Absolute entropy gap. We additionally compare teacher and student predictive uncertainty using

$$\Delta H(s) = |H(p_T^s) - H(p_S^s)|\,.$$

A smaller entropy gap indicates more similar teacher and student predictive uncertainty.

Aggregation protocol. Training-time panels plot the metric logged at each optimizer step. A valid position is a student-visited response position selected by the response mask and, when applicable, the sample-level self-distillation mask; prompt tokens and right-padding are excluded. All-position overlap and entropy panels, as well as token-mean JSD, are masked means over these positions, so longer responses contribute proportionally more token states. For sequence-mean JSD, we first average valid positions within each rollout and then take the arithmetic mean across rollouts with at least one valid position. The Q1–Q4 panels partition each rollout by its realized response length before applying the same masked mean within each partition. These position-resolved views test whether this agreement is concentrated near the beginning or end of a response.

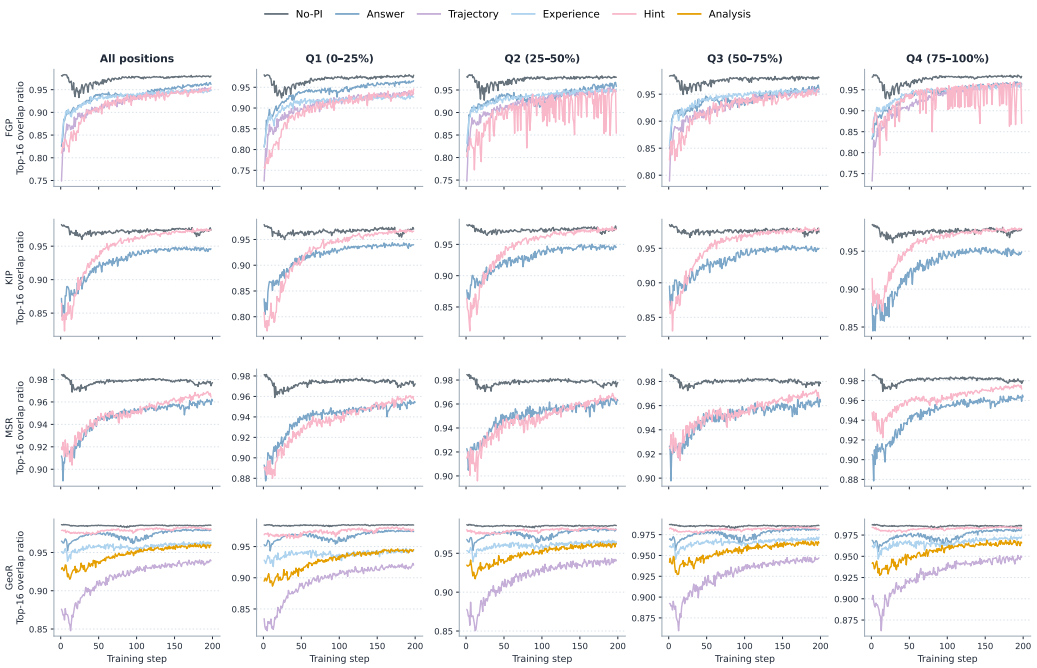


Figure 8: Training-time teacher–student Top-16 overlap across FGP, KIP, MSR, and GeoR.

D.2 Training Dynamics

We track teacher–student agreement throughout the individual FGP, KIP, MSR, and GeoR runs. Figures 8–10 describe agreement between the evolving student and its training-time teacher at token states reached by the student.

D.3 Top-K Robustness

The preceding analysis fixes the overlap metric at $K = 16$. To assess sensitivity to this choice, we evaluate support sizes $K \in \{1, 4, 16, 64, 128\}$ for each task and PI condition. At every training step, the height of a surface is $\mathrm{Overlap@}\,K - \mathrm{Overlap@}\,16$, computed over valid response-token states. The $K = 16$ slice is consequently zero. Values close to zero indicate that the Top-16 view is representative of agreement at the corresponding support size. Departures from zero show how the measured agreement changes with K. Figures 11, 12, 13, and 14 show the resulting Top-K robustness profiles for FGP, KIP, MSR, and GeoR, respectively.

D.4 Task-Level PI Routing Diagnostics

Mixed-task routing follows the validation-based selection protocol in Section 5.3. Building on the diagnostics defined above, Figures 15, 16, 17, and 18 report training-time teacher–student diagnostics for the uniform No-PI, Answer, and Hint runs, together with the task-routed Selected run. All routing panels show steps 0–100.

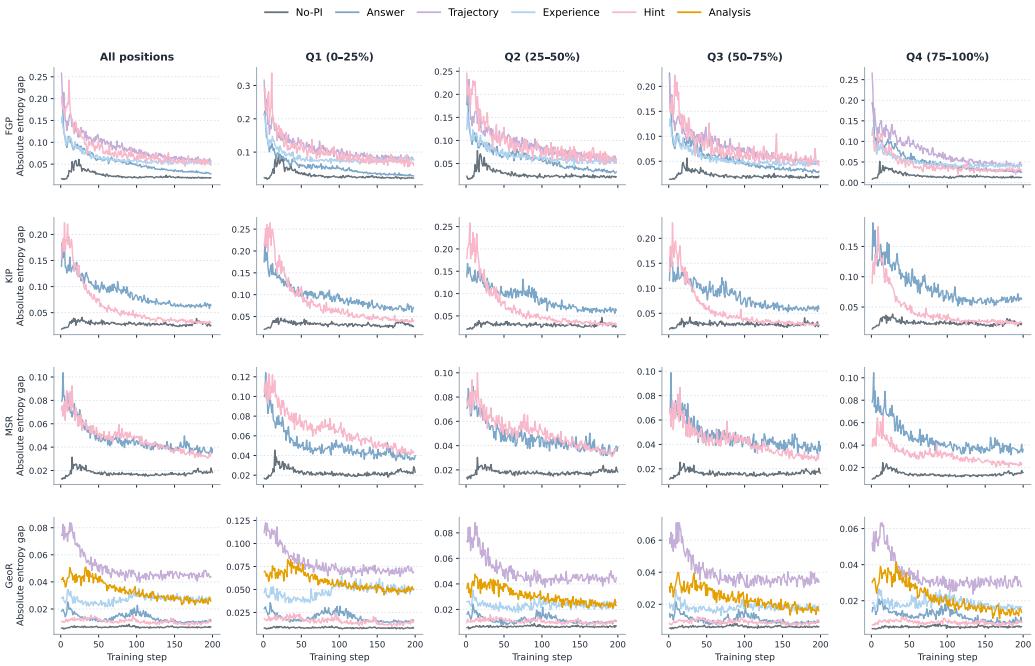


Figure 9: Training-time absolute entropy gap across FGP, KIP, MSR, and GeoR.

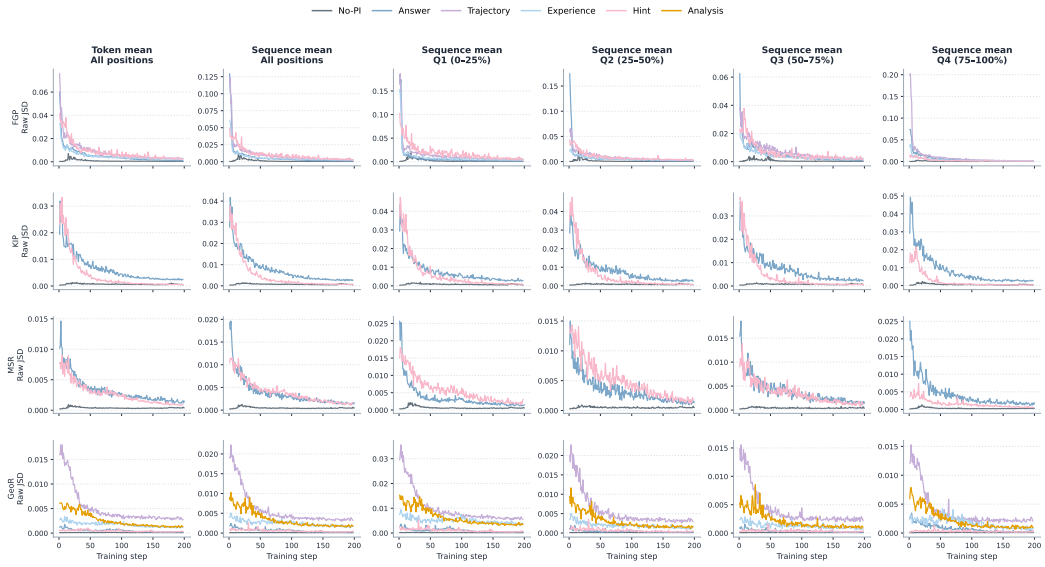


Figure 10: Training-time raw JSD across FGP, KIP, MSR, and GeoR. The first column shows token-level JSD and the remaining columns show sequence-mean Top-16-plus-tail JSD. “Raw” omits importance-sampling and rollout-correction weights.

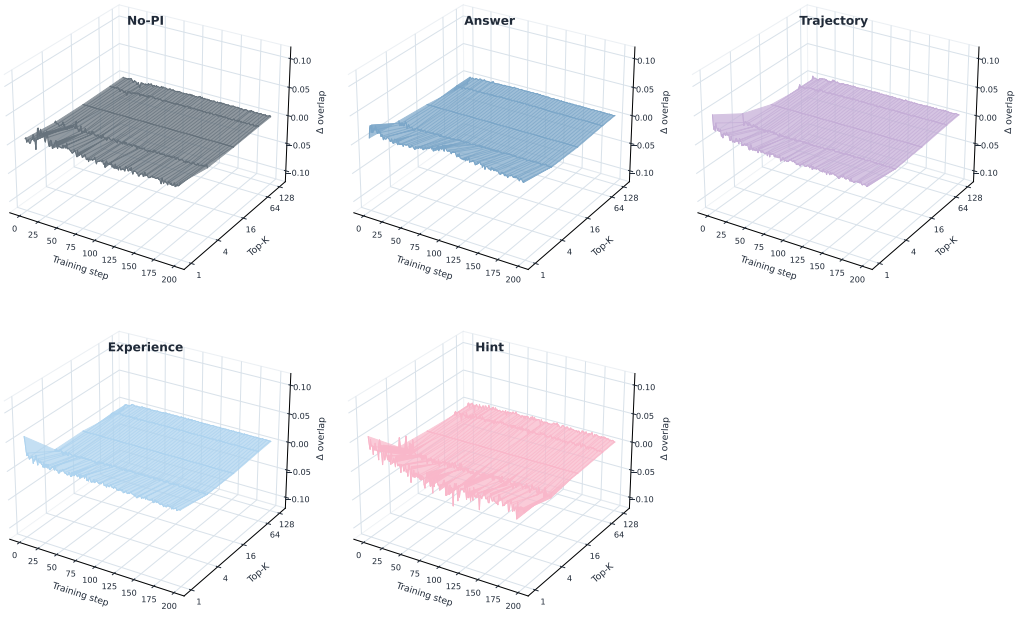


Figure 11: Top-K overlap robustness for FGP.

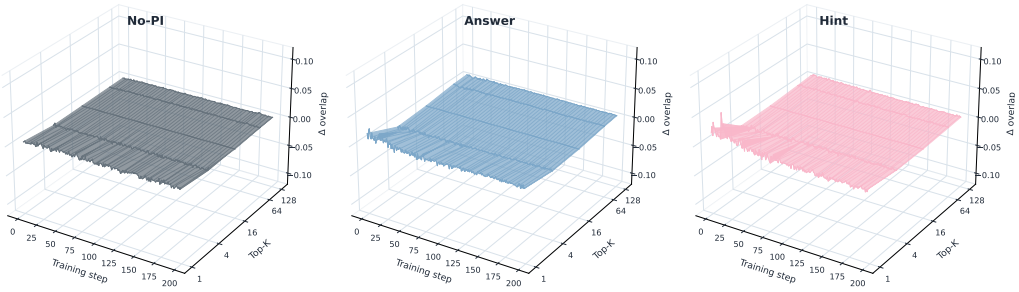


Figure 12: Top-K overlap robustness for KIP.

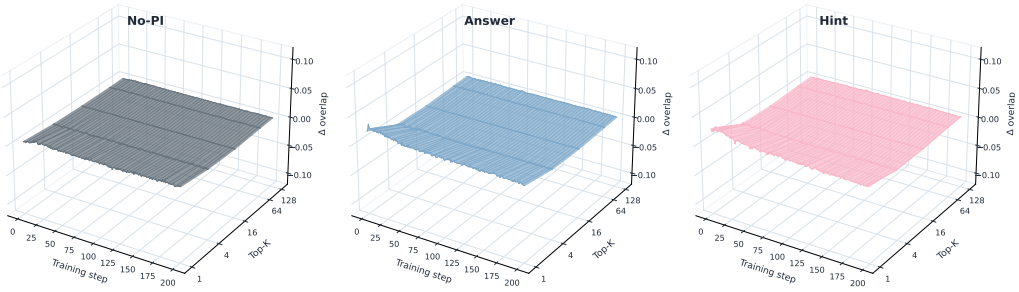


Figure 13: Top-K overlap robustness for MSR.

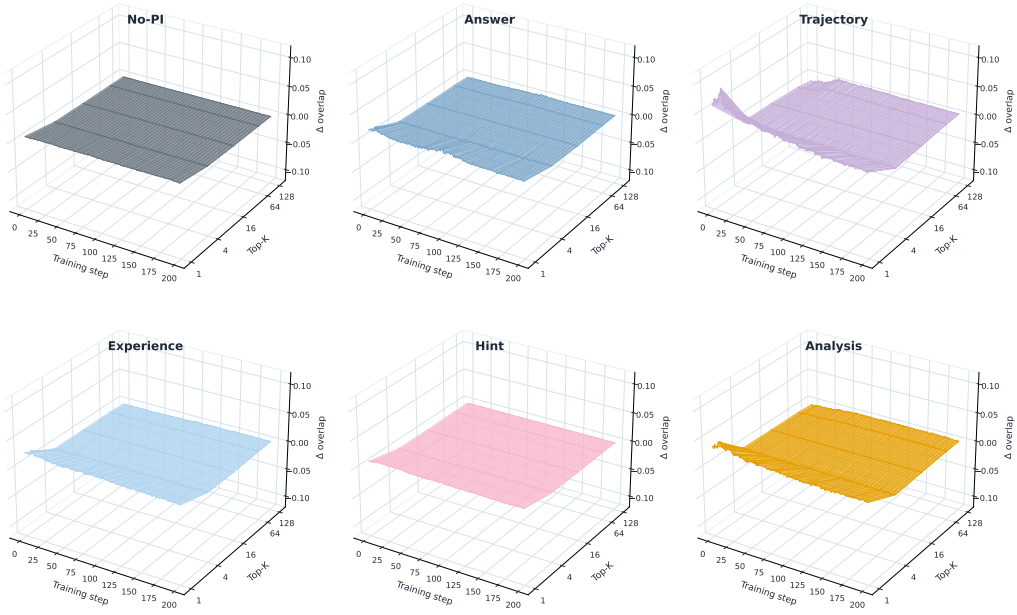


Figure 14: Top-K overlap robustness for GeoR.



Figure 15: Training-time Top-16 overlap for the mixed-task routing runs.



Figure 16: Training-time absolute entropy gap for the mixed-task routing runs.

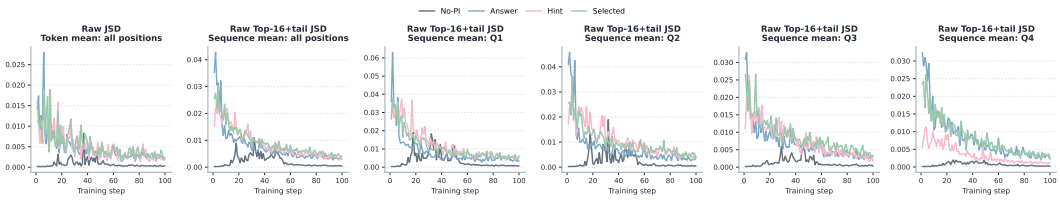


Figure 17: Training-time raw JSD for the mixed-task routing runs. The first panel shows token-level JSD and the remaining panels show sequence-mean Top-16-plus-tail JSD. "Raw" omits importance-sampling and rollout-correction weights.

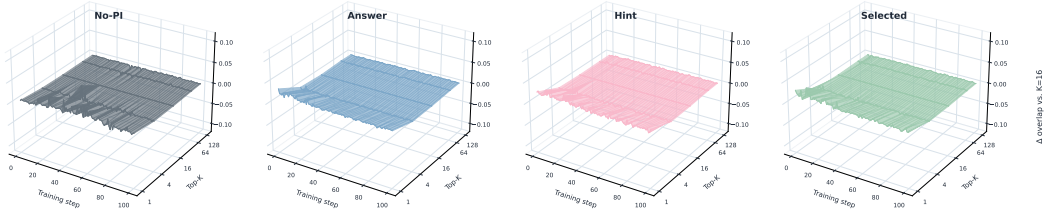


Figure 18: Top-K robustness of teacher–student overlap in the mixed-task routing runs.