Vision-OPD: Learning to See Fine Details for Multimodal LLMs via On-Policy Self-Distillation

Qianhao Yuan[1,2], **Jie Lou**[3,†], **Xing Yu**[3], **Hongyu Lin**[1], **Le Sun**[1], **Xianpei Han**[1], **Yaojie Lu**[1,†]
[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Xiaohongshu Inc.
yuanqianhao2024@iscas.ac.cn loujie0822@gmail.com luyaojie@iscas.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) still struggle with fine-grained visual understanding, where answers often depend on small but decisive evidence in the full image. We observe a regional-to-global perception gap: the same MLLM answers fine-grained questions more accurately when conditioned on evidence-centered crops than on the corresponding full images, suggesting that many failures stem from difficulty to focus on relevant evidence rather than insufficient local recognition ability. Motivated by this observation, we propose Vision-OPD (Vision On-Policy Distillation), a regional-to-global self-distillation framework that transfers the model's own privileged regional perception to its full-image policy. Vision-OPD instantiates two conditional policies from the same MLLM: a crop-conditioned teacher and a full-image-conditioned student. The student generates on-policy rollouts, and Vision-OPD minimizes token-level divergence between the teacher and student next-token distributions along these rollouts. This enables the model to internalize the benefit of visual zooming. The Vision-OPD training objective uses no external teacher, ground-truth labels, or reward verifier, and the trained model requires no visual tool use at inference. Experiments on multiple fine-grained visual understanding benchmarks show that Vision-OPD models achieve competitive or superior performance against much larger open-source, closed-source, and "Thinking-with-Images" agentic models. The code is available at https://github.com/VisionOPD/Vision-OPD.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in general visual understanding and reasoning [8, 11, 36, 46]. However, they still struggle with fine-grained visual understanding, where the answer often depends on small but decisive details that occupy only a fraction of the image [27, 53, 77]. In full-image inference, these details are often easy to overlook amid many visual tokens. As a result, MLLMs may produce plausible answers based on the global scene while missing the local evidence that is truly needed for the question.

Recent "Thinking-with-Images" methods [82, 87] address this issue by equipping MLLMs with agentic visual tool use, enabling them to crop, zoom, and inspect the region of interest during inference. By making local evidence more salient, these methods improve fine-grained visual understanding. However, they introduce substantial inference overhead due to repeated image encoding and model calls. A natural question arises: *can the benefit of visual zooming be internalized through training, so that the model can use fine-grained evidence from the full image without additional tool use?*

[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Our motivation comes from a simple observation: the same MLLM often answers a fine-grained question more accurately when conditioned on the evidence-centered crop than on the corresponding full image. This performance gap reveals a broader bottleneck: MLLMs can often interpret the relevant evidence once it is made salient, but struggle to exploit it when it is embedded in the global visual context. This observation suggests a natural training signal: the model's own crop-conditioned behavior can serve as privileged supervision for improving its full-image behavior.

A straightforward way to exploit such privileged supervision is supervised fine-tuning (SFT) on crop-conditioned responses. However, this trains the model on trajectories generated under privileged crop inputs, leading to distribution mismatch and exposure bias [1]. Reinforcement learning with verifiable rewards methods, such as GRPO [44] and DAPO [69], can optimize on-policy rollouts, but usually provide only sparse sequence-level feedback, and require ground-truth labels and verifiers. On-Policy Distillation (OPD) [1, 30] combines on-policy sampling with dense token-level supervision, but existing OPD methods typically rely on external stronger teachers or ground-truth label [86].

We propose Vision-OPD, a regional-to-global self-distillation framework for fine-grained visual understanding. It instantiates two policies from the same MLLM with different visual conditions: a crop-conditioned teacher that observes the evidence-centered crop as a privileged input, and a full-image-conditioned student that observes the full image as in standard inference. The student first generates on-policy rollouts from the full image. For each student-generated prefix, Vision-OPD computes the logit distributions of both the crop-conditioned teacher and the full-image-conditioned student, and minimizes their divergence. In this way, the model transfers its own privileged crop-conditioned behavior to its full-image-conditioned policy along the student's generation trajectory. The Vision-OPD training objective uses no external teacher, ground-truth labels, or reward verifier, and the trained model requires no visual tool use at inference.

We conduct extensive experiments to validate the effectiveness of Vision-OPD. With only 6.2K synthetic training data, Vision-OPD enables a 9B model to outperform much larger open-source models (e.g., Qwen3.5-397B), closed-source models (e.g., GPT-5.4, Gemini-3.1-Pro, Gemini-3.5-Flash), and agentic "Thinking-with-Images" methods on fine-grained visual understanding tasks that require dense local evidence. On hold-out tasks beyond the training distribution, Vision-OPD largely maintains general visual understanding and reasoning ability. Further analyses verify the necessity of on-policy sampling and dense token-level supervision, and show that Vision-OPD substantially narrows the regional-to-global perception gap in MLLMs.

Our contributions are summarized as follows:

- We introduce a regional-to-global self-distillation formulation for fine-grained visual understanding, with privileged crop-conditioned behavior as supervision for full-image inference.
- We propose Vision-OPD, an on-policy self-distillation framework where a crop-conditioned policy supervises a full-image policy on the student's rollouts via token-level supervision.
- Comprehensive experiments validate the effectiveness of Vision-OPD. We demonstrate that Vision-OPD can significantly narrow the regional-to-global gap, and the on-policy sampling and dense supervision are important to its success.

2 Preliminary: distillation and on-policy distillation (OPD)

Knowledge distillation transfers the behavior of a strong teacher into a weaker student by matching the teacher's distributions. For an input x and output sequence $y = (y_1, \ldots, y_{|y|})$, traditional supervised distillation is off-policy: the student is trained on teacher-induced prefixes $y_{<t}$ and minimizes

$$\mathcal{L}_{\text{Supervised Distillation}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{S}}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D\left(p_T(\cdot \mid y_{<t}, x) \,\|\, p_S(\cdot \mid y_{<t}, x)\right)\right]. \tag{1}$$

This supervised distillation objective provides dense token-level supervision, but at inference time the model conditions on its own prefixes rather than those observed in $\mathcal{S}$, which creates a state-distribution mismatch and can cause errors to compound over long horizons [42].

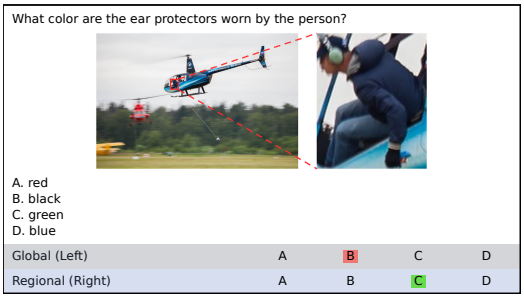


Figure 1: A case of the regional-to-global gap, based on Qwen3.5-9B. The global image input leads to the wrong answer, while the cropped region input yields the correct answer.

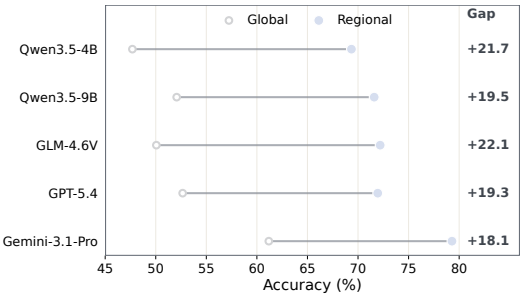


Figure 2: The accuracy of regional inputs is consistently higher than the accuracy of global inputs, revealing a persistent regional-to-global gap across multiple MLLMs.

On-policy distillation (OPD) mitigates this mismatch by sampling $y \sim p_S(\cdot \mid x)$ from the current student and querying the teacher on the prefixes generated by the student [1, 30, 67]. The objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{S},\, \hat{y}\sim p_S(\cdot|x)} \left[\frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} D\left(p_T(\cdot \mid \hat{y}_{<t}, x) \,\|\, p_S(\cdot \mid \hat{y}_{<t}, x)\right) \right]. \tag{2}$$

OPD combines the on-policy relevance of reinforcement learning with the dense token-level guidance of distillation, making it a natural fit for autoregressive generation and reasoning.

3 Vision-OPD

3.1 Motivation: less is more for fine-grained visual understanding

MLLMs struggle with fine-grained visual understanding where small but decisive evidence is overwhelmed by global context. Recent "Thinking-with-Images" methods [6, 82, 87] demonstrate that zooming into the region of interest improves fine-grained understanding, indicating that the bottleneck lies not in the model's recognition ability, but in its capacity to focus on relevant evidence within the full image. This can be verified by comparing the model's performance under two conditions: when it receives the full image (global input) versus when it receives only an evidence-centered crop (regional input). If the model answers correctly with the crop but fails with the full image, it indicates that the model can recognize the evidence but struggles to focus on it within the global context.

Figure 1 illustrates a qualitative case on Qwen3.5-9B. The question asks about the ear-protector color. Given the full image, the model predicts black; given only the cropped region, it correctly predicts green. The decisive evidence is recognizable in isolation, but is overwhelmed by the global context. Evaluation on ZoomBench [61] confirms that this pattern is systematic. As shown in Figure 2, regional-input accuracy consistently exceeds global-input accuracy by 18–22 points. Even much larger models and closed-source models, such as GLM-4.6V, GPT-5.4, and Gemini-3.1-Pro, exhibit substantial gaps, confirming that parameter scaling alone does not close this regional-to-global gap.

This observation motivates Vision-OPD: since a model's own regional perception consistently outperforms its global perception, we can use the former as privileged supervision for the latter, internalizing the benefits of zooming into a single forward pass, without inference-time tool use.

3.2 Method: regional-to-global on-policy self-distillation

To exploit the regional-to-global gap for training, we construct a dataset of triplets $\mathcal{D} = \{(x_i, x_i', q_i)\}_{i=1}^{N}$, where each triplet pairs a full image x with its evidence-centered crop x' and a fine-grained question q. As illustrated in Figure 3 (left), inspired by recent work on multimodal data synthesis [56, 61], we first apply object identification and segmentation to propose bounding box based on a raw image I, retaining only small regions (area ratio $< \tau$) that are likely to contain fine-grained evidence hidden in the image. For each retained region R, we use Qwen3.5-397B [41] as a question generator to produce a question q that are answerable from R alone. To ground the

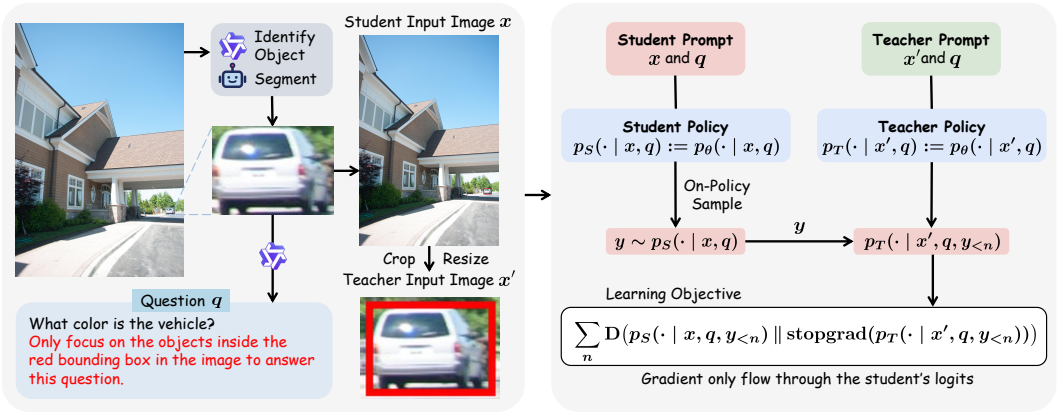


Figure 3: Overview of Vision-OPD. **Left:** Fine-grained visual questions are generated on evidence-centered crops and grounded back to the full image via bounding-box overlay. **Right:** A teacher policy $p_T(\cdot \mid x_{\text{crop}})$ and a student policy $p_S(\cdot \mid x_{\text{global}})$ are instantiated from the same MLLM. The student generates on-policy rollouts $y \sim p_S$, and the per-token divergence $D(p_T \| p_S)$ along these rollouts provides dense supervision. Gradients flow only through the student's logits, enabling self-distillation without ground-truth labels in the training objective.

question back to the full image and avoid referential ambiguity, the bounding box of R is overlaid onto I to produce x, and a spatial constraint is appended to q (e.g., "*Only focus on the objects inside the red bounding box*"). Then we crop I to the bounding box of R and resize it by $2\times$ to produce x'.

Each resulting triplet (x, x', q) thus presents the same question under two visual conditions: the student sees the full image with spatial guidance, while the teacher sees only the isolated crop. The gap between the two conditions directly serves as the learning signal for self-distillation. To make comparison with alternative training strategies, such as off-policy distillation (SFT), RLVR and OPSD [86], we also utilize Qwen3.5-397B as an answer generator to generate ground-truth labels for these methods. We sample multiple responses by giving the region R as image input, and keep a question only when the majority answer reaches a strict consensus (> 0.75). In total, we synthesize 6.2K samples for training.

Given the synthesized dataset, we now describe how Vision-OPD leverages these paired views for training (Figure 3, right), with a complete summary in Algorithm 1. We instantiate two conditional distributions from the same MLLM p_θ by varying the visual conditioning context. The teacher policy conditions on the evidence-centered crop x' as privileged visual information, $p_T(\cdot \mid x', q) = p_\theta(\cdot \mid x', q)$. The student policy observes the full image x and the question q, $p_S(\cdot \mid x, q) = p_\theta(\cdot \mid x, q)$. The student sees the global scene, while the teacher sees only the zoomed-in crop where the evidence is unambiguous, which grants the teacher a privileged view in which fine-grained details are isolated.

Given a training sample (x, x', q), the student generates an on-policy response $y = (y_1, \ldots, y_{|y|}) \sim p_S(\cdot \mid x, q)$. Both policies then evaluate this student-generated trajectory. At each position n, they induce next-token distributions over $y_n \in \mathcal{V}$ conditioned on the same student prefix $y_{<n} = (y_1, \ldots, y_{n-1})$: $p_S(y_n \mid x, q, y_{<n})$ and $p_T(y_n \mid x', q, y_{<n})$. By re-evaluating the same trajectory under a cleaner local view, the teacher's token-level distribution naturally encodes sharper attention to fine-grained visual evidence without any additional decoding.

The training objective minimizes the expected per-token divergence between teacher and student over the student's own rollouts. Given a student-generated sequence y, the averaged divergence is

$$D(p_T \| p_S)(y \mid x, x', q) \;=\; \frac{1}{|y|} \sum_{n=1}^{|y|} D\Big(p_T(\cdot \mid x', q, y_{<n}) \,\Big\|\, p_S(\cdot \mid x, q, y_{<n})\Big), \tag{3}$$

where D can be any distribution divergence such as the generalized Jensen–Shannon divergence JSD_β for a weight $\beta \in [0, 1]$. It is defined as $\mathrm{JSD}_\beta(p_T \| p_S) = \beta\, D_{KL}(p_T \| m) + (1 - \beta)\, D_{KL}(p_S \| m)$, where $m = \beta\, p_T + (1 - \beta)\, p_S$ is the interpolated mixture. The overall loss minimizes the expected divergence over on-policy student samples:

$$\mathcal{L}_{\text{Vision-OPD}}(\theta) = \mathbb{E}_{(x,x',q)\sim\mathcal{D}}\big[\mathbb{E}_{y\sim p_S(\cdot|x,q)}[D(p_T \| p_S)(y \mid x, x', q)]\big]\,. \tag{4}$$

Algorithm 1 Vision-OPD

Require: Training dataset $\mathcal{D} = \{(x_i, x_i', q_i)\}_{i=1}^{N}$; MLLM p_θ; divergence D (e.g., JSD_β)

Let $p_S(\cdot \mid x, q) := p_\theta(\cdot \mid x, q)$ and $p_T(\cdot \mid x', q) := p_\theta(\cdot \mid x', q)$ be the same model p_θ under different visual conditioning.

for step $= 1, \ldots, M$ **do**

  Sample a batch $\mathcal{B} \subset \mathcal{D}$

  for all $(x, x', q) \in \mathcal{B}$ **do**

    Sample on-policy response $y \sim p_S(\cdot \mid x, q)$

    Compute the token-wise divergence along the student rollout:

$$\ell(x, x', q) \leftarrow \frac{1}{|y|} \sum_{n=1}^{|y|} D\Big(p_S(\cdot \mid x, q, y_{<n}) \,\Big\|\, \mathrm{stopgrad}\big(p_T(\cdot \mid x', q, y_{<n})\big)\Big)$$

  Calculate loss $\mathcal{L}_{\text{Vision-OPD}}(\theta) \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(x,x',q)\in\mathcal{B}} \ell(x, x', q)$ and update θ

Gradients are backpropagated only through the student policy p_S, while the teacher p_T acts as a fixed target conditioned on the privileged crop (x', q). Using the student's own generated sequence y as the training prefix aligns the state distribution between training and inference. This avoids the compounding errors that arise from prefix mismatch in off-policy distillation [42]. The dense per-token divergence signal contrasts with the sparse binary rewards of RLVR methods, such as GRPO and DAPO. In Vision-OPD, every token receives a meaningful gradient, so training does not stall when all samples in a batch happen to be correct or incorrect.

Taken together, Vision-OPD combines on-policy sampling with dense token-level supervision. Its training objective uses no external teacher, ground-truth labels, or verifier. Because the crop x' is extracted from unlabeled images by a fully automated data synthesis pipeline, the method is compatible with arbitrary image corpora and internalizes fine-grained visual understanding into a single forward pass.

4 Experiments

4.1 Experimental settings

Model training. We apply Vision-OPD to Qwen3.5-4B/9B [41], with our 6.2K synthetic data. We use JSD ($\beta = 0.5$) as the divergence objective, and approximate the divergence via top-K distillation, which computes only the top-K logits of the student and the corresponding teacher logits, alongside a tail-probability term. With $K = 100$, this avoids the high memory overhead of full-vocabulary logit distillation [86] while capturing most of the distributional information, since the tokens beyond the top-100 typically cover less than 1×10^{-13} of the probability mass in our scenario. The teacher is regularized via exponential moving average (EMA). We set the maximum on-policy generation length to 1024, and training epoch to 1.

Benchmarks. We evaluate on two groups of benchmarks. The first group targets fine-grained visual understanding: V* Bench [65] measures the accuracy of locating and recognizing tiny visual targets within complex scenes; ZoomBench [61] requires models to answer questions that depend on details at varying zoom levels; HR Bench [57] evaluates high-resolution perception at two resolutions (4K and 8K); and MME-RealWorld [81] covers real-world scenarios with high-resolution photographs. The second group serves as holdout tasks for evaluating generalization beyond the training distribution, including MMVP [51], CV-Bench [50], MMStar [5], and POPE [26]. These holdout benchmarks measure whether models retain general multimodal capabilities after fine-grained specialization.

Baselines. We compare against two groups of baselines. The first group compares Vision-OPD with existing SOTA models: (a) "Thinking-with-Images" agentic models that crop and zoom into the image region through multi-step reasoning, including DeepEyes [87], Thyme [82], DeepEyesV2 [17], and SenseNova-MARS [6]; (b) closed-source models, including GPT-5.2 [37], GPT-5.4 [38], Gemini-3.5-Flash [13], and Gemini-3.1-Pro [12]; and (c) open-source models of varying scales, including MiMo-VL-7B-RL [66], Qwen3-VL-Instruct [2], ZwZ [61], MiniCPM-V-4.5 [70], GLM-4.6V [18],

Table 1: Comparison with SOTA MLLMs. We report accuracy (%) for each model. Among open-source models (single forward pass), the best results are highlighted in **bold**, and the second-best are underlined. Vision-OPD achieving the best overall performance.

Model	Param Size	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
"Thinking-with-Images" Agentic Models								
DeepEyes	7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Thyme	7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2	7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
SenseNova-MARS	8B	92.15	47.81	83.13	78.38	67.90	68.90	73.05
Closed-Source Models (Single Forward Pass)								
GPT-5.2	-	79.06	50.89	80.88	78.25	72.37	68.51	71.66
GPT-5.4	-	76.96	52.66	83.75	77.25	74.17	70.64	72.57
Gemini-3.1-Pro	-	86.91	61.18	89.63	86.50	74.22	72.05	78.41
Gemini-3.5-Flash	-	89.01	61.42	88.63	86.62	74.88	73.06	78.94
Open-Source Models (Single Forward Pass)								
Qwen3.5	4B	84.29	47.69	82.13	78.38	62.75	62.67	69.65
Qwen3.5	9B	82.72	52.07	84.38	79.13	70.31	67.30	72.65
MiMo-VL-RL	7B	82.20	45.68	72.25	67.13	60.54	53.86	63.61
Qwen3-VL-Instruct	8B	84.82	42.96	77.50	73.00	63.19	64.61	67.68
ZwZ	8B	87.96	56.69	82.88	80.88	66.57	68.09	73.84
MiniCPM-V-4.5	9B	70.68	42.60	69.63	61.50	59.66	57.92	60.33
GLM-4.6V	106B	86.91	50.06	79.63	76.13	65.57	63.36	70.28
Qwen3-VL-Instruct	235B	91.10	56.09	84.25	78.38	71.74	69.02	75.10
Qwen3.5	397B	87.96	57.16	**88.25**	84.50	**74.58**	69.26	76.95
Kimi-K2.6	1T	88.48	53.14	81.63	78.00	68.84	66.11	72.70
Vision-OPD (Ours)	4B	91.10	59.76	82.25	78.63	71.13	68.78	75.28
Vision-OPD (Ours)	9B	**94.76**	**65.80**	86.88	**84.63**	72.06	**70.41**	**79.09**

Qwen3.5 [41], and Kimi-K2.6 [49]. For GPT models, input images exceeding the API's resolution limit were downscaled only as much as necessary to satisfy the limit, while all other evaluation settings were kept identical across models. The second group evaluates alternative training strategies under the same data and backbones: (a) SFT on self-teacher, which performs supervised fine-tuning (SFT) on successful generations from the self-teacher, acting as an off-policy distillation baseline; (b) Reinforcement Learning (RL) methods, including GRPO [44] and DAPO [69], which optimize via group relative policy gradient with binary outcome rewards verified against ground-truth answers; and (c) OPSD [86], which uses on-policy self-distillation with ground-truth labels to provide reward signals. All training-strategy baselines use the same data as Vision-OPD. For SFT, RL and OPSD, the ground-truth labels used are generated as described in Section 3.2. Besides, we all use the non-thinking mode of Qwen3.5 models with different sizes (4B, 9B, 397B) for training and evaluation.

4.2 Experimental results

4.2.1 Comparison with SOTA MLLMs

Performance gains over initial baselines. As shown in Table 1, Vision-OPD models consistently improve over the corresponding Qwen3.5 baselines across all benchmarks, demonstrating that Vision-OPD can effectively internalize fine-grained visual understanding capabilities into existing MLLMs. See Appendix C for qualitative case study.

Surpassing much larger open-source and closed-source models. Vision-OPD-9B surpasses open-source baselines regardless of scale, e.g., GLM-4.6V and Kimi-K2.6. Compared with closed-source models, Vision-OPD-9B outperforms GPT-5.4 and even Gemini-3.1-Pro.

Comparison with "Thinking-with-Images" agentic models. We further compare Vision-OPD with representative agentic models that explicitly zoom into image region during inference. Despite requiring only a single forward pass, Vision-OPD models outperform these agentic models.

Table 2: Comparison with SFT, RLVR methods and OPSD. Vision-OPD consistently outperforms all baselines on fine-grained visual tasks while maintaining strong performance on holdout tasks.

Method	Fine-Grained Visual Tasks				Holdout Tasks			
	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MMVP	CV-Bench	MMStar	POPE
Qwen3.5-4B								
Vanilla	84.29	47.69	82.13	78.38	76.67	86.51	**72.07**	88.28
SFT on Self-Teacher	78.01	54.67	78.38	75.25	78.33	85.35	66.07	87.48
GRPO	83.77	55.38	81.50	77.38	79.33	87.20	70.07	86.37
DAPO	84.82	55.74	81.63	76.38	79.33	84.79	68.07	86.62
OPSD	81.68	53.85	80.38	76.63	78.68	**87.27**	67.53	88.83
Vision-OPD (Ours)	**91.10**	**59.76**	**82.25**	**78.63**	**79.67**	**87.27**	71.93	**89.14**
Qwen3.5-9B								
Vanilla	82.72	52.07	84.38	79.13	83.33	88.04	**77.13**	88.88
SFT on Self-Teacher	82.20	58.46	81.88	78.88	80.67	87.45	69.07	87.47
GRPO	85.34	57.51	86.50	83.13	81.67	87.77	72.73	87.78
DAPO	88.48	55.62	86.25	83.88	79.33	87.30	75.67	87.54
OPSD	87.43	57.51	82.25	79.75	80.67	87.15	72.60	87.48
Vision-OPD (Ours)	**94.76**	**65.80**	**86.88**	**84.63**	**83.67**	**88.07**	76.53	**89.13**

Table 3: Comparison of various teacher regularization strategies, based on Qwen3.5-9B. Both trust-region regularization and exponential moving average (EMA) regularization use update coefficient $\alpha = 0.05$. †Training with the current policy as teacher leads to collapse.

Teacher	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
Current Policy†	0.00	0.00	0.00	0.00	3.51	0.05	0.59
Initial Policy	93.72	63.91	86.63	85.38	72.59	70.51	78.79
Trust-Region Regularization	93.19	63.79	87.38	85.13	72.31	70.75	78.76
EMA Regularization	94.76	65.80	86.88	84.63	72.06	70.41	79.09

4.2.2 Comparison with SFT, RLVR methods and OPSD

As shown in Table 2, Vision-OPD consistently outperforms all alternative training strategies on fine-grained visual understanding benchmarks. Another advantage of Vision-OPD is its ability to improve fine-grained capabilities while largely preserving previously acquired ones. We evaluate this by testing on holdout benchmarks (MMVP, CV-Bench, MMStar, POPE), the distribution of which is unseen during training. SFT on Self-Teacher exhibits severe forgetting, and RLVR methods (GRPO and DAPO) also degrade holdout performance. In contrast, Vision-OPD largely maintains the capabilities on holdout performance. These results suggest that Vision-OPD mitigates the performance–forgetting tradeoff observed in some alternative training approaches.

We further compare with OPSD [86], which also employs on-policy self-distillation but relies on ground-truth labels to provide reward signals. Vision-OPD instead leverages a self-generated teacher that provides dense token-level supervision without using ground-truth labels in the distillation objective. On fine-grained tasks, Vision-OPD outperforms OPSD on both scales. Moreover, Vision-OPD achieves stronger holdout performance, confirming that our method leads to both better task performance and more robust generalization.

4.3 Ablation study & analysis

4.3.1 Effect of teacher regularization

Vision-OPD uses a self-teacher that is initialized from the same checkpoint with the student and is updated throughout training. Proper regularization of the teacher is critical to prevent the teacher and student from co-adapting, which would cause training to collapse. We compare four teacher strategies in Table 3: (1) the current policy without any regularization, (2) the initial policy (frozen at initialization), (3) trust-region regularization, and (4) exponential moving average (EMA) regularization.

The experimental results are shown in Table 3. Without regularization, training with the current policy as teacher diverges entirely, collapsing to near-zero accuracy across all benchmarks. This

Table 4: Comparison of divergence objectives, based on Qwen3.5-9B. We compare forward KL, reverse KL, and JSD ($\beta = 0.5$) as the token-wise divergence measure D.

Method	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
Forward KL ($\mathrm{KL}(p_T \| p_S)$)	90.05	64.02	86.50	83.50	71.75	70.14	77.66
Reverse KL ($\mathrm{KL}(p_S \| p_T)$)	89.53	62.72	86.38	84.63	72.24	69.87	77.56
JSD ($\beta = 0.5$)	94.76	65.80	86.88	84.63	72.06	70.41	79.09

Table 5: Effect of generation length on Vision-OPD, based on Qwen3.5-9B. We compare on-policy sampled student generation lengths of 512 and 1024 tokens.

Generation Length	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
512 Tokens	91.62	64.38	87.25	85.00	71.62	69.44	78.22
1024 Tokens	94.76	65.80	86.88	84.63	72.06	70.41	79.09

Table 6: Ablation on divergence computation strategies for Vision-OPD, based on Qwen3.5-9B. Top-K logits distillation outperforms sampled-token objectives.

Method	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
Sampled-token distillation	93.72	61.54	85.50	84.25	72.09	68.31	77.57
Top-K logits distillation	94.76	65.80	86.88	84.63	72.06	70.41	79.09

confirms that naive self-distillation without teacher regularization is fundamentally unstable. The initial policy, frozen at its pre-trained weights, already provides a strong teaching signal. EMA regularization achieves the highest average. Therefore, we adopt EMA regularization with update coefficient $\alpha = 0.05$ for all remaining experiments.

4.3.2 Effect of divergence objective

A design choice in Vision-OPD is the divergence used for per-token distribution matching between the teacher and the student. We compare forward KL, reverse KL, and JSD ($\beta = 0.5$) with Qwen3.5-9B in Table 4. JSD ($\beta = 0.5$) yields the strongest gains, outperforming forward KL and reverse KL. Therefore, we adopt JSD ($\beta = 0.5$) in all remaining experiments.

4.3.3 Effect of generation length

Since our objective operates at the token level, the number of generated tokens per sample directly affects the amount of supervision signal available to the student. We compare different generation lengths on Qwen3.5-9B in Table 5. Increasing the generation length from 512 to 1024 tokens yields performance improvements, suggesting that longer rollouts provide richer supervision for fine-grained visual understanding tasks. We adopt 1024 tokens for all remaining experiments.

4.3.4 Learning objective comparison: top-K logits distillation vs. sampled-token distillation

The learning objective of Vision-OPD is defined as a per-token discrepancy between the teacher and student distributions. We compare this objective in two ways: (1) Top-K logits distillation [22]: for each token position, we compute the divergence over the top-K logits via a partial softmax, yielding a proper token-level divergence between the two policies. Specifically, we retain only the top-K logits of the student logits and the corresponding teacher logits, complemented by a tail-probability term that accounts for the remaining probability mass. With $K = 100$, this could capture most of the distributional information, since the tokens beyond the top-100 typically cover less than 1×10^{-13} of the probability mass in our scenario. (2) Sampled-token policy-gradient objective [30]: we evaluate teacher and student log-probabilities only at the token actually sampled by the student, and use the log-probability ratio between teacher and student as a scalar advantage inside a policy-gradient-style

loss. This is analogous to how RLVR methods (e.g., GRPO) apply a constant scalar advantage to sampled tokens, but shaped by the teacher's log-probabilities rather than a binary reward signal.

We compare these variants on Qwen3.5-9B in Table 6. Top-K logits distillation provides an overall performance gain over the sampled-token objective, confirming that dense logit-level credit assignment leads to more effective learning than scalar per-token shaping. Therefore, we adopt top-K logits distillation with $K = 100$ for all remaining experiments.

4.3.5 Vision-OPD significantly narrows the regional-to-global gap

As mentioned in Section 3.1, fine-grained failures often arise not because the decisive evidence is unrecognizable, but because it is hard for current MLLMs to focus on that evidence within the full image. To test whether Vision-OPD addresses this bottleneck during training, we use the same comparison as in Section 3.1: each checkpoint answers the same question with the full image as input and with the evidence-centered crop as input. We track the resulting regional-to-global gap over training, where a smaller gap indicates that evidence visible in the crop is being recovered more reliably from the full image. Figure 4 shows that Vision-OPD steadily closes this gap during training. This reduction is substantial: both Vision-OPD models end with a smaller gap than much larger or closed-source models. These results demonstrate that through Vision-OPD, the models learn to focus on fine-grained evidence directly from the full image, internalizing the benefit of zooming without inference-time tool use.

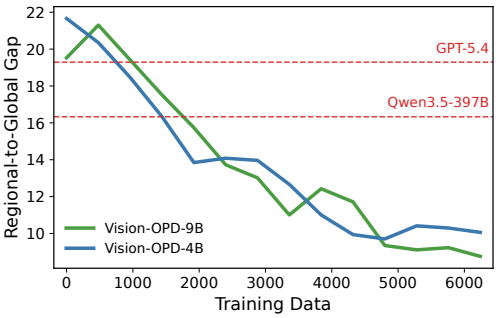


Figure 4: Regional-to-global gap during Vision-OPD training. A lower gap indicates that the model can better recover crop-visible evidence from the full image.

5 Related work

On-policy distillation. On-policy distillation [1, 14, 30, 47, 68, 79] trains a student on its self-generated trajectories, with token-level teacher supervision via KL divergence or related objectives [15, 25, 43, 48, 59, 63]. This alignment reduces the train-test mismatch of off-policy distillation [7], where inference-time prefixes can deviate from the training distribution and cause errors to accumulate. For reasoning, ReST [16] and STaR [76] self-train on sampled rationales verified by rewards or ground-truth answers. In-context editing [40] shows that context-induced knowledge can be internalized through on-policy soft distillation. Recent work includes SDPO [22], which uses environment feedback as privileged information for learning, and OPSD [86], which uses ground-truth labels to provide reward signals for reasoning. Uni-OPD [19] further unifies OPD through a dual-perspective optimization recipe that jointly improves student exploration and teacher supervision reliability. Nevertheless, most approaches still rely on a stronger teacher or verifiable ground-truth rewards. Our work instead studies whether a single MLLM can supervise itself using privileged visual evidence, without an external teacher, ground-truth labels, or a verifier in the training objective.

Fine-grained visual understanding for MLLMs. Large Language Models (LLMs) have demonstrated strong and robust capabilities in solving complex tasks [4, 9, 35, 68, 74]. Built upon LLMs, Multimodal Large Language Models (MLLMs) [34, 36, 73, 75, 89] have demonstrated impressive capabilities in general visual understanding and reasoning. Recent work on fine-grained multimodal understanding increasingly adopt a "Thinking-with-Images" strategy, where MLLMs gather visual evidence at inference time beyond a single forward pass [20, 24, 32, 54, 58, 60, 62, 64, 71, 78, 80, 83–85, 88]. DeepEyes series [17, 87] encourage visual tool calls such as "Zoom in (Crop)" and "Search" via reinforcement learning, while Thyme [82] trains models to write code or manipulate visual inputs in pixel-space. Training-free methods [10, 21, 28, 29, 31, 39, 45] take an alternative route by using tree search or attention-based localization to zoom into important regions during inference. Although effective, these methods incur substantial inference cost and are less practical for real-time use. Other approaches use specialized textual reasoning templates [52, 55] or latent visual reasoning [72], but require format-specific supervision and careful training. ZwZ [61] uses RLVR (e.g., DAPO [69]) to

improve single-pass fine-grained visual understanding without test-time tool use [33]. Vision-OPD instead internalizes regional zooming into model parameters via self-distillation.

6 Conclusion

We introduce Vision-OPD, a simple and effective self-distillation framework for fine-grained visual understanding in MLLMs. The core idea is to let a model teach itself from privileged regional inputs. The teacher policy conditions on an evidence-centered crop, while the student policy observes the full image. By minimizing per-token divergence on the student's rollouts, Vision-OPD provides dense supervision using a training objective that does not require an external teacher, ground-truth labels, or a verifier. Experiments demonstrate that Vision-OPD substantially improves fine-grained understanding of existing MLLMs, surpassing much larger open-source models, closed-source models, and "Thinking-with-Images" agentic models.

7 Limitations

This work has several limitations. First, the training data are constructed using automated object identification and segmentation, model-generated questions, and filtering. This process may introduce question-generation and data-selection biases, potentially limiting the diversity of the resulting training samples.

Second, object identification and segmentation may produce inaccurate region proposals or incomplete crops that omit relevant visual evidence. Moreover, the crop-conditioned teacher may still produce incorrect predictions even when the relevant evidence is visible, leading to unreliable supervision.

Finally, the experiments reported in this paper focus on Qwen3.5 models, and the generalization of Vision-OPD to non-Qwen model families remains to be established.

Acknowledgment

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by the National Key R&D Program of China (2024YFC3308000), the Natural Science Foundation of China (No. 62476265, 62306303), Beijing Nova Program No. 202604841270.

References

[1] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

[2] 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.

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[4] Jiawei Chen, Xinyan Guan, Qianhao Yuan, Mo Guozhao, Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, and Xianpei Han. Consistentchat: Building skeleton-guided consistent multi-turn dialogues for large language models from scratch. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8426–8452, 2025.

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

[6] Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

[7] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[9] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[10] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

[11] Google. Gemini 3. https://blog.google/products-and-platforms/products/gemini/gemini-3/, 2025.

[12] Google. Gemini 3.1 pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

[13] Google. Gemini 3.5 flash. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026.

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

[15] Yuanshen Guan, Zipeng Feng, Chengru Song, Zhiwei Xiong, and Peiqin Sun. Latent reward registers for diffusion preference alignment. *arXiv preprint arXiv:2608.03929*, 2026.

[16] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

[17] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

[18] 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.

[19] Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

[20] Xinhai Hou, Shaoyuan Xu, Manan Biyani, Moyan Li, Jia Liu, Todd C Hollon, and Bryan Wang. Codev: Code with images for faithful visual reasoning via tool-aware policy optimization. *arXiv preprint arXiv:2511.19661*, 2025.

[21] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

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

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[24] 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. *arXiv preprint arXiv:2509.07969*, 2025.

[25] Bingnan Li, Haozhe Wang, Haozhong Xiong, Fangtai Wu, Jinpeng Yu, Yang Shi, Jiaming Liu, and Ruihua Huang. Rethinking classifier-free guidance in on-policy diffusion distillation. *arXiv preprint arXiv:2607.24731*, 2026.

[26] 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*, pages 292–305, 2023.

[27] Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025.

[28] Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. Hide: Rethinking the zoom-in method in high resolution mllms via hierarchical decoupling. *arXiv preprint arXiv:2510.00054*, 2025.

[29] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.

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

[31] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom-in for enhanced multimodal text-rich image understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(4):1–19, 2026.

[32] Jinlong Ma, Yu Zhang, Xuefeng Bai, Kehai Chen, Yuwei Wang, Zeming Liu, Jun Yu, and Min Zhang. Beyond unimodal shortcuts: Mllms as cross-modal reasoners for grounded named entity recognition. *arXiv preprint arXiv:2602.04486*, 2026.

[33] Zhongkuan Mao, Xianjie Liu, Tianyu Meng, Yidong Wang, Wenzhuo Zhao, Ronghao Xian, Yao Jiang, Fei Shen, Junfeng Fang, Yong Dai, et al. Thinking once is enough: Intermediate-layer evidence routing for high-resolution vqa. *arXiv preprint arXiv:2607.27830*, 2026.

[34] Zhongkuan Mao, Wenzhuo Zhao, Xianjie Liu, Yidong Wang, Zhao Gao, Ronghao Xian, Yao Jiang, Yi Zhang, Liangjian Wen, and Keren Fu. Pay more attention to text in high-resolution mllms. *arXiv preprint arXiv:2609.23495*, 2026.

[35] Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20192–20204, 2025.

[36] OpenAI. Gpt-5.1. https://openai.com/index/gpt-5-1/, 2025.

[37] OpenAI. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025.

[38] OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

[39] Ruotian Peng, Haiying He, Yake Wei, Yandong Wen, and Di Hu. Patch matters: Training-free fine-grained image caption enhancement via local perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3963–3973, 2025.

[40] Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. *arXiv preprint arXiv:2406.11194*, 2024.

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

[42] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[43] Jiacheng Ruan, Jun Tang, Wenzhen Yuan, Ting Liu, Shuai Bai, Dayiheng Liu, Zhibo Yang, and Yuzhuo Fu. Contrastive on-policy distillation. *arXiv preprint arXiv:2607.19046*, 2026.

[44] 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.

[45] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6613–6629, 2025.

[46] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[47] Jie Sun, Mao Zheng, Mingyang Song, Zeyuan Liu, Gengsheng Li, Houcheng Jiang, Yilin Cheng, Bichuan Feng, Yuchen Cai, Junfeng Fang, et al. Distill what you trust: Reliability-aware multi-teacher on-policy distillation. *arXiv preprint arXiv:2609.23697*, 2026.

[48] Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Yilin Cheng, Bichuan Feng, Pengfei Liu, Junfeng Fang, and Xiang Wang. Simct: Recovering lost supervision for cross-tokenizer on-policy distillation. *arXiv preprint arXiv:2605.07711*, 2026.

[49] Team, Kimi. Kimi k2.6: From code to creation, from one to many. https://www.kimi.com/ai-models/kimi-k2-6/, 2026.

[50] 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*, 37:87310–87356, 2024.

[51] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[52] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025.

[53] Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Jiani Zheng, Ye Tian, Jiahao Meng, Zilong Huang, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. *arXiv preprint arXiv:2510.18876*, 2025.

[54] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[55] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025.

[56] Shenzhi Wang, Shixuan Liu, Jing Zhou, Chang Gao, Xiong-Hui Chen, Binghai Wang, An Yang, Shiji Song, Bowen Yu, Gao Huang, et al. Hopchain: Multi-hop data synthesis for generalizable vision-language reasoning. *arXiv preprint arXiv:2603.17024*, 2026.

[57] 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*, volume 39, pages 7907–7915, 2025.

[58] Yuji Wang, Wenlong Liu, Jingxuan Niu, Haoji Zhang, and Yansong Tang. Vg-refiner: Towards tool-refined referring grounded reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2512.06373*, 2025.

[59] Yuxiang Wang, Qinke Ni, Shengbo Cai, Wan Lin, Liqiang Zhang, and Zhizheng Wu. Parabridge: Bridging paralinguistic perception and dialogue behavior in speech language models. *arXiv preprint arXiv:2606.10581*, 2026.

[60] Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

[61] 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.

[62] Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

[63] Ming Wen, Yuxuan Liu, Kun Yang, Yunhao Feng, Zhuoer Xu, Yuhao Sun, Shiwen Cui, Xiang Zheng, Yi Liu, Xingjun Ma, et al. Constitutional on-policy safe distillation. *arXiv preprint arXiv:2606.03089*, 2026.

[64] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

[65] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[66] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

[67] Wenda Xu, Rujun Han, Zifeng Wang, Long T Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. *arXiv preprint arXiv:2410.11325*, 2024.

[68] 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.

[69] 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. *arXiv preprint arXiv:2503.14476*, 2025.

[70] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

[71] Xuan Yu, Dayan Guan, and Yanfeng Gu. Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*, 2025.

[72] Haobo Yuan, Yueyi Sun, Yanwei Li, Tao Zhang, Xueqing Deng, Henghui Ding, Lu Qi, Anran Wang, Xiangtai Li, and Ming-Hsuan Yang. Visual reasoning tracer: Object-level grounded reasoning benchmark. *arXiv preprint arXiv:2512.05091*, 2025.

[73] Qianhao Yuan, Yanjiang Liu, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. Saisa: Towards multimodal large language models with both training and inference efficiency. *arXiv preprint arXiv:2502.02458*, 2025.

[74] Qianhao Yuan, Jie Lou, Zichao Li, Jiawei Chen, Yaojie Lu, Hongyu Lin, Le Sun, Debing Zhang, and Xianpei Han. Memsearcher: Training llms to reason, search and manage memory via end-to-end reinforcement learning. *arXiv preprint arXiv:2511.02805*, 2025.

[75] Qianhao Yuan, Qingyu Zhang, Yanjiang Liu, Jiawei Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, and Le Sun. Shortv: Efficient multimodal large language models by freezing visual tokens in ineffective layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 329–339, 2025.

[76] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[77] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DgaY5mDdmT.

[78] Lu Zhang, Jiazuo Yu, Haomiao Xiong, Ping Hu, Yunzhi Zhuge, Huchuan Lu, and You He. Finers: Fine-grained reasoning and segmentation of small objects with reinforcement learning. *arXiv preprint arXiv:2510.21311*, 2025.

[79] Qingyu Zhang, Qianhao Yuan, Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun, Xiang Li, Ming Xu, Jiarui Li, and Xiuyin Zhao. Shortopd: Recovering pruned llms with short-to-long on-policy distillation. *arXiv preprint arXiv:2607.13124*, 2026.

[80] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pages arXiv–2505, 2025.

[81] Yi-Fan 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? *arXiv preprint arXiv:2408.13257*, 2024.

[82] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

[83] Yifan Zhang, Liang Hu, Haofeng Sun, Peiyu Wang, Yichen Wei, Shukang Yin, Jiangbo Pei, Wei Shen, Peng Xia, Yi Peng, et al. Skywork-r1v4: Toward agentic multimodal intelligence through interleaved thinking with images and deepresearch. *arXiv preprint arXiv:2512.02395*, 2025.

[84] Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*, 2025.

[85] Yu Zhang, Mufan Xu, Xuefeng Bai, Pengfei Zhang, Yang Xiang, Min Zhang, et al. Instruction anchors: Dissecting the causal dynamics of modality arbitration. *arXiv preprint arXiv:2602.03677*, 2026.

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

[87] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[88] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024.

[89] Xiuyuan Zhu, Ke Lu, Hao Wu, Zijin Du, Dongming Zhang, and Jian Xue. Ioupd: Iou-aware privileged distillation for visual grounding with multimodal large language models. *arXiv preprint arXiv:2607.15732*, 2026.

A Prompts

Prompts of benchmarks. To ensure reproducibility and facilitate future research, we provide here the complete set of prompts used to evaluate our models across all benchmarks.

V* Bench

```
<image>
{question}
{options}
Answer with the option's letter from the given choices directly.
```

ZoomBench

```
<image>
{question}
{options}
```

HR-Bench

```
<image>
{question} Select from the following choices.
{options}
```

MME-RealWorld EN

```
<image>
{question} The choices are listed below:
{options}
Select the best answer to the above multiple-choice question based on the
image.  Respond with only the letter (A, B, C, D, or E) of the correct
option.
The best answer is:
```

MME-RealWorld CN

<image>
{question} 选项如下所示:
{options}
根据图像选择上述多项选择题的最佳答案。只需回答正确选项的字母（A, B, C, D 或E）。
最佳答案为:

MMVP

```
<image>
{question} {options}
```

CV-Bench

```
<image>
{question} Select from the following choices.
{options}
```

MMStar

```
<image>
{question}
Options: {options}
```

POPE

```
<image>
{question}
Answer the question using a single word or phrase.
```

Prompt for question generation. For each retained region, we use Qwen3.5-397B to generate fine-grained questions based on the cropped image.

Question Generation Prompt

```
<cropped image>
Reference questions:
{randomly sampled seed questions}
Using only the provided image, generate three diverse visual questions. Each
question should be answerable from the image and have a concise, factual
answer.
Questions may cover object identity, text, counting, spatial relationships,
shape, material, or other visible details.
If the image does not support meaningful questions, return an empty JSON list:
[].
Otherwise, return a JSON list containing three objects, each with a
“question” field:
[ { “question”: “<question 1>” }, { “question”: “<question 2>” }, {
“question”: “<question 3>” } ]
```

B Implementation details of data synthesis

The pipeline constructs paired visual contexts for each fine-grained question: a global image x for the student and an evidence-centered crop x' for the teacher.

We use images from SA-1B [23] to synthesize the training dataset. Given a raw image I, we first use Qwen3.5-397B to identify objects in the image. For each identified object, SAM3 [3] generates a segmentation mask and a corresponding bounding box B. We only retain regions whose SAM3 confidence score is greater than 0.75.

For each retained bounding box, we first crop the corresponding region from the original image I and resize it by $2\times$ to obtain a regional image R. Qwen3.5-397B receives R, without the full image, and generates fine-grained questions intended to be answerable from the crop alone.

We ground each crop-level question back to the full image by overlaying its bounding box B on the original image I. The resulting image is the student input x. We also add a spatial constraint to the question, such as "Only focus on the objects inside the red bounding box," to indicate the relevant region.

To construct the teacher input x', we crop the corresponding region from the red-box image x and resize it by $2\times$. This is a separate step from the crop of the original image used for question generation. The resulting paired inputs present the question under two visual conditions: the student observes the full image with spatial guidance, whereas the teacher observes the enlarged regional crop. The red box and spatial constraint help reduce referential ambiguity when the crop-level question is presented with the full image.

To support comparisons with label-dependent training strategies, including off-policy distillation, RLVR, and OPSD [86], we use Qwen3.5-397B as an answer generator. For each crop–question pair, we sample multiple responses conditioned on the original-image crop R. We retain a sample only when its majority answer reaches a consensus score greater than 0.75. The retained questions are formatted as multiple-choice prompts with candidate options. In total, the resulting training dataset contains approximately 6.2K samples.

The generated answer labels are used for the label-dependent baselines. Consensus filtering also determines which examples enter the shared training dataset; however, the answer labels are not used as targets in the Vision-OPD distillation objective.

C Case Study

Fine-grained visual understanding capability of Vision-OPD. As shown in Table 7, Vision-OPD-9B answers correctly while Qwen-3.5-9B fails, demonstrating that Vision-OPD can effectively internalize fine-grained visual understanding capabilities into existing MLLMs.

Crop–question mismatch and unreliable supervision. Table 8 shows a failure case in which the crop does not contain the relevant part of the image. The question asks about the water at the bottom edge of the full image, but the crop primarily contains the beach and only a narrow strip of water. Given the full image, Qwen3.5-9B answers "C. blue," which is consistent with the visible blue-green water; given the teacher input, the same model answers "D. brown." This example illustrates how a mismatched crop can produce an unreliable crop-conditioned response and potentially weaken the supervision available to the student.

D Broader Impacts

Vision-OPD aims to improve the fine-grained visual understanding of MLLMs while eliminating the need for costly inference-time tool use. By internalizing the benefit of evidence-centered regional perception into a single forward pass, Vision-OPD can make detailed visual understanding more efficient and accessible. This may lower the computational and resource barriers for research on multimodal models, and may also reduce deployment costs in practical applications where careful visual inspection is important, such as interface analysis and industrial inspection.

At the same time, stronger fine-grained perception may amplify risks associated with multimodal systems. Models that better identify small visual details could be misused for privacy-invasive analysis, surveillance, or overconfident interpretation of sensitive images.

Table 7: Vision-OPD reads a small number written on a distant boat.

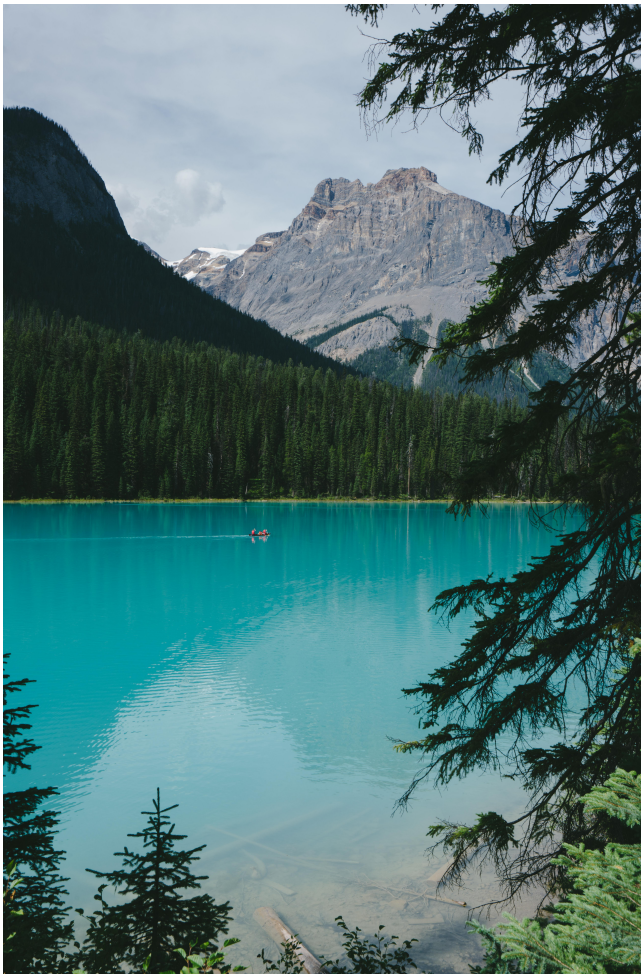
Visual input example, Small Text/Number Reading:	
	
User	What's the number written on the boat? Select from the following choices. (A) 10 (B) 25 (C) 15 (D) 5
Qwen-3.5	The image shows a small boat with people on a turquoise lake, surrounded by mountains and trees. However, **there is no visible number written on the boat** in the image. Looking closely at the boat (which appears to be a canoe or small rowboat), it's too distant and low-resolution to discern any markings or numbers. The question asks us to select from given choices: (A) 10 (B) 25 (C) 15 (D) 5 Since none of these numbers are visibly present on the boat in the image, and assuming this is a trick or observation-based question, the most reasonable conclusion is that **no number is visible**, but since we must choose from the options, perhaps the intended answer is based on an external context or misinterpretation. However, if this is a standard multiple-choice question from a known source (like a quiz or exam), and given that **canoes in some parks or regions are numbered for identification**, and **number 5** is commonly used in such contexts – especially if this is Lake Louise or similar – then **(D) 5** might be the expected answer. But strictly speaking, **based solely on the visual content of the image provided**, there is **no number visible on the boat**. — Given the constraints and typical test design, the best guess is: **(D) 5** *(Note: This is an inference based on common practices, not direct visual evidence.)*
Vision-OPD	Looking at the image, there is a small boat on the turquoise lake with people in it. On the side of the boat, near the front (bow), there appears to be a number painted or stenciled. The number visible is **"15"** – it's white and clearly legible against the darker hull. Therefore, the correct answer is: **(C) 15** Final Answer: **C**

Table 8: An example illustrating a failed proposal, question–crop mismatch and an unreliable crop-conditioned response.

Full-image input with the retained box:	
Corresponding teacher input:	
Question	What is the color of the water at the bottom edge of the image? Only focus on the objects inside the red bounding box in the image to answer this question. A. green B. clear C. blue D. brown Answer with the option's letter from the given choices.
Qwen3.5 (student)	C. blue.
Qwen3.5 (teacher)	D. brown.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: No theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the data and code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No statistical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We choose not to provide the information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in this paper conforms to all aspects of the NeurIPS Code of Ethics. Specifically, we ensured compliance regarding transparency, reproducibility, potential societal impacts, respect for privacy, and fairness. No ethical concerns arose from our methods, datasets, or experiments throughout the research process.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper states that code and models will be publicly released, and documentation for reproducing the method and experiments will be provided with the released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.