VEGO: VISUAL EVIDENCE-GATED ON-POLICY LEARNING FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models still struggle with fine-grained visual understanding, particularly when answering questions that depend on small or inconspicuous image regions. Existing remedies either invoke cropping tools at inference, incurring considerable overhead from repeated tool calls and visual processing, or internalize region-focused perception through on-policy self-distillation, where a teacher observes a question-relevant crop while the student sees the full image. However, such distillation treats every generated token uniformly, although privileged visual evidence is informative for only a small subset of content-bearing tokens. Moreover, responses often continue unnecessarily after the answer has already been determined, resulting in redundant generation and inefficient inference. To address these limitations, we propose VeGo, which contrasts the teacher's next-token predictions under an evidence-rich crop and a degraded, evidence-blind view to estimate each token's reliance on fine-grained visual evidence. This enables distillation to prioritize evidence-sensitive tokens while preserving the overall loss magnitude. To reduce redundant generation, VeGo also identifies when the necessary visual evidence has been sufficiently encoded in the generated text and applies a length reward only to the subsequent region, leaving the evidence-writing prefix unaffected. Experiments across V*, ZoomBench, Visual Probe, HR-Bench 4K and HR-Bench 8K show that VeGo achieves the highest average accuracy among open-source, agentic, and distillation baselines while producing the shortest responses.

1 INTRODUCTION

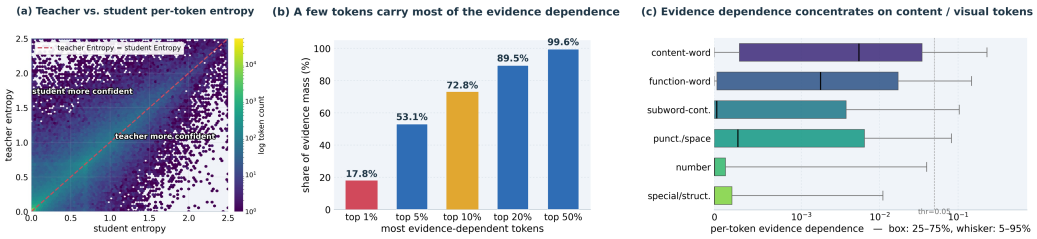


Figure 1: **Dependence on visual evidence is sparse and concentrated.** (a) Per-token predictive entropy of the teacher against the student: most tokens lie near the diagonal with comparable confidence, and only a minority fall below it. (b) Cumulative share of total per-token evidence dependence — the divergence between an evidence-seeing and an evidence-blind teacher — carried by the most evidence-dependent tokens. (c) Distribution of per-token evidence dependence by token category: dependence concentrates on content-bearing and visually grounded words, is lower on function words and punctuation, and is negligible for numbers and structural tokens.

Despite rapid advances in multimodal large language models (MLLMs) Grattafiori et al. (2024); Yang et al. (2024); Liu et al. (2024); Bai et al. (2025); GLM-V Team et al. (2026); Qwen Team (2026), fine-grained visual understanding remains a persistent challenge Wu & Xie (2024); Wang et al. (2025d), particularly when answering questions that depend on small or inconspicuous image regions. To address this limitation, recent approaches equip MLLMs with visual cropping tools,

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.

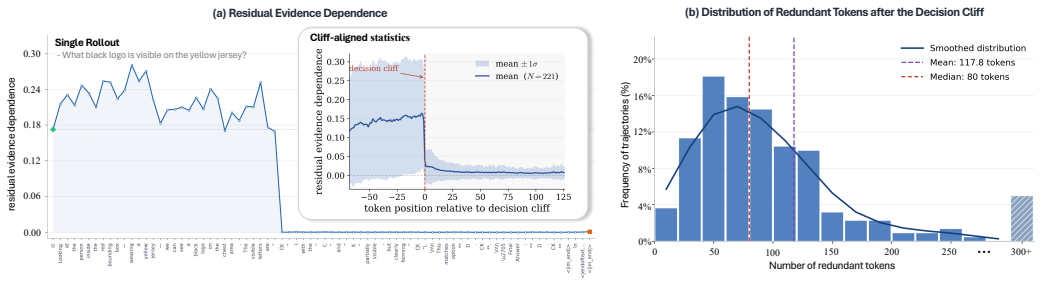


Figure 2: **Residual evidence dependence reveals substantial post-decision redundancy.** (a) in a representative rollout, residual evidence dependence collapses when the answer is committed and remains near zero throughout the subsequent tail. Trajectories aligned at their decision cliffs exhibit the same sharp transition. (b) the resulting post-cliff tails contain a median of 80 redundant tokens, demonstrating considerable room for length reduction after the visual evidence has been written.

allowing them to iteratively localize and inspect task-relevant regions at higher resolution during inference Zheng et al. (2025); Lai et al. (2025). While effective, these methods rely on multi-step tool interactions and lengthy reasoning trajectories, resulting in substantial token overhead and inefficient inference. More recently, Vision-OPD Yuan et al. (2026) offers a tool-free alternative that internalizes region-focused perception through on-policy self-distillation Zhao et al. (2026): a teacher model observes a question-relevant crop, whereas the student receives only the full image, enabling the student to acquire fine-grained visual reasoning capabilities without explicitly invoking cropping tools at inference time.

Despite its effectiveness, Vision-OPD overlooks the uneven value of privileged visual evidence across a reasoning trajectory. Only a small fraction of generated tokens depend on the cropped region, and only at these tokens does the teacher hold a clear advantage. Uniform distillation thus weakens supervision on visually critical tokens while introducing noise elsewhere. Moreover, once sufficient evidence has been incorporated into the generated prefix, the answer is already determined and further reasoning no longer benefits from the crop. Yet full-trajectory distillation continues to imitate this redundant suffix, increasing inference overhead. Both limitations arise because existing distillation methods identify neither where visual evidence matters nor when it has been fully conveyed.

To solve these two issues, we measure when the teacher's privileged visual input is actually informative. Specifically, we compare its next-token predictions under a clean crop and a degraded, evidence-blind version of the same region. Their divergence provides a measurement of dependence on fine-grained visual evidence (evidence dependence): viewed token by token, it reveals where distillation should be concentrated; tracked along the trajectory, it indicates when the relevant evidence has been fully incorporated and further reasoning becomes redundant.

Figure 1 (a) shows that most tokens lie near the diagonal, indicating comparable predictive confidence between the student and teacher. Only a minority exhibit substantially greater teacher confidence, suggesting that the benefit of the privileged crop may be localized rather than uniform. Motivated by this pattern, we examine the distribution of the evidence-dependence divergence introduced above. Figure 1 (b) reveals a strong concentration: the top 5% of tokens account for 53.1% of total evidence dependence, and the top 10% account for 72.8%. As shown in Figure 1 (c), This concentration also follows a semantic pattern: content-bearing and visually grounded tokens exhibit greater evidence dependence, whereas function words and punctuation generally show lower dependence, with numbers and structural tokens contributing little. Together, these observations motivate concentrating distillation on evidence-critical tokens, strengthening supervision where the teacher's privileged view is informative while limiting noise elsewhere.

Read cumulatively along the trajectory, the same diagnostic reveals when the visual evidence has been exhausted. As shown in Figure 2 (a), residual evidence dependence in an individual response remains high while the model is resolving the answer and collapses to nearly zero once the answer is committed. Aligning trajectories at this transition, Figure 2 (b) reveals a consistent decision cliff separating an evidence-active prefix from a near-zero tail. Figure 2 (c) further shows that this tail is

substantial in length, containing a median of 80 tokens and a mean of 117.8 tokens, while accounting for a median of 65.6% of each trajectory. Beyond the cliff, the evidence-seeing and evidence-blind teachers agree, indicating that no additional visual evidence is being incorporated. We therefore leave the evidence-writing prefix untouched and apply an accuracy-gated length objective only to the post-cliff region.

Both mechanisms follow directly from the evidence-dependence diagnostic, which determines how each stage of a trajectory should be supervised. Before the decision cliff, where the teacher still carries visual evidence, we distill, and reweight this distillation toward the tokens the diagnostic identifies as evidence-critical. After the cliff, where the two teachers agree and no visual evidence remains to transfer, we instead reward brevity. In this way, the same signal guides both stages: it indicates where evidence should be distilled and when generation can be shortened.

We summarize our contributions as follows:

- We introduce an evidence-dependence diagnostic for VLM on-plolicy learning and show that it is concentrated on content/vision tokens and cumulatively marked by a decision cliff.
- Using the per-token reading of the diagnostic, we propose an evidence-reweighted on-policy distillation that focuses supervision on the tokens that truly require fine-grained perception.
- Using the cumulative reading of the diagnostic, we locate the decision cliff and compress the evidence-free tail with an accuracy-gated length reward, motivated by the observation that distillation is vacuous once the two teachers agree.

2 Related Work

Fine-Grained Visual Perception in MLLMs. MLLMs extend the strong reasoning capabilities of language models to visual understanding Grattafiori et al. (2024); Yang et al. (2024); Li et al. (2024); Zhu et al. (2025). Recent approaches improve fine-grained perception by actively acquiring visual evidence during reasoning Zheng et al. (2025); Lai et al. (2025); Wang et al. (2025b); Wei et al. (2026a; 2025); Wu et al. (2025); Zhang et al. (2025a;b). DeepEyes learns to invoke cropping and searching tools Zheng et al. (2025), while Thyme generates programs for pixel-level image manipulation Zhang et al. (2025c). Training-free methods localize and enlarge salient regions through search or attention Fu et al. (2025); Liu et al. (2025); Peng et al. (2025); Shen et al. (2025), but often introduce substantial inference overhead. Other methods employ structured templates Wang et al. (2025a;c), latent visual reasoning Yuan et al. (2025), or reinforcement learning for more efficient tool use Wei et al. (2026a). In contrast, Vision-OPD Yuan et al. (2026) internalizes regional visual enhancement through self-distillation. Our method further concentrates this supervision on tokens whose predictions depend on fine-grained visual evidence.

On-Policy Distillation. Knowledge distillation transfers teacher knowledge to a student, while privileged information provides additional training-only signals. Conventional off-policy distillation may suffer from trajectory mismatch, whereas on-policy distillation supervises student-generated states and better aligns teacher feedback with the current policy Agarwal et al. (2024); DeepSeek-AI (2026); Hou et al. (2026); Lu & Thinking Machines Lab (2025). Related approaches include reward-verified self-training Gulcehre et al. (2023); Zelikman et al. (2022), context-induced soft distillation Qi et al. (2024), SDPO Hübotter et al. (2026), OPDS Zhao et al. (2026), and Uni-OPD Hou et al. (2026). OPSD constructs dense token-level targets from privileged context without requiring a separate stronger teacher. In the multimodal setting, Vision-OPD Yuan et al. (2026) follows this paradigm by distilling predictions conditioned on enhanced regional evidence into an original-view student. Our method extends it by contrasting evidence-preserving and evidence-degraded teachers to estimate token-level visual relevance and selectively allocate on-policy supervision.

3 Method

3.1 Preliminaries: On-Policy Self-Distillation for Vision-Language Models

In OPSD, the student p_S and the teacher p_T share one architecture; the teacher is an exponential moving average (EMA) copy of the student, updated slowly so that the distillation target stays stable

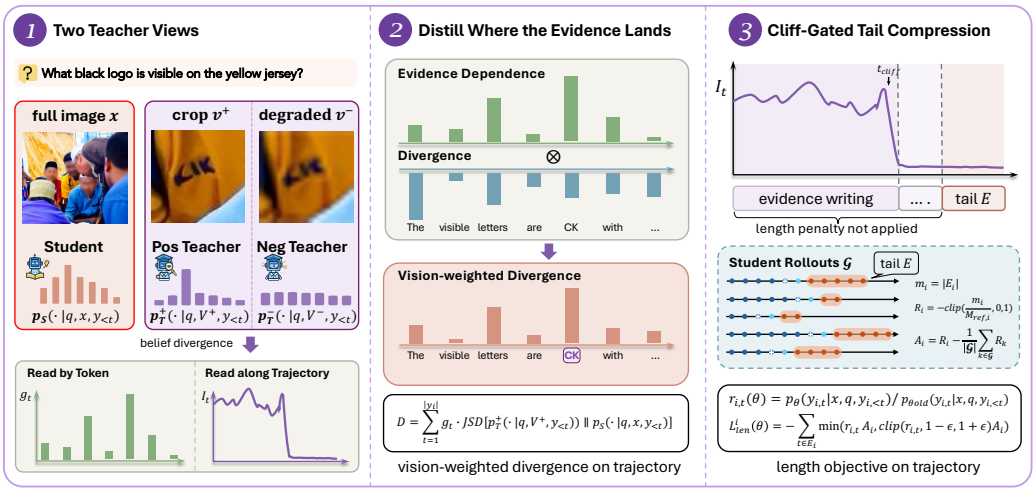


Figure 3: **Method overview.** (1) A student sees the full image and EMA teacher sees a clean or degraded crop. (2) Per-token dependence reweights the teacher–student divergence toward evidence-critical tokens. (3) Residual dependence marks the decision cliff ; the tail beyond it gets no distillation gradient and is compressed by a GRPO length objective.

yet tracks the student's ability. The two receive asymmetric visual inputs of the question q: the student sees the full image x, whereas the teacher sees a clean crop v^+ of the question-relevant region. We additionally define an evidence-blind teacher view v^-, obtained by down-sampling v^+ so that the fine detail is destroyed while the global layout is preserved. We denote the teacher's next-token distributions under the two views as $p_T^+(\cdot \mid v^+, q, y_{<t})$ and $p_T^-(\cdot \mid v^-, q, y_{<t})$.

For each (x, v^+, q) and the trajectory y generated by student, On-Policy Self-Distillation Zhao et al. (2026) transfers the teacher's fine-grained perception with

$$D(p_T^+ \| p_S)(y \mid x, v^+, q) = \frac{1}{|y|} \sum_{t=1}^{|y|} d_t(y \mid x, v^+, q), \tag{1}$$

where

$$d_t(y \mid x, v^+, q) = JSD_{0.5}\Big[sg\big(p_T^+(\cdot \mid v^+, q, y_{<t})\big) \,\Big\|\, p_S(\cdot \mid x, q, y_{<t}) \Big] \tag{2}$$

is the per-token generalized Jensen–Shannon divergence. This applies equal supervision to every token and inherits the full length of each trajectory. The two parts below remove these two inefficiencies, each driven by one reading of a single evidence-dependence signal: the disagreement between the evidence-seeing teacher p_T^+ and the evidence-blind teacher p_T^-.

3.2 EVIDENCE-SENSITIVITY REWEIGHTED DISTILLATION

Definition 1 (Per-token evidence dependence g_t). Given the sampled prefix $y_{<t}$, the per-token evidence dependence is the divergence between the two teachers' next-token beliefs,

$$g_t \;=\; JSD_{0.5}\Big[p_T^+(\cdot \mid v^+, q, y_{<t}) \,\Big\|\, p_T^-(\cdot \mid v^-, q, y_{<t}) \Big]. \tag{3}$$

g_t is large when producing the next token genuinely requires resolving the fine detail, and $g_t \approx 0$ when the two teachers agree, i.e. seeing the evidence confers no advantage at that position.

As Fig 1 shows, g_t is highly concentrated: a small fraction of content- and vision-grounded tokens carries most of it. We therefore use g_t as a per-token weight so that supervision concentrates on evidence-critical tokens. The vision-weighted divergence on trajectory y is

$$D(p_T^+ \| p_S)(y \mid x, v^+, q) = \frac{\gamma}{|y|} \sum_{t=1}^{|y|} g_t \cdot d_t(y \mid x, v^+, q), \tag{4}$$

where $sg(\cdot)$ is the stop-gradient operator. Naively normalizing the token weights by their sum $\sum_t g_t$ can excessively amplify individual token-level distillation terms when the total weight mass is small, thereby destabilizing training. We instead adopt a divergence-sum normalization that preserves the total loss magnitude while redistributing it across tokens

$$\gamma = sg\Big[\sum_{t=1}^{|y|} d_t(y \mid x, v^+, q) / \Big(\sum_{t=1}^{|y|} g_t \cdot d_t(x, v^+, q, y) + \varepsilon\Big)\Big], \tag{5}$$

where ε is a small positive constant added for numerical stability. The overall loss minimizes the expected weighted divergence over student rollouts $\{y\}_{i=1}^{G}$,

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{(x,v^+,q)\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G} D\big(p_T^+ \| p_S\big)(y_i \mid x, v^+, q)\right]. \tag{6}$$

3.3 EVIDENCE-COMPLETION LENGTH COMPRESSION

Definition 2 (Residual evidence dependence I_t). Let ϕ be a fixed decision probe—a short suffix that elicits the answer (e.g. *"The correct answer is: **"*)—and let $\oplus$ denote concatenation. The residual evidence dependence measures the two teachers' disagreement about the decision after the prefix $y_{<t}$ has been committed,

$$I_t = JSD_{\alpha=0.5}\Big[\, p_T^+(\cdot \mid v^+, q,\, y_{<t} \oplus \phi) \;\Big\|\; p_T^-(\cdot \mid v^-, q,\, y_{<t} \oplus \phi) \,\Big]. \tag{7}$$

I_t acts as a ruler for how much decisive evidence remains to be written: while I_t is high the decision still hinges on seeing the crop; once $I_t \to 0$ the answer is already determined by the text generated so far and later tokens are irrelevant to "seeing clearly".

Decision cliff. As Figure 2 shows, along a trajectory I_t stays high while the model is resolving the answer and then collapses sharply. We locate the *decision cliff* t_{cliff} as the position of steepest sustained descent of the I_t profile—the point beyond which I_t stays near zero. Everything up to and including t_{cliff} is the evidence-writing prefix on which the answer is being determined, whereas the tokens after it form a decision-redundant suffix. We define the compression region as the suffix only,

$$E_i = \big\{\, t \,:\, t > \text{sent-end}(t_{\text{cliff}}) \text{ in } y_i \,\big\}, \tag{8}$$

i.e. from the end of the sentence containing t_{cliff} to the end of y_i. Beyond the cliff the evidence-seeing and evidence-blind teachers agree, so the teacher carries no evidence to distill there and the only remaining preference is brevity. We optimize this region with reinforcement rather than distillation. Crucially, the objective below is defined only for $t \in E_i$: the evidence-writing prefix $t \le t_{\text{cliff}}$ receives no gradient from it, so the tokens that establish the answer are left untouched and accuracy is preserved while the redundant tail is shortened.

Group-relative length reward. We apply the length signal exclusively to tokens in the compression region E_i. For each trajectory, let $m_i = |E_i|$ denote the number of tokens in this region. Thus, the scalar reward is determined by the length of E_i, rather than by the length of the full response. We define the clipped, reference-normalized regional length reward as

$$R_i = -\operatorname{clip}\left(\frac{m_i}{M_{\text{ref},i}}, 0, 1\right), \tag{9}$$

where $M_{\text{ref},i}$ is a per-example reference length. The reference length may be a constant or difficulty-adaptive, relaxing the budget only for longer or harder questions:

$$M_{\text{ref},i} = m_{\text{base}} \cdot \max\left(1, \frac{L_{\text{ref},i}}{\text{median}\{L_{\text{ref},j}\}_{j=1}^{D}}\right), \tag{10}$$

where $L_{\text{ref},i}$ is the reference length for question i. We then convert this scalar reward into a group-relative advantage, which is broadcast to every suffix token via the indicator $\mathbf{1}[t \in E_i]$. For the group G of trajectories generated from the same prompt,

$$A_i = R_i - \frac{1}{|G|}\sum_{k\in G} R_k, \qquad \hat{A}_{i,t} = A_i\, \mathbf{1}[t \in E_i]. \tag{11}$$

Table 1: Comparison on visual perception benchmarks. We report accuracy (%) for each model. Our method achieved the best performance among all methods except the closed-source models.

Model	Param. Size	V*	ZoomBench	Visual Probe	HR-Bench 4K	HR-Bench 8K	Average
Closed-Source Models							
GPT-5.2	–	79.06	50.89	–	81.12	78.38	–
GPT-5.4	–	76.96	52.66	–	84.00	77.88	–
Gemini-3.1-Pro	–	87.96	61.18	–	89.63	86.88	–
Gemini-3.5-Flash	–	89.01	61.42	–	89.12	86.62	–
Open-Source Baselines							
Qwen3.5-9B	9B	85.34	51.60	39.87	84.25	79.75	68.16
Qwen3.5-27B	27B	82.20	57.63	31.26	82.00	76.50	65.92
Qwen3.5-Flash	35B	80.63	51.95	32.30	80.88	74.38	64.03
Qwen3.5-Plus	397B-A3B	85.86	58.58	33.94	83.12	75.50	67.40
Kimi-K3	2.8T	87.43	57.16	42.52	82.00	78.25	69.47
"Thinking-with-Images" Agentic Models							
DeepEyes	7B	85.86	46.51	43.25	75.13	72.63	64.68
SenseNova-MARS	8B	92.15	47.81	54.74	83.13	78.38	71.24
Distillation Models							
ZwZ	8B	91.62	57.51	50.79	84.38	82.38	73.34
Vision-OPD-9B	9B	93.72	61.89	54.61	84.38	84.25	75.77
OPD-V	9B	91.62	60.47	**55.78**	**86.12**	83.62	75.52
VeGo (Ours)	9B	**94.24**	**65.56**	55.19	85.75	**86.00**	**77.35**

Group centering is performed at the trajectory level, whereas the indicator $\mathbf{1}[t \in E_i]$ restricts the resulting advantage to tokens inside E_i. Consequently, tokens outside E_i, including those in the evidence-writing prefix, have zero length advantage and contribute no direct token-level term to $\mathcal{L}_{\text{len}}$.

x dLength objective. With the importance ratio $r_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, q, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, q, y_{i,<t})$, we optimize a clipped policy-gradient objective restricted to E_i:

$$\mathcal{L}_{\text{len}}(\theta) = -\mathbb{E}\left[\frac{1}{\sum_i |E_i|} \sum_i \sum_{t \in E_i} \min\left(r_{i,t}\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\right)\right], \tag{12}$$

with a dual-clip lower bound applied when $\hat{A}_{i,t} < 0$ for stability. Since the length signal is restricted to E_i, $\mathcal{L}_{\text{len}}$ favors concise continuations and timely termination after sufficient visual evidence has been written, without directly penalizing the evidence-writing prefix.

3.4 Overall objective

The two parts are combined into a single objective in which the same evidence-dependence signal, read per token (g_t) and cumulatively (I_t), governs where we distill and where we compress:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{OPSD}}(\theta) + \lambda\, \mathcal{L}_{\text{len}}(\theta), \tag{13}$$

where λ weights the length-compression channel against evidence-focused distillation.

4 Experiments

4.1 Setup

Implementation Details. We apply VegGo to Qwen3.5-9B Qwen Team, Alibaba (2026) on the 6.5K synthetic data from Vision-OPD Yuan et al. (2026), training for 1 epoch with a maximum on-policy generation length of 1024 and 8 rollouts per prompt. The distillation objective is the generalized Jensen–Shannon divergence ($\alpha = 0.5$), approximated via top-K distillation ($K = 100$)

Table 2: Average number of output tokens across benchmarks. Lower values indicate more concise responses. Bold values denote the best results.

Model	Param. Size	V⋆	ZoomBench	Visual Probe	HR-Bench 4K	HR-Bench 8K	Average
DeepEyes	7B	159.00	196.39	208.92	168.13	178.28	182.14
SenseNova-MARS	8B	257.02	281.18	325.65	282.98	290.55	287.48
ZwZ	8B	58.76	**84.07**	95.51	198.45	212.30	129.82
Vision-OPD-9B	9B	50.16	105.87	104.33	180.49	142.83	116.74
OPD-V	9B	71.38	129.86	127.52	175.64	148.85	130.65
VeGo (Ours)	**9B**	**40.17**	101.14	**55.35**	**132.06**	**103.25**	**86.39**

that keeps only the top-K student logits and matching teacher logits plus a tail-probability term; this avoids full-vocabulary logit distillation while losing negligible mass, as tokens beyond the top-100 cover less than 1×10^{-13} of the probability in our scenario. The teacher is an EMA of the student. The ε in equation 5 is set to be 1×10^{-8}. For length compression, m_{base} is set to be 256 and $\{L_{ref}\}_{j=1}^{\mathcal{D}}$ is obtained by sampling in the training datasets offline. The two channels combine as $\mathcal{L} = \mathcal{L}_{\text{reweight}} + \lambda \mathcal{L}_{\text{len}}$ with $\lambda = 0.5$.

Evaluation Benchmarks. We evaluate VeGo on five benchmarks that connect visually demanding inputs to broader multimodal problem solving. V⋆ Bench Wu & Xie (2024) tests the localization and recognition of small targets in high-resolution, crowded scenes; ZoomBench Wei et al. (2026b) probes whether models recover subtle regional evidence across fine-grained VQA dimensions; Visual Probe Lai et al. (2025) targets fine-grained perception under strong distractors; and HR-Bench Wang et al. (2025d) extends understanding to native 4K and 8K resolutions. Beyond answer accuracy, we further report the average number of output tokens consumed at inference on each benchmark, an efficiency metric that reflects whether a model reaches the correct answer without spending unnecessary generation budget.

Compared Baselines. We compare VeGo against four baseline groups: (a) closed-source models: GPT-5.2 OpenAI (2026a), GPT-5.4 OpenAI (2026b), Gemini-3.1-Pro Google DeepMind (2026a), and Gemini-3.5-Flash Google DeepMind (2026b); (b) open-source models of different scales: Qwen3.5-9B, Qwen3.5-27B, Qwen3.5-Flash, Qwen3.5-Plus Qwen Team, Alibaba (2026), and Kimi-K3 Team (2026); (c) "Thinking-with-Images" agents: DeepEyes Zheng et al. (2025) and SenseNova-MARS Chng et al. (2025), which use cropping and zooming for multi-step reasoning; and (d) distillation models: ZwZ Wei et al. (2026b), Vision-OPD Yuan et al. (2026), and OPD-V Bi et al. (2026). Vision-OPD and OPD-V are particularly relevant as they also perform on-policy self-distillation with privileged visual information. We report answer accuracy for all four groups and inference-time output token counts for the agentic and distillation models to assess the accuracy–efficiency trade-off.

4.2 Main Results

As shown in Tables 1 and 2, VeGo achieves the highest average accuracy among the evaluated open-source, agentic, and distillation models, while producing the shortest average responses among models with reported output lengths. It consistently outperforms larger open-source models, surpasses closed-source models on V⋆ and ZoomBench, and remains competitive on HR-Bench. Compared with "Thinking-with-Images" agents, VeGo achieves higher accuracy across all benchmarks using substantially fewer tokens, showing that fine-grained perception need not require lengthy responses. It also improves both average accuracy and generation efficiency over Vision-OPD and OPD-V. These results demonstrate the effectiveness of combining evidence-sensitive reweighting with post-decision compression.

4.3 Abaltion Studies

Ablation on the Distillation Reweighting Strategy. Table 3 compares token-level distillation weighting strategies using the same backbone and training data. *Uniform Weighting* reproduces the standard Vision-OPD objective. *Top-20 Student–Teacher JSD Selection* retains only

Table 3: Ablation of token-weighting strategies for distillation. We report accuracy (%) for each strategy. Bold values indicate the best and best results

Weighting Strategy	V*	ZoomBench	HR-Bench 4K	HR-Bench 8K	Average
Uniform Weighting (Vision-OPD)	93.72	61.89	84.38	84.25	81.06
Top-20 Student–Teacher JSD Selection	92.67	61.42	83.38	82.50	79.99
Entropy-Gap Weighting ($k = 5$)	94.24	63.20	85.12	83.75	81.58
Entropy-Gap Weighting ($k = 8$)	92.67	63.45	83.88	83.88	80.97
Entropy-Gap Weighting ($k = 10$)	93.19	62.84	85.12	83.88	81.26
Evidence Weighting (Weight-Sum Norm.)	93.72	**63.79**	85.62	86.25	82.35
Evidence Weighting (Loss-Sum Norm., Ours)	**94.76**	63.67	**86.38**	**86.75**	**82.89**

Table 4: Ablation on the length-compression strategy. We report accuracy (%, ↑) and the average number of output tokens (Len, ↓) for each benchmark.

Length-Compression Strategy	V*		ZoomBench		HR-Bench 8K		Average	
	Acc (↑)	Len (↓)	Acc (↑)	Len (↓)	Acc (↑)	Len (↑)	Acc (↑)	Len (↑)
Evidence only (no length reward)	94.76	45.29	63.67	144.37	86.75	144.17	81.73	111.27
post-decision, adaptive (Ours)	94.24	40.17	65.56	101.14	86.00	103.25	81.93	81.52
w/o adaptive reference	94.76	39.15	64.14	82.49	85.62	103.02	81.51	74.88
w/o post-decision masking	95.29	42.17	61.78	115.28	85.62	107.08	80.90	88.17

the 20 tokens with the largest student–teacher divergence. *Entropy-Gap Weighting* assigns $w_t = \sigma(k(H_{\text{student}} - H_{\text{teacher}}))$, giving larger weights to tokens for which the student is more uncertain, with k controlling the gate sharpness. Our *Evidence Weighting* instead uses the divergence between the evidence-rich and evidence-blind teachers to measure each token's dependence on visual evidence. We compare weight-sum and loss-sum normalization for this weighting.

Top-20 selection reduces the average accuracy to 79.99%, below uniform weighting, suggesting that focusing only on the hardest tokens discards useful supervision. Entropy-Gap Weighting provides marginal and unstable gains (80.97%–81.58%), since predictive uncertainty does not necessarily reflect visual-evidence dependence. Both Evidence variants outperform the difficulty- and uncertainty-based alternatives, confirming the importance of evidence-sensitive weighting. Loss-sum normalization performs best overall, achieving the highest average accuracy and the best results on V*, HR-Bench 4K, and HR-Bench 8K. It preserves the total loss magnitude while redistributing it across tokens, whereas weight-sum normalization makes the effective loss scale depend on the weight distribution.

Ablation on the Length-Compression Strategy. We next isolate the effect of our length-compression design, which builds on the evidence-reweighted model (top block of Table 4) by adding a length penalty with two components: restricting the penalty to the *post-decision* region (post-decision masking) and setting a per-sample *adaptive* reference length. Starting from the evidence-only model, which already reaches an average accuracy of 81.73% but generates 111 tokens on average, our full method **post-decision, adaptive** reduces the average length to 94 tokens—a 27% reduction—while slightly raising accuracy to at 81.93%. This confirms that the post-decision region is genuinely redundant: it can be substantially shortened without affecting the answer.

Removing either component degrades this trade-off. Without the adaptive reference (*w/o adaptive reference*), a single fixed budget is applied to all samples; this compresses length further to 75 tokens but lowers accuracy to 81.51%, as harder questions that require longer reasoning are over-penalized. Without post-decision masking (*w/o post-decision masking*), the length penalty is applied to all tokens of the response, which shifts part of the penalty signal forward from the post-decision tokens to the tokens before the decision point. On the one hand, this leads to longer outputs: before the decision point the model is still writing in visual evidence, so the strong distillation signal there dilutes the length penalty and weakens its compression effect. On the other hand, it yields the lowest accuracy: the misplaced penalty interferes with the evidence-writing process and thereby harms perception. Overall, the full configuration offers the best trade-off, achieving the largest length

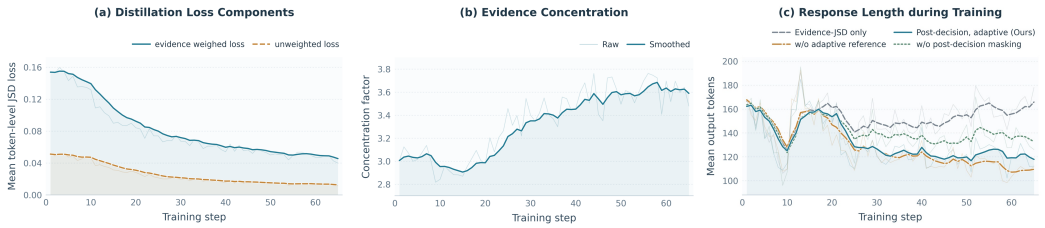


Figure 5: **Training dynamics of the proposed method.** (a) Mean token-level distillation losses with uniform weighting and evidence weighting (Weight-Sum Norm.). (b) Evidence concentration factor, defined as the ratio between the two losses in (a). (c) Mean response lengths under different length-compression configurations.

reduction that preserves accuracy, and both the post-decision restriction and the adaptive reference are necessary to reach it.

Gradient comparison. Table 4 shows that penalizing the whole trajectory lowers accuracy. We trace this to gradient interference: at a single optimizer step we back-propagate the distillation loss $\mathcal{L}_{\text{distill}}$ and the length objective $\mathcal{L}_{\text{len}}$ separately and measure the per-tensor cosine between their gradients (Figure 4). With the penalty confined to the post-decision region E_i (Ours), the two gradients are near-orthogonal (median $|\cos|$=0.024, 88% of tensors below 0.1); removing the mask shifts the distribution far right (median $|\cos|$=0.273) with a heavy mode near 0.4. Since the pre-cliff prefix is where the teacher still carries visual evidence, an all-span length gradient opposes distillation on evidence-critical tokens. Masking confines compression to the redundant tail, so the two objectives optimize along nearly independent directions—shortening outputs without hurting accuracy.

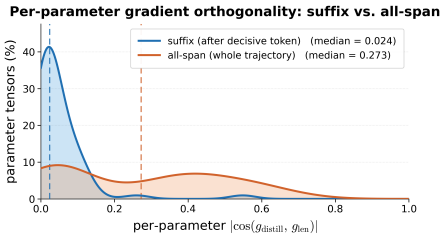


Figure 4: **Gradient orthogonality between distillation and length objectives.** Distribution of the per-tensor cosine similarity over model parameters, comparing post-decision masking (suffix) with whole-trajectory penalization (all-span). Dashed lines mark the medians.

4.4 Dynamics Analysis

Figure 5 (a) shows that both the unweighted and evidence-weighted (Weight-Sum Norm.) distillation losses decrease steadily, indicating stable optimization. Meanwhile, the increasing concentration factor in Figure 5 (b) suggests that the remaining student–teacher discrepancy becomes progressively concentrated on evidence-sensitive tokens. As shown in Figure 5 (c), Evidence alone does not shorten responses, whereas all length-regularized variants reduce output length. The fixed-reference variant compresses most aggressively. Note that the evidence-weighted (Weight-Sum Norm.) loss here is just for a direct comparison with the uniformly weighted loss, the model is actually trained with evidence-weighted (Loss-Sum Norm.) loss.

5 Conclusion

We present VeGo, a unified framework for accurate and efficient fine-grained visual perception. It uses the discrepancy between evidence-seeing and evidence-blind teachers to focus on-policy distillation on evidence-sensitive tokens, strengthening the transfer of visually grounded knowledge. The same signal identifies a decision cliff beyond which the visual evidence has been sufficiently expressed, enabling VeGo to compress redundant post-decision generation while preserving accuracy. Experiments demonstrate that VeGo improves fine-grained perception while producing substantially more concise responses.

AI USE STATEMENT

In this work, we used generative AI tools for literature search, code assistance, and language polishing. We have not used generative AI tools for generating data, experimental results, or final scientific claims, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for writing refinement. We have reviewed all AI-assisted work. Specifically, all references suggested by AI tools were manually verified against the original sources, AI-assisted code was reviewed, executed, and tested by the authors for correctness, and AI-assisted text was carefully checked and edited by the authors for factual accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

R. Agarwal, N. Vieillard, Y. Zhou, et al. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

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

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.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026.

X. Fu, M. Liu, Z. Yang, et al. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

GLM-V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, et al. Glm-5v-turbo: Toward a native foundation model for multimodal agents, Apr 2026. URL https://arxiv.org/abs/2604.26752. ///GUI .

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/technologies/gemini/, 2026a. Large multimodal model.

Google DeepMind. Gemini 3.5 Flash. https://deepmind.google/technologies/gemini/, 2026b. Large multimodal model.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

C. Gulcehre, T. L. Paine, S. Srinivasan, et al. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

W. Hou, S. Peng, W. Wang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

J. Hübotter, F. Lübeck, L. Behric, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

X. Lai, J. Li, W. Li, et al. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

X. Liu, Y. Hu, Y. Zou, et al. Hide: Rethinking the zoom-in method in high-resolution MLLMs via hierarchical decoupling. *arXiv preprint arXiv:2510.00054*, 2025.

K. Lu and Thinking Machines Lab. On-policy distillation, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation.

OpenAI. GPT-5.2. https://openai.com/, 2026a. Large language model.

OpenAI. GPT-5.4. https://openai.com/, 2026b. Large language model.

R. Peng, H. He, Y. Wei, Y. Wen, and D. 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*, pp. 3963–3973, 2025.

S. Qi, B. Yang, K. Jiang, et al. In-context editing: Learning knowledge from self-induced distributions. *arXiv preprint arXiv:2406.11194*, 2024.

Qwen Team. Qwen3.5-omni technical report, Apr 2026. URL https://arxiv.org/abs/2604.15804. Hybrid Attention MoE256k 215 / SOTA ARIA -.

Qwen Team, Alibaba. Qwen3.5 Technical Report. https://qwenlm.github.io/, 2026. Large language model.

H. Shen, K. Zhao, T. Zhao, et al. Zoomeye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of EMNLP*, pp. 6613–6629, 2025.

Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

H. Wang, X. Li, Z. Huang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.

H. Wang, A. Su, W. Ren, F. Lin, and W. Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025b.

J. Wang, Z. Kang, H. Wang, et al. VGR: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025c.

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, 2025d.

L. Wei, L. He, J. Lan, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026a.

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*, 2026b.

Y. Wei, L. Zhao, K. Lin, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

J. Wu, J. Guan, K. Feng, et al. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

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 (CVPR)*, pp. 13084–13094, 2024.

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

H. Yuan, Y. Sun, Y. Li, et al. Visual reasoning tracer: Object-level grounded reasoning benchmark. *arXiv preprint arXiv:2512.05091*, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

E. Zelikman, Y. Wu, J. Mu, and N. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.

X. Zhang, Z. Gao, B. Zhang, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via reinforcement learning. *arXiv preprint*, 2025a.

Y. Zhang, L. Hu, H. Sun, et al. Skywork-R1V4: Toward agentic multimodal intelligence through interleaved thinking with images and deepresearch. *arXiv preprint arXiv:2512.02395*, 2025b.

Y. Zhang, X. Lu, S. Yin, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025c.

S. Zhao, Z. Xie, M. Liu, et al. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Z. Zheng, M. Yang, J. Hong, et al. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Sensitive Analysis

We evaluate the sensitivity of VeGo to the length-penalty strength $\lambda \in \{0.6, 0.8, 1.0, 1.2\}$ on V*, Visual Probe, HR-Bench 4K, and HR-Bench 8K. As shown in Figure 5, increasing λ generally leads to shorter responses: the mean output length decreases from 92.1 tokens at $\lambda = 0.6$ to 78.1 tokens at $\lambda = 1.2$, a reduction of approximately 15.2%. In contrast, the mean accuracy remains nearly unchanged, ranging only from 80.2% to 80.4%. This indicates that the post-decision length objective can adjust response conciseness without noticeably affecting predictive performance. We use $\lambda = 1.0$ by default, as it provides a balanced trade-off between response length and accuracy.

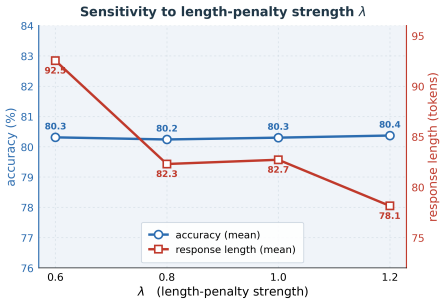


Figure 6: Sensitivity analysis of the length-penalty strength λ.

B Case Study

Figure 7 presents a challenging fine-grained perception case, where the two arrows occupy only a small region of a visually complex image. Although Vision-OPD successfully identifies the relevant sign, it incorrectly interprets the arrows as pointing in opposite directions and selects option (D). VeGo instead captures the subtle directional evidence that both arrows point left and selects the correct option (A). This example suggests that localizing the relevant region alone is insufficient; the model must also preserve fine-grained visual evidence when producing its answer. The result is consistent with our Evidence-JSD weighting, which emphasizes tokens whose predictions depend strongly on the visual evidence.

Figure 8 highlights the complementary benefit of response compression. Both methods correctly identify that the volleyball is to the left of the man with the white cap. However, Vision-OPD repeatedly describes the same spatial relation before giving the answer, whereas VeGo states the necessary evidence once and terminates promptly. This behavior is consistent with our post-decision compression objective, which encourages concise continuations after sufficient evidence has been expressed. Together, these examples illustrate that VeGo can improve fine-grained visual reasoning while reducing redundant post-decision generation.

What direction arrows are shown on the black sign? (A) Two left arrows (B) Two right arrows (C) No arrows (D) Both left and right arrows



Vision-OPD:

The black sign in the image displays **two yellow arrows**, one pointing to the **left**
and one pointing to the **right**.

Looking at the options:

- (A) Two left arrows → ❌ Incorrect
- (B) Two right arrows → ❌ Incorrect
- (C) No arrows → ❌ Incorrect
- (D) Both left and right arrows → ✅ Correct

Answer: (D) Both left and right arrows

VeGo:

The black sign in the image displays two yellow arrows, both pointing to the **left**.

Therefore, the correct answer is:
(A) Two left arrows

Figure 7: Accuracy comparison on fine-grained visual evidence.

What black logo is visible on the yellow jersey? (A) left (B) right



Vision-OPD:

The volleyball is on the **left** side of the man with the white cap.

Looking at the image:
- The man with the white cap is on the **right** side of the frame.
- The volleyball (yellow and blue) is on the **left** side of the frame, relative to the man.

So, from the viewer's perspective, the volleyball is to the **left** of the man.

Answer: **(A) left**

VeGo:

The volleyball is located to the **left** of the man with the white cap in the image.

Answer: (A)

Figure 8: Response-efficiency comparison.