DISTILLING ONE TOKEN IS ENOUGH: RETHINKING ON-POLICY DISTILLATION FOR VISUAL PERCEPTION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has recently emerged as an effective approach for improving visual perception in multimodal large language models (MLLMs) by aligning teacher and student predictive distributions over student-generated trajectories. In visual perception tasks, these trajectories typically contain descriptive tokens that verbalize visual evidence and decision tokens that produce the final answer. This raises a fundamental question: is dense token-level supervision across the entire trajectory necessary? In this work, we find substantial redundancy in OPD supervision for visual perception. Distilling only a single decision token or only the descriptive tokens on multiple-choice visual perception data achieves downstream performance comparable to full-response distillation. Moreover, replacing the teacher distribution at the decision position with a hard ground-truth label fails to reproduce the improvement, highlighting the importance of the teacher's soft predictive distribution. Motivated by these findings, we propose **D**ecision-**O**nly **D**istillation (**DOD**), a simple, rollout-free distillation strategy that uses a direct-answer prompt and matches teacher and student distributions only at the first output position. We further introduce target-vocabulary distillation, which restricts distribution matching to the candidate decision tokens (e.g., A, B, C, and D) rather than the full vocabulary. Consequently, each training example requires distillation at only one token position over only a handful of vocabulary entries, eliminating rollout and reducing distillation costs while achieving better overall visual perception performance than standard OPD. Code is available at the anonymous link.

1 INTRODUCTION

Knowledge distillation improves a student model by learning from the supervision provided by a more capable teacher (Hinton et al., 2015). The teacher's advantage may arise from greater model capacity or access to privileged information unavailable to the student (Lopez-Paz et al., 2016), such as verified solution traces for reasoning (Zhao et al., 2026) or question-relevant image crops for visual perception (Yuan et al., 2026). In autoregressive models, conventional offline distillation typically trains on fixed reference or teacher-generated responses (Kim & Rush, 2016). However, these training prefixes can differ from those encountered during the student's own generation, creating a distribution mismatch that limits effective transfer (Agarwal et al., 2024). On-policy distillation (OPD) addresses this mismatch by sampling responses from the current student and querying the teacher on the same generated prefixes, providing guidance on states the student actually visits (Agarwal et al., 2024; Gu et al., 2024). This approach has enabled effective training for both LLM reasoning and MLLM visual perception (Hou et al., 2026; Yu et al., 2026; Liu et al., 2026a).

Recent studies suggest that the dense token-level supervision commonly used in OPD is often redundant. For reasoning tasks, distilling a small subset of informative tokens or a truncated response prefix can match or even surpass full-response distillation (Xu et al., 2026; Liu et al., 2026b; Zhang et al., 2026; Ziheng et al., 2026). These approaches nevertheless retain intermediate reasoning text within partially or fully generated trajectories as the context or target for supervision, as illustrated in Figure 1(a). Such text can serve as a computational scaffold, decomposing a problem into manageable subproblems and making intermediate results available for subsequent deductions (Wei et al.,

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.

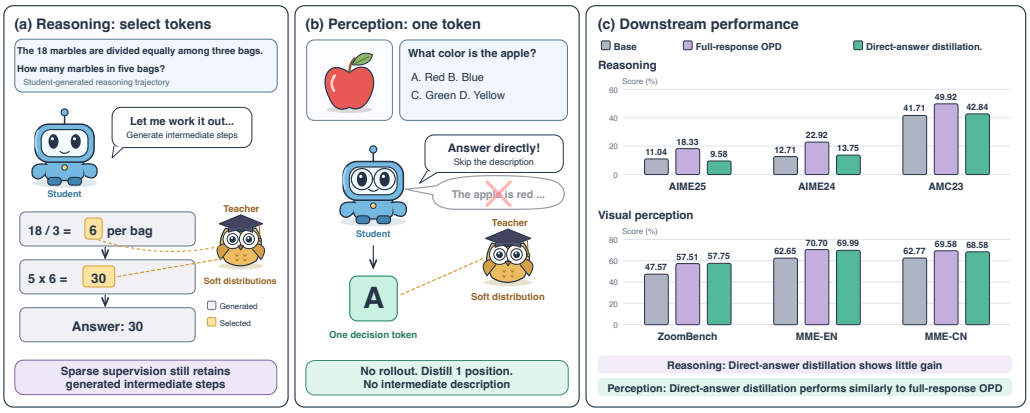

Figure 1: **Intermediate generation plays different roles in OPD for reasoning and visual perception.** **(a)** Token-selective OPD for reasoning applies teacher supervision at selected positions while retaining generated intermediate steps. **(b)** Direct-answer distillation for visual perception matches teacher and student predictive distributions at the first answer position under a direct-answer prompt, without requiring token generation. **(c)** Direct-answer distillation yields little consistent improvement over the base model on reasoning benchmarks, but closely matches full-response OPD on visual perception benchmarks. The reasoning experiments distill Qwen3-4B-Instruct into Qwen3-1.7B (Yang et al., 2025) using DAPO-Math-17K (Yu et al., 2025a) for 100 training steps. The perception experiments use on-policy self-distillation (OPSD) with Qwen3.5-4B (Team, 2026b) on Vision-OPD-6K (Yuan et al., 2026) for one epoch.

2022). For example, solving a mathematical word problem may require defining variables, formulating equations, and performing calculations before reaching an answer.

Visual perception, by comparison, centers on extracting and identifying evidence from visual input. For example, identifying an object's color requires recognizing a visible attribute rather than deriving an answer through a sequence of reasoning steps. In perception tasks, intermediate text often verbalizes visual observations that support the final answer. Such descriptions may therefore offer less benefit than intermediate steps in reasoning tasks (Jin et al., 2026; Kanade & Ganu, 2026), leaving their necessity for effective distillation unclear. This motivates our central question: *Are full-trajectory generation and supervision necessary for improving visual perception through OPD?*

We investigate this question by varying both response generation and supervision in multiple-choice visual perception tasks. Our prompting and loss strategies compare supervision over descriptive tokens, a directly generated decision token, and the entire response. Supervising either component alone retains most of the gains from full-response supervision, indicating substantial redundancy in full-response supervision and showing that intermediate descriptions can be omitted during distillation. Figure 1(c) highlights the contrast with reasoning: directly generating and distilling only the final answer yields little consistent improvement over the base model, whereas full-response OPD improves performance across all three benchmarks. These results suggest that the role of intermediate text in OPD differs between reasoning and visual perception. Moreover, replacing the teacher's soft predictive distribution at the decision position with hard ground-truth labels reduces the perceptual gains, suggesting that the teacher's distribution conveys useful dark knowledge beyond the correct answer label alone (Hinton et al., 2015).

Motivated by these findings, we propose Decision-Only Distillation (DOD), a simple and efficient distillation strategy for improving visual perception. For multiple-choice training questions, DOD uses a prompt requesting a single answer-option token and matches teacher and student predictive distributions only at the first output position. Because no generated prefix is needed at this position, DOD is rollout-free. We further introduce target-vocabulary distillation, which restricts distribution matching to candidate answer tokens, such as A, B, C, and D. By retaining soft supervision over the available choices, this approach reduces the distillation objective to a single position and a handful of vocabulary entries. Experiments across multiple visual perception benchmarks demonstrate that DOD achieves better overall perception performance than full-response OPD while eliminating rollout and reducing distillation costs.

2 Related Work

2.1 On-Policy Distillation for Visual Perception in MLLMs

Recent work has explored on-policy distillation as a means of improving visual perception in MLLMs by transferring stronger visual evidence into student-generated responses. A common principle is to construct informative asymmetry between teacher and student through evidence-centered crops (Yuan et al., 2026), asymmetric image augmentations (Li et al., 2026), or counterfactually identified visual regions (Venkatraman et al., 2026). By matching predictions from a teacher given more informative visual inputs, the student learns to recognize details that are harder to identify under its own input conditions.

However, stronger teacher conditioning does not necessarily ensure that the resulting supervision improves reliance on visual evidence (Liu et al., 2026a). Contrasting predictions under different visual conditions helps identify visually dependent signals, which can guide token weighting or the construction of teacher targets (Liu et al., 2026a; Aniri et al., 2026; Liang et al., 2026). Other approaches explicitly distinguish visual grounding from language modeling or separate perceptual supervision from reasoning supervision (Yoon et al., 2026; Wang et al., 2026a). Adaptive supervision further accounts for the distinction between transferable capabilities and discrepancies induced by privileged information (Hou et al., 2026; Yu et al., 2026). These studies explore how to construct and identify visual supervision signals that facilitate OPD for visual perception.

2.2 Efficient On-Policy Distillation

Selective token-level OPD for efficiency builds on the observation that teacher supervision varies in usefulness across response positions (Xu et al., 2026; Lee et al., 2026). Token-level uncertainty and teacher–student disagreement provide signals for allocating supervision, while student compatibility helps distinguish large discrepancies from locally learnable corrections (Xu et al., 2026; Lee et al., 2026; Wang et al., 2026b). Gradient-estimation reliability provides a complementary criterion by accounting for whether a token-level correction can be accurately estimated from sampled tokens (Sheng et al., 2026). These criteria motivate selective supervision and reweighting instead of treating every response position as equally informative. Liu et al. (2026b) further show that supervising as few as one or two tokens per generated reasoning trajectory can match or surpass full-token distillation, highlighting substantial redundancy in dense supervision for reasoning tasks.

Efficiency can also be improved by shortening the trajectories used for distillation. Progressive rollout schedules, fixed truncation, and early stopping reduce generation and teacher-scoring costs while retaining useful supervision from response prefixes (Zhang et al., 2026; Ziheng et al., 2026). Both token selection and rollout truncation retain partially or fully generated intermediate text as the context or target for distillation. DOD instead prompts the MLLM to directly generate a single answer-option token for multiple-choice visual perception tasks and applies soft distribution matching at that position. With no generated prefix required, DOD removes the need for rollout altogether.

3 Preliminaries

3.1 Off-policy Distillation

Let π_θ and π_{T} denote the student and teacher policies with a shared token vocabulary $\mathcal{V}$. Their inputs, x_{S} and x_{T}, may differ when the teacher receives privileged information. Off-policy distillation trains the student on trajectories obtained independently of its current policy, such as reference responses or teacher-generated outputs (Kim & Rush, 2016; Agarwal et al., 2024).

A common hard-target approach performs supervised fine-tuning (SFT) on teacher-generated responses. Under token-wise argmax decoding, the teacher constructs a response $\hat{y}$ autoregressively:

$$\hat{y}_t \in \arg\max_{v \in \mathcal{V}} \pi_{\mathrm{T}}\left(v \mid x_{\mathrm{T}}, \hat{y}_{<t}\right). \tag{1}$$

Given the resulting fixed dataset $\mathcal{D}_{\text{off}} = \{(x_{\text{S}}, x_{\text{T}}, \hat{y})\}$, the student minimizes the token-averaged negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x_{\text{S}}, x_{\text{T}}, \hat{y}) \sim \mathcal{D}_{\text{off}}} \left[\frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \log \pi_\theta \left(\hat{y}_t \mid x_{\text{S}}, \hat{y}_{<t} \right) \right]. \tag{2}$$

This objective treats each teacher-generated token as a hard target. Off-policy distillation can also retain soft teacher distributions and match them at the same fixed response:

$$\mathcal{L}_{\text{KD}}(\theta) = \mathbb{E}_{(x_{\text{S}}, x_{\text{T}}, \hat{y}) \sim \mathcal{D}_{\text{off}}} \left[\frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} D_{\text{KL}} \Big(\pi_{\text{T}} \left(\cdot \mid x_{\text{T}}, \hat{y}_{<t} \right) \Big\| \, \pi_\theta \left(\cdot \mid x_{\text{S}}, \hat{y}_{<t} \right) \Big) \right], \tag{3}$$

where D_{KL} denotes Kullback-Leibler (KL) divergence.

3.2 On-policy Distillation

On-policy distillation (OPD) provides teacher supervision on student-generated trajectories (Agarwal et al., 2024; Gu et al., 2024). The objective is

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{(x_{\text{S}}, x_{\text{T}}) \sim \mathcal{D}_{\text{in}}} \mathbb{E}_{y \sim \pi_\theta(\cdot \mid x_{\text{S}})} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D_{\text{KL}}(\pi_\theta(\cdot \mid x_{\text{S}}, y_{<t}) \Big\| \, \pi_{\text{T}}(\cdot \mid x_{\text{T}}, y_{<t})) \right], \tag{4}$$

where $\mathcal{D}_{\text{in}}$ is the input-pair dataset. The Jensen–Shannon divergence (JSD) can also be used (see Appendix C.1 for details).

To mitigate the computational and memory overhead of evaluating the full vocabulary $\mathcal{V}$, the tail-bucket approximation can be adopted for the Top-k OPD objective. Let $\mathcal{V}_t^{(k)} \subset \mathcal{V}$ denote the set of the k most probable tokens under the student at step t. The tail-bucket KL divergence is defined as:

$$\begin{aligned} &D_{\text{KL}}^{\mathcal{V}_t^{(k)}} \Big(\pi_\theta(\cdot \mid x_{\text{S}}, y_{<t}) \Big\| \, \pi_{\text{T}}(\cdot \mid x_{\text{T}}, y_{<t}) \Big) \\ &\quad = \sum_{v \in \mathcal{V}_t^{(k)}} \pi_\theta(v \mid x_{\text{S}}, y_{<t}) \log \frac{\pi_\theta(v \mid x_{\text{S}}, y_{<t})}{\pi_{\text{T}}(v \mid x_{\text{T}}, y_{<t})} + \pi_\theta^{\text{tail}} \log \frac{\pi_\theta^{\text{tail}}}{\pi_{\text{T}}^{\text{tail}}}. \end{aligned} \tag{5}$$

where the tail probabilities are given by the remaining mass:

$$\pi^{\text{tail}}(\cdot) = 1 - \sum_{v \in \mathcal{V}_t^{(k)}} \pi(v \mid \cdot). \tag{6}$$

4 Rethinking Generation and Supervision in Visual OPD

We investigate the question raised in Section 1: *Are full-trajectory generation and supervision necessary for improving visual perception through OPD?* We separate a response into a decision token and descriptive tokens, and compare three training strategies that vary which components are generated and supervised. We then examine whether any benefit from distilling the decision position can be reproduced using a hard ground-truth label.

4.1 Separating Decisions from Descriptions

For a multiple-choice question, let $d \in \mathcal{A} \subset \mathcal{V}$ denote a single answer-option token, where $\mathcal{A}$ contains the candidate options, and let $z = (z_1, \ldots, z_m)$ denote the remaining descriptive tokens. A conventional response often takes the form $y = (z, d)$, describing visual evidence before giving the final answer. Under OPD, however, the teacher and student are conditioned on the same student-generated prefix. At the final decision position, sharing an already informative description can make their answer distributions similar, potentially weakening the supervision attributable to the teacher's privileged visual input.

To evaluate decision supervision without a preceding description, we use an *answer-first-then-explain* prompt, p_{exp}, that requests the selected option followed by an explanation. The resulting trajectory is $y = (d, z)$, with $y_1 = d$. We operationally refer to all subsequent explanatory tokens as descriptive tokens. By replacing the KL divergence in Equation 5 with the JSD, the token-wise distillation loss is

$$\ell_t^p(y) = D_{\text{JSD}}^{\mathcal{V}_t^{(k)}}\left(\pi_\theta(\cdot \mid x_{\text{S}}^p, y_{<t}) \,\middle\|\, \pi_{\text{T}}(\cdot \mid x_{\text{T}}^p, y_{<t})\right), \tag{7}$$

where x_{S}^p and x_{T}^p denote student and teacher inputs under prompt p. For an answer-first trajectory with m descriptive tokens, its decision and description losses are

$$\ell_{\text{dec}}(y) = \ell_1^{p_{\text{exp}}}(y), \qquad \ell_{\text{desc}}(y) = \frac{1}{m}\sum_{t=2}^{m+1} \ell_t^{p_{\text{exp}}}(y). \tag{8}$$

The first term matches distributions before any description is available; the second matches distributions along the student's explanatory continuation.

4.2 Experimental Design

We use Qwen3.5-4B (Team, 2026b) in the OPSD setting of Vision-OPD (Yuan et al., 2026). The student receives the original image and multiple-choice question, while the teacher is constructed from the same model with access to question-relevant image crops as privileged visual information. Training uses Vision-OPD-6K for one epoch.

Direct-answer distillation. A direct-answer prompt, p_{dir}, requests only the selected answer-option token. We match teacher and student predictive distributions at this first output position:

$$\mathcal{L}_{\text{direct}}(\theta) = \mathbb{E}_{(x_{\text{S}}, x_{\text{T}})\sim\mathcal{D}_{\text{in}}}\left[D_{\text{JSD}}^{\mathcal{V}_1^{(k)}}\left(\pi_\theta(\cdot \mid x_{\text{S}}^{p_{\text{dir}}}) \,\middle\|\, \pi_{\text{T}}(\cdot \mid x_{\text{T}}^{p_{\text{dir}}})\right)\right]. \tag{9}$$

At this position, the response prefix is empty, and the loss is independent of any sampled answer token. Thus, direct-answer distillation can be computed without rollout.

Full-response OPD. Using p_{exp}, the student generates an answer followed by its explanation, and both components receive teacher supervision. We average the losses within each component and assign the two groups equal weight:

$$\mathcal{L}_{\text{full}}(\theta) = \mathbb{E}_{p_{\text{exp}}}\left[\frac{1}{2}\ell_{\text{dec}}(y) + \frac{1}{2}\ell_{\text{desc}}(y)\right]. \tag{10}$$

This prevents the many descriptive positions from overwhelming the single decision position solely because of response length. This balanced, answer-first variant is distinguished from the original Vision-OPD baseline in our comparison.

Description-only OPD. We retain the answer-first generation protocol but mask the loss at the initial decision position:

$$\mathcal{L}_{\text{desc}}(\theta) = \mathbb{E}_{p_{\text{exp}}}[\ell_{\text{desc}}(y)]. \tag{11}$$

4.3 Are Full Responses Necessary?

Figure 2 compares the three strategies with the base model and original Vision-OPD (Yuan et al., 2026). We report accuracy on HR-Bench 4K/8K (Wang et al., 2025), MME-RealWorld EN/CN (Zhang et al., 2025a), V*Bench (Wu & Xie, 2024), and ZoomBench (Wei et al., 2026).

Both decision-only and description-only supervision retain most of the aggregate improvement. Full-response OPD achieves the highest aggregate accuracy of 70.65% and performs best among the three strategies on three of the six benchmarks. Nevertheless, direct-answer distillation and description-only OPD achieve comparable aggregate accuracies of 69.99% and 70.08%, respectively. Direct-answer distillation also achieves the highest accuracy among the three strategies on ZoomBench and HR-Bench 4K/8K. These results suggest substantial redundancy in full-response supervision in OPD for visual perception: supervising either the decision token or the descriptive tokens alone retains most of the aggregate improvement, while direct-answer distillation additionally eliminates the need to generate descriptive text.

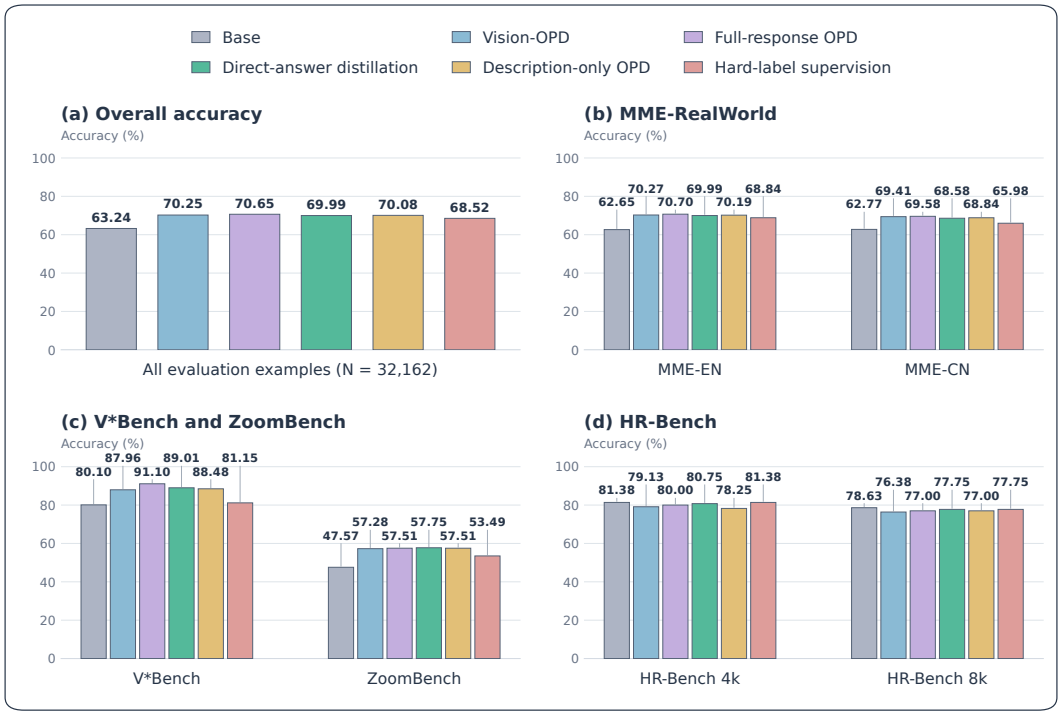

Figure 2: **Comparing generation and supervision strategies for visual OPD. (a)** Sample-weighted accuracy over all 32,162 evaluation examples. **(b–d)** Accuracy on the six visual perception benchmark evaluations.

4.4 Is a Hard Answer Label Sufficient?

To distinguish soft distribution matching from hard answer label supervision, we retain the direct-answer setting and replace the teacher's predictive distribution with the ground-truth option $d^\star$. The corresponding objective is

$$\mathcal{L}_{\text{hard}}(\theta) = -\mathbb{E}_{(x_S, d^\star)\sim\mathcal{D}_{\text{GT}}} \log \pi_\theta(d^\star \mid x_S^{p_{\text{dir}}}), \tag{12}$$

where $\mathcal{D}_{\text{GT}}$ contains the training inputs paired with their ground-truth options.

Correct answer labels provide useful supervision, but do not fully reproduce the improvement from soft teacher targets. Hard-label supervision reaches 68.52%, improving over the base model by 5.28 percentage points but falling 1.47 points below direct-answer distillation. The gap is particularly pronounced on V*Bench (81.15% vs. 89.01%) and ZoomBench (53.49% vs. 57.75%). These results support the value of the teacher's predictive distribution beyond the identity of the correct option.

5 Method

Motivated by the findings in Section 4, we propose **Decision-Only Distillation (DOD)**, which retains soft teacher supervision while reducing distillation to a single response position and a small set of candidate answer tokens. As illustrated in Figure 3, DOD combines rollout-free first-position supervision with target-vocabulary distillation (TVD).

Remark. DOD requires explicit candidate answers during training, but does not restrict inference to multiple-choice questions. Our evaluation on open-ended perception tasks demonstrates that DOD generalizes beyond the multiple-choice format. Open-ended visual questions with reference answers can be converted into multiple-choice format by constructing plausible distractors (Zhang et al., 2025b), offering a potential route to broader training data.

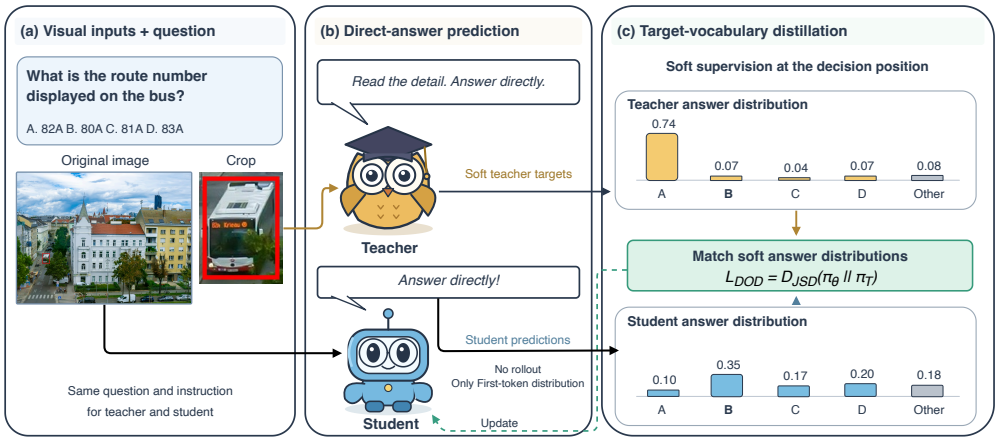

Figure 3: **Overview of DOD.** The student uses the original image, while the teacher receives a question-relevant crop. DOD aligns their first-token soft predictions without rollout, preserving candidate probabilities and a tail bucket. Gradients update only the student.

5.1 Rollout-Free Direct-Answer Distillation

DOD adopts the direct-answer protocol in Section 4.2. Given a multiple-choice question, the prompt p_{dir} requests only the decision token $d \in \mathcal{A}$. Teacher and student predictions are matched at this first output position. The student receives the original image, while the teacher obtains privileged visual information such as a question-relevant crop.

At the first output position, $y_{<1} = \emptyset$, so both predictive distributions depend only on their respective inputs. The loss requires neither a sampled decision token nor a student-generated trajectory. DOD is thus rollout-free.

5.2 Target-Vocabulary Distillation

Rollout-free direct-answer distillation restricts supervision to one position, but the objective in Equation 9 still matches predictions over the general token vocabulary. DOD further concentrates supervision on the question's candidate answer tokens. Specifically, we replace the student-selected Top-k set $\mathcal{V}_t^{(k)}$ in Equation 5 with $\mathcal{A}$, such as $\{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$. Each candidate retains its original probability, and all non-candidate tokens are aggregated into the tail bucket defined in Equation 6.

We use JSD to match the resulting distributions. Denoting its tail-bucket form by $D_{\text{JSD}}^{\mathcal{A}}$, the DOD objective is

$$\mathcal{L}_{\text{DOD}}(\theta) = \mathbb{E}_{(x_{\text{S}}, x_{\text{T}}) \sim \mathcal{D}_{\text{in}}} \left[D_{\text{JSD}}^{\mathcal{A}} \left(\pi_\theta(\cdot \mid x_{\text{S}}^{p_{\text{dir}}}) \,\middle\|\, \pi_{\text{T}}(\cdot \mid x_{\text{T}}^{p_{\text{dir}}}) \right) \right]. \tag{13}$$

The expanded JSD formulation is provided in Appendix C.1. Unlike Top-k selection, target-vocabulary selection explicitly focuses on candidate answer tokens, providing more targeted supervision during distillation (Appendix C.2).

6 Experiments

6.1 Experimental Setup

Model Training. We conduct experiments with Qwen3.5-4B and Qwen3.5-9B (Team, 2026b) under the OPSD setting. The teacher and student are initialized from the same pretrained model, and the teacher parameters are updated using an exponential moving average (EMA) of the student parameters. The student receives the original image, while the teacher receives a question-relevant image crop as privileged visual information. We train on Vision-OPD-6K for one epoch.

Table 1: **Comparison on visual perception benchmarks.** We report accuracy (%) for each model. Average denotes the arithmetic mean across the six evaluations. The best results are highlighted in **bold**, and the second-best are underlined. The results of Open-Source MLLMs are sourced from Yuan et al. (2026).

Model	Param Size	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
Open-Source MLLMs								
MiMo-VL-RL	7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
MiniCPM-V-4.5	9B	70.68	42.60	69.63	61.50	62.65	61.64	61.45
GLM-4.6V	106B	86.91	50.06	**82.13**	78.88	65.57	65.62	71.53
Kimi-K2.6	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Base Models								
Qwen3.5	4B	80.10	47.57	81.38	78.63	62.65	62.77	68.85
Qwen3.5	9B	85.86	51.60	81.38	77.62	70.16	67.04	72.28
Vision-OPD	4B	87.96	57.28	79.13	76.38	70.27	69.41	73.41
Vision-OPD	9B	85.86	**60.59**	82.00	78.25	70.08	69.31	74.35
DOD (Ours)								
DOD	4B	88.48	59.76	80.50	**79.88**	70.99	68.18	74.63
DOD	9B	**92.15**	59.76	81.38	78.88	**71.40**	**70.14**	**75.62**

Table 2: **Results on Do You See Me (DYSM).** We report accuracy (%) for each model. Avg. Acc. denotes micro-averaged accuracy over all examples within the corresponding 2D or 3D group, and Overall denotes micro-averaged accuracy over the entire benchmark. Within each model size, the best and second-best results are highlighted in **bold** and underlined, respectively.

	2D Visual Perception Tasks								3D Visual Perception Tasks						
Model	Figure Ground	Visual Spatial	Color Disamb.	Shape Disamb.	Letter Disamb.	Form Const.	Visual Closure	Avg. Acc.	Color Disamb.	Letter Disamb.	Shape Disamb.	Form Const.	Visual Spatial	Avg. Acc.	Overall
Qwen3.5-4B	**34.44**	66.87	**89.95**	27.50	15.56	**85.93**	55.36	63.72	72.50	9.38	85.83	**28.75**	25.00	48.79	60.89
Vision-OPD-4B	25.56	68.11	85.78	29.58	15.56	84.81	65.48	63.91	**92.50**	2.08	88.33	**28.75**	32.50	54.03	62.04
DOD-4B	27.78	**75.06**	88.73	**30.83**	**18.52**	84.07	**71.43**	**67.93**	**92.50**	**16.67**	**89.17**	**28.75**	**36.25**	**57.66**	**65.98**
Qwen3.5-9B	**38.89**	71.22	93.14	29.58	**25.93**	85.93	68.45	68.12	78.33	**13.54**	89.17	41.25	31.25	54.84	65.60
Vision-OPD-9B	32.22	80.02	**95.83**	32.50	23.70	**94.07**	72.02	73.22	83.33	4.17	90.00	67.50	**40.00**	60.08	70.72
DOD-9B	32.22	**80.27**	95.34	**37.08**	**25.93**	93.70	**75.60**	**74.11**	**84.17**	8.33	**90.83**	**72.50**	30.00	**60.48**	**71.53**

Baselines. We compare DOD with open-source models of various sizes, including MiMo-VL-7B-RL (Team, 2025), MiniCPM-V-4.5 (Yu et al., 2025b), GLM-4.6V (Team, 2026a), and Kimi-K2.6 (Moonshot AI, 2026). To assess the gains from distillation, we include the corresponding base models, Qwen3.5-4B and Qwen3.5-9B, and their counterparts trained with Vision-OPD, denoted as Vision-OPD-4B and Vision-OPD-9B (Yuan et al., 2026).

Benchmarks. We evaluate visual perception on V* Bench and ZoomBench, which focus on recognizing fine-grained visual details, and HR-Bench and MME-RealWorld, which assess perception in high-resolution images. We further evaluate on Do You See Me (Kanade & Ganu, 2026), an open-ended benchmark that comprehensively assesses visual perception abilities across multiple dimensions beyond the multiple-choice format.

6.2 Experimental results

Performance on visual perception benchmarks. As shown in Table 1, DOD improves average accuracy from 68.85% to 74.63% for Qwen3.5-4B and from 72.28% to 75.62% for Qwen3.5-9B, outperforming their Vision-OPD counterparts by 1.22 and 1.27 percentage points, respectively. DOD-9B achieves the highest average accuracy among the compared models, including much larger open-source MLLMs, e.g., GLM-4.6V and Kimi-K2.6, and obtains the best results on V* Bench and MME-RealWorld EN/CN. DOD-4B also achieves the highest accuracy on HR-Bench 8K. These results demonstrate that distilling a single decision position can effectively improve visual perception without requiring rollout during training.

Generalization to diverse perceptual tasks. Table 2 further evaluates DOD on DYSM. DOD improves overall accuracy from 60.89% to 65.98% at 4B and from 65.60% to 71.53% at 9B, exceeding Vision-OPD at both scales. It also achieves the highest average accuracy within each model size on both 2D and 3D tasks. These results suggest that the benefits extend beyond fine-grained recognition in natural images to a broader range of perceptual abilities.

Table 3: **Training resource comparison.** Qwen3.5-4B with batch size 96. Step times are measured using four H800 GPUs. Vision-OPD uses 8 rollouts per prompt and top-100 distillation.

Metric	Vision-OPD	DOD
H800 GPUs required	≥ 4	1
Time/step (s)	$\sim$730	$\sim$66
Peak CPU RAM (GB)	$\sim$590	$\sim$110

Computational efficiency. As shown in Table 3, DOD can run on a single H800 GPU, whereas Vision-OPD requires at least four in our setting. By eliminating rollouts and on-policy teacher scoring, and restricting vocabulary matching to candidate tokens and an aggregated tail bucket, DOD achieves an approximately $11.1\times$ training-step speedup and reduces peak CPU memory usage by 81.4%.

6.3 Ablation Study

Why target-vocabulary distillation? We examine the student distribution at the decision position under the direct-answer prompt. As shown in Figure 4, candidate answers A–D account for 75.92% of the probability mass on average, while generic tokens such as *The*, *Looking*, and *To* alone account for 22.46%. Conventional top-k distillation thus spends supervision on distinctions among tokens unrelated to the final decision. In contrast, target-vocabulary distillation retains candidate tokens individually and merges all others into a single tail bucket. When all candidate tokens are retained in the top-k set, this removes the conditional matching term over non-candidate entries (Appendix C.2).

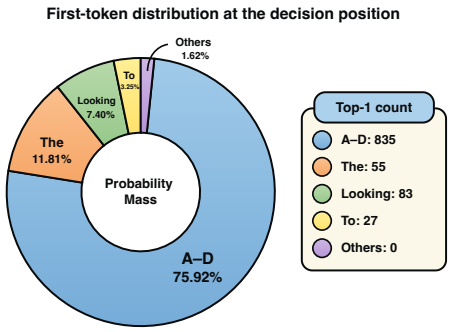

Figure 4: **First-token probability analysis.** Statistics are computed over 1,000 training examples using Qwen3.5-4B.

Effect of target-vocabulary distillation. Table 4 compares direct-answer distillation with and without target-vocabulary distillation on Qwen3.5-4B. Target-vocabulary distillation achieves comparable performance on HR-Bench 4K and MME-RealWorld CN, while improving ZoomBench and HR-Bench 8K by approximately 2 percentage points. It also improves the overall average from 73.97% to 74.63%. Meanwhile, it reduces the number of individually matched vocabulary entries from k to the number of candidate answer tokens, and concentrates fine-grained distillation on decision-relevant alternatives.

Table 4: **Ablation study on target-vocabulary distillation.** Both variants use Qwen3.5-4B. The best results are highlighted in **bold**.

Method	TVD	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average
Direct-answer Distillation	✗	**89.01**	57.75	**80.75**	77.75	69.99	**68.58**	73.97
DOD	✓	88.48	**59.76**	80.50	**79.88**	**70.99**	68.18	**74.63**

7 Conclusion

In this work, we revisit on-policy distillation for visual perception and find that, when training on multiple-choice perception tasks, supervising a single decision token achieves downstream perception performance comparable to full-response supervision, while hard-label supervision fails to reproduce these gains. Motivated by these findings, we propose decision-only distillation (DOD), a rollout-free approach that combines direct-answer distillation with target-vocabulary distillation. DOD outperforms standard OPD across diverse perception benchmarks while substantially reducing training time and resource requirements.

AI USE STATEMENT

Generative AI tools were used to assist in checking the correctness of code and mathematical derivations, as well as to improve the readability and clarity of the manuscript. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide a link to our code and detailed implementation settings in the paper to facilitate reproduction of our experiments.

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*, volume 2024, pp. 21246–21263, 2024.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. Opd-v: Visual on-policy self-distillation with modality balance, 2026. URL https://arxiv.org/abs/2608.05131.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

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.

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8567–8608, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.387. URL https://aclanthology.org/2026.acl-long.387/.

Aditya Sanjiv Kanade and Tanuja Ganu. Do you see me : A multidimensional benchmark for evaluating visual perception in multimodal LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7285–7326, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.343. URL https://aclanthology.org/2026.eacl-long.343/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. Sead: Competence-aware on-policy distillation via entropy-guided supervision. *arXiv preprint arXiv:2606.28562*, 2026.

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

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

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

Zhishuai Liu, Xingzi Xu, Mehmet Saygin Seyfioglu, Pan Xu, and Karim Bouyarmane. Extremely sparse supervision incentivizes reasoning ability, 2026b. URL https://arxiv.org/abs/2609.04565.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Moonshot AI. Kimi k2.6. https://huggingface.co/moonshotai/Kimi-K2.6, 2026.

Huanxin Sheng, Zhiling Ye, Haonan Wang, Jian Wang, Jinjie Gu, and Jian Kang. 1% of tokens can be enough: On gradient estimation in on-policy distillation, 2026. URL https://arxiv.org/abs/2609.24432.

GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026a. URL https://arxiv.org/abs/2507.01006.

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

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

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception before supervision: Self-contained visual distillation from counterfactual blind spots. *arXiv preprint arXiv:2608.09931*, 2026.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026a.

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, pp. 7907–7915, 2025.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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)*, pp. 13084–13094. IEEE, 2024.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025a. URL https://arxiv.org/abs/2503.14476.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025b. URL https://arxiv.org/abs/2509.18154.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 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, 2026. URL https://arxiv.org/abs/2605.18740.

Yaocheng Zhang, Jiajun Chai, Yuqian Fu, Songjun Tu, Xiaohan Wang, Wei Lin, Guojun Yin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Are full rollouts necessary for on-policy distillation? *arXiv preprint arXiv:2605.31490*, 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*, volume 2025, pp. 89655–89701, 2025a.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29580–29590. IEEE, 2025b.

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, 2026. URL https://arxiv.org/abs/2601.18734.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

Appendix

A Implementation Details

Training. All experiments were conducted on NVIDIA H800 GPUs. Both the student and teacher were initialized from the official Qwen3.5 checkpoint and trained on Vision-OPD-6K for one epoch with batch size 96. We optimized the student using AdamW with a learning rate of 2×10^{-6}, linearly warmed up over the first 10 optimization steps and held constant thereafter. The teacher was updated using an exponential moving average (EMA) of the student parameters: $\theta_{\mathrm{T}} \leftarrow (1-\alpha)\theta_{\mathrm{T}} + \alpha\theta$, with $\alpha = 0.05$. For a fair comparison, we retrained the Vision-OPD baseline using the same training hyperparameters as DOD.

Evaluation. For all benchmarks except Do You See Me, we used rule-based answer extraction and matched the extracted answers against the ground-truth answers. For Do You See Me, we used an LLM-as-a-judge protocol with GPT-OSS-120B to assess response correctness.

B Prompt

We provide the training instructions and evaluation prompt templates used in our experiments.

B.1 Training Instructions

The following instructions are appended to the training question and answer choices.

Direct-answer instruction (p_{dir})

```
Answer with the option's letter from the given choices directly.
```

Answer-then-explain instruction (p_{exp})

```
Answer with the option's letter from the given choices directly.
Then explain your answer.
```

B.2 Evaluation Prompts

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:
```

Do You See Me

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

C Distillation Objective and Theoretical Analysis

C.1 Jensen–Shannon Divergence

Let q and p denote the teacher and student distributions, respectively, over the same set of vocabulary entries, including the aggregated tail bucket. The weighted Jensen–Shannon divergence is defined as

$$\mathrm{JSD}_\beta(q\|p) = \beta D_{\mathrm{KL}}(q\|m) + (1-\beta)D_{\mathrm{KL}}(p\|m), \qquad m = \beta q + (1-\beta)p, \tag{14}$$

where

$$D_{\mathrm{KL}}(q\|p) = \sum_j q_j \log \frac{q_j}{p_j}. \tag{15}$$

We use $\beta = 0.5$ throughout our experiments, yielding

$$\mathrm{JSD}_{0.5}(q\|p) = \frac{1}{2}\sum_j q_j \log \frac{2q_j}{q_j+p_j} + \frac{1}{2}\sum_j p_j \log \frac{2p_j}{q_j+p_j}. \tag{16}$$

C.2 Target-Vocabulary Distillation as Distribution Coarsening

We analyze a single decision position and omit the conditioning input for clarity. Let $\mathcal{V}$ be the full vocabulary, $\mathcal{C}$ the candidate answer tokens, and $\mathcal{K}$ the entries retained by conventional top-k distillation. Both teacher and student probabilities are evaluated on the same retained set. We assume $\mathcal{C} \subseteq \mathcal{K}$ and define

$$\mathcal{I} = \mathcal{K} \setminus \mathcal{C}, \qquad \mathcal{B} = \mathcal{V} \setminus \mathcal{K}, \qquad \mathcal{R} = \mathcal{V} \setminus \mathcal{C} = \mathcal{I} \cup \mathcal{B}. \tag{17}$$

Here, $\mathcal{I}$ contains individually retained non-candidate tokens, $\mathcal{B}$ is the original top-k tail, and $\mathcal{R}$ is the target-vocabulary tail. For any group $\mathcal{A}$, write $q_{\mathcal{A}} = \sum_{v\in\mathcal{A}} q_v$ and $p_{\mathcal{A}} = \sum_{v\in\mathcal{A}} p_v$.

KL decomposition. With tail aggregation, the two KL objectives are

$$\mathcal{L}^{\mathrm{KL}}_{\text{top-}k} = \sum_{c\in\mathcal{C}} q_c \log \frac{q_c}{p_c} + \sum_{i\in\mathcal{I}} q_i \log \frac{q_i}{p_i} + q_{\mathcal{B}} \log \frac{q_{\mathcal{B}}}{p_{\mathcal{B}}}, \tag{18}$$

$$\mathcal{L}^{\mathrm{KL}}_{\mathrm{TVD}} = \sum_{c\in\mathcal{C}} q_c \log \frac{q_c}{p_c} + q_{\mathcal{R}} \log \frac{q_{\mathcal{R}}}{p_{\mathcal{R}}}. \tag{19}$$

Define conditional distributions over the non-candidate entries distinguished by top-k, namely $\mathcal{I} \cup \{\mathcal{B}\}$:

$$\bar{q} = \left(\left\{ \frac{q_i}{q_\mathcal{R}} \right\}_{i \in \mathcal{I}}, \frac{q_\mathcal{B}}{q_\mathcal{R}} \right), \qquad \bar{p} = \left(\left\{ \frac{p_i}{p_\mathcal{R}} \right\}_{i \in \mathcal{I}}, \frac{p_\mathcal{B}}{p_\mathcal{R}} \right). \tag{20}$$

Expanding their conditional KL gives

$$q_\mathcal{R} D_{\mathrm{KL}}(\bar{q} \| \bar{p}) = \sum_{i \in \mathcal{I}} q_i \log \frac{q_i / q_\mathcal{R}}{p_i / p_\mathcal{R}} + q_\mathcal{B} \log \frac{q_\mathcal{B} / q_\mathcal{R}}{p_\mathcal{B} / p_\mathcal{R}} \tag{21}$$

$$= \sum_{i \in \mathcal{I}} q_i \log \frac{q_i}{p_i} + q_\mathcal{B} \log \frac{q_\mathcal{B}}{p_\mathcal{B}} - q_\mathcal{R} \log \frac{q_\mathcal{R}}{p_\mathcal{R}}. \tag{22}$$

Therefore,

$$\boxed{\mathcal{L}^{\mathrm{KL}}_{\text{top-}k} = \mathcal{L}^{\mathrm{KL}}_{\mathrm{TVD}} + q_\mathcal{R} D_{\mathrm{KL}}(\bar{q} \| \bar{p}).} \tag{23}$$

Extension to JSD. The same decomposition applies to both KL components of JSD. Let

$$m_\mathcal{R} = \beta q_\mathcal{R} + (1 - \beta) p_\mathcal{R}, \qquad \alpha = \frac{\beta q_\mathcal{R}}{m_\mathcal{R}}, \qquad \bar{m} = \alpha \bar{q} + (1 - \alpha) \bar{p}. \tag{24}$$

It follows that

$$\mathcal{L}^{\mathrm{JSD}}_{\text{top-}k} - \mathcal{L}^{\mathrm{JSD}}_{\mathrm{TVD}} = \beta q_\mathcal{R} D_{\mathrm{KL}}(\bar{q} \| \bar{m}) + (1 - \beta) p_\mathcal{R} D_{\mathrm{KL}}(\bar{p} \| \bar{m}) \tag{25}$$

$$= m_\mathcal{R} \, \mathrm{JSD}_\alpha(\bar{q} \| \bar{p}) \geq 0. \tag{26}$$

Thus, target-vocabulary distillation removes the conditional matching term within the non-candidate group while preserving candidate-specific matching and supervision of the total non-candidate probability mass.