Evidence-RL: Towards Evidence-intensive Visual Reasoning

Haojie Huang[1,2,*] **Xinlei Yu**[1,*] **Chengming Xu**[3,*] **Zhangquan Chen**[4] **Cheng Yang**[5]
Qingdong He[5] **Yu Yang**[2] **Jiangning Zhang**[2] **Xiaobin Hu**[1,†]
greatesthhj@zju.edu.cn ben0xiaobin0hu1@nus.edu.sg
[1]National University of Singapore [2]Zhejiang University
[3]Fudan University [4]Tsinghua University [5]Tencent
[*]Equal contribution. [†]Corresponding author.

Abstract

Vision-Language Models (VLMs) should answer from concrete image evidence rather than language priors, dataset shortcuts, or irrelevant visual context. Existing perception-aware post-training methods encourage image use through global perturbations or attention proxies, but they do not test whether a sampled answer causally depends on the local evidence that supports it. We propose **Counterfactual Evidence Disentanglement** (CED), a training-time evidence audit for VLM grounding. For each response, CED neutralizes an object-centric Evidence Region and compares the resulting support drop against matched non-evidence Regions. We combine this signal with answer correctness inside GRPO, rewarding correct answers that rely on the evidence path rather than shortcut or nuisance paths. CED uses weak object-level proposals, requires no question-specific evidence annotations, and adds no inference-time overhead. Across nine public benchmarks and four backbones, CED outperforms prior RL-based post-training methods, with targeted analyses verifying its object-centric signal. Project page including code and checkpoint is available at `https://evidencerl.github.io`.

1 Introduction

Vision-Language Models (VLMs) are increasingly used as general-purpose visual reasoners [Bai et al., 2025b,a, Wang et al., 2025b, Li et al., 2024], but correct-looking answers need not be visually grounded. A model may answer a counting or spatial question by inspecting the relevant evidence, or by relying on language priors, dataset regularities, and scene-level common sense, a failure pattern documented in hallucination, counting, spatial, and shortcut-reasoning evaluations [Li et al., 2023, Paiss et al., 2023, Wang et al., 2024, Guan et al., 2024, Rahmanzadehgervi et al., 2024, Yin et al., 2026b]. Post-training can amplify either behavior depending on what the training signal can observe.

Recent perception-aware methods attempt to reintroduce the image into the post-training loop. PAPO [Wang et al., 2025d] compares model behavior under the original image and a corrupted image, while VPPO [Huang et al., 2025] uses attention-derived visual dependency to adjust the training signal. These methods are important because they move beyond text-only supervision. However, they still target *coarse visual dependence* as a weaker visual anchor rather than *causal evidence dependence*. In general, the image is not a single cause. It contains the target evidence that supports the answer, irrelevant context, visually salient distractors, and artifacts introduced by masking, zeroing, or noise. A global image perturbation can show that the model is sensitive to the image, but it cannot tell whether the current answer specifically depends on the local evidence that would falsify it if removed. Attention-based dependency is also an internal routing proxy, not a direct test of whether the answer is supported by the relevant evidence.

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

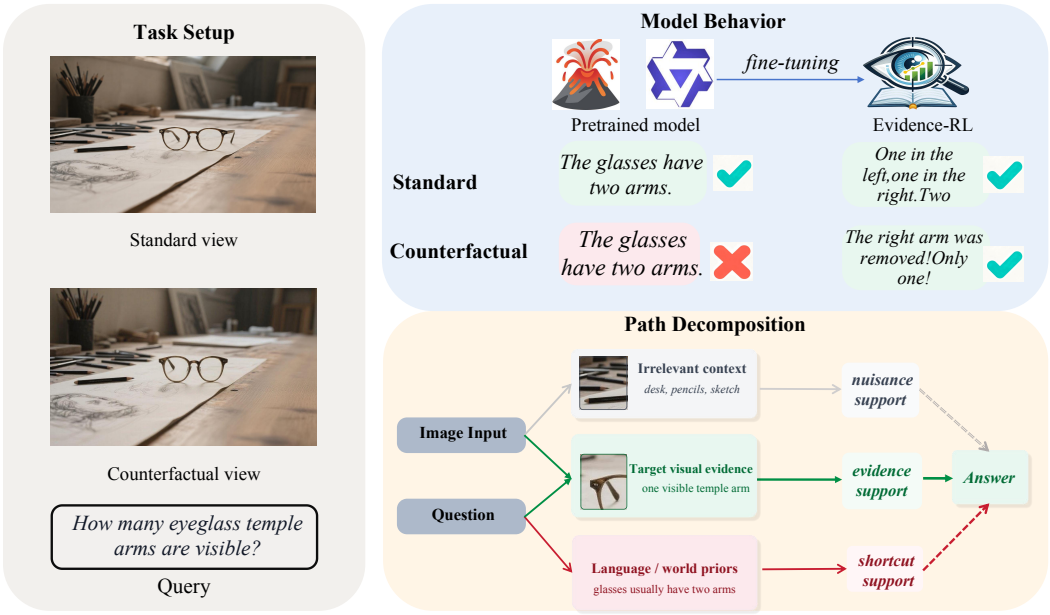


Figure 1: **Left:** Task setup with a counting query over two views of the same scene. **Upper right:** A pretrained VLM gives the same prior-based answer for both views; EVIDENCE-RL inspects the visual evidence and adjusts its answer when the scene changes. **Lower right:** Causal path decomposition: a candidate answer can be supported by the target visual evidence, by irrelevant context, or by language priors, which motivates CED to test whether the response depends specifically on the evidence path.

The path decomposition in Figure 1 causally formalizes this interference, following the structural-causal view that interventions should isolate the path being tested [Pearl, 2009]. A candidate answer can be produced by language and world priors, by irrelevant visual context or distractors, or by the target evidence that actually supports the answer. The shortcut problem arises because these paths meet at the same observed output: the same correct answer may be caused by the evidence path or by a non-evidence path. This suggests a more general requirement for grounded VLM post-training. Instead of asking only whether the model uses the image, we should ask whether the answer depends on the particular visual evidence that supports it, while discounting nuisance sensitivity to other regions and to the intervention itself. Without an image-conditional probe, a training signal computed on the model's text output alone cannot distinguish answers that traverse the target-evidence path from those that traverse a shortcut path; we make this consequence precise as *evidence-closed self-evolution* in Section 5 where grounding becomes formally unidentifiable.

Based on this causal view, we propose **Counterfactual Evidence Disentanglement** (CED), a novel post-training framework that trains VLMs to rely on evidence-specific counterfactual dependence rather than coarse image sensitivity. The core idea is to turn the causal graph into a training-time counterfactual test: for each sampled response, CED asks whether the model's support for that response decreases more when the object-centric Evidence Region is neutralized than when matched non-evidence Regions are neutralized. A response is therefore favored only when its correctness is tied to the evidence path, rather than preserved through language priors, irrelevant context, or generic sensitivity to visual perturbations. We implement this test by intervening on visual tokens with mean replacement, which removes region-specific information while preserving the local feature baseline, making the intervention a structured local removal of evidence rather than zeroing, random noise, or global corruption. The resulting evidence signal is combined with answer correctness inside GRPO [Shao et al., 2024], encouraging correct answers that are causally grounded in the image without requiring question-specific region annotations or adding inference-time overhead.

We evaluate CED on public VLM benchmarks spanning VQA, hallucination detection, counting, spatial reasoning, and perception-heavy visual reasoning, in which CED outperforms prior RL-based post-training methods, with clear gains on evaluations that require resisting language priors and grounding answers in visual evidence. Targeted analyses confirm that CED captures evidence-specific dependence, that mean replacement provides a cleaner counterfactual intervention than

common alternatives, and that the signal follows plausible target evidence rather than arbitrary regions. Throughout the paper we use CED for the counterfactual evidence diagnostic itself (the reward signal) and EVIDENCE-RL for the full post-training pipeline that combines CED with GRPO (Section 3). Our contributions can be summarized as follows:

- We articulate the gap between *coarse visual dependence* and *counterfactual evidence dependence*, and formalize visual grounding as a structural causal graph in which the same observed answer can arise from the target-evidence path or from shortcut and nuisance paths. We further characterize why text-only self-evolution can remain evidence-closed.
- We propose CED, a counterfactual evidence audit that tests whether a candidate answer depends on a proposed Evidence Region relative to matched non-evidence Regions. The audit is defined on the causal graph rather than tied to a specific evidence type; we instantiate it for spatial-region evidence and outline how the same recipe extends to attribute-level and relational evidence by varying the intervention.
- We integrate CED into GRPO and validate it through signal diagnostics, controlled ablations (correctness-only, matched baselines, intervention type, proposal robustness), and nine public benchmarks across four backbones. The results show that targeting counterfactual evidence dependence transfers across backbones, recovers gains on perception-heavy benchmarks, and adds no inference-time overhead.

2 Related Work

VLM reasoning. Vision-language models built on LLM backbones [Li et al., 2024, Bai et al., 2025b,a, Wang et al., 2025b] have evolved from task-specific perception systems into general-purpose multimodal reasoners, supporting visual QA, counting, spatial reasoning, diagrammatic reasoning, and scientific problem solving [Paiss et al., 2023, Wang et al., 2024, Lu et al., 2022, 2023, Yue et al., 2024]. This progress exposes a persistent gap between answer correctness and visual grounding: a model can produce a plausible answer while losing the sparse visual evidence that should justify it [Yu et al., 2026, Chen et al., 2026]. Benchmarks such as HallusionBench [Guan et al., 2024], VLMsAreBlind [Rahmanzadehgervi et al., 2024], and FREAK [Yin et al., 2026b] make this failure measurable through hallucination, shortcut reasoning, and fine-grained visual errors. We address this from the post-training side, rewarding answers whose likelihood depends on relevant visual evidence.

Post-training VLMs. Reinforcement learning has become a central post-training mechanism for improving reasoning, from PPO-style optimization [Schulman et al., 2017, Shao et al., 2024] to verifiable-reward and self-evolution recipes [Guo et al., 2025, Acikgoz et al., 2026], and now extends to low-level vision [Xu et al., 2026]. Recent VLM methods, including VLM-R1 [Shen et al., 2025], ViCrit [Wang et al., 2025c], Perception-R1 [Xiao et al., 2025], SophiaVL-R1 [Fan et al., 2025], and VAPO [Yue et al., 2025], bring this paradigm to multimodal reasoning. The key difficulty is that visual tasks lack cheap per-sample verifiers analogous to code execution or arithmetic checking. Text-only judges can score the plausibility of an answer or rationale but cannot tell whether it was caused by the image, so grounding-oriented RL needs rewards that observe visual evidence.

Grounded Rewards. Existing attempts remain limited. Image-conditional rewards compare rollouts under original and perturbed images [Wang et al., 2025d, Huang et al., 2025, Du et al., 2026] or distill from crop-conditioned teachers [Yuan et al., 2026], yet neither tests whether the decisive region supports the answer. Process-level and self-evolving rewards [Fan et al., 2025, Zhang et al., 2025b, Acikgoz et al., 2026, Zhao et al., 2026] score the reasoning trace, but remain text-closed and can reward language-prior shortcuts with consistent explanations; we formalize this as *evidence-closed self-evolution* in Section 5. Counterfactual and test-time visual interventions expose or correct grounding failures at inference time [Leng et al., 2024, Liu et al., 2025, Li et al., 2025a, Wu et al., 2026, Yin et al., 2026a], but keep the model frozen. Our intervention is related to perturbation-based visual explanations [Zeiler and Fergus, 2014, Fong and Vedaldi, 2017, Petsiuk et al., 2018], but we use the score as a training reward. Closest training-time methods require region annotations in reasoning traces [Sun et al., 2026], process/checklist supervision [Li et al., 2026, Zhang et al., 2025a, Wang et al., 2025a, Qin et al., 2025], or adversarial text/preference pairs [Chen et al., 2025, Yang et al., 2026]. In contrast, CED brings region-level counterfactual intervention into RL as a reward, using off-the-shelf object proposals instead of question-specific annotations.

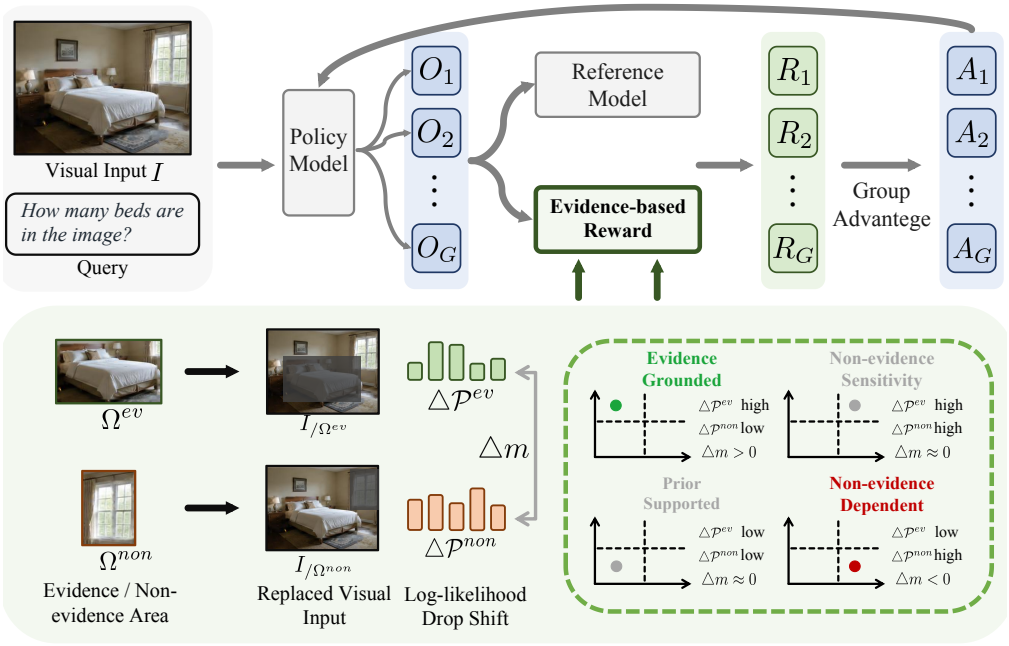


Figure 2: Overview of EVIDENCE-RL during post-training. A policy samples candidate answers, and CED scores each answer by applying the same feature-space intervention to the proposed Evidence Region and to matched non-evidence Regions. The contrastive support drop yields an evidence margin, which is combined with correctness to form the GRPO reward. The counterfactual audit is used only during training; inference uses the trained VLM normally.

3 Method

Figure 2 summarizes EVIDENCE-RL as a training-time pipeline: sample a candidate answer, contrast the proposed Evidence Region with matched non-evidence Regions, intervene in feature space, compute an answer-conditioned evidence margin, and use the resulting reward in GRPO. We now define the counterfactual probe and its integration into post-training.

3.1 CED: Counterfactual Evidence Margin

Correctness alone leaves open whether an answer is supported by the image or by language priors. We therefore use "grounding" in a narrow sense: an answer is grounded when its support depends on a specific locus of visual evidence. CED turns this evidence sensitivity into a training signal complementary to answer correctness.

Given image I, question q, and candidate answer y, we identify an Evidence Region Ω^{ev} hypothesized to contain answer-relevant evidence and K non-evidence Regions $\{\Omega_k^{\text{non}}\}_{k=1}^{K}$. For a region Ω, we intervene after spatial merging by replacing its visual tokens with the mean $\boldsymbol{\mu}_T$ of neighboring tokens:

$$\tilde{\mathbf{h}}_i = \begin{cases} \boldsymbol{\mu}_T & \text{if } i \in \mathcal{T}(\Omega), \\ \mathbf{h}_i & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{T}(\Omega)$ maps a spatial region to visual-token indices. Feature-space mean replacement removes region-specific information while keeping the local representation manifold largely intact, reducing artifacts compared with zeroing or noise injection. We evaluate these alternatives in Section 4.3.3 and give a mechanistic account in Section A.5.

The evidence sensitivity of y to region Ω is defined as the counterfactual log-likelihood drop:

$$s(\Omega) \;=\; \log \pi_\theta(y \mid I, q) \;-\; \log \pi_\theta(y \mid \tilde{I}_{\setminus\Omega}, q). \tag{2}$$

If the model can support y without inspecting Ω, then $s(\Omega)$ remains close to zero. To obtain a bounded and contrastive score, we compare the evidence-region sensitivity with reference-region

sensitivities:

$$m(I, q, y) \;=\; \tanh\!\left(\frac{s(\Omega^{\mathrm{ev}}) - \mu(s^{\mathrm{non}})}{\sigma(s^{\mathrm{non}}) + \epsilon}\right), \tag{3}$$

where $s^{\mathrm{non}} = \{s(\Omega_k^{\mathrm{non}})\}_{k=1}^{K}$, and $\mu(s^{\mathrm{non}})$ and $\sigma(s^{\mathrm{non}})$ are the mean and standard deviation over non-evidence Regions. A positive margin indicates that the candidate answer depends more on the proposed Evidence Region than on area-matched non-evidence Regions.

Non-evidence Regions as a local null. Non-evidence Regions are not treated as annotated negatives. They define a sample-local null distribution for how much the candidate answer changes under the same feature-space intervention applied to regions that are area-matched and spatially separated from the proposed Evidence Region. Thus, if a response is sensitive mainly to generic masking artifacts or irrelevant visual context, the Evidence and non-evidence interventions should induce comparable support drops. A high margin is obtained only when the Evidence Region produces a larger drop than this non-evidence distribution. In implementation, the K non-evidence Regions jointly estimate this baseline rather than forming one-to-one evidence–non-evidence pairs.

Response-conditioned local interventions. CED is designed to isolate the target-evidence path in Figure 1. It scores the sampled answer y itself, measuring the likelihood drop of that response rather than a distribution-level shift. It also intervenes locally, targeting the region that should carry the relevant evidence instead of perturbing the whole image. Finally, the intervention is applied in feature space after spatial merging, which preserves the representation outside the intervened region and reduces pixel-level encoder artifacts. These choices make the score answer-specific, evidence-localized, and less confounded by intervention noise.

Weakly-supervised evidence proposals. The Evidence Region Ω^{ev} is a weak spatial proposal resolved from dataset metadata, using object-level COCO annotations [Lin et al., 2014] in our experiments. Such boxes provide spatial priors: the same object box may or may not contain evidence depending on the question. Reference Regions are sampled to match the intervention scale while avoiding the proposed Evidence Region. CED resolves this proposal ambiguity by comparing the counterfactual sensitivity of Ω^{ev} against matched Reference Regions, so the evidence signal is determined by model behavior which are expected to transfer to various downstream tasks. Proposal-layer reliability is evaluated in Section 4.3.3.

3.2 EVIDENCE-RL: CED-Guided GRPO

Having defined the evidence margin m, we integrate it into GRPO as a correctness-anchored reward. The resulting pipeline, including CED reward shaping and binary-action routing, is denoted by EVIDENCE-RL.

Correctness-anchored reward. The training reward is:

$$R_{\text{train}} = R_{\text{ans}} \cdot g(m) + \varepsilon_{\text{tie}}\, m, \qquad g(m) = \tfrac{1}{2}\left(1 + \tanh\!\left(\tfrac{m}{\tau_g}\right)\right), \tag{4}$$

where τ_g=0.20, ε_{tie}=0.10, and R_{ans} is the answer-correctness score defined in Appendix A.1. When the answer is correct, $g(m)$ favors responses with stronger Evidence Region support. When the answer is wrong and $R_{\text{ans}} \to 0$, the evidence term becomes a bounded tiebreaker, with $|\varepsilon_{\text{tie}} m| \le \varepsilon_{\text{tie}}$. Thus, after within-group normalization $A_i = (R_{\text{train},i} - \mu)/\sigma$, correctness remains the dominant gradient driver, while CED ranks equally correct rollouts by evidence support.

4 Experiments

The experiments are designed to validate four properties of our method:(1)the counterfactual reward provides a usable evidence-dependence signal;(2)RL with this signal improves benchmark performance against matched-backbone baselines;(3)the gains transfer across backbones,;(4)the signal remains stable under intervention and proposal perturbations. Unless explicitly stated, ours denotes the Answer-CED variant. The main head-to-head comparison uses Qwen2.5-VL-7B [Bai et al., 2025b] as the base model for ours; cross-backbone validation further covers Qwen2.5-VL-3B/7B [Bai et al., 2025b], Qwen3-VL-8B [Bai et al., 2025a], and Qwen3.5-9B [Team, 2026]. For reference,

Table 1: Signal-validation statistics on counting and presence tasks.

Task	Non-constant raw reward	Zero variance	Mean reward std	Same answer, different reward	Mean negation
Counting	**99.5%**	24.0%	0.1299	**90.0%**	0.003
Presence	21.1%	79.0%	0.0497	5.0%	0.777

Table 2: Nine-benchmark evaluation using Qwen2.5-VL-7B as the base model for Ours. Backbone, baseline, and benchmark sources are cited in the setup paragraph. Best **bold**; second-best underlined. Each Δ row reports absolute gain over the corresponding base.

	Grounding					General Reasoning				
Model	CountBench	SpatialEval	Hallusion	VLMs AreBlind	FREAK	MathVista	MMBench	MMMU	ScienceQA	**Avg**
VLM backbones										
LLaVA-v1.6-7B	55.60	23.60	51.00	28.23	9.50	22.80	71.30	29.00	57.20	38.69
InternVL3.5-8B	86.90	24.00	69.60	**54.92**	12.80	51.90	83.40	50.80	**89.90**	58.25
Qwen2.5-VL-3B	70.71	53.94	65.37	42.46	11.14	52.10	80.40	50.29	74.93	55.70
Qwen2.5-VL-7B	81.82	59.21	68.64	46.44	12.18	63.10	84.73	50.63	83.98	61.19
RL-based methods										
VAPO-Thinker-7B	86.90	60.80	**71.20**	48.62	13.10	51.10	81.80	44.60	82.30	60.05
Δ	+5.10	+1.59	+2.60	+2.18	+0.92	−12.00	−2.90	−6.00	−1.70	−1.14
VPPO-7B	85.90	61.80	68.20	49.35	12.70	67.90	84.90	52.10	88.40	63.47
Δ	+4.10	+2.59	−0.40	+2.91	+0.52	+4.80	+0.20	+1.50	+4.40	+2.28
Perception-R1-7B	84.90	59.70	66.70	45.22	12.20	67.10	81.70	48.10	82.30	60.88
Δ	+3.10	+0.49	−1.90	−1.22	+0.02	+4.00	−3.00	−2.50	−1.70	−0.31
PAPO-G-H	**90.90**	**64.80**	69.50	47.92	12.80	**69.40**	82.80	50.50	85.60	63.80
Δ	+9.10	+5.59	+0.90	+1.48	+0.62	+6.30	−1.90	−0.10	+1.60	+2.61
VLM-R1	74.80	56.40	66.60	44.15	11.00	60.30	79.90	48.60	73.60	57.26
Δ vs. Qwen2.5-VL-3B	+4.10	+2.50	+1.20	+1.69	−0.14	+8.20	−0.50	−1.70	−1.30	+1.56
SophiaVL-R1	82.83	61.38	66.87	47.09	24.18	66.30	87.41	52.18	87.64	63.99
Δ	+1.01	+2.17	−1.77	+0.65	+12.00	+3.20	+2.68	+1.55	+3.66	+2.80
Ours	88.89	63.34	70.08	54.02	**26.40**	**69.40**	**87.53**	**57.47**	87.08	**67.13**
Δ	+7.07	+4.13	+1.44	+7.58	+14.22	+6.30	+2.80	+6.84	+3.10	**+5.94**

the backbone rows also include LLaVA-v1.6-7B [Li et al., 2024] and InternVL3.5-8B [Wang et al., 2025b]. Matched baselines include recent RL-based VLM post-training methods [Yue et al., 2025, Huang et al., 2025, Xiao et al., 2025, Wang et al., 2025d, Shen et al., 2025, Fan et al., 2025]. Signal-validation diagnostics use Qwen3-VL-8B-Instruct as a frozen probe. Evaluation covers nine public benchmarks spanning grounding, faithfulness, and general reasoning: CountBench [Paiss et al., 2023], SpatialEval [Wang et al., 2024], HallusionBench [Guan et al., 2024], VLMsAreBlind [Rahmanzadehgervi et al., 2024], FREAK [Yin et al., 2026b], MathVista [Lu et al., 2023], MMBench [Liu et al., 2024], MMMU [Yue et al., 2024], and ScienceQA [Lu et al., 2022], with no overlap between training images and evaluation benchmarks.

4.1 Validating the Evidence-Dependence Signal

We first check whether the reward separates evidence dependence from correctness-only. On counting tasks, 99.5% of groups have non-constant rewards and 90.0% contain same-answer trajectories with different rewards, showing that the signal preserves within-group discrimination beyond answer-string matching. Presence tasks show higher zero variance, consistent with their two-action yes/no structure discussed in Section 3.2. Cross-model validation and blindfold tests are reported in Appendix A.4.

4.2 Main Results on Nine Benchmarks

Table 2 compares ours with VLM backbones and RL-based baselines on nine benchmarks. Benchmarks are grouped into *Grounding* and *General Reasoning*. Each Δ reports the absolute gain over the corresponding base model, which is Qwen2.5-VL-7B unless the row label specifies another base.

Ours obtains the highest average score and the largest mean improvement in the matched-backbone comparison. It is also the only RL method with non-negative Δ on all nine benchmarks. The largest gains appear on perception-heavy benchmarks, including VLMsAreBlind, FREAK, and MMMU, while MathVista, MMBench, and ScienceQA remain positive. This pattern supports the intended role of the reward: increasing reliance on visual evidence without sacrificing general transfer.

Table 3: Controlled comparison on Qwen3.5-9B with matched data, router, and compute. Δ is the mean gain over the frozen base; Δ_{CED} is Answer-CED minus correctness-only.

	Grounding					General Reasoning					
				VLMs							
Method	CountBench	SpatialEval	Hallusion	AreBlind	FREAK	MathVista	MMBench	MMMU	ScienceQA	**Avg**	Δ
λ=0 (corr.-only)	89.90	40.45	57.93	72.90	18.73	73.40	79.44	36.00	82.24	61.22	−1.24
VPPO (retrained)	86.87	38.32	58.37	72.97	19.07	72.50	79.62	37.33	81.09	60.68	−1.78
PAPO (retrained)†	86.87	8.59	58.46	69.96	19.40	14.00	44.87	33.22	34.24	41.07	−21.39
Answer-CED	**93.90**	**54.35**	**78.10**	**73.10**	**23.12**	**83.75**	**88.58**	**75.67**	**93.61**	**73.80**	**+11.34**
Δ_{CED}	+4.00	+13.90	+20.17	+0.20	+4.39	+10.35	+9.14	+39.67	+11.37	**+12.58**	–

† PAPO's global-perturbation KL reward induces repetitive reasoning loops on Qwen3.5-9B (25% repeated-segment rate, 58% budget exhaustion at 24k tokens).

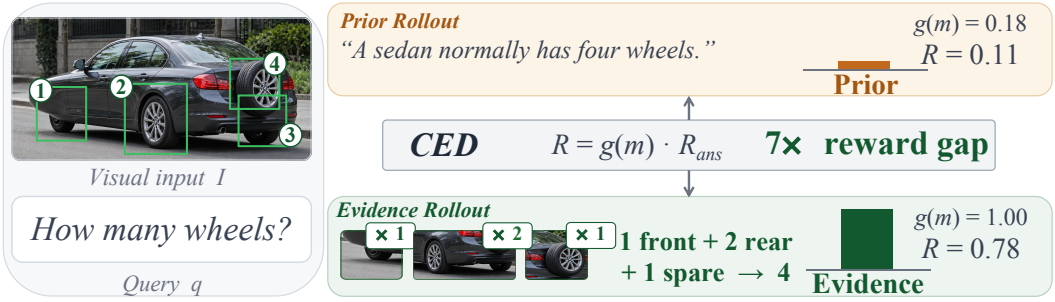


Figure 3: A case showing why correctness-only reward is insufficient. Two rollouts from the same GRPO group both answer "4" correctly, but the CED gate $g(m)$ creates a 7× reward gap between a prior-based rollout ($g(m)$=0.18, R=0.11) and an evidence-grounded rollout ($g(m)$=1.00, R=0.78), enabling GRPO to select the grounded trajectory.

4.3 Ablations and Diagnostics

Having established the aggregate gains, we now diagnose their source. We first isolate the CED contribution under strictly controlled conditions (Section 4.3.1), then examine cross-backbone generalization and the Answer-vs-CoT design choice (Section 4.3.2), and finally stress-test the reward signal against proposal and intervention perturbations (Section 4.3.3).

4.3.1 Isolating the CED Contribution

Is the CED reward necessary? Table 3 compares correctness-only, retrained VPPO, retrained PAPO, and Answer-CED under matched conditions: the same Qwen3.5-9B backbone, 15,314-sample training set, router, 2,000-step schedule, and LoRA configuration. Without CED, the same data and compute do not produce transferable improvements. Correctness-only and VPPO produce gains on several grounding benchmarks but substantial negative transfer to general reasoning, yielding net-negative average deltas. PAPO becomes unstable under its global-perturbation KL objective, with 25% repeated-segment rate and 58% budget exhaustion at 24k tokens. Answer-CED improves all nine benchmarks over correctness-only (+12.58 average) and exceeds matched VPPO by +13.12. CED uniquely converts the same data and compute into transferable improvements by restoring evidence-specific within-group discrimination among equally correct rollouts.

Figure 3 illustrates this mechanism. Two rollouts from the same GRPO group both answer "4" correctly: one relies on the prior *sedan* = 2+2 without inspecting the image, while the other detects the unusual wheel configuration (1 front +2 rear +1 spare) through visual evidence. Correctness-only reward produces no ranking signal while the CED gate $g(m)$ multiplies R_{ans} by the evidence margin, creating a 7× reward gap that GRPO uses to select the grounded trajectory.

4.3.2 Generalization and Variant Analysis

Does CED transfer across backbones? We apply the same recipe to four from two model families. Table 4 reports frozen-base scores, ours, and per-benchmark gains. Our method yields positive mean improvement on every backbone and non-negative gains on all 36 benchmark–backbone cells.

Table 4: Cross-backbone validation of Ours. Each block reports the frozen base, Ours, and absolute Δ against that block's base. The Qwen3.5-9B block additionally reports 3-seed mean$\pm$std, with its mean Δ in the last column.

	Grounding					General Reasoning				
				VLMs						
Model	CountBench	SpatialEval	Hallusion	AreBlind	FREAK	MathVista	MMBench	MMMU	ScienceQA	**Avg. Δ**
Backbone: Qwen2.5-VL-3B										
Qwen2.5-VL-3B	**70.71**	53.94	65.37	42.46	11.14	52.10	80.40	50.29	74.93	–
Ours	**70.71**	**55.56**	**67.40**	**49.51**	**25.51**	**60.40**	**82.55**	**50.57**	**80.43**	–
Δ	+0.00	+1.62	+2.03	+7.05	+14.37	+8.30	+2.15	+0.28	+5.50	**+4.59**
Backbone: Qwen2.5-VL-7B										
Qwen2.5-VL-7B	81.82	59.21	68.64	46.44	12.18	63.10	84.73	50.63	83.98	–
Ours	**88.89**	**63.34**	**70.08**	**54.02**	**26.40**	**69.40**	**87.53**	**57.47**	**87.08**	–
Δ	+7.07	+4.13	+1.44	+7.58	+14.22	+6.30	+2.80	+6.84	+3.10	**+5.94**
Backbone: Qwen3-VL-8B-Instruct										
Qwen3-VL-8B	94.90	65.26	72.86	67.01	28.85	67.54	89.15	57.82	91.99	–
Ours	**95.92**	**67.75**	**74.30**	**68.88**	**31.18**	**68.04**	**89.58**	**58.85**	**92.82**	–
Δ	+1.02	+2.49	+1.44	+1.87	+2.33	+0.50	+0.43	+1.03	+0.83	**+1.33**
Backbone: Qwen3.5-9B										
Qwen3.5-9B	80.80	52.73	47.20	46.44	9.95	80.00	88.03	66.21	90.78	–
Ours	**93.90**	**54.35**	**78.10**	**73.10**	**23.12**	**83.75**	**88.58**	**75.67**	**93.61**	–
Δ	+13.10	+1.62	+30.90	+26.66	+13.17	+3.75	+0.55	+9.46	+2.83	**+11.34**
	93.60	**56.04**	**78.35**	**73.37**	**23.11**	**84.40**	**89.87**	**75.39**	**94.46**	
Ours (3 seeds)	$\pm$0.48	$\pm$0.27	$\pm$0.21	$\pm$0.17	$\pm$0.40	$\pm$0.50	$\pm$0.04	$\pm$0.22	$\pm$0.13	**+11.83**

Table 5: Answer vs. CoT variants on Qwen3.5-9B; Avg. Δ is the mean gain over the base.

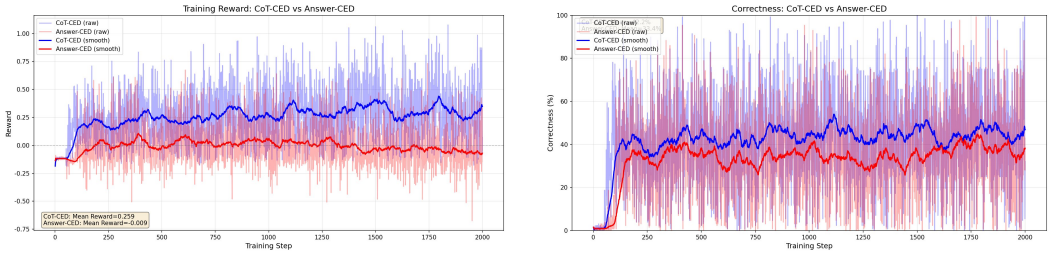
	Grounding					General Reasoning				
				VLMs						
Variant	CountBench	SpatialEval	Hallusion	AreBlind	FREAK	MathVista	MMBench	MMMU	ScienceQA	**Avg. Δ**
Answer	**93.90**	**54.35**	**78.10**	**73.10**	23.12	**83.75**	**88.58**	**75.67**	93.61	**+11.34**
CoT	92.90	53.80	**78.10**	69.09	**30.35**	81.88	88.53	68.23	**94.66**	+10.60

Which CED variant should be used: Answer or CoT? The default Answer variant directly rewards image-to-answer dependence. To isolate the effect of where the counterfactual margin is applied, Table 5 compares Answer and CoT variants on the same Qwen3.5-9B base. Answer obtains a slightly higher mean gain with stronger general transfer on downstream tasks like MMMU even though CoT obtains larger reward. This discrepancy reflects a structural mismatch between dense per-token rewards and GRPO's group-relative optimization. CoT-CED averages evidence margins over the full reasoning trace, and because trace length is policy-controlled, the model can raise the per-token average by truncating low-margin connective tokens rather than by strengthening visual grounding which is a reward-hacking pathway also documented in the process-reward literature Guo et al. [2025], Shao et al. [2024]. Tables A.13–A.14 confirm this collapse: CoT-CED's mean chain shrinks to 3.6 tokens on counting, degenerating to object-cue shorthand. Answer-CED avoids this by scoring only the final-answer span, a fixed-length outcome interface that eliminates the chain-length degree of freedom. Full length and generation diagnostics are in Appendix A.11.

Figure 4 shows why the default remains Answer-CED despite CoT-CED's stronger training-side signal. CoT-CED reaches higher reward and a denser positive margin during training, but this margin advantage does not yield stronger average transfer in Table 5. The likely failure mode is that CoT-CED can raise a per-token margin by shortening the chain around visually salient object cues, whereas Answer-CED applies the counterfactual directly to the final answer. We therefore use Answer-CED as the default variant; additional length and generation diagnostics are in Section A.11.

4.3.3 Reward Signal Robustness

Is the reward signal robust to proposal quality? Figure 5 evaluates whether CED's signal depends on proposal relevance or simply on masking magnitude. Panel (a) compares the original COCO-based proposal with a random object box as Ω^{ev} across seven task types. The evidence margin drops from $\bar{m}$=0.268 (COCO) to $\bar{m}$=0.015 (random box), and the relevant-beats-random rate drops from 0.636 to 0.487 (chance level). Panel (b) shows IoU-graded degradation on the count-exclusion diagnostic: $\bar{m}$ decreases monotonically from 0.501 at high IoU to 0.219 at low IoU, approaching the random floor at 0.147. Scale and shift perturbations are reported in Section A.7. These results confirm two properties: CED does not reward masking magnitude, and the COCO-based proposal, despite being a weak spatial prior without question-specific annotation, provides sufficient evidence localization for a discriminative signal.



(a) Training reward (b) Training correctness

Figure 4: Training dynamics for Answer-CED and CoT-CED on Qwen3.5-9B. CoT-CED produces larger training-side rewards, but the downstream comparison in Table 5 favors Answer-CED on average; we therefore use Answer-CED as the default variant.

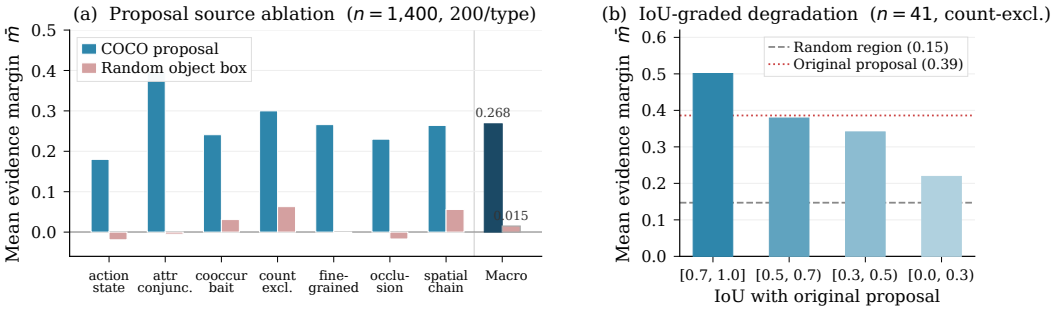


Figure 5: CED probe robustness under proposal perturbation. **(a)** Replacing the COCO proposal with a random object box as Ω^{ev} collapses the evidence margin to near zero across all seven task types, confirming that the signal tracks proposal relevance rather than masking magnitude. **(b)** IoU-graded degradation on the count-exclusion diagnostic: $\bar{m}$ decreases monotonically as spatial overlap with the original proposal drops, approaching the random-region floor at lowest IoU.

Which intervention type produces the strongest discrimination? Within local interventions, mean replacement gives the strongest discrimination, followed by zero replacement and Gaussian-noise replacement (AUC $0.669 > 0.641 > 0.629$ averaged across keying modes). Global random masking performs below chance, with an above-baseline rate of 0.490, showing that ours depends on structured local counterfactuals rather than corruption magnitude alone. With the frozen pre-RL checkpoint, the Evidence Region signal is 1.534, compared with 0.520 for shifted regions, 0.319 for wrong-category regions, and 0.566 for random regions, consistent with the proposal-robustness results in Section 4.3.3. Full perturbation results and mechanistic analysis are in Section A.5.

5 Discussion

Visual evidence over language priors on FREAK. The FREAK gains make the same mechanism visible at the trajectory level. The representative cases in Figure 6 contain images whose visual evidence contradicts a common language prior. PAPO and LLaVA-v1.6-7B follow the prior-consistent answer, whereas the CED-trained model recovers the image-grounded answer. Thus, the qualitative rollouts instantiate the rerouting predicted by Figure 1: answers that would otherwise be driven by shortcut priors are redirected toward visual evidence. The selected cases are typical members of this failure regime rather than engineered edge cases.

Behavior in the absence of visual evidence. Blank-image ScienceQA items provide the complementary test. In examples whose image is only a blank 448×448 placeholder, such as "What is the mass of a cement truck?" with options 20 tons / 20 pounds / 20 ounces, the frozen base selects the prior-consistent answer and is scored correct, while CED responds "Cannot determine from image" and is scored incorrect. This is precisely the downstream tasks' behavior targeted by the reward: when no answer is grounded in image evidence, the model should avoid substituting language priors. This also explains part of the modest ScienceQA gains in Table 4: CED improves the visual-reasoning portion as in other benchmarks, while the blank-image subset rewards the prior-following behavior that CED suppresses. The net effect remains positive on every backbone.



FREAK-548
Q: point or tail?
GT: tail
Other: point/point/head
ours: **tail**

FREAK-1683
Q: rolling-pin handles?
GT: 1
Other: 2/2/2
ours: **1**

FREAK-1694
Q: toothbrush rows?
GT: 1
Other: 3/2/3
ours: **1**

FREAK-1713
Q: eyeglass arms?
GT: 1
Other: 2/2/2
ours: **1**

FREAK-1751
Q: rickshaw tires?
GT: 3
Other: 2/2/2
ours: **3**

Figure 6: Five prior-conflict FREAK cases. Each card reports the ground truth, the outputs of PAPO, LLaVA-v1.6-7B, and the Qwen2.5-VL-7B base model in that order, and the output of ours. The non-CED models follow a plausible object prior, whereas ours follows the local visual evidence.

No text-only degradation. A potential concern is that rewarding evidence dependence may harm general text reasoning. We evaluate the Qwen3.5-9B Answer-CED checkpoint on eight text-only benchmarks (ARC-Easy, ARC-Challenge, OpenBookQA, CommonsenseQA, WinoGrande, PIQA, BoolQ, TruthfulQA-MC2) spanning commonsense, science, and factual reasoning. The mean accuracy change is -0.27 pp, with no single benchmark dropping more than 1.5 pp. CED improves visual grounding without degrading the language backbone.

Evidence-closed text-only self-evolution. The following analysis makes this intuition precise: an image-conditional signal is structurally necessary to distinguish grounded from shortcut trajectories. In VLM RL, model-judged or self-evolving audits often collapse back to text because the cheap verifiable targets used in LLM self-evolution, such as executable code or checkable arithmetic, have no per-sample-cheap visual analogue. Let S_t be the reward, I_t the image, $\hat{Y}_t$ the answer, and $G_t \in \{0, 1\}$ the latent grounding indicator. This conditional independence is a direct consequence of the path structure in Figure 1: the evidence path and the shortcut path both terminate at the answer node, so a reward that observes only the answer cannot determine which upstream path produced it.A text-only reward satisfies $S_t \perp\!\!\!\perp I_t \mid (\hat{Y}_t, Q_t)$, and therefore $I(S_t; G_t \mid \hat{Y}_t, Q_t) = 0$: two trajectories that reach the same correct answer through grounding versus a language-prior shortcut induce identical reward distributions. We call this regime *evidence-closed self-evolution*. As long as the audit observes only text, judge capacity, annealing, or trustworthiness reweighting cannot resolve the ambiguity, since the limitation comes from the conditional independence structure. CED breaks this closure by injecting an image-conditional signal. The evidence margin $m(I, q, y)$ in Equation (3) cannot be computed from $(\hat{Y}_t, Q_t)$ alone, and the strict inequality $H(G_t \mid \mathcal{B}_t, A_t) < H(G_t \mid \mathcal{B}_t)$ formalizes the empirical fact that CED reorders trajectories that correctness rewards cannot. The formal argument is developed in Section A.12. This yields a unified diagnosis of recent self-evolving multimodal RL pipelines: stable shortcut correction depends less on judge capacity than on whether the reward can observe the visual evidence path.

6 Conclusion

We introduced EVIDENCE-RL, a post-training framework for VLMs built around CED, a counterfactual evidence diagnostic motivated by a causal view of visual grounding. CED moves beyond rewarding answer correctness alone by asking whether a sampled response depends on the Evidence Region causally. With spatial-region interventions, this diagnostic can be directly integrated into GRPO as an evidence-aware reward, producing grounded reasoning behavior with no additional inference cost. Across four backbones and nine benchmarks, EVIDENCE-RL consistently improves performance, with the largest gains on benchmarks where shortcut behavior is most exposed.

Post-training should reward correct answers for depending on the visual evidence relevant to the question, rather than for matching the answer distribution alone. Though our experiments use object-centric interventions, gains generalize to various downstream tasks: the CED signal supervises the reasoning process and instills faithful evidence dependence, the capability most absent from current VLMs. Achieving high accuracy without depending on the relevant evidence is not visual reasoning; making that dependence a prioritized training objective is what evidence-intensive reasoning requires.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and area chair for their constructive feedback.

Funding. This work received no specific third-party funding.

Competing Interests. The authors declare no competing interests.

References

Emre Can Acikgoz, Cheng Qian, Jonas Hübotter, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. Tool-r0: Self-evolving llm agents for tool-learning from zero data. *arXiv preprint arXiv:2602.21320*, 2026.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025.

Tianhao Chen, Yuheng Wu, Kelu Yao, Xiaogang Xu, Xiaobin Hu, and Dongman Lee. Last but not least: Boundary attention calibration for multimodal kv cache compression. *arXiv preprint arXiv:2606.14782*, 2026.

Yuetian Du, Yucheng Wang, Rongyu Zhang, Zhijie Xu, Boyu Yang, Ming Kong, Jie Liu, and Qiang Zhu. Linking perception, confidence and accuracy in mllms. *arXiv preprint arXiv:2603.12149*, 2026.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *arXiv preprint arXiv:2505.17018*, 2025.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, pages 3449–3457, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*, 2025a.

Yantao Li, Qiang Hui, Chenyang Yan, Kanzhi Cheng, Fang Zhao, Chao Tan, Huanling Gao, Jianbing Zhang, Kai Wang, Xinyu Dai, et al. Palmr: Towards faithful visual reasoning via multimodal process alignment. *arXiv preprint arXiv:2603.06652*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3170–3180, 2023.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, and Cees GM Snoek. Regionreasoner: Region-grounded multi-round visual reasoning. *arXiv preprint arXiv:2602.03733*, 2026.

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

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

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

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

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, et al. Vicrit: A verifiable reinforcement learning proxy task for visual perception in vlms. *arXiv preprint arXiv:2506.10128*, 2025c.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025d.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. Revis: Sparse latent steering to mitigate object hallucination in large vision-language models. *arXiv preprint arXiv:2602.11824*, 2026.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

Haoxuan Xu, Yi Liu, Tianfu Li, Ruolin Shen, Boyuan Jiang, Jinlong Peng, Donghao Luo, Xiaobin Hu, Shuicheng Yan, and Haoang Li. Irpo: Boosting image restoration via post-training grpo. *arXiv preprint arXiv:2512.00814*, 2026.

Yilin Yang, Zhenghui Guo, Yuke Wang, Omprakash Gnawali, Sheng Di, and Chengming Zhang. Hii-dpo: Eliminate hallucination via accurate hallucination-inducing counterfactual images. *arXiv preprint arXiv:2602.10425*, 2026.

Bo Yin, Xiaobin Hu, Chengming Xu, Ruolin Shen, Mo Yang, Jiangning Zhang, Peng-Tao Jiang, Cheng Tan, and Shuicheng Yan. Spot-e: Test-time entropy shaping with visual spotlights for frozen vlms. *arXiv preprint arXiv:2606.20244*, 2026a.

Zhihan Yin, Jianxin Liang, Yueqian Wang, Yifeng Yao, Huishuai Zhang, and Dongyan Zhao. Freak: A fine-grained hallucination evaluation benchmark for advanced mllms. *arXiv preprint arXiv:2603.19765*, 2026b.

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31544–31555, June 2026.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. Perceptual-evidence anchored reinforced learning for multimodal reasoning. *arXiv preprint arXiv:2511.18437*, 2025a.

Juntian Zhang, Song Jin, Chuanqi Cheng, Yuhan Liu, Yankai Lin, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, et al. Viper: Empowering the self-evolution of visual perception abilities in vision-language model. *arXiv preprint arXiv:2510.24285*, 2025b.

Rosie Zhao, Anshul Shah, Xiaoyu Zhu, Xinke Deng, Zhongyu Jiang, Yang Yang, Joerg Liebelt, and Arnab Mondal. On robustness and chain-of-thought consistency of rl-finetuned vlms. *arXiv preprint arXiv:2602.12506*, 2026.

Table A.1: Complete hyperparameter listing. "Paper notation" gives the corresponding symbol in the main text when applicable.

Category	Parameter	Value	Paper notation
RL & optimization			
Learning rate	`lr`	1×10^{-5}	—
KL penalty coefficient	`kl_coeff`	0.01	—
GRPO group size	`group_size`	32	—
Training steps	`n_steps`	2,000	—
Trainable layers	`n_trainable_layers`	4 (last)	—
Precision	`dtype`	bfloat16	—
Generation & sampling			
Temperature	`temperature`	1.0	—
Top-p	`top_p`	0.95	—
Max new tokens (evidence)	`max_new_tokens`	32	—
Reward function (Equation (4)*)*			
Response reward weight	`alpha_resp`	0.70	see note[†]
Answer reward weight	`alpha_ans`	0.30	see note[†]
Response logprob temp.	`tau_resp`	0.20	—
Answer logprob temp.	`tau_ans`	1.00	—
Gate temperature	`tau_resp` (shared)	0.20	τ_g
Gate baseline at m=0	(derived)	0.50	$g(0)$
Evidence tie-breaker	`evidence_eps`	0.10	ε_{tie}
Additive evidence weight	`evidence_eps` (shared)	0.10	λ
Reward clipping	`[min, max]`	$[-1.25,\ 1.00]$	—
Data			
Rerank candidates	`rerank_num_candidates`	6	—
non-evidence Regions per sample	K	3	K

[†] In Equation (4), R_{ans} is a composite: $R_{\text{ans}} =$ `alpha_ans` $\cdot$ answer_score $+$ `alpha_resp` $\cdot$ response_score, where each score is computed from log probabilities scaled by its temperature.

A Supplementary Experiments and Implementation Details

A.1 Reproducibility Details

Dataset split and task-family distribution. The training data are derived from COCO val2017 [Lin et al., 2014] after quality filtering and format normalization. This yields 21,758 raw samples, split into train (17,502 valid), validation (2,352 valid), and probe (1,904 valid) subsets. After RL-specific filtering, 15,314 samples are loaded for online training. The loaded training set contains three task families: attribute (6,562), counting (2,188), and spatial (6,564). Training images are drawn from COCO val2017 [Lin et al., 2014]. CountBench [Paiss et al., 2023], SpatialEval [Wang et al., 2024], and MathVista [Lu et al., 2023] use independently sourced evaluation images and share no image-level overlap with the training set; the remaining benchmarks are cited in Section 4 and likewise contain no COCO val2017 images.

Training configuration. Training uses `balanced_no_replacement` sampling for uniform coverage across task families, GRPO group size 32, and 2,000 optimization steps. The trainable part of the model is restricted to the last four layers. Full training and reward hyperparameters are listed in Table A.1.

Broader impacts. CED is a training-time diagnostic and reward component for improving visual grounding in VLMs. It does not directly interact with end users or generate user-facing content. Improved grounding may reduce shortcut-driven errors in safety-relevant visual reasoning tasks. The main risk is that counterfactual probing can also reveal model vulnerabilities, although CED exposes only a scalar evidence margin rather than a detailed attack surface.

Table A.2: Methodological comparison with perception-aware VLM RL methods. The *Causal audit* column indicates whether the method tests causal evidence dependence for a candidate answer.

Method	Signal source	Granularity	Causal audit	Stage
PAPO [Wang et al., 2025d]	Global mask KL	Global	No	Training
VPPO [Huang et al., 2025]	Attention weights	Token-level	No	Training
Vision-SR1[†]	Text self-verification	Description-level	No	Training
PaLMR[‡]	LLM-as-Judge	Trajectory-level	No	Training
SophiaVL-R1 [Fan et al., 2025]	Thinking-reward RM	Trajectory-level	No	Training
PEARL[§]	Perception checklist	Sample-level	No	Training
VAPO[¶]	Trajectory anchoring	Token-level	No	Training
VLM-R1[*]	Self-verification	Response-level	No	Training
Perception-R1[††]	Proxy localization	Region-level[a]	No	Training
CED (Ours)	Counterfactual intervention	**Region-level**	**Yes**	Training

[†]Li et al. [2025b]; [‡]Li et al. [2026]; [§]Zhang et al. [2025a]; [¶]Yue et al. [2025]; [*]Shen et al. [2025]; [††]Xiao et al. [2025].
[a]Perception-R1 reaches region-level granularity through proxy localization and requires additional supervision.

A.2 Intervention Implementation Details

Replacement strategies. We evaluate three replacement strategies: *mean replacement* (default), *zero replacement*, and *Gaussian-noise replacement*. All interventions are applied to the last-layer vision-encoder output after spatial merging. Mean replacement substitutes Evidence Region visual tokens with the mean of neighboring visual tokens outside $\mathcal{T}(\Omega)$; zero replacement sets selected tokens to $\mathbf{0}$; Gaussian-noise replacement substitutes zero-mean Gaussian samples.

Spatial-merge-aware token mapping. Modern vision encoders apply a 2×2 spatial merge before feeding visual tokens to the LLM. Given merge dimensions $(s_w, s_h) = (2, 2)$, an original grid coordinate (g_x, g_y) maps to the merged-token index

$$(g_y \mathbin{//} s_h) \times w_{\text{merged}} + (g_x \mathbin{//} s_w).$$

For non-evidence Region construction, the mapping module supports Chebyshev-distance ring expansion on the merged feature map.

Proposal construction pipeline. Positive proposals are sourced from COCO instance annotations [Lin et al., 2014]. They are object-level bounding boxes, not question-specific region annotations. The Evidence Region Ω^{ev} is resolved from sample metadata using the priority `proposal_bbox` > `target_bbox` > `argument_bbox` > `metadata.bbox` > center-box fallback. non-evidence Regions $\{\Omega_k^{\text{non}}\}$ are sampled as area-matched random boxes with IoU ≤ 0.05 to Ω^{ev}. If box sampling fails, random visual-token subsets of matching size are used as fallback. Quality diagnostics filter proposals with extreme area fractions or insufficient token coverage before training.

A.3 Comparison with Perception-Aware RL Methods

Table A.2 compares CED with representative perception-aware VLM RL methods. The key distinction is whether the training signal tests candidate-answer dependence on local visual evidence, rather than coarse image dependence or text-only consistency.

A.4 Additional Evidence-Signal Diagnostics

Relevant-vs-random intervention. On a held-out set of 500 samples per task family, we compare Evidence Region sensitivity $s(\Omega^{\text{ev}})$ with mean non-evidence Region sensitivity $\mu(s^{\text{non}})$. On counting tasks, the Evidence Region elicits higher sensitivity in 63.4% of samples, with mean Evidence Region $s = 2.566$ versus mean non-evidence Region $s = 0.638$. On presence tasks, the rate drops to 55.0%, with mean Evidence Region $s = 1.082$ versus mean non-evidence Region $s = 0.302$. This supports the regime separation used by the routing rule in Section 3.2.

Cross-model validation. We also validate the signal on InternVL3.5-8B [Wang et al., 2025b] and LLaVA-v1.6-Mistral-7B [Li et al., 2024] using pixel-level intervention. Table A.3 shows that

Table A.3: Cross-model evidence sensitivity validation (N=50 per task per model).

Model	Task	Relevant > Random rate	Mean margin
InternVL3.5-8B	Counting	0.68	0.0070
	Presence	0.62	0.0737
LLaVA-v1.6-7B	Counting	0.78	0.0074
	Presence	0.60	0.1625

Table A.4: Blindfold test results (N=500 per task family).

Condition	Task	Rel. > Rand. rate	Mean Delta
Normal	Counting	0.634	$+1.928$
Normal	Presence	0.550	$+0.780$
Blindfold	Counting	0.404	-0.114
Blindfold	Presence	0.448	$+0.070$

counting tasks maintain higher relevant-vs-random rates across both models, while presence tasks remain weaker.

Blindfold test. After replacing the image with a blank input, the relevant-vs-random rate drops to 0.404 on counting and 0.448 on presence, both below the 0.50 chance level. Mean delta collapses from $+1.928/+0.780$ to $-0.114/+0.070$, confirming that the signal originates from real image content.

A.5 Intervention-Type Ablations

Why mean replacement reduces intervention artifacts. The perturbation ablation shows that mean replacement gives the strongest discrimination within the local intervention family: AUC 0.669 for mean replacement, compared with 0.641 for zero replacement and 0.629 for Gaussian-noise replacement. This ordering follows from the role of the replacement as a structural baseline. Write the Evidence Region token representation as

$$h_T = b_T + e_T, \tag{5}$$

where b_T is a context-consistent baseline component and e_T is the answer-relevant evidence component. The ideal counterfactual removes e_T while preserving b_T. Replacement strategies differ in how much they perturb this baseline:

$$h_T^{\text{mean}} = \mu_T = b_T + \delta_T, \qquad h_T^{\text{zero}} = \mathbf{0}, \qquad h_T^{\text{noise}} = b_T + \xi_T. \tag{6}$$

Let F be the downstream logit assigned to the candidate answer and J_T its local Jacobian with respect to h_T. A first-order expansion gives

$$\Delta F^{\text{mean}} \approx J_T(\delta_T - e_T), \qquad \Delta F^{\text{zero}} \approx -J_T(b_T + e_T), \qquad \Delta F^{\text{noise}} \approx J_T(\xi_T - e_T). \tag{7}$$

All three contain the desired signal term $-J_T e_T$. Mean replacement adds only the neighborhood-offset artifact $J_T\delta_T$, zero replacement adds the systematic baseline-removal artifact $-J_T b_T$, and Gaussian replacement adds sample-specific noise. Since context mixing makes the neighborhood offset smaller than the full baseline norm at the intervention layer, mean replacement preserves the non-evidence part of the representation more faithfully.

AUC ablation and global masking. Table A.5 reports the local perturbation-family ablation. Global masking performs worse: it does not support stable answer-conditioned discrimination, with above-baseline rate 0.490 below chance. Local counterfactual intervention retains discriminative utility with AUC 0.620.

A.6 Reward Hyperparameter Sensitivity

Tables A.7 and A.8 show that the default reward hyperparameters lie in stable regions. Across $\lambda \in [0.05, 0.20]$, the non-constant reward rate remains 96–100%. For τ_g, values around 0.15–0.30 yield balanced selection behavior; we use $\tau_g = 0.20$ by default.

Table A.5: Perturbation-family ablation within the local intervention family, evaluated by logits-JS AUC.

Key mode	Mean repl.	Zero repl.	Gaussian-noise repl.
prompt_last	**0.6780**	0.6602	0.6299
answer_first	**0.6592**	0.6215	0.6273
Average	**0.6686**	0.6409	0.6286

Table A.6: Gaussian-noise profile under `prompt_last`. Stability drops under `logits_only`.

Config	Fixed metric	Best metric	Best paired AUC
logits24	0.6299	0.6299	0.7108
logits_only	0.5753	0.5920	0.6407

Table A.7: Evidence weight (λ) sensitivity. $\lambda = 0.00$ is correctness-only.

λ	Non-const. rate	Reward std	Zero-var. rate
0.00	100%	0.0528	0%
0.05	100%	0.0508	0%
0.10	**96%**	**0.0597**	**4%**
0.15	100%	0.0415	4%
0.20	100%	0.0491	0%

Table A.8: Gate temperature (τ_g) sensitivity.

τ_g	Non-const. rate	Reward std	Zero-var. rate
0.10	100%	0.0568	4%
0.20	**96%**	**0.0597**	**4%**
0.30	96%	0.0469	4%
0.50	100%	0.0378	0%

A.7 Proposal Robustness to Spatial Perturbations

This experiment tests whether the CED probe remains useful when the target proposal is imperfect. It is orthogonal to Section A.5: that subsection varies the intervention type while holding the region fixed; here we hold the intervention type fixed as mean replacement and vary the region.

Protocol. Starting from the true Evidence Region box of each sample, we generate perturbed proposals and re-run the full CED pipeline with K=3 non-evidence Regions. We test three perturbation axes: IoU stratification with the original box, scale changes around the box center, and translations by a fixed fraction of the image short side. A `random` region baseline gives the expected floor without spatial targeting. We report mean evidence margin $\bar{m}$, evidence-non-evidence gap $\Delta = \mu(s^{\text{ev}}) - \mu(s^{\text{non}})$, and evidence-beats-random rate $\Pr(s^{\text{ev}} > s^{\text{rand}})$. All numbers use the Qwen3.5-9B CoT-CED checkpoint on n=41 count-exclusion samples.

Interpreting the results. IoU gives the cleanest trend: $\bar{m}$ and Δ both decrease as overlap with the original proposal drops. The lowest IoU bin approaches the random floor, indicating that CED does not merely reward masking magnitude. Scale and shift perturbations do not show systematic degradation within the tested ranges. Their evidence-reference gaps stay well above the random baseline, suggesting that moderate detector imprecision is a second-order effect for this diagnostic. All measurements here are probe diagnostics rather than downstream accuracy estimates.

A.8 Shared-Candidate Baselines and Reward Diagnostics

Unified-condition protocol. Recent perception-aware methods differ in backbone and data. We therefore separate checkpoint transfer from signal design by evaluating lightweight method variants under a shared-candidate protocol with the same backbone, candidate pool, and benchmark manifest.

Table A.9: Proposal-perturbation results on count-exclusion (n=41). Higher values indicate stronger evidence dependence; the `random` row gives the floor without spatial targeting.

Axis	Condition	$\bar{m}$	Δ	Rel. rate
IoU stratification	$[0.7, 1.0]$	**0.501**	**2.032**	**0.732**
	$[0.5, 0.7)$	0.379	1.282	0.683
	$[0.3, 0.5)$	0.341	1.386	0.683
	$[0.0, 0.3)$	0.219	0.555	0.561
Scale	$0.50\times$	0.377	1.113	0.732
	$0.75\times$	0.481	1.750	0.732
	$1.25\times$	0.433	2.230	0.683
	$1.50\times$	0.295	2.003	0.634
	$2.00\times$	0.382	1.880	0.707
Shift	5%	0.304	1.602	0.634
	10%	0.428	1.457	0.707
	20%	0.532	1.899	0.732
original proposal (reference)		0.386	1.852	0.683
`random` region (floor)		0.147	0.277	0.537

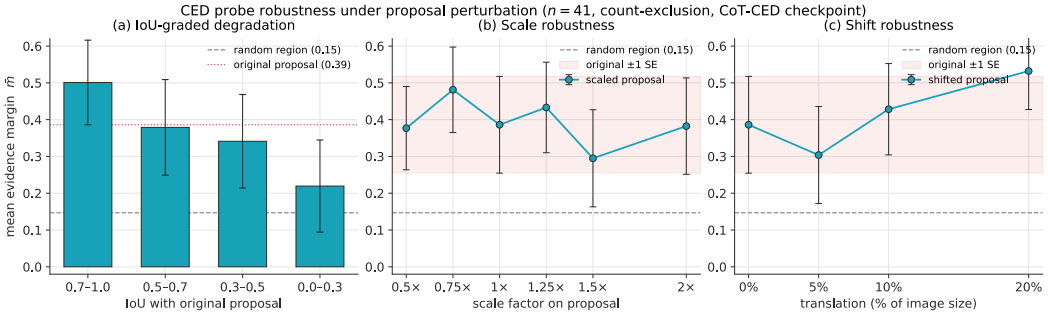


Figure A.1: CED probe robustness under proposal perturbation on the count-exclusion diagnostic. (a) Mean evidence margin $\bar{m}$ across IoU bins, with the random-region baseline as a grey dashed reference. (b,c) $\bar{m}$ under scale and translation perturbations, with the original ± 1 SE band in red. Error bars are ± 1 SE.

Table A.10: Unified-condition baseline comparison on the offline rerank dataset (N=1,353).

Method	Reward mean	Signal strength	Signal-corr
Vanilla GRPO	-0.149	N/A	N/A
PAPO-Lite	-0.156	KL $= 0.024$	0.034
CED (Ours)	$+0.448$	Margin $= -0.007$	Selective

Shared-candidate comparison. Table A.11 reports Lite versions of representative methods on the counting shared-candidate slice. The key readout is not final accuracy alone, which is saturated on this slice, but whether the reward can produce non-constant within-group signal and flip candidate preference beyond log-probability ranking.

Within-group variance telemetry. Correctness-only has within-group zero-variance rate 0.949, non-constant rate 0.051, and same-answer-different-reward discrimination 0.0. Additive and ours reduce the zero-variance rate to 0.035 and restore the non-constant rate to 0.965. Ours gives slightly higher discrimination, 0.555 versus 0.535. Re-scoring the pre-RL checkpoint gives a yes/no shortcut bias gap of 0.077 for Correctness-only, 0.075 for Additive, and 0.015 for ours. The intervention area-reward Pearson correlation is 0.080, ruling out area-driven reward hacking. Reward-length correlations are consistent across methods: -0.226, -0.222, and -0.230.

Table A.11: Unified-condition method comparison on the counting shared-candidate slice: 94 samples with at least two distinct candidates.

Method	Acc.	Strong inf.	$r_{\mathrm{nc}} \uparrow$	flip $\uparrow$
LogProb	94.7	75.0	–	–
Correctness-only	97.9	100.0	12.8	4.3
CED (Ours)	**97.9**	**100.0**	**100.0**	**76.6**
Perception-R1-Lite	97.9	100.0	13.8	5.3
VLM-R1-Lite*	94.7	75.0	0.0	0.0
VPPO-Lite*	94.7	75.0	0.0	0.0
PAPO-Lite*	94.7	75.0	0.0	0.0

* No reward-based reranking recovered; falls back to log probability.

A.9 Reward Discrimination and Failure Modes

The same-answer-different-reward statistic in Table 1 is visualized in Figure A.2. Rollouts with the same final answer receive different rewards once the evidence margin enters the signal, reflecting differences in the visual evidence each rollout uses.

Under Correctness-only, evidence-intensive tasks exhibit low within-group discrimination, with non-constant raw reward rate $r_{\mathrm{nc}} \approx 0.05$ and zero-variance rate ≈ 0.95. Additive and Routed rewards restore discrimination, with $r_{\mathrm{nc}} \approx 0.97$ and zero-variance rate ≈ 0.04. On the presence probe, all formulations remain limited by the binary-task variance bound, supporting the routing of presence tasks to the audit pathway.

A.10 Scope and Extensions of the Evidence Interface

Behavioral target. CED is designed to reward a model behavior: a candidate answer should be supported by the visual evidence that justifies it. In the causal view of grounding, this means that the model's support for an answer should decrease more under intervention on the target-evidence path than under matched interventions on non-evidence paths. This requirement is not specific to bounding boxes or object regions.

Region-level instantiation. In this paper, we instantiate the audit with object-centric region interventions because regions provide a scalable interface for constructing Evidence Regions and matched non-evidence Regions without question-specific evidence annotations. This choice aligns with the benchmark families studied in our experiments, including counting, spatial reasoning, hallucination detection, and object-centric visual reasoning, where many failures arise when the model answers from language priors, scene context, or salient distractors rather than the relevant local evidence.

Beyond object boxes. In the causal graph of Figure 1, CED rewards a path-level property: the answer should depend more on the target-evidence path than on shortcut or nuisance paths. Object-region interventions query this property by probing one spatial locus of the evidence path without question-specific evidence annotations, making the audit scalable to open-ended VQA training. The behavioral target itself is defined on the graph structure, not on the intervention format. Richer evidence interfaces such as attribute-preserving interventions, multi-region proposals, or text-span units query the same path at finer granularity or for different evidence types; they refine how the evidence path is probed, not what behavior is rewarded.

A.11 Answer-CED vs. CoT-CED Details

The two variants differ in where the evidence sensitivity is computed: CoT-CED applies the counterfactual to the full reasoning trace, while Answer-CED applies it only to the final answer. Both variants are trained for $2{,}000$ steps on Qwen3.5-9B with LoRA (r=16, α=32). The main paper reports the training trajectories in Figure 4; here, Table A.12 summarizes the final training state and Tables A.13 and A.14 provide the behavior-level diagnostics.

CoT-CED achieves higher training reward (Table A.12, Figure 4), yet transfers worse on average (Table 5). This reflects a structural mismatch between dense per-token rewards and GRPO's group-relative optimization. CoT-CED averages evidence rewards over the full reasoning trace, and because



Figure A.2: Qualitative examples of same-answer-different-reward under the CED method. All rollouts share the same answer, while CED produces reward spans of 0.50–0.67 according to visual evidence quality.

Table A.12: CoT-CED vs. Answer-CED: training-end summary over 2,000 steps.

Metric	**CoT-CED**	**Answer-CED**
Mean training reward	$+0.259$	-0.009
Final training correctness	43.4%	34.2%

trace length is policy-controlled, the model can raise the per-token average by truncating low-reward tokens rather than by improving visual grounding. Tables A.13–A.14 confirm this collapse: CoT-CED's mean chain shrinks to 3.6 tokens on counting (vs. 49.3 for Answer-CED), degenerating to object-cue shorthand. This parallels the process-reward hacking reported by Guo et al. [2025], who abandon dense process rewards in favor of outcome-level signals for the same reason. Answer-CED removes this degree of freedom by scoring only the final-answer span, a fixed-length interface whose reward cannot be manipulated through generation length.

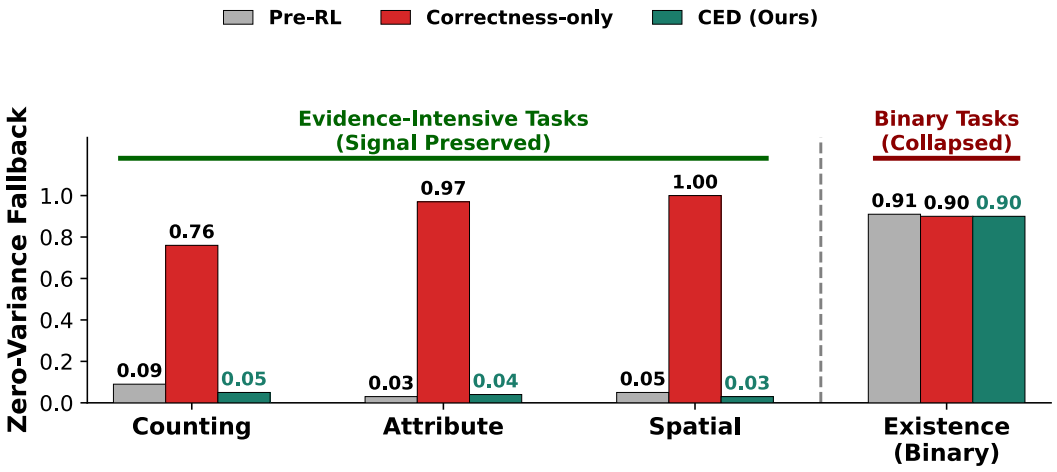


Figure A.3: Within-group zero-variance rate across task families. Evidence-intensive tasks show a large reduction from correctness-only reward to our reward; binary-dominated groups remain near high zero-variance because of their two-action structure.

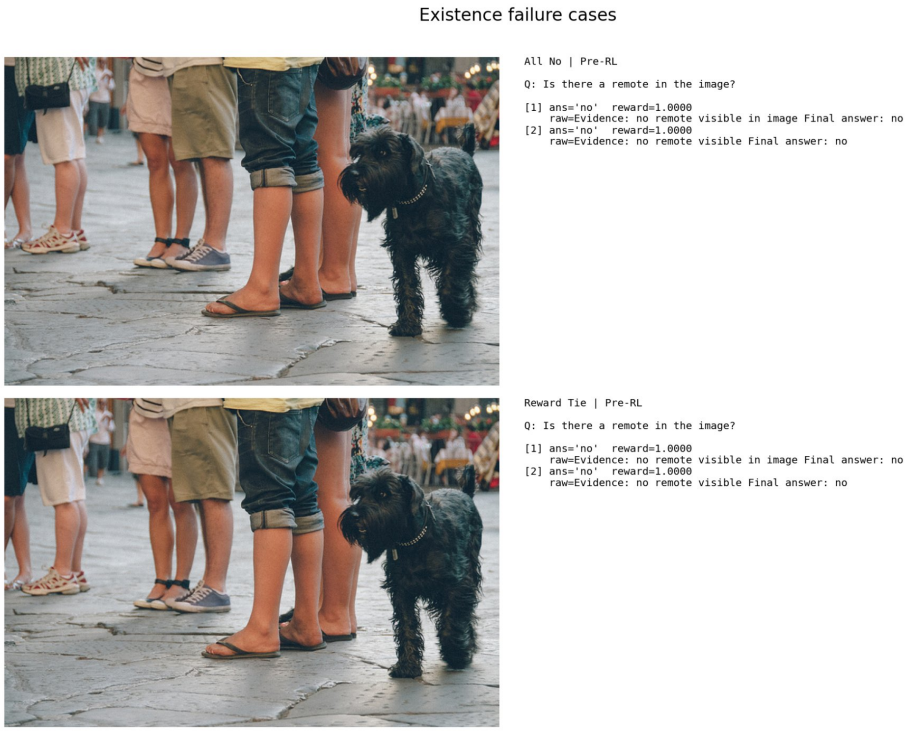


Figure A.4: Failure cases on presence tasks: identical negation answers with identical rewards yield zero within-group variance.

A.12 Formalizing Evidence-Closed Self-Evolution

This appendix provides the complete formalization of the evidence-closed self-evolution regime introduced in Section 5.

Notation. Let $O_t = (I_t, Q_t)$ denote the image-question pair, $\hat{Y}_t$ the prediction, $V_t = \nu(I_t)$ the task-relevant visual evidence, $G_t \in \{0, 1\}$ the grounding indicator, and S_t the endogenous reward.

Table A.13: End-of-training CoT behavior on held-out prompts. CoT-CED collapses toward an object-cue answer format, while Answer-CED preserves natural multi-sentence reasoning.

Metric	Benchmark	CoT-CED	Answer-CED
Mean CoT length (tokens)	Counting	3.6	49.3
Median CoT length (tokens)	Counting	3	47
p_{90} CoT length (tokens)	Counting	6	78
Mean CoT length (tokens)	HallusionBench	3.8	82.1
Median CoT length (tokens)	HallusionBench	3	61
p_{90} CoT length (tokens)	HallusionBench	6	173
Image-referential opener rate	Counting	1.0%	79.8%
Image-referential opener rate	HallusionBench	5.2%	64.2%
Unique-token ratio	Counting	99.1%	75.8%
Unique-token ratio	HallusionBench	96.8%	68.1%

Table A.14: Representative side-by-side generations. Both variants produce the correct final answer, but CoT-CED reduces the chain to an object cue, while Answer-CED keeps an explicit visual check.

Prompt	CoT-CED	Answer-CED
How many beds are there in the image?	`The beds / 2`	*Got it, let's look at the image. There's a loft bed with a lower bed. The upper bunk is one, the lower bed is another. So that's two beds.*
How many weights are there in the image?	`The weights / Two`	*Got it, let's look at the image. The man is holding two dumbbells, one in each hand. There's one in the left hand and one in the right hand. So that's two weights.*

Definition 1 (Evidence-closed self-evolution)**.** *A training step is **evidence-closed** if its signal T_t^{cl} is measurable with respect to $\mathcal{B}_t = \sigma(\mathcal{F}_t, I_t, Q_t, \hat{Y}_t, S_t)$.*

Lemma 1 (Endogenous reward bottleneck)**.** *If $S_t \perp\!\!\!\perp I_t \mid (\hat{Y}_t, Q_t, \mathcal{F}_t)$, then $I(S_t; V_t \mid \hat{Y}_t, Q_t, \mathcal{F}_t) = 0$.*

Proof. Since $V_t = \nu(I_t)$, the conditional independence implies $S_t \perp\!\!\!\perp V_t \mid (\hat{Y}_t, Q_t, \mathcal{F}_t)$, which gives the result. □

Proposition 1 (No incremental grounding information)**.** *If T_t^{cl} is evidence-closed, then $I(T_t^{\mathrm{cl}}; G_t \mid \mathcal{B}_t) = 0$.*

Proof. T_t^{cl} is $\mathcal{B}_t$-measurable, so conditioning on $\mathcal{B}_t$ leaves no remaining uncertainty in T_t^{cl} that can carry additional information about G_t. □

Proposition 2 (Closed-loop reward does not identify grounding)**.** *If two predictors induce the same distribution over $(I_t, Q_t, \hat{Y}_t, S_t)$ but differ in grounding, then any objective of the form $J(h) = \mathbb{E}_h[\phi(I_t, Q_t, \hat{Y}_t, S_t)]$ cannot distinguish them.*

Proof. The objective depends only on the joint distribution of $(I_t, Q_t, \hat{Y}_t, S_t)$. If two predictors have the same joint distribution over these variables, they have the same value of $J(h)$ regardless of their grounding behavior. □

Assumption 1 (Audit informativeness)**.** *On evidence-intensive tasks, $I(A_t; G_t \mid \mathcal{B}_t) > 0$, where A_t denotes the evidence audit signal.*

Proposition 3 (Information increment from evidence auditing)**.** *Under Assumption 1, $H(G_t \mid \mathcal{B}_t, A_t) < H(G_t \mid \mathcal{B}_t)$.*

Proof. By the definition of conditional mutual information, $I(A_t; G_t \mid \mathcal{B}_t) = H(G_t \mid \mathcal{B}_t) - H(G_t \mid \mathcal{B}_t, A_t)$. The assumption makes the left-hand side strictly positive, giving the stated strict inequality. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the scope and contributions of EVIDENCE-RL and CED, and these claims are supported by the method, experiments, and analysis sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses scope and limitations in Section A.10. The discussion clarifies that CED targets a general evidence-dependent model behavior, while the experiments instantiate this audit with object-centric region interventions. The appendix also analyzes reward discrimination, failure modes, and robustness.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The appendix states the assumptions used in the theoretical analysis and provides proofs for the formal claims. These results clarify sufficient conditions for evidence-based rewards to improve evidence reliance rather than characterize arbitrary VLM training dynamics.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The appendix reports dataset construction, splits, training configuration, reward details, and hyperparameters needed to reproduce the main experiments. Code and model weights are publicly available at `https://github.com/evidencerl/code` and `https://huggingface.co/hhj-ai/Evidence-RL-9B`.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The training data are based on publicly available datasets. Code is released at `https://github.com/evidencerl/code` and model weights are released at `https://huggingface.co/hhj-ai/Evidence-RL-9B`, with instructions to reproduce the main experimental results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setup is described in the experiments section and appendix, including datasets, backbones, baselines, evaluation protocols, training configuration, reward details, and hyperparameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: For the primary backbone (Qwen3.5-9B), we report 3-seed mean and standard deviation across all nine benchmarks (Table 4). The remaining backbone and baseline comparisons are single-run due to the computational cost of post-training large VLMs, but are evaluated on a broad set of benchmarks and backbones to reduce over-interpretation of stochastic variation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The anonymous manuscript reports reproducibility-relevant training settings and hyperparameters, but does not report hardware-specific worker type, memory, or wall-clock time. These details will be documented in the public release.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and confirm that the work complies with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The appendix includes a broader impacts discussion. The method is intended to improve visual grounding and reduce unsupported VLM answers, while the discussion notes that better grounding does not eliminate failures or guarantee safe deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The work does not release a high-risk dataset or a model with a special dual-use capability beyond standard VLM post-training. Potential societal impacts are discussed separately in the broader impacts section.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The public datasets, benchmarks, and base models used in this work are cited in the paper. The released code repository documents the licenses and usage terms for all assets.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Code is released at `https://github.com/evidencerl/code` and model weights are released at `https://huggingface.co/hhj-ai/Evidence-RL-9B`, with documentation. The training data are based on publicly available datasets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, human-subject experiments, or human preference annotation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used only for writing, editing, formatting, or code-debugging assistance and did not affect the core methodology, experimental design, scientific claims, or originality of the research.

    Guidelines:

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