WHERE TO APPLY PRIVILEGE: ON-POLICY SELF-DISTILLATION WITH TEXTUAL PRIVILEGE FOR MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) has emerged as an effective post-training paradigm for multimodal large language models (MLLMs) that enhances visual understanding without external teacher models or reward verifiers. While vanilla OPSD adopts reference answers as textual privilege, recent variants for visual understanding primarily rely on visual privilege. However, we observe that MLLMs leverage answer-supporting evidence more effectively from targeted textual descriptions than from the original images, even when both inputs are individually sufficient for deriving the answer. This suggests substantial potential for textual privilege in self-distillation. Nevertheless, naive use of textual privilege can introduce context-specific artifacts, such as "based on the description", even when no description is available to the student. This behavior reflects a failure to disentangle genuine improvements from traces of privileged text, which we term the *privileged-context attribution problem*. To address this problem, we propose **Role-OPD**. It adopts caption-based textual privilege to enhance the model's visual understanding. Furthermore, instead of presenting the privileged caption as user-provided context, Role-OPD places it in an intermediate assistant turn, framing it as the teacher's own visual understanding. This role-aware design suppresses the transfer of context-specific artifacts while preserving the genuine improvements from privileged captions. Experiments on six fine-grained visual understanding benchmarks show that Role-OPD achieves consistent gains and outperforms existing post-training methods.

1 INTRODUCTION

Multimodal large language models (MLLMs) support a broad range of visual understanding tasks Zhao et al. (2026b); Ma et al. (2026); Zhang et al. (2023); Huang et al. (2025), yet they still struggle to effectively leverage fine-grained visual evidence for complex reasoning. Recently, on-policy self-distillation (OPSD) has emerged as a promising post-training paradigm for improving visual understanding without external teacher models or reward verifiers. In this paradigm, the student policy samples trajectories, while a teacher that shares the same architecture and initialization leverages privileged context to provide dense token-level targets for student-visited prefixes Agarwal et al. (2024); Lu (2025); Zhao et al. (2026a).

While OPSD adopts reference answers as textual privilege Zhao et al. (2026a), recent variants aimed at enhancing the visual understanding of MLLMs primarily rely on visual privilege, such as evidence-centered image crops Yuan et al. (2026), leaving fine-grained textual privilege largely underexplored. However, we observe that MLLMs leverage answer-supporting evidence more effectively from targeted textual descriptions than from the original images. In controlled comparisons, the same model consistently achieves higher task accuracy and approaches its empirical performance ceiling when provided with textual evidence, even when both inputs are individually sufficient for deriving the answer. This reveals a clear evidence-utilization gap between original images and targeted textual descriptions and suggests the substantial potential of textual privilege for self-distillation.

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

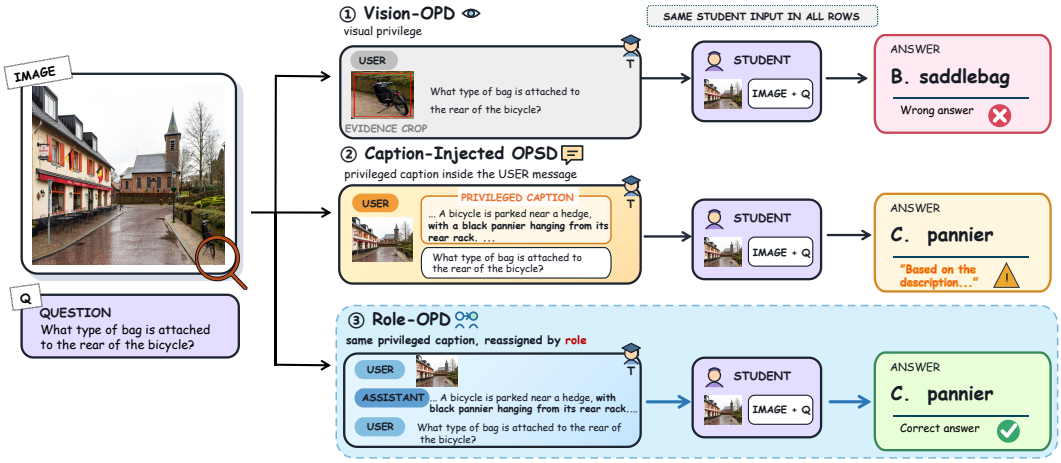

Figure 1: **Comparison of privilege types and placement strategies.** All responses are generated by the student given only the original image and question. ① When distilled with Vision-OPD Yuan et al. (2026), which adopts evidence-centered image crops as visual privilege, the student fails to yield correct answers. ② Distilling with the privileged caption that is naively incorporated as user-provided context enables the student to answer correctly but introduces a context-specific artifact (e.g., "based on the description"). ③ **Role-OPD** presents the same privileged caption as the teacher's own visual understanding in an intermediate assistant turn. This role-aware design enables the student to output the correct answers cleanly without hallucinating unavailable context.

Despite substantial potential, naively incorporating textual privilege as user-provided context introduces a critical flaw: the generation of student-side context-specific artifacts. For example, even when no description of the visual evidence is available to the student model at inference time, it may still generate phrases such as "based on the description". This occurs because feeding textual privilege as user-provided context not only improves the teacher's utilization of answer-supporting evidence but also alters its output distribution, leaving traces such as linguistic cues and context-dependent response patterns. Standard token-level distillation indiscriminately transfers both effects to the student Li et al. (2026). Consequently, the student fails to disentangle genuine improvements from the traces of privileged text, leading to responses that hallucinate unavailable context. We formally term this failure mode the *privileged-context attribution problem*, a crucial bottleneck that severely undermines the efficacy of textual privilege.

To address this problem, we propose **Role-OPD**, an on-policy self-distillation framework that uses caption-based textual privilege and rethinks where to apply privilege. Crucially, instead of presenting the privileged caption as user-provided context, Role-OPD places it in an intermediate assistant turn, framing it as the teacher's own visual understanding. As shown in Figure 1, this role-aware design delivers more reliable supervision, retaining the genuine improvements from the privileged caption while preventing the student from generating context-specific artifacts. Experiments on six fine-grained visual understanding benchmarks show that Role-OPD achieves consistent gains and outperforms existing post-training methods. Our main contributions are summarized as follows:

- We identify an evidence-utilization gap: MLLMs leverage answer-supporting evidence more effectively from targeted textual descriptions than from the original images. This gap highlights the substantial potential of textual privilege for self-distillation.
- We uncover and formalize the privileged-context attribution problem, showing that naively incorporating textual privilege as user-provided context induces context-specific artifacts in student outputs. To tackle this problem, we propose **Role-OPD**, a role-aware design that presents the privileged caption as the teacher's own visual understanding in an intermediate assistant turn, effectively preventing artifacts while preserving genuine improvements.
- Comprehensive experiments across six fine-grained visual understanding benchmarks demonstrate that Role-OPD consistently outperforms existing post-training methods. We will open-source our code and models to advance visual understanding research in MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) combine large language models (LLMs) with visual encoders through modality connectors, enabling the unified processing of visual and textual inputs Alayrac et al. (2022); Li et al. (2023a); Liu et al. (2023); Bai et al. (2023). Large-scale multimodal pretraining and instruction tuning further equip these models with broad capabilities in visual understanding and reasoning Liu et al. (2023); Dai et al. (2023). Recent MLLMs, including Qwen3-VL Bai et al. (2025), InternVL3.5 Wang et al. (2025a), and GLM-4.5V Hong et al. (2025), have advanced visual question answering, document understanding, and complex multimodal reasoning through improved vision–language alignment. Despite this progress, MLLMs still struggle to effectively utilize fine-grained visual evidence, especially when such details are subtle, localized, or densely presented Wu & Xie (2024); Wang et al. (2025b); Zhang et al. (2025). To preserve these visual details, prior work has explored techniques including high-resolution processing, resolution-adaptive visual tokenization, and multi-resolution feature fusion Liu et al. (2024); Guo et al. (2024); Chen et al. (2024b); Wang et al. (2024); Luo et al. (2025). A complementary line of work interleaves step-by-step reasoning with active visual operations (e.g., dynamic zooming or cropping) to gather answer-supporting evidence on demand. While effective, such iterative visual interaction significantly increases inference overhead and latency Wu & Xie (2024); Zheng et al. (2026); Su et al. (2026); Wang et al. (2026); Zhang et al. (2026). This bottleneck motivates the development of post-training paradigms that internalize fine-grained visual understanding as an inherent capability.

2.2 On-Policy Self-Distillation

Unlike conventional distillation which suffers from distribution shifts on offline rollouts Hinton et al. (2015); Kim & Rush (2016), on-policy distillation queries the teacher directly at student-visited prefixes Agarwal et al. (2024). Building on this, On-policy self-distillation (OPSD) further eliminates the need for an external teacher by sharing the same architecture and initialization between the two roles. Inspired by Learning Under Privileged Information (LUPI) Vapnik & Izmailov (2015), the teacher constructs an advantage using training-time privileged context, such as reference answers Zhao et al. (2026a). Recent multimodal variants explore two primary forms of privilege. One line leverages *reasoning privilege*. For example, OmniOPSD conditions the teacher on rationale trajectories from frontier models Cheng et al. (2026). Another line exploits *visual privilege*, where methods like Vision-OPD Yuan et al. (2026) and Imagine-OPD Cai et al. (2026) supply the teacher with evidence-centered image crops or zoomed views to transfer regional perception to the student. Despite these advances in utilizing reasoning paths and visual privilege, the potential of fine-grained descriptive text remains largely underexplored. However, we identify that naively incorporating textual privilege as user-provided context introduces a critical flaw: the privileged-context attribution problem, which yields context-specific artifacts in student outputs. Role-OPD explores the effective utilization of textual privilege through a role-aware design. It retains genuine improvements in visual understanding while suppressing the transfer of context-specific artifacts to the student.

3 Method

In this section, we first introduce the preliminaries to understand our method in Section 3.1. Then, in Section 3.2, we find that MLLMs leverage answer-supporting evidence more effectively from targeted textual descriptions than from the original images, and further analyze the "privileged-context attribution problem" that arises when textual privileged information is directly incorporated into the teacher model's user prompt. Accordingly, we introduce our **Role-OPD** method in Section 3.3. Role-OPD introduces a role-aware teacher context mechanism that places textual privileged information in the teacher model's assistant turn, preserving the teacher's advantage in evidence utilization while mitigating spurious attribution on the student side.

3.1 Preliminaries

We investigate on-policy self-distillation for MLLMs. Each example consists of an original image x, a question q, and a reference answer a available only during training. The original image x

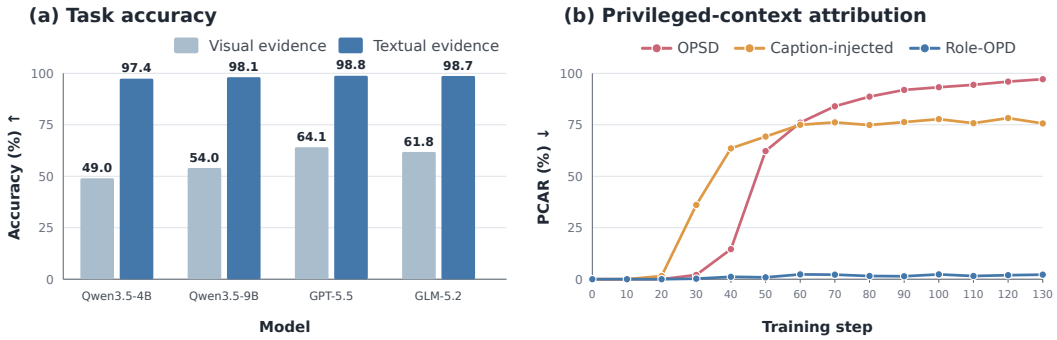

Figure 2: **Evidence utilization with original images and targeted textual descriptions, and the privileged-context attribution problem. (a)** The model achieves substantially higher task accuracy given textual rather than visual evidence, even when both inputs are individually sufficient for deriving the answer. **(b)** OPSD and caption-injected OPSD suffer from the persistent privileged-context attribution problem, whereas **Role-OPD** maintains the Privileged-Context Attribution Rate (PCAR) below 2.34% throughout training, as evaluated by an LLM judge.

contains sufficient answer-supporting visual evidence for question q. The trainable multimodal large language model p_θ receives only the original image x and question q, and samples an on-policy trajectory:

$$y = (y_1, \ldots, y_T) \sim p_\theta(\cdot \mid x, q). \tag{1}$$

Let $y_{<t} = (y_1, \ldots, y_{t-1})$ denote the student-generated prefix up to position t. The student's predictive distribution at position t is defined as:

$$p_{S,t}(\cdot) = p_\theta(\cdot \mid x, q, y_{<t}). \tag{2}$$

Let $\bar{\theta}$ denote the detached teacher parameters. At token position t, the teacher receives the privilege R and scores the same student prefix:

$$p_{T,t}(\cdot) = p_{\bar{\theta}}\big(\cdot \mid x, q, R, y_{<t}\big). \tag{3}$$

Here, R denotes the privileged reference answer. OPSD distills this privileged teacher distribution into the student model along its on-policy trajectory:

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{y \sim p_\theta(\cdot|x,q)} \left[\frac{1}{T} \sum_{t=1}^{T} D_{dist}\left(p_{T,t} \,\|\, p_{S,t}\right) \right]. \tag{4}$$

Here, D_{dist} denotes the token-level distillation objective. As the prefixes $y_{<t}$ are sampled from the current student policy, supervision remains on-policy, and gradients propagate only through θ.

3.2 Motivation: Placement Matters for Textual Privilege

Textualizing visual evidence for a privileged teacher. Modern multimodal large language models are typically built around a large language model and incorporate visual information through a visual encoder Liu et al. (2023; 2024); Wang et al. (2025a); OpenAI (2026; 2025); Bai et al. (2025); Google (2026b;a). Given this language-model-centric architecture, models may utilize visual and textual evidence differently. To compare their ability to leverage these two forms of evidence under controlled conditions, we construct a textual-description counterpart for each image–question pair, where both modalities alone support correct answers. Specifically, given an image x, question q, and reference answer a, we prompt a strong MLLM with the triplet (x, q, a) to generate a caption c that explicitly verbalizes the visual evidence supporting a. Conditioning this generation on a guarantees that the correct answer can be fully deduced from the caption c alone. Consequently, we deem x and c to be answer-sufficient with respect to q: both provide sufficient information to derive a, even though x inherently encompasses substantially more task-irrelevant signals. Following manual inspection to ensure caption fidelity, this process yields our dataset $\mathcal{D} = \{(x_i, q_i, a_i, c_i)\}_{i=1}^{N}$. The full prompt can be found in Appendix C.

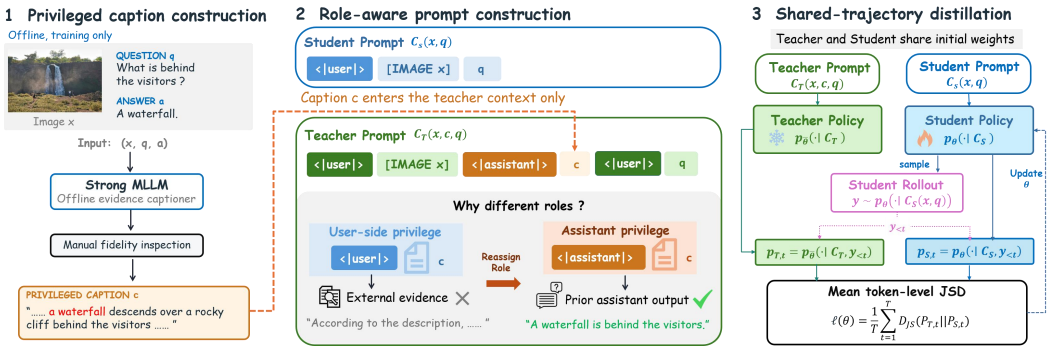

Figure 3: **Overview of the Role-OPD framework. (1) Privileged caption construction**: An offline MLLM generates an answer-supporting caption c conditioned on the image, question, and reference answer (x, q, a). **(2) Role-aware prompt construction**: The privileged caption c is placed in an intermediate assistant turn to form the teacher prompt $C_T(x, c, q)$, while the student prompt $C_S(x, q)$ contains only the image and question. **(3) Shared-trajectory distillation**: Rollouts y sampled from the student policy p_θ provide shared prefixes $y_{<t}$ evaluated by both the frozen teacher $p_{\bar{\theta}}$ and the student p_θ to compute the token-level Jensen–Shannon divergence loss for updating θ.

We construct 6.2K image-caption-question triplets (x, c, q) built upon the source data from Vision-OPD Yuan et al. (2026) as our experimental data to evaluate models' ability to leverage answer-sufficient visual inputs and their targeted textualizations. For each triplet, we evaluate models under two controlled settings. In the **visual-evidence setting**, the model receives the image x and question q. In the **textual-evidence setting**, the image x is replaced by caption c. We conduct this comparison on Qwen3.5-4B, Qwen3.5-9B, GPT-5.5 and GLM-5.2. As shown in Figure 2(a), task accuracy remains below 65% for all models in the visual-evidence setting, whereas it approaches 100% in the textual-evidence setting. This substantial gap indicates that, although both inputs are sufficient to answer the question, models leverage answer-supporting evidence more effectively when it is provided as a targeted textual description than when it must be extracted from the original image.

Privileged-context attribution. A teacher provided with a reference answer or caption typically generates more accurate responses. However, its outputs may include attributions to this privileged source. The student then learns to reproduce these attribution patterns during on-policy distillation despite never observing the source itself. For example, under standard OPSD, the student may state, *"According to the reference solution, the answer is X"*, whereas under caption-injected OPSD, it may claim, *"Based on the description, I infer that ..."*. We term this phenomenon *privileged-context attribution*. This problem arises when the student imitates the teacher's attribution patterns and attributes its response to contextual information that it never received.

We quantify this problem using the *Privileged-Context Attribution Rate* (PCAR), defined as the fraction of model-generated responses that explicitly attribute their conclusions to privileged context not included in the model's input. Given a set of student responses $\mathcal{Y}$, we use an LLM judge J to identify such attributions. Specifically, $J(y) = 1$ if response y explicitly refers to or credits the privileged context as a source of its answer, and $J(y) = 0$ otherwise. We compute PCAR as

$$\mathrm{PCAR}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \mathbb{I}[J(y) = 1]\,. \tag{5}$$

The full prompt for the LLM judge J can be found in Appendix C. To track this problem throughout training, we use our 6.2K examples built upon source data from Vision-OPD as the training dataset and compute PCAR on the student's rollouts. As shown in Figure 2(b), PCAR reaches 97.14% for OPSD and 75.65% for caption-injected OPSD, indicating severe privileged-context attribution in both settings. Notably, relocating the image caption from the user prompt to an intermediate assistant turn frames it as the teacher's own visual understanding. Under this configuration, PCAR remains below 2.34% at all evaluation steps. This sharp contrast suggests that the message role affects whether the teacher explicitly attributes its response to the privileged text. Such attribution patterns may then be transferred to the student. This motivates us to propose **Role-OPD**, which

introduces textualized visual evidence as privileged information into on-policy distillation while avoiding the privileged-context attribution problem.

3.3 ROLE-OPD

Role-aware privilege. Privileged information improves the teacher's answers, but when presented as part of the user input, it reveals its external origin. In particular, when the caption c is supplied directly by the user, the teacher may treat it as user-provided evidence and explicitly reference it in its response. During on-policy distillation, these attribution patterns are transferred to the student even though the caption c is absent from the student's input. Thus, our goal is to retain the distillation benefits of the information in c while eliminating any signal that it originates from an external source.

Role-OPD achieves this by relocating the privileged caption c within the conversational flow. The student receives the standard inference context:

$$C_S(x, q) = \texttt{[user:}\quad x, q\texttt{]}. \tag{6}$$

For the teacher, we insert caption c in an intermediate assistant turn between the image and the question:

$$C_T(x, c, q) = \texttt{[user:}\quad x\texttt{]} \rightarrow \texttt{[assistant:}\quad c\texttt{]} \rightarrow \texttt{[user:}\quad q\texttt{]}. \tag{7}$$

By placing c in an intermediate assistant turn, Role-OPD frames it as the teacher's self-generated visual understanding from x, not as user-supplied evidence. This allows the teacher to leverage the privileged information in c when answering q while reducing explicit attributions to external user-provided context. Role-OPD thus preserves the teacher's informational advantage while blocking the transfer of such attribution artifacts to the student.

Role-guided self-distillation. We optimize the student p_θ via on-policy self-distillation from a frozen teacher $p_{\bar{\theta}}$, with both initialized from the same pretrained checkpoint. For each image-caption-question triple $(x, c, q) \in \mathcal{D}$, the student first samples a completion $y \sim p_\theta(\cdot \mid C_S(x, q))$. At each token position t, the student and teacher condition on the same student-generated prefix $y_{<t}$ but use their own contexts, producing the next-token distributions $p_{S,t} = p_\theta(\cdot \mid C_S(x, q), y_{<t})$ and $p_{T,t} = p_{\bar{\theta}}(\cdot \mid C_T(x, c, q), y_{<t})$, respectively. We then minimize the average token-level Jensen–Shannon divergence between these distributions:

$$\mathcal{L}_{\text{Role-OPD}}(\theta) = \mathbb{E}_{(x,c,q)\sim\mathcal{D}}\, \mathbb{E}_{y\sim p_\theta(\cdot|C_S(x,q))} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\text{JS}}(p_{T,t} \,\|\, p_{S,t}) \right]. \tag{8}$$

Here, the Jensen–Shannon divergence is defined as

$$D_{\text{JS}}(p_{T,t} \,\|\, p_{S,t}) = \beta D_{\text{KL}}(p_{T,t} \,\|\, m) + (1 - \beta) D_{\text{KL}}(p_{S,t} \,\|\, m)\,, \tag{9}$$

where $m = \beta p_{T,t} + (1 - \beta) p_{S,t}$ is the interpolated mixture. By aligning the student with a teacher equipped with textualized visual evidence along student-generated trajectories, Role-OPD helps the model more effectively leverage task-relevant visual semantics, thereby enhancing its visual understanding. The role-aware teacher context mitigates privileged-context attribution artifacts on the student side, while textualized visual evidence is used only during training, leaving inference unchanged and introducing no additional overhead.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Training Settings. We instantiate Role-OPD on Qwen3.5-4B and Qwen3.5-9B Team (2026). Our training set comprises 6.2K image-caption-question triplets (x, c, q), constructed by augmenting the samples from Vision-OPD Yuan et al. (2026) with captions generated automatically by Qwen3.5-397B and validated by human annotators. Both the teacher and student are initialized from the same pretrained checkpoint, with the teacher frozen and only the student optimized throughout training. We perform policy alignment using the Jensen–Shannon divergence (JSD) with $\beta = 0.5$. To overcome the high memory overhead inherent to full-vocabulary distillation, we adopt a top-K distillation strategy. This strategy evaluates only the top $K = 100$ student logits together with

Table 1: Comparison of Role-OPD with state-of-the-art MLLMs and post-training methods across six benchmarks, with mean performance reported. Results for a subset of the baseline models are taken from Vision-OPD. Bold denotes the best result within each matched post-training group.

Model	Param Size	V* Bench	ZoomBench	HR-Bench		MME-RW		Average
				4K	8K	EN	CN	
Closed-Source Models (Single Forward Pass)								
GPT-5.2	–	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4	–	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini-3.1-Pro	–	87.96	61.18	89.63	86.88	76.53	73.31	79.25
Gemini-3.5-Flash	–	89.01	61.42	89.12	86.62	75.31	73.97	79.24
Open-Source Models (Single Forward Pass)								
GLM-4.6V	106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Qwen3-VL-Instruct	235B	91.10	56.09	86.13	80.38	71.74	69.04	75.75
Qwen3.5	397B	87.96	57.16	89.38	85.50	74.82	69.82	77.44
Kimi-K2.6	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Qwen3.5 (4B) Post-Training Methods								
Qwen3.5	4B	81.15	49.59	84.38	80.13	64.21	63.75	70.54
+ SFT on self-teacher	4B	84.29	54.91	78.88	76.38	71.39	70.12	72.66
+ GRPO	4B	86.39	60.36	79.50	78.63	71.42	70.26	74.43
+ DAPO	4B	86.91	54.44	80.38	76.25	74.94	72.15	74.18
+ OPSD	4B	86.39	54.32	82.38	79.62	74.81	72.52	75.01
+ Vision-OPD	4B	89.01	59.88	83.62	80.62	74.86	70.53	76.42
+ Role-OPD (Ours)	**4B**	**92.15**	**61.18**	**85.75**	**81.62**	**75.61**	**72.62**	**78.16**
Qwen3.5 (9B) Post-Training Methods								
Qwen3.5	9B	84.29	52.31	85.50	81.38	71.38	67.79	73.78
+ SFT on self-teacher	9B	88.48	58.82	83.50	79.88	73.21	71.74	75.94
+ GRPO	9B	90.05	58.11	84.88	80.25	75.68	71.57	76.76
+ DAPO	9B	89.53	58.70	86.63	80.50	75.08	71.27	76.95
+ OPSD	9B	91.62	57.63	85.75	81.63	75.42	72.66	77.45
+ Vision-OPD	9B	91.62	**65.33**	87.75	85.38	73.42	70.56	79.01
+ Role-OPD (Ours)	**9B**	**96.34**	62.60	**88.13**	**85.75**	**76.62**	**72.86**	**80.38**

the teacher logits for the corresponding tokens, supplemented by a tail-probability correction. We cap the maximum generation length at 1024 tokens and train all models for one epoch. For PCAR evaluation, we employ Qwen3.5-397B as an LLM judge. All experiments were conducted using eight NVIDIA H20 GPUs.

Evaluation Benchmarks. We evaluate Role-OPD on six benchmarks that span visually rich inputs and diverse multimodal problem-solving scenarios. Each benchmark requires models to identify answer-supporting evidence and leverage it for accurate question answering. V* Bench Wu & Xie (2024) evaluates the localization and recognition of small targets in high-resolution, visually cluttered scenes, whereas ZoomBench Wei et al. (2026) evaluates the ability to identify subtle regional cues from full images across multiple fine-grained VQA dimensions. HR-Bench 4K and HR-Bench 8K Wang et al. (2025b) extend image understanding to native 4K and 8K resolutions, respectively, where critical visual details may be lost through conventional downsampling. MME-RW EN and MME-RW CN Zhang et al. (2025) provide broader evaluation settings across diverse high-resolution tasks and real-world scenarios in English and Chinese, requiring both detailed visual perception and comprehensive understanding. Together, these benchmarks offer a holistic evaluation of Role-OPD's ability to handle diverse visual understanding challenges under varying conditions of scale, resolution, language, and realism.

Baselines. We compare Role-OPD with two broad baseline categories. First, we evaluate it alongside state-of-the-art MLLMs drawn from two distinct groups: (i) leading closed-source models, such as GPT-5.2 OpenAI (2025), GPT-5.4 OpenAI (2026), Gemini-3.1-Pro Google (2026a), and Gemini-3.5-Flash Google (2026b), and (ii) representative open-source models, including GLM-4.6V Hong et al. (2025), Qwen3-VL-Instruct Yang et al. (2025), Qwen3.5 Qwen Team (2026), and Kimi-K2.6 Kimi Team (2026). Second, we evaluate Role-OPD against competing post-training approaches

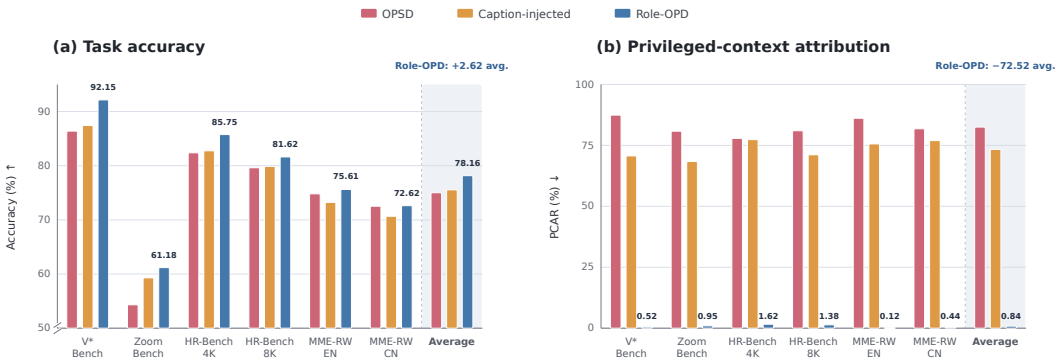

Figure 4: Ablation study of privileged-text form and role-aware design. We compare Role-OPD with standard OPSD and caption-injected OPSD in terms of **(a)** accuracy (higher is better) and **(b)** Privileged-Context Attribution Rate (PCAR, lower is better).

under the same backbone and training data constraints. These include *SFT on self-teacher*, an off-policy distillation approach using successful generations from the self-teacher, as well as *GRPO* Shao et al. (2024) and *DAPO* Yu et al. (2026), which optimize group-relative policy objectives using binary outcome rewards verified against reference answers. We also include *OPSD* Zhao et al. (2026a), an on-policy self-distillation method guided by reference answers as textual privilege, and *Vision-OPD* Yuan et al. (2026), which employs evidence-centered image crops as visual privilege.

4.2 Main Results

Table 1 compares Role-OPD with diverse post-training methods as well as state-of-the-art MLLMs. When applied to Qwen3.5-4B, Role-OPD raises the average performance from 70.54 to 78.16, an improvement of 7.62 points. Under the same backbone and training data constraints, it achieves the best overall results among the evaluated post-training methods, including SFT, GRPO, and DAPO. Compared to standard OPSD, which adopts reference answers as textual privilege, Role-OPD leverages privileged captions and a role-aware design to yield an average gain of $+3.15$ points. Furthermore, Role-OPD secures a 1.74-point advantage over Vision-OPD, indicating that our role-aware textual privilege is more effective than the visual privilege adopted by Vision-OPD under this setting. This improvement also holds on the larger Qwen3.5-9B backbone, where Role-OPD lifts the average score by 6.60 points, maintaining a consistent lead over OPSD and Vision-OPD. In addition to outperforming post-training methods, Role-OPD remains competitive against substantially larger models. Notably, Role-OPD-4B outperforms all larger open-source models listed in Table 1 on our six visual benchmarks, while also exceeding closed-source models like GPT-5.2 and GPT-5.4. Moreover, despite having only 9B parameters, Role-OPD-9B surpasses top-tier closed-source models such as Gemini-3.1-Pro and Gemini-3.5-Flash. Beyond the targeted fine-grained benchmarks, we further evaluate generalization and multimodal capability retention on five held-out benchmarks. Role-OPD also achieves the highest average performance under both backbone settings. See Appendix G for benchmark descriptions and complete results.

4.3 Ablation Study

Ablation of Privileged-Text Form. We investigate how the form of privileged text affects OPSD. Standard OPSD supplements the teacher's input with reference answers, whereas caption-injected OPSD leverages descriptive captions that contain answer-supporting evidence instead. As shown in Figure 4, replacing reference answers with detailed captions improves the average accuracy from 75.01% to 75.54%. This indicates that fine-grained textual descriptions provide more effective supervision than answer-only labels. However, modifying only the form of privileged text cannot resolve the privileged-context attribution problem. While the average PCAR drops from 82.52% for standard OPSD to 73.36% for caption-injected OPSD, both ratios remain substantially high. Thus, although fine-grained captions deliver richer supervision, context-specific artifacts are still consistently transferred to the student model, limiting the full benefits of privileged captions.

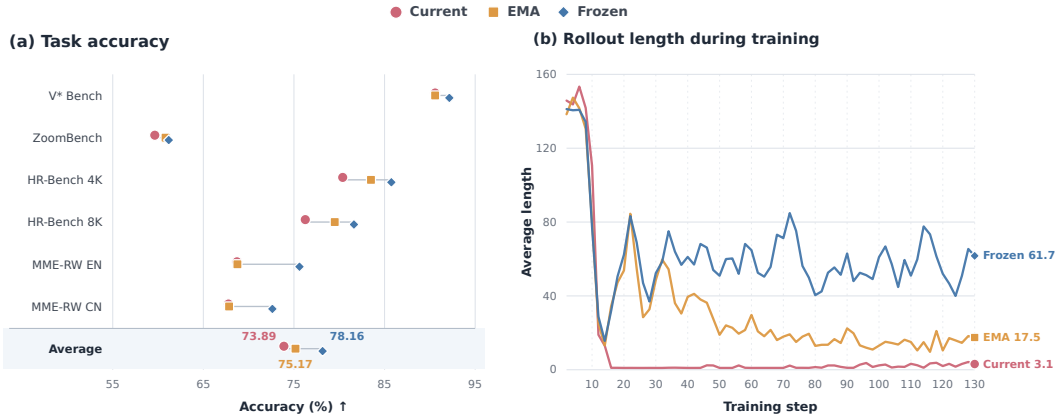

Figure 5: Effect of the teacher update strategy. **(a)** Task accuracy on six fine-grained visual benchmarks and their average. **(b)** Average rollout length for student samples during training.

Ablation of the Role-Aware Design. We isolate the effect of our role-aware design by comparing caption-injected OPSD and Role-OPD, both of which adopt the same captions as privileged text. Caption-injected OPSD naively treats the captions as user-provided context, whereas Role-OPD reframes them as the teacher's own visual understanding in an intermediate assistant turn. As shown in Figure 4, this modification improves the average accuracy from 75.54% to 78.16%, a gain of 2.62 points, with consistent improvements across all six benchmarks. More importantly, the average PCAR drops sharply from 73.36% to only 0.84%. By framing these privileged captions as self-derived visual understanding, Role-OPD enables the teacher to integrate this information more naturally into its output distribution rather than treating it as a user-provided cue. Consequently, this role-aware design not only preserves the genuine improvements enabled by privileged captions but also prevents context-specific artifacts from propagating to the student model. Additional controls test whether the gains arise merely from explicit attribution suppression or from the multi-turn context structure. The first adds a no-attribution instruction, while the second uses a three-turn all-user context (see Appendices H and I for details).

4.4 Further Analysis

Effect of the Teacher Update Strategy. We investigate how the teacher update strategy influences Role-OPD. As shown in Figure 5(a), freezing the teacher yields the best performance, with an average accuracy of 78.16%, compared with 75.17% for EMA and 73.89% for the current policy. Figure 5(b) offers a plausible explanation for this gap. Under the current or EMA update schemes, generated trajectories progressively shorten toward brief answer-only outputs, reaching an average of only 3.1 and 17.5 tokens, respectively, by the end of training. Such shortened sequences expose fewer student-visited prefixes for token-level distillation, leading to increasingly sparse supervision. In contrast, a frozen teacher serves as a stable learning anchor. Although generated trajectories also shrink under this setting, they remain substantially longer throughout training, averaging 61.7 tokens by the end. This avoids the severe collapse toward answer-only outputs seen with current and EMA updates, preserving the dense token-level supervision required for effective self-distillation.

5 Conclusion

We identify an evidence-utilization gap between original images and targeted textual descriptions, revealing the potential of textual privilege for on-policy self-distillation. However, naively using textual privilege as user-provided context causes the privileged-context attribution problem. To address this, we propose Role-OPD, which frames privileged captions as the teacher's own visual understanding in an intermediate assistant turn. This suppresses context-specific artifacts on the student side while preserving genuine improvements. Extensive experiments show that Role-OPD improves fine-grained visual understanding and outperforms existing post-training methods, highlighting the importance of where textual privilege is applied during self-distillation.

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 *International Conference on Learning Representations*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 2024b.

Zebang Cheng, Shuimu Chen, Boxue Yang, Yuanshen Guan, Jingyi Chen, Zheng Lian, Xiaojiang Peng, Fei Ma, LaiZhong Cui, and Qi Tian. Omniopsd: Rationale-privileged on-policy self-distillation for affective computing. *arXiv preprint arXiv:2606.15920*, 2026.

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

Google. Gemini 3.1 Pro, 2026a. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Google. Gemini 3.5 Flash, 2026b. URL `https://deepmind.google/models/model-cards/gemini-3-5-flash/`.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, 2024.

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

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Xingyue Huang, Gregor Franke, Ziyi Yang, Jiamu Bai, Weijie Bai, Jinhe Bi, Zifeng Ding, Yiqun Duan, Chengyu Fan, Wendong Fan, et al. Loong: Synthesize long chain-of-thoughts at scale through verifiers. *arXiv preprint arXiv:2509.03059*, 2025.

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

Kimi Team. Kimi K2.6: From code to creation, from one to many, 2026. URL `https://www.kimi.com/ai-models/kimi-k2-6/`.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, 2026.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023b.

Yunhe Li, Hao Shi, Wenhao Liu, Mengzhe Ruan, Hanxu Hou, Zhongxiang Dai, Shuang Qiu, and Linqi Song. Demopsd: Disagreement-modulated policy self-distillation. *arXiv preprint arXiv:2607.02502*, 2026.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

Kevin Lu. Thinking machines lab. on-policy distillation. *Thinking Machines Lab: Connectionism*, 2025.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *International Conference on Learning Representations*, 2025.

Xiaowen Ma, Yunpu Ma, Chenyang Lin, Sikuan Yan, Jinhe Bi, Zixuan Cao, Yijun Tian, Volker Tresp, and Hinrich Schuetze. Self-evolving multi-agent systems via textual backpropagation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

OpenAI. Introducing GPT-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

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

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.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 2026.

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

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 2024a.

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

Vladimir Vapnik and Rauf Izmailov. Learning with intelligent teacher: Similarity control and knowledge transfer: In memory of alexey chervonenkis. In *International symposium on statistical learning and data sciences*, 2015.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Gengyuan Zhang, Jinhe Bi, Jindong Gu, Yanyu Chen, and Volker Tresp. Spot! revisiting video-language models for event understanding. *arXiv preprint arXiv:2311.12919*, 2023.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *International Conference on Learning Representations*, 2026.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025.

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*, 2026a.

Xuanle Zhao, Qiushi Sun, Jingyu Xiao, Xuexin Liu, Haoyue Yang, Qiaosheng Chen, Xianzhen Luo, Jing Huang, Yufeng Zhong, Lei Chen, et al. Beyond nl2code: A structured survey of multimodal code intelligence. *arXiv preprint arXiv:2606.15932*, 2026b.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, 2026.

Table 2: Detailed training settings of Role-OPD.

Setting	Value	Description
Hardware and optimization		
GPU allocation	8 NVIDIA H20 GPUs	Training worker allocation
Training precision	BF16	Numerical precision
Trainable parameters	All	Parameters updated during training
Optimizer	AdamW	Student model optimizer
Learning rate	2×10^{-6}	Peak learning rate
Adam coefficients	$(0.9, 0.999)$	Adam betas
Weight decay	0.01	AdamW weight decay
Gradient clipping	1.0	Maximum gradient norm
Linear warmup	10 steps	Learning rate warmup
Training epochs	1	Number of passes over the training set
Training batch size	48	Mini-batch size per training step
Actor update mini-batch size	48	Mini-batch size for actor update
On-policy generation and self-distillation		
Rollouts per prompt	8	On-policy responses per prompt
Maximum prompt length	8192 tokens	Prompt truncation limit
Maximum response length	1024 tokens	On-policy generation limit
Rollout correction mode	Token	Token-level rollout correction
Rollout correction threshold	2.0	Clipping threshold for rollout correction
Distillation objective	Jensen–Shannon divergence	Token-level distribution matching
Top-K size	100	Distillation candidate set
Teacher parameter update	Frozen	Detached teacher update rule
Teacher construction	Always on	Per-step teacher construction
Student image field	images	Original image
Teacher image field	images	Original image
Teacher privilege field	caption	Source of privileged text
Privilege role	assistant	Role assigned to privilege
Reproducibility		
Checkpoint	Final step	Checkpoint used for evaluation
Data / generation seed	42 / 42	Data and generation seeds

APPENDIX

A THE USE OF LLM

We used LLM-based tools as writing aids during this paper's preparation, primarily to polish language and phrasing. All generated suggestions were reviewed, edited, and verified by the authors before incorporation. We also used LLMs to generate the synthetic datasets in our experiments. The generation process and resulting data were manually inspected and filtered to ensure correctness and relevance. The authors are solely responsible for the content of this paper.

B DETAILED TRAINING SETTINGS

We detail Role-OPD's training settings in Table 2, organized into three categories: hardware and optimization, on-policy generation and self-distillation, and reproducibility. We implement Role-OPD using FSDP and vLLM, with Python 3.12, PyTorch 2.10.0, Transformers 5.5.0, vLLM 0.18.0, and Ray 2.53.0. The perception chat template prepends an empty block to each prompt, and no explicit reasoning trace is requested or supervised during training.

Privileged Caption Generation

```
<image>
Instruction:
Given an image and a multiple-choice question with the correct answer, write a
detailed, natural caption. The caption should mention the visual detail relevant
to the question and describe it accurately based on the given answer.

Question: {question} Answer: {answer}

Output the caption enclosed in <text></text> tags.
```

Figure 6: Prompt template for generating privileged captions that contain answer-supporting evidence, from the image, question, and ground-truth answer.

Reference Answer Attribution Judgment

```
Instruction:
Given a model response, determine whether it attributes its conclusion to a reference answer.

Label YES if the response refers to the reference answer as the source of its conclusion,
and label NO otherwise.

Model response: {response}

Output only <label>YES</label> or <label>NO</label>.
```

Caption Attribution Judgment

```
Instruction:
Given a model response, determine whether it attributes its conclusion to a provided description.

Label YES if the response refers to the provided description as the source of its conclusion,
and label NO otherwise.

Model response: {response}

Output only <label>YES</label> or <label>NO</label>.
```

Figure 7: Prompt templates for privileged-context attribution judgment.

C PROMPTS

This section presents three prompt templates adopted in our experiments. Figure 6 shows the prompt for generating privileged captions that contain answer-supporting evidence. Figure 7 presents separate prompts for detecting privileged-context attribution problem under two forms of textual privilege: reference answers and captions. Figure 8 shows the pairwise comparison prompt used to judge which image caption is of higher quality. Following Vision-OPD Yuan et al. (2026), we use their benchmark prompts for model evaluation.

D EFFECT OF DISTILLATION OBJECTIVE

We compare forward KL, reverse KL, and JSD ($\beta = 0.5$) using Qwen3.5-4B. As shown in Table 3, JSD achieves the highest average score of 78.16, outperforming forward KL and reverse KL by 0.83 and 0.99 points, respectively. These results support our choice of JSD as the distillation objective.

Caption Preference Judgment

```
<image>
Instruction:
Given an image and two candidate captions, select the caption that provides a better
description of the image. Consider visual accuracy, specificity, and naturalness.

Caption A: {caption_a}
Caption B: {caption_b}

Output only <label>A</label> or <label>B</label>.
```

Figure 8: Prompt template for caption preference judgment.

Table 3: Comparison of distillation objectives on Qwen3.5-4B. We evaluate forward KL, reverse KL and JSD ($\beta = 0.5$). Bold indicates the best result in each column.

Setting	V* Bench	ZoomBench	HR-Bench		MME-RW		Average
			4K	8K	EN	CN	
Forward KL	91.10	60.59	85.50	81.25	75.25	70.27	77.33
Reverse KL	90.58	59.64	85.13	81.25	75.29	71.10	77.17
JSD ($\beta = 0.5$)	**92.15**	**61.18**	**85.75**	**81.62**	**75.61**	**72.62**	**78.16**

E Teacher Context Construction

To generate supervision signals for on-policy self-distillation, we condition the teacher on textual privilege in the form of detailed image captions. As illustrated in Figure 9(a), a direct approach termed caption-injected OPSD places the image, privileged captions, and the target question within a single user turn. In contrast, Role-OPD reorganizes the teacher context into a multi-turn dialogue, as shown in Figure 9(b). Specifically, an initial user turn provides the input image, an intermediate assistant turn delivers the privileged captions as the teacher's own visual understanding, and a later user turn poses the target question.

F Human and LLM Preference Evaluation

Since Role-OPD leverages privileged captions to boost visual understanding, evaluating the distilled model's autonomous image-description capability without privilege serves as a direct test of the successful internalization of fine-grained visual perception. To verify this, we conduct pairwise preference evaluations on the V* benchmark. Specifically, we prompt Role-OPD, Vision-OPD, OPSD, and the base Qwen3.5-4B model with the same instruction "Describe the image in detail". We then construct randomized pairwise comparisons between Role-OPD and the three baselines. The comparisons are independently evaluated by three human annotators and three LLM judges (Qwen3.5-397B, GPT-5.5, and GLM-5.2), with evaluation workload evenly distributed across judges in each group. As shown in Figure 10, Role-OPD achieves an overwhelming preference rate, securing 84.99% of the human votes and 81.85% of the LLM votes, substantially exceeding the aggregated preference share pooled across all three baseline methods.

For human evaluation, we use a custom web-based platform for strict blind evaluation. As illustrated in Figure 11, candidate descriptions are presented in randomized pairs without revealing their source models. In this representative case comparing Vision-OPD (Description A) against Role-OPD (Description B), the human annotator preferred Description B. Although Vision-OPD notes the child, it misses fine-grained visual cues like the polka-dot shirt, blue motorcycle, and picnic table while hallucinating non-existent corrugated-metal buildings. In contrast, Role-OPD accurately captures both localized details and legible scene text, yielding faithful descriptions that align far

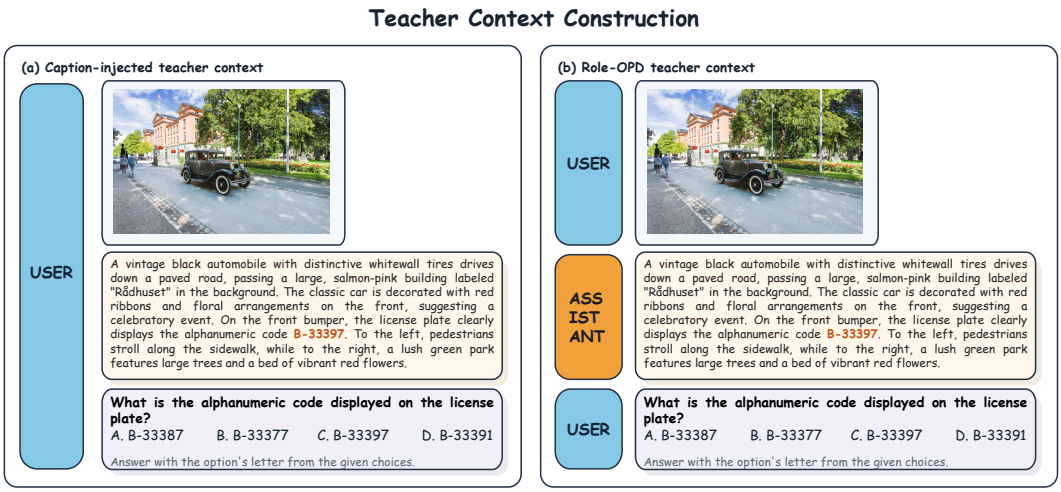

Figure 9: **Comparison of teacher context construction strategies. (a) Caption-injected OPSD** naively incorporates privileged captions as user-provided context, combining the image, caption, and question within a single user turn. **(b) Role-OPD** reorganizes the input into a multi-turn dialogue. It presents the same captions in an intermediate assistant turn, framing these privileged captions as the teacher's own visual understanding.

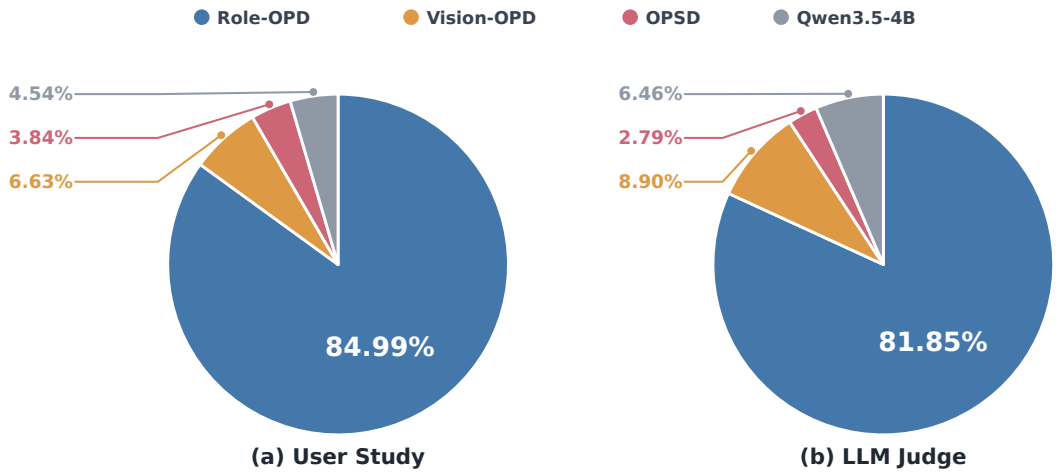

Figure 10: **Pairwise preference vote distribution.** Human and LLM judge preference shares for autonomous image description between Role-OPD and baselines.

better with human perception. Finally, for the automated evaluations, the LLM judges are guided by a standardized evaluation prompt detailed in Figure 8.

G Generalization and Capability Retention

To examine whether the improvements in fine-grained visual understanding come at the cost of broader multimodal capabilities, we further assess Role-OPD on five held-out benchmarks beyond the training distribution. We compare it against the corresponding Qwen3.5 base models, OPSD, and Vision-OPD at matched model sizes of 4B and 9B. Together, these benchmarks span three complementary aspects: visual perception, core multimodal reasoning, and hallucination. Specifically, MMVP Tong et al. (2024b) probes fine-grained visual perception through image pairs designed to expose patterns that CLIP-based encoders struggle to distinguish, while Visual Probe Lai et al.



Description A

The image depicts an outdoor market or festival set in a lush, tropical environment. In the foreground, a gravel path and grassy area host several vendor stalls under white, maroon, and blue canopies. The stalls display colorful goods including clothing, textiles, and crafts. People are browsing, shopping, and interacting including a small child walking alone in the center of the scene.

Behind the market, traditional thatched-roof huts and modern buildings with corrugated metal roofs are visible. Lush green trees frame the scene, and in the background, a large, forested mountain rises under a partly cloudy sky. Overhanging branches with vibrant red flowers (likely flamboyant trees) add a natural canopy at the top of the frame.

Overall, the image captures a lively, culturally rich outdoor gathering nestled in a scenic, mountainous tropical setting.

Description B

The image depicts an outdoor market or festival set in a lush, tropical environment. In the foreground, a small child in a polka-dot shirt walks across a gravel path. Behind, several market stalls are arranged under white, thatched, and branded canopies (including one for onusportswear), displaying colorful textiles, clothing, and crafts. People browse the stalls, some seated, others standing. To the left, a blue motorcycle is parked near a picnic table. The background features dense green trees and a prominent, forested mountain range under a cloudy sky. Overhanging branches with bright red flowers frame the top of the scene, adding vibrancy to the natural setting.

Which description is better?

☐ Description A is better

☐ Description B is better

Figure 11: **User study web interface.** An example of the web-based evaluation platform used in our user study.

(2026) provides an additional held-out test of fine-grained perception on queries drawn from outside the training distribution. CV-Bench Tong et al. (2024a) assesses vision-centric 2D and 3D understanding, covering counting, spatial relationships, depth ordering, and relative distance. MMStar Chen et al. (2024a) measures six core multimodal capabilities using vision-indispensable samples selected to reduce data leakage. POPE Li et al. (2023b) quantifies object hallucination through binary questions about object presence. As shown in Table 4, Role-OPD achieves the highest average performance at both scales (78.15 with Qwen3.5-4B and 80.45 with Qwen3.5-9B), outperforming Vision-OPD by 0.72 and 0.52 points and surpassing its base model on every held-out benchmark. These results indicate that the fine-grained specialization introduced by Role-OPD does not come at the expense of broader multimodal capabilities.

Table 4: Comparison of Role-OPD with post-training methods across five held-out benchmarks for evaluating generalization beyond the training distribution and multimodal capability retention, with mean performance reported. Bold denotes the best result within each matched post-training group.

Model	MMVP	CV-Bench	MMStar	POPE	Visual Probe	Average
Qwen3.5 (4B) Post-Training Methods						
Qwen3.5	76.67	87.11	78.53	88.31	42.82	74.69
+ OPSD	78.67	88.13	76.47	88.70	49.23	76.24
+ Vision-OPD	78.33	87.36	78.73	**89.08**	53.64	77.43
+ Role-OPD (Ours)	**79.67**	**88.76**	**79.20**	88.83	**54.31**	**78.15**
Qwen3.5 (9B) Post-Training Methods						
Qwen3.5	81.00	88.23	82.73	89.34	41.22	76.50
+ OPSD	81.33	88.59	81.40	89.10	50.68	78.22
+ Vision-OPD	82.33	88.26	83.27	**89.69**	56.12	79.93
+ Role-OPD (Ours)	**82.67**	**89.40**	**83.67**	89.57	**56.92**	**80.45**

Table 5: Task accuracy and Privileged-Context Attribution Rate (PCAR) across three training settings. The upper block reports task accuracy (higher is better) and the lower block reports PCAR (lower is better). We compare caption-injected OPSD, caption-injected OPSD with the No-Attribution Instruction, and Role-OPD. Bold denotes the best result in each block.

Method	V* Bench	ZoomBench	HR-Bench		MME-RW		Average
			4K	8K	EN	CN	
Task Accuracy							
Caption-Injected OPSD	87.43	59.29	82.75	79.88	73.22	70.66	75.54
+ No-Attribution Instruction	85.34	59.05	82.25	78.63	74.97	71.24	75.25
Role-OPD (Ours)	**92.15**	**61.18**	**85.75**	**81.62**	**75.61**	**72.62**	**78.16**
Privileged-Context Attribution Rate (PCAR)							
Caption-Injected OPSD	70.68	68.40	77.38	71.13	75.61	76.98	73.36
+ No-Attribution Instruction	8.38	6.39	8.13	7.88	9.42	5.09	7.55
Role-OPD (Ours)	**0.52**	**0.95**	**1.62**	**1.38**	**0.12**	**0.44**	**0.84**

H Controlling for Explicit Privileged-Context Attribution

The role-aware design of Role-OPD substantially reduces privileged-context attribution while improving task accuracy. However, one possible explanation is that this reduction mainly reflects the suppression of explicit attribution phrases such as "based on the description". To examine this possibility, we construct a control based on caption-injected OPSD. It retains the same privileged captions in the teacher's user turn and appends the explicit instruction, "Answer the question directly without mentioning or referring to the provided description". The instruction is visible only to the teacher, while the student context and all other training settings remain unchanged. This control directly tests whether an explicit no-attribution instruction can reproduce the benefits of assigning the privileged caption to the assistant role. As shown in Table 5, adding the no-attribution instruction substantially reduces the average PCAR of caption-injected OPSD from 73.36% to 7.55%, yet leaves task accuracy essentially unchanged, from 75.54% to 75.25%. In contrast, Role-OPD further reduces the average PCAR to 0.84% while achieving a substantially higher average accuracy of 78.16%. Explicit attribution suppression alone therefore cannot reproduce the benefits of Role-OPD. Constraining the teacher's output phrasing suppresses overt attribution without producing the corresponding accuracy gain, whereas Role-OPD simultaneously improves accuracy and reduces privileged-context attribution to a negligible level. These findings support our interpretation that the role assigned to the privileged caption affects more than the surface form of the teacher's response. By placing the caption in an intermediate assistant turn, Role-OPD frames the privileged evidence as part of the teacher's own visual understanding rather than as externally supplied context, allowing the teacher to integrate this evidence more naturally into its predictive distribution and thereby provide more effective supervision to the student.

Table 6: Effect of role assignment under a controlled multi-turn context structure. We compare Caption-Injected OPSD, a three-turn all-user control, and Role-OPD. The upper block reports task accuracy (higher is better), while the lower block reports PCAR (lower is better). Bold denotes the best result in each block.

Method	V* Bench	ZoomBench	HR-Bench		MME-RW		Average
			4K	8K	EN	CN	
Task Accuracy ↑							
Caption-Injected OPSD	87.43	59.29	82.75	79.88	73.22	70.66	75.54
+ Three-Turn User-Role Context	86.39	58.93	83.00	80.12	75.02	72.40	75.98
Role-OPD (Ours)	**92.15**	**61.18**	**85.75**	**81.62**	**75.61**	**72.62**	**78.16**
Privileged-Context Attribution Rate (PCAR) ↓							
Caption-Injected OPSD	70.68	68.40	77.38	71.13	75.61	76.98	73.36
+ Three-Turn User-Role Context	24.61	75.50	65.75	56.63	51.74	45.31	53.26
Role-OPD (Ours)	**0.52**	**0.95**	**1.62**	**1.38**	**0.12**	**0.44**	**0.84**

I Controlling for Multi-Turn Context Structure

To disentangle the effect of the role-aware design from that of the multi-turn context structure, we construct a three-turn control that uses the same image–caption–question ordering as Role-OPD but assigns all three messages to the user role. In contrast to Role-OPD, where the privileged caption is placed in an intermediate assistant turn, this control keeps the caption as user-provided context while exactly matching the number of turns. As shown in Table 6, the three-turn user-role control achieves an average task accuracy of 75.98%, remaining below the 78.16% achieved by Role-OPD. Although the multi-turn structure reduces the average PCAR from 73.36% to 53.26%, attribution remains prevalent, in sharp contrast to the 0.84% PCAR of Role-OPD. These results indicate that merely reorganizing the teacher context into multiple turns is insufficient to reproduce the benefits of Role-OPD. When the privileged caption remains assigned to the user role, the teacher continues to treat it as externally provided evidence, allowing the associated attribution patterns to propagate to the student. In contrast, assigning the caption to the assistant role is critical for simultaneously improving task accuracy and suppressing privileged-context attribution.