Attend to Evidence: Evidence-Anchored Spatial Attention Supervision for Multimodal RLVR

Ruina Hu[1,2,†], Chen Wang[2,4,†], Lai Wei[2,5], Jionghao Bai[2,6], Bin Yu[1,2,3], Weiran Huang[5], Kai Wang[1,*], Yue Wang[2,7,*]

[1]Harbin Institute of Technology [2]Zhongguancun Academy
[3] Harbin Institute of Technology Qingdao Research Institute
[4]Nankai University [5]Shanghai Jiaotong University [6]Zhejiang University [7]XinZhuAI
s-hrn25@bza.edu.cn, dr.wangkai@hit.edu.cn, yuewang@bza.edu.cn

Abstract

Reinforcement learning with verifiable rewards (RLVR) improves vision-language models (VLMs) by optimizing outcome rewards derived from final answers. However, such outcome-only rewards do not tell the model which image regions justify an answer. For questions that require visual grounding, these rewards cannot distinguish responses supported by relevant visual evidence from those produced by language-prior shortcuts or lucky guesses. We introduce **EASE** (Evidence-Anchored Spatial Attention), which augments multimodal RLVR with visual-evidence process supervision. EASE converts annotated evidence regions into a smoothed visual-token target and uses it to guide response-to-image attention during RL training, but only on high-reward trajectories. The annotations are used solely as privileged training labels, while inference requires only the original image and question. Across Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B, EASE raises average scores over DAPO by 2.5 to 3.1 points on perception, hallucination, visual math, and multimodal reasoning benchmarks. Diagnostics and ablations show that EASE better aligns visual attention with annotated evidence regions. Code are available at https://github.com/Nrich-sunny/Attend-to-Evidence/.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a common way to improve language-model reasoning and is increasingly used in vision-language model (VLM) post-training (Shao et al., 2024; Yu et al., 2026; Yang et al., 2025a; Liu et al., 2025b; Tan et al., 2026). Outcome rewards are attractive in multimodal tasks because many answers can be checked automatically. Recent VLM systems show that these verifiable outcome signals can improve multimodal reasoning (Wang et al., 2025c,a; Meng et al., 2025; Liu et al., 2025a; Wang et al., 2025d; Huang et al., 2025; Wang et al., 2026; Cao et al., 2025; Ni et al., 2026). However, outcome-only reward does not tell the model where the answer should come from in the image. A VLM may need to find a small object, compare two regions, or combine clues across separated regions before it can answer reliably. If this evidence-acquisition step is left unsupervised, RL can reward a correct answer without knowing whether the model actually used the relevant visual evidence rather than a shortcut.

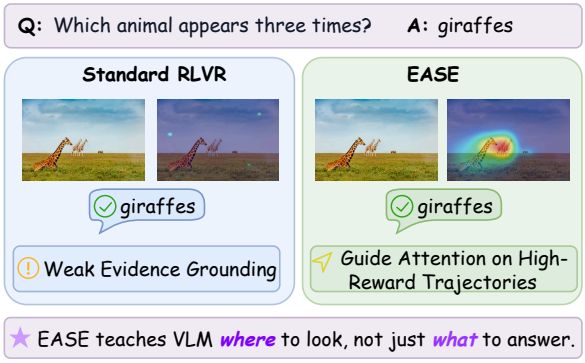


Figure 1: **From outcome rewards to evidence acquisition.** Standard RLVR may answer correctly while attending weakly to the key evidence region. EASE encourages stronger attention to the supporting region.

This limitation creates a *perception credit assignment* problem. An outcome reward tells the policy whether the response is correct, but not which image regions supported that response. A correct answer may reflect relevant visual evidence, but it may also come from language priors, dataset bias, a label-correlated shortcut, or guessing. An incorrect answer is also ambiguous because the model may have looked at the wrong region, or it may have found the right evidence and then reasoned incorrectly. This ambiguity matters most

[†]Equal contribution.
[*]Equal corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 4295–4313
October 24-29, 2026 ©2026 Association for Computational Linguistics

for fine-grained and multi-evidence questions that require object comparison, counting, localized attributes, spatial relations, or clues from separated regions. Without acquiring these regions, a model can produce plausible text while remaining weakly grounded and vulnerable to hallucination (Li et al., 2023a; Guan et al., 2023; Asadi et al., 2026; Shi et al., 2026; Xu et al., 2026; Fu et al., 2025).

Process supervision can reduce this ambiguity by making visual evidence acquisition part of the training signal. Evidence boxes and human attention annotations in visual question answering data indicate which image regions support an answer and can serve as training-only labels for where the model should look (Hudson and Manning, 2019; Das et al., 2017). Such supervision has been used to improve grounding in VLMs (Selvaraju et al., 2019). In RL, the same idea can be used to check more than the outcome. A high-reward response should also be supported by the visual evidence that justifies it. Recent multimodal RL and visual-thinking methods use stronger rollouts, perturbation objectives, visual-dependency signals, or textual traces augmented with grounding tags and spatial coordinates (Liu et al., 2025a; Huang et al., 2025; Yuan et al., 2025; Yang et al., 2025b; Zhu et al., 2026). EASE instead uses evidence boxes only to supervise internal attention during training, leaving the VLM's inference format unchanged.

We propose **EASE** (**E**vidence-**A**nchored **S**patial Att**E**ntion), a framework that adds visual evidence-anchored process supervision to multimodal RLVR. Figure 1 shows the intended contrast. Outcome-only RL may answer correctly while attending weakly to the image region that justifies the answer, whereas EASE encourages the model to keep the answer correct and place more attention on the supporting visual evidence.

EASE maps annotated evidence regions to a smoothed target distribution over visual tokens. During RL training, this target guides where generated response tokens attend within the image. The guidance is applied only to trajectories that receive high verifier rewards, since low-reward responses can have unreliable attention patterns. The evidence annotations are used only as privileged training labels, and inference uses the same image-question pathway as the underlying VLM without evidence metadata. The training data includes both examples with one supporting region and examples with multiple supporting regions, so the model learns to focus on a local cue when one region is sufficient and to cover several regions when the answer depends on multiple clues.

Empirically, EASE consistently improves outcome-only RL baselines across Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B. When trained and evaluated under the same protocol as DAPO, EASE improves the average score by 2.9, 3.1, and 2.5 points on these backbones, respectively, with gains on perception-heavy reasoning, hallucination robustness, visual math, and logic benchmarks. EASE is also competitive with recent 7B-scale multimodal RL methods on shared benchmarks. Diagnostics and ablations link these gains to stronger evidence-aligned visual attention, reward-aware gating, the smoothed target, and mixed single-/multi-evidence training setup.

Our contributions are summarized as follows.

- We identify visual evidence acquisition as a missing process signal in multimodal RLVR and introduce EASE, which adds visual evidence-anchored process supervision without requiring evidence metadata at inference.
- We map annotated evidence regions to a smoothed target distribution over visual tokens and apply this guidance only to high-reward trajectories.
- We validate EASE across multiple VLM backbones and benchmarks for perception-heavy reasoning, hallucination robustness, visual math, and multimodal reasoning, with diagnostics showing better alignment between visual attention and annotated evidence regions.

2 Related Work

RLVR for multimodal reasoning. RLVR has become a common VLM post-training paradigm, improving multimodal reasoning through task verifiers, rollout refinement, reflection training, chart-specific verifiable rewards, and policy optimization (Yang et al., 2025a; Wang et al., 2025c,a; Meng et al., 2025; Sinha et al., 2025; Zhang et al., 2026; Liu et al., 2025b; Tan et al., 2026). However, the main training signal remains answer-level. It does not specify how the response should acquire supporting image regions.

Visual perception signals in multimodal RL. Several multimodal RL methods introduce perception-oriented signals, such as clean and noisy image rollouts (Liu et al., 2025a), perception-aware

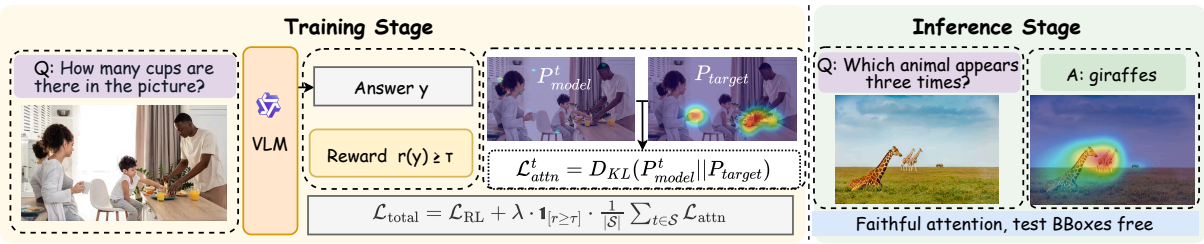


Figure 2: **Overview of EASE.** During RL training, EASE maps dataset evidence boxes to a smoothed spatial target over visual tokens. The auxiliary objective guides response-to-vision attention on high-reward trajectories toward this target, and inference uses the standard image and question inputs.

objectives (Wang et al., 2025d), visual dependency or focus measures (Huang et al., 2025; Wang et al., 2026), perceptual evidence checklists (Zhang et al., 2025), and grounded traces or pointing supervision (Cao et al., 2025; Ni et al., 2026). These signals encourage visual reliance through perturbations, reward-derived proxies, checklist verification, textual traces, or point-level targets, but they do not directly match response-to-vision attention to localized answer evidence. EASE instead supervises this spatial attention distribution with localized evidence regions.

Evidence grounding and hallucination. Grounded VLMs connect language generation with localized visual evidence through boxes, coordinates, or grounded rationales (Peng et al., 2024; Chen et al., 2023; You et al., 2024; Xia et al., 2025). Hallucination studies show that strong VLMs can produce plausible answers or reasoning traces without reliably using the image (Asadi et al., 2026; Shi et al., 2026; Wu et al., 2025; Xu et al., 2026; Fu et al., 2025). Attention and representation analyses further show that visual information can be concentrated in middle layers, steered to reduce hallucination, or misallocated during generation (Jiang et al., 2025; Yin et al., 2025; Huang et al., 2024; Leng et al., 2024; Bu et al., 2025; Xi et al., 2026). EASE brings this grounding perspective into RL by using answer-relevant regions as training supervision for spatial attention.

3 Method: EASE

Figure 2 gives an overview of EASE, which augments outcome-based multimodal RL with training-time evidence supervision. EASE converts evidence boxes into a soft visual-token target, regularizes response-to-vision attention toward this target for high-reward trajectories during actor updates.

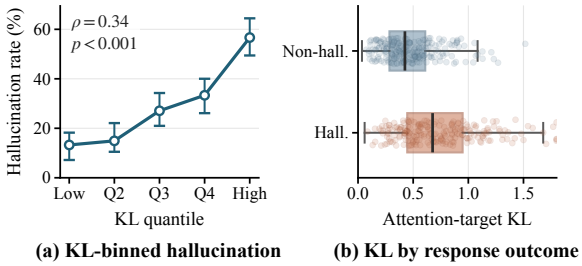


Figure 3: **Motivating diagnostic for outcome-only RL.** Higher attention-target mismatch corresponds to increased hallucination risk in (a) and a right-shifted KL distribution for hallucinated responses in (b). Details are in Appendix A.

3.1 Motivating Observation

We first examine whether RL with only outcome rewards still leaves failures in visual evidence acquisition. We analyze evidence-localizable examples from HallusionBench-Image with a Qwen3-VL-4B policy trained by GRPO using final-answer rewards. For each response, we compute the KL divergence between response-to-vision attention normalized over visual tokens and an evidence target constructed from annotated supporting regions. This score measures attention-target mismatch. Evidence boxes are used only to construct the diagnostic target, not fed as input to the policy.

Figure 3 shows that this mismatch is statistically associated with hallucination. Figure 3(a) shows that hallucination risk increases as KL divergence grows, and Figure 3(b) shows that hallucinated responses have higher KL overall than responses that do not hallucinate. Together, these trends show that outcome rewards can improve final accuracy while leaving a measurable mismatch between response-to-vision attention and answer-relevant evidence. EASE therefore uses evidence-attention alignment as an operational process signal for visual evidence acquisition during RL.

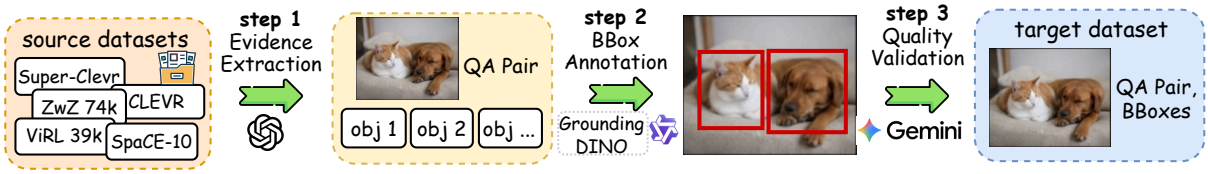


Figure 4: **Evidence annotation pipeline.** Given an image-question-answer triple, Step 1 extracts answer-relevant evidence phrases, Step 2 localizes each phrase with grounding models, and Step 3 validates the proposed boxes with semantic and geometric checks. The annotated pool contains single-evidence examples for local grounding and multi-evidence examples for cross-region reasoning. Appendix B gives implementation details and dataset statistics.

3.2 Evidence Annotation

EASE uses evidence annotations that identify the image regions supporting an answer. Each training example is represented as $(I, q, a, \mathcal{B})$, where I is the original image, q is the question, a is the reference answer, and $\mathcal{B} = \{B_k\}_{k=1}^{K}$ is the bbox field for answer-relevant regions. These boxes are stored as training metadata and used only to construct the auxiliary attention target, leaving the image and prompt unchanged. Only training examples are annotated, and evaluation benchmarks are used without evidence metadata.

Figure 4 summarizes the annotation pipeline. Step 1 extracts minimal answer-relevant evidence phrases from each image-question-answer triple. Step 2 localizes each phrase with grounding models and stores the resulting coordinates as box metadata. Step 3 validates the proposed boxes with semantic and geometric checks. Existing evidence boxes are normalized and passed through the same validation stage. Appendix B gives implementation details and source-level statistics.

The resulting annotated pool is split into single-evidence and multi-evidence examples for localized grounding and cross-region evidence acquisition. The default training distribution samples a balanced mixture from the two pools.

3.3 Evidence-Anchored Attention Target

Let $\mathcal{V}$ denote the set of visual tokens and let c_v be the image-plane center of visual token v, expressed in the same coordinate system as the evidence boxes. EASE converts the evidence boxes into a soft target distribution over $\mathcal{V}$. A box $B_k = (x_k, y_k, w_k, h_k)$ defines a Gaussian component centered at

$$\mu_k = \left(x_k + \frac{w_k}{2}, y_k + \frac{h_k}{2}\right), \tag{1}$$

with diagonal covariance

$$\Sigma_k = \mathrm{diag}\left(\left(\frac{w_k}{4}\right)^2, \left(\frac{h_k}{4}\right)^2\right). \tag{2}$$

This parameterization makes two standard deviations match half the box width and height, which concentrates most probability mass inside the annotated region while retaining tolerance to boundary noise and coarse boxes.

Rather than treating a box as a hard binary mask, we evaluate each Gaussian on the visual-token grid and normalize it into a distribution $G_k(v)$:

$$G_k(v) = \frac{\mathcal{N}(c_v; \mu_k, \Sigma_k)}{\sum_{v' \in \mathcal{V}} \mathcal{N}(c_{v'}; \mu_k, \Sigma_k)}. \tag{3}$$

Multi-evidence examples are represented by an equal-weight mixture:

$$P_{\text{evid}}(v) = \frac{1}{K} \sum_{k=1}^{K} G_k(v). \tag{4}$$

Equal weighting assigns the same total supervision mass to each supporting region, encouraging coverage across all evidence boxes rather than allowing large boxes to dominate the target.

We smooth the target with a uniform distribution to avoid assigning zero probability to all background tokens:

$$P_{\text{target}}(v) = (1 - \alpha) P_{\text{evid}}(v) + \alpha U(v), \tag{5}$$

where $U(v) = 1/|\mathcal{V}|$ and α controls background smoothing.

3.4 Response-to-Vision Attention

To connect evidence supervision with generation, EASE measures how response tokens attend to visual tokens during the actor update. We normalize attention logits only over the visual-token set, yielding a conditional spatial distribution over image regions. This captures where the response looks

within the image without constraining its total attention mass to vision versus text. Let L denote the number of transformer layers in the language model. We extract response-to-vision attention from a semantic middle layer, setting $\ell^\star = \lfloor 2L/3 \rfloor$. This targets a stage where visual information has been semantically refined but is not yet dominated by output-adjacent lexical prediction (Jiang et al., 2025). For response token t, attention head h, and head dimension d at layer $\ell^\star$,

$$p_h^t(v) = \frac{\exp((\mathbf{q}_{h,t}^{\ell^\star})^\top \mathbf{k}_{h,v}^{\ell^\star}/\sqrt{d})}{\sum_{v' \in \mathcal{V}} \exp((\mathbf{q}_{h,t}^{\ell^\star})^\top \mathbf{k}_{h,v'}^{\ell^\star}/\sqrt{d})}. \quad (6)$$

Here, p_h^t is the head-wise visual-token-normalized response-to-vision distribution. Averaging over the H heads gives the model's conditional spatial attention for token t:

$$P_{\text{model}}^t(v) = \frac{1}{H} \sum_{h=1}^{H} p_h^t(v). \quad (7)$$

When the token belongs to trajectory g, we write the same distribution as $P_{\text{model}}^{g,t}$.

To reduce memory overhead, we compute the attention loss on at most 64 sampled valid response tokens per sequence by default, rather than on the full generation.

3.5 Reward-Aware RL Objective

Let π_θ denote the VLM policy. For each prompt $x = (I, q)$, outcome-only multimodal RL samples a group of responses $\{y^{(g)}\}_{g=1}^G$ and obtains raw verifier rewards $r^{(g)} = R(y^{(g)}, a)$. The underlying GRPO or DAPO actor loss computes group-normalized advantages from these per-response rewards. We denote this base actor objective by $\mathcal{L}_{\text{RL}}$ and add EASE as an auxiliary process-supervision term. In the main experiments, we instantiate $\mathcal{L}_{\text{RL}}$ with DAPO because it provides the strongest outcome-only baseline.

For a response token t from trajectory g, the default attention loss uses model-to-target KL.

$$\mathcal{L}_{\text{attn}}^{g,t} = D_{\text{KL}}\left(P_{\text{model}}^{g,t} \,\|\, P_{\text{target}}\right). \quad (8)$$

This direction penalizes the visual-token-normalized attention probability assigned outside the evidence-anchored target. We also evaluate the reverse direction, $D_{\text{KL}}(P_{\text{target}} \| P_{\text{model}}^{g,t})$, as a coverage-oriented ablation.

EASE applies attention supervision only to trajectories whose raw verifier reward meets a threshold τ. Let $m^{(g)} = \mathbb{1}[r^{(g)} \geq \tau]$ denote this reward gate. The total objective is

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{RL}} + \lambda_{\text{attn}} \mathcal{L}_{\text{EASE}}, \\ \mathcal{L}_{\text{EASE}} &= \frac{\sum_{g=1}^{G} m^{(g)} \frac{1}{|\mathcal{S}^{(g)}|} \sum_{t \in \mathcal{S}^{(g)}} \mathcal{L}_{\text{attn}}^{g,t}}{\sum_{g=1}^{G} m^{(g)} + \epsilon}. \end{aligned} \quad (9)$$

where $\mathcal{S}^{(g)}$ is the sampled response-token set for trajectory g, λ_{attn} is the attention-loss weight, and ϵ prevents division by zero when no trajectory in the group passes the gate. The gate uses raw verifier rewards rather than group-normalized advantages because it selects trajectories that can serve as reliable process demonstrations.

3.6 Training and Inference

Evidence annotations are privileged training labels. They are used to construct P_{target} and the auxiliary actor loss during training. After training, EASE uses the same image-question pathway and decoding procedure as the underlying VLM.

4 Experiments

4.1 Experimental Setup

Models and training data. We evaluate Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-4B, and Qwen3-VL-8B (Bai et al., 2025a). All controlled variants use the VQA corpus annotated in Section 3.2, with additional statistics in Appendix B. EASE trains on a balanced mixture of examples grounded in one evidence region and examples requiring multiple evidence regions.

Baselines. For controlled comparisons, we compare EASE with the supervised base model, GRPO, and DAPO under the same training data, backbone, and evaluation protocol. Table 2 further places EASE in context with public 7B-scale multimodal reasoning methods, including ThinkLite-VL-7B (Wang et al., 2025c), VL-Rethinker-7B (Wang et al., 2025a), MM-Eureka-7B (Meng et al., 2025), NoisyRollout-7B (Liu et al., 2025a), PAPO$_D$-7B (Wang et al., 2025d), VPPO-RL-7B (Huang et al., 2025), and VGPO-7B (Wang et al., 2026). These public rows use reported results and are intended for contextual comparison rather than strict protocol-controlled evaluation.

Benchmarks. We evaluate perception with HR-Bench (Wang et al., 2025b), V* (Wu and Xie, 2024), and CV-Bench (Tong et al., 2024),

Table 1: **Generalization across backbones.** Controlled comparison of Base, GRPO, DAPO, and EASE across three VLM backbones under the same training and evaluation protocol. Hall. denotes HallusionBench-Image. Bold and underlined scores mark the best and second-best results in each column.

Method	Perception				Hallucination		Math/Geometry				Logic	Avg.
	HR4K	HR8K	V*	CV-B.	POPE	Hall.	MathVerse_V	MathVista	WeMath	MMK12	LogicVista	
Qwen3-VL-4B												
Base	78.3	72.9	80.1	85.0	88.5	68.2	28.0	73.7	70.7	58.9	45.3	68.1
GRPO	79.1	<u>76.3</u>	86.4	<u>85.5</u>	89.3	71.0	60.7	<u>75.7</u>	<u>77.8</u>	78.9	50.4	75.6
DAPO	<u>79.9</u>	74.1	<u>87.4</u>	85.1	<u>89.5</u>	<u>71.3</u>	<u>62.2</u>	74.3	77.2	<u>80.2</u>	<u>51.8</u>	<u>75.7</u>
EASE (ours)	**81.3**	**79.5**	**90.1**	**87.3**	**90.8**	**74.1**	**65.1**	**77.8**	**82.9**	**81.3**	**56.9**	**78.8**
Qwen2.5-VL-7B												
Base	71.6	67.9	78.5	75.3	85.9	52.9	36.6	68.5	47.8	49.4	45.1	61.8
GRPO	74.3	70.3	83.2	76.9	<u>87.6</u>	55.6	62.6	<u>70.1</u>	<u>68.4</u>	72.9	46.9	69.9
DAPO	<u>74.8</u>	<u>70.4</u>	<u>84.3</u>	<u>77.0</u>	87.3	<u>56.0</u>	<u>64.5</u>	68.7	68.0	<u>77.0</u>	<u>47.3</u>	<u>70.5</u>
EASE (ours)	**75.3**	**73.0**	**88.5**	**78.3**	**87.9**	**57.2**	**67.8**	**75.2**	**72.9**	**81.6**	**49.6**	**73.4**
Qwen3-VL-8B												
Base	78.9	74.6	86.4	85.4	87.0	71.6	54.6	<u>77.2</u>	71.7	67.0	53.0	73.4
GRPO	80.5	<u>80.6</u>	<u>89.0</u>	<u>86.3</u>	88.7	71.9	65.4	75.9	79.7	79.7	58.5	77.8
DAPO	<u>81.1</u>	79.8	88.0	85.5	<u>89.1</u>	<u>73.6</u>	<u>65.7</u>	76.0	<u>82.9</u>	<u>80.9</u>	<u>60.0</u>	<u>78.4</u>
EASE (ours)	**83.8**	**81.6**	**91.6**	**87.1**	**89.6**	**77.8**	**69.7**	**78.1**	**86.4**	**82.2**	**62.3**	**80.9**

hallucination with POPE (Li et al., 2023a) and HallusionBench-Image (Guan et al., 2023), visual math and geometry with MathVerse_V (Zhang et al., 2024), MathVista (Lu et al., 2024), WeMath (Qiao et al., 2024), and MMK12 (Meng et al., 2025), and broader multimodal reasoning with LogicVista (Xiao et al., 2024). MathVerse_V denotes the vision-centric subset of MathVerse.

Implementation details. Unless otherwise noted, models are trained for 2 epochs with learning rate 1×10^{-6}. EASE uses the smoothed Gaussian-mixture target with $\alpha = 0.1$, model-to-target KL, $\lambda_{\text{attn}} = 0.001$, reward-1 gating under the 0/1 verifier, and at most 64 valid response tokens for the attention loss. Additional implementation details are provided in Appendix C.4.

4.2 Main Results

Consistent Gains under Controlled Training. The controlled comparison in Table 1 more directly isolates the effect of the proposed training signal, since Base, GRPO, DAPO, and EASE are evaluated under the same protocol within each backbone. On Qwen2.5-VL-7B, EASE improves DAPO on every reported benchmark and raises the average from 70.5 to 73.4. The largest gains appear on vision-centric mathematical reasoning and fine-grained perception, including MathVista (+6.5), WeMath (+4.9), MMK12 (+4.6), and V* (+4.2). This pattern aligns with the purpose of EASE, where the auxiliary loss serves as a process signal that encourages the policy's response-to-vision attention to concentrate on answer-relevant visual evidence before generating the response.

Scalability across Model Scales and Families. The same trend is visible across Qwen3-VL backbones. For Qwen3-VL-4B, EASE improves the DAPO average from 75.7 to 78.8 and gives gains across the full benchmark suite, including fine-grained perception, hallucination, math/geometry, and LogicVista. For Qwen3-VL-8B, EASE reaches the best average score among the compared variants, improving DAPO from 78.4 to 80.9 and again improving every reported benchmark. These results suggest that evidence-focused attention regularization transfers across model generations and scales when applied under the same training and evaluation protocol.

Contextual Comparison with Public RL Methods. Table 2 situates EASE among recent 7B-scale multimodal reasoning methods on shared benchmarks from prior work. EASE-7B achieves the best average score in this comparison and obtains the strongest results on MathVista, MathVerse_V, WeMath, and LogicVista. It also ranks second on MMK12, showing that evidence-anchored attention supervision remains effective across several forms of vision-centric reasoning. These results indicate that EASE is competitive with leading 7B-scale multimodal RL methods.

Table 2: **Contextual comparison with prior 7B-scale multimodal reasoning methods.** Prior scores are taken from the VGPO report, where available. "*" indicates that NoisyRollout is trained on Geo3K. Bold and underlined scores indicate the best and second-best results in each column.

Method	MathVista	MathVerse$_V$	WeMath	MMK12	Geo3K	LogicVista	Avg.
ThinkLite-VL-7B	73.6	30.9	43.7	50.5	45.3	44.3	48.1
VL-Rethinker-7B	65.2	64.8	67.5	66.0	44.4	47.0	59.2
MM-Eureka-7B	66.3	63.6	66.6	61.7	41.1	47.9	57.9
NoisyRollout-7B	72.9	51.7	64.4	51.2	**55.2***	47.0	57.1
PAPO$_D$-7B	72.3	66.6	69.4	80.8	48.8	45.9	64.0
VPPO-RL-7B	70.5	<u>67.6</u>	70.6	**81.8**	46.9	48.8	64.4
VGPO-7B	<u>74.1</u>	<u>67.6</u>	<u>72.5</u>	81.5	45.8	<u>49.4</u>	<u>65.2</u>
EASE-7B (ours)	**75.2**	**67.8**	**72.9**	<u>81.6</u>	<u>48.9</u>	**49.6**	**66.0**

Table 3: **Core ablations on Qwen3-VL-4B.** We ablate reward gating, evidence-target construction and validity, evidence composition, and the attention extraction layer. L denotes the number of language-model blocks, and the default EASE layer is $\ell^\star = \lfloor 2L/3 \rfloor$.

Variant	V*	Hall.	MMK12	Avg.
EASE (ours)	**90.1**	**74.1**	**81.3**	**81.8**
Reward-aware supervision				
w/o reward gating	86.4	71.0	77.9	78.4
Evidence target construction				
hard box target	87.8	71.6	78.5	79.3
target-to-model KL	89.5	72.8	80.6	81.0
w/o smoothing	88.2	72.0	79.3	79.8
Evidence target validity and robustness				
50% box jitter (IoU ≈ 0.7)	89.3	73.5	80.7	81.2
random targets	81.9	69.3	77.5	76.2
non-overlapping targets	82.2	68.5	78.2	76.3
Evidence composition				
single-only	89.5	72.7	78.9	80.4
multi-only	88.0	73.2	80.7	80.6
Attention extraction granularity				
early layer ($L/4$)	86.8	71.3	77.8	78.6
central middle layer ($L/2$)	88.7	72.8	79.6	80.4
final layer (L)	86.2	70.8	77.2	78.1

4.3 Ablation Studies

Ablation study on reward-aware supervision. We ablate the reward gate that restricts attention supervision to trajectories with positive outcome rewards. Removing this gate applies the attention loss to all sampled responses and causes consistent drops across the representative benchmarks in Table 3. This suggests that evidence boxes alone are insufficient when the generated response is wrong, since low-reward rollouts can provide misleading response-to-evidence alignments. Reward-aware gating therefore helps EASE treat only successful trajectories as reliable process demonstrations.

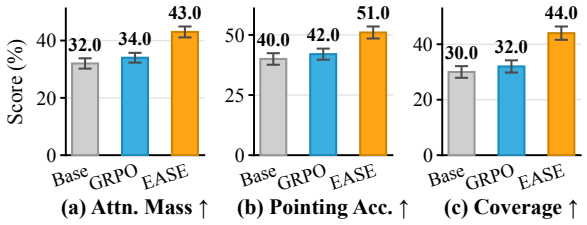


Figure 5: **Evidence-acquisition diagnostics.** On Qwen3-VL-4B, EASE increases response-to-vision attention on annotated evidence regions compared with Base and GRPO, as measured by attention mass, pointing accuracy, and multi-evidence coverage on a held-out validation set. Error bars denote 95% CIs.

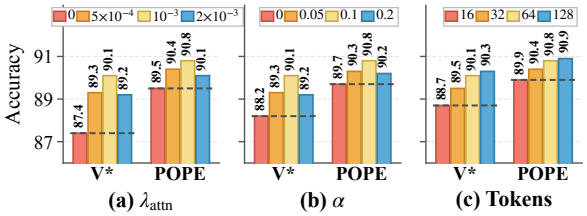
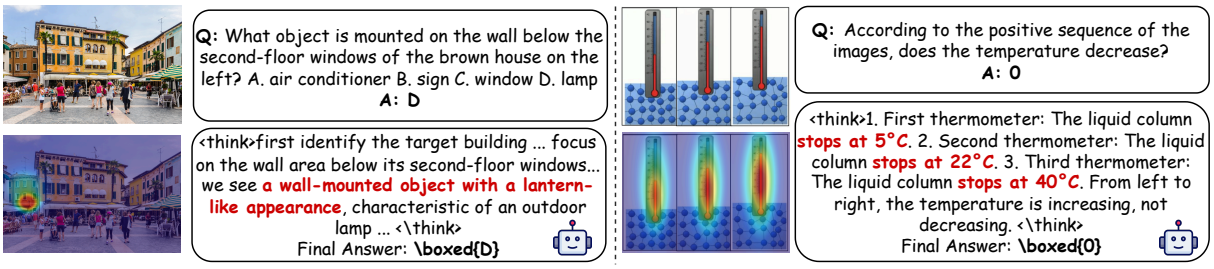


Figure 6: **Hyperparameter sensitivity.** On the Qwen3-VL-4B backbone, we vary the attention-loss weight, background smoothing, and sampled response-token budget around the default EASE configuration. Marked bars indicate default settings.

Ablation study on evidence target construction. We ablate three components of the evidence target, including Gaussian smoothing, background smoothing, and the KL direction. Replacing the Gaussian target with a hard box mask causes the largest drop, suggesting that binary supervision is brittle when evidence spans object parts, coarse boxes, or multiple visual tokens. Removing background smoothing also hurts performance, indicating that a small amount of background mass helps



(a) Single Evidence (b) Multi Evidences

Figure 7: **Qualitative evidence analysis.** Representative examples with EASE reasoning traces, final answers, and response-to-vision attention maps. Attention maps are used only for analysis.

avoid over-constraining contextual visual tokens. Reversing the KL direction performs closer to the default but remains weaker. These results support the smoothed Gaussian target with KL from model attention to the evidence target, which guides attention toward evidence regions while preserving tolerance to annotation noise and useful context.

Ablation study on evidence target validity and robustness. We evaluate target validity and robustness by jittering boxes in 50% of training examples to an IoU of approximately 0.7, replacing targets with randomly placed boxes of the same size, and translating targets to non-overlapping regions. As shown in Table 3, moderate jitter causes only a 0.6-point average decrease and EASE remains 1.6 points above DAPO, whereas random and non-overlapping targets both fall below DAPO. These results show that EASE tolerates moderate localization error but requires approximately correct spatial supervision.

Ablation study on evidence composition. We compare training on one-region examples, multi-region examples, and their balanced mixture. The one-region setting performs well on V*, which emphasizes fine-grained localization, but drops more on MMK12, where many examples require multiple visual cues. Conversely, the multi-region setting better preserves MMK12 but loses more on V*. The balanced mixture gives the best overall result by combining local grounding with cross-region evidence acquisition.

Ablation study on attention extraction granularity. We ablate the layer used to extract response-to-vision attention. Early-layer supervision performs poorly, suggesting that shallow visual representations are not yet sufficiently aligned with answer-level evidence. The central middle layer improves over the early layer but remains below the default setting. The final layer also degrades performance, likely because late attention is more affected by lexical prediction and response formatting than by evidence acquisition. These results support extracting attention from $\ell^\star = \lfloor 2L/3 \rfloor$, where visual tokens have been semantically refined while still retaining useful spatial evidence signals.

4.4 Further Analysis

Evidence-acquisition diagnostics. We measure response-to-vision attention on a held-out evidence-annotated validation set, using evidence boxes only for offline analysis. Figure 5 reports attention mass, pointing accuracy, and multi-evidence coverage, which respectively measure attention inside evidence regions, whether the strongest attention peak lands on evidence, and whether multiple supporting regions are covered. EASE improves all three metrics over both the base model and outcome-only GRPO, suggesting that the auxiliary objective improves attention-based evidence-acquisition diagnostics rather than only final-answer accuracy. Detailed metric definitions are provided in Appendix E.

Functional validation of evidence use. Attention alignment alone does not establish whether the model functionally relies on the aligned visual evidence. On 1,000 held-out examples, we mask annotated evidence or an area- and shape-matched non-evidence region. We also suppress response-token attention to evidence tokens at the supervised layer and redistribute the removed mass over the remaining visual tokens, with a matched control that suppresses equal mass from non-evidence tokens. Table 4 shows larger evidence-specific gaps for EASE than DAPO under both interventions. Evidence masking also produces more correct-to-incorrect flips for EASE, whereas matched masking has little effect. These results support stronger functional reliance on annotated evidence, while

Table 4: **Evidence interventions.** Accuracy on 1,000 held-out evidence-annotated examples after masking annotated evidence or suppressing evidence-token attention. The matched control masks an area- and shape-matched non-evidence region or suppresses equal attention mass from non-evidence tokens. Question-only applies to masking. Gap is matched-control accuracy minus evidence-intervention accuracy.

Method	Original	Evidence intervention	Matched control	Question only	Gap
Masking					
DAPO	79.0	19.6	73.4	12.8	53.8
EASE (ours)	82.3	10.9	76.1	12.5	65.2
Attention suppression					
DAPO	79.0	74.8	77.6	–	2.8
EASE (ours)	82.3	74.4	80.5	–	6.1

not treating the supervised attention layer as a complete causal account of visual information use.

Hyperparameter sensitivity. We further test whether EASE relies on narrow hyperparameter choices. Figure 6 varies the attention-loss weight λ_{attn}, background smoothing α, and the sampled token budget around the default configuration. Setting $\lambda_{\text{attn}} = 0$ recovers outcome-only DAPO, while moderate attention supervision improves both V* and POPE. Very large weights reduce accuracy, indicating that the auxiliary loss should guide rather than dominate policy optimization. A small amount of background smoothing is also beneficial, with $\alpha = 0.1$ giving the best or near-best scores. Finally, sampling 64 response tokens matches or nearly matches the 128-token setting, supporting the memory-efficient default.

Qualitative evidence analysis. Figure 7 provides qualitative examples of how EASE changes response-to-vision attention during generation. In these cases, the model produces correct reasoning traces and final answers while its attention concentrates on regions that visually support the answer. These examples illustrate the intended effect of the auxiliary objective, which encourages high-reward trajectories to align response tokens with answer-relevant evidence regions. Appendix H further analyzes representative failure cases to clarify the boundary of this effect.

The quantitative reasoning-quality evaluation is reported in Appendix F.

5 Conclusion

We presented EASE, a framework that adds evidence-anchored process supervision to multimodal RLVR. EASE converts annotated evidence regions into soft spatial targets and regularizes response-to-vision attention for high-reward trajectories, encouraging successful responses to align with visually grounded evidence. Across three VLM backbones and diverse reasoning and hallucination benchmarks, EASE improves outcome-only RL baselines and strengthens measured evidence acquisition. These results suggest that supervising visual evidence acquisition is a practical complement to final-answer rewards for more grounded multimodal reasoning.

Limitations

EASE relies on evidence annotations during training. Although these annotations are used only for the auxiliary loss and are not provided at inference, obtaining reliable boxes introduces additional data construction cost. The method is also most suitable for tasks whose supporting evidence can be localized to one or more image regions, and box-based supervision may be less suitable for holistic, stylistic, affective, or globally distributed cues. Finally, EASE mainly supervises where response-to-vision attention is allocated within the image, rather than directly optimizing how much the generation depends on visual input. We therefore interpret the attention diagnostics as evidence of improved spatial grounding, not as a complete causal measure of visual reliance.

Acknowledgments

This work is supported by the Zhongguancun Academy, (Grant No.s C20250203).

Ethical Considerations

This work is intended for research on multimodal RLVR and grounded reasoning, not for direct deployment in safety-critical settings. Additional details on artifact use, data characteristics, model-assisted annotation, computation, and AI-assistant use are provided in Appendix G.

References

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. 2026. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Weijue Bu, Guan Yuan, and Guixian Zhang. 2025. Conscious gaze: Adaptive attention mechanisms for hallucination mitigation in vision-language models. *arXiv preprint arXiv:2512.05546*.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. 2025. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Zhaokai Wang, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. 2025. Space-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. *Preprint*, arXiv:2506.07966.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

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

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023b. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. 2026. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *Advances in Neural Information Processing Systems*, 38:20538–20559.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600.

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

Chufan Shi, Cheng Yang, Yaokang Wu, Linhao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. 2026. Are vlms seeing or just saying? uncovering the illusion of visual re-examination. *Preprint*, arXiv:2605.15864.

Sanchit Sinha, Oana Frunza, Kashif Rasul, Yuriy Nevmyvaka, and Aidong Zhang. 2025. Chart-rvr: Reinforcement learning with verifiable rewards for explainable chart reasoning. *arXiv preprint arXiv:2510.10973*.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2026. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Advances in neural information processing systems*, 38:5772–5822.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025c. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. 2026. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*.

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

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang,

Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. 2026. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*.

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

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. 2025. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*.

Gongli Xi, Ye Tian, Mengyu Yang, Huahui Yi, Liang Lin, Xiaoshuai Hao, Kun Wang, and Wendong Wang. 2026. Large vision-language models get lost in attention. *arXiv preprint arXiv:2605.05668*.

Jiaer Xia, Bingkui Tong, Yuhang Zang, Rui Shao, and Kaiyang Zhou. 2025. Bootstrapping grounded chain-of-thought in multimodal llms for data-efficient model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 208–217.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *Preprint*, arXiv:2407.04973.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. 2026. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025a. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2376–2385.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025b. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pages 57153–57180.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

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

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

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Xin Zhang, Xingyu Li, Rongguang Wang, Ruizhong Miao, Zheng Wang, Dan Roth, and Chenyang Li. 2026. Chart-rl: Generalized chart comprehension via reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.06958*.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. 2026. Analyzing reasoning consistency in large multimodal models under cross-modal conflicts. *arXiv preprint arXiv:2601.04073*.

A Motivating Diagnostic Details

The diagnostic in Figure 3 tests whether hallucinations after RL with only outcome rewards are associated with evidence attention mismatch. We use a Qwen3-VL-4B policy trained with GRPO and final-answer rewards on 903 HallusionBench-Image examples that remain after applying the evidence annotation and filtering pipeline in Appendix B. For each generated response, we compute the KL divergence between the response-to-vision attention distribution normalized over visual tokens and the Gaussian mixture evidence target defined in Section 3.3. Responses judged incorrect by the official HallusionBench protocol are treated as hallucinated for this analysis. Figure 3(a) groups examples into equal-sized KL quantile bins and reports the hallucination rate in each bin, showing whether risk increases as mismatch grows. Spearman correlation is computed over individual examples. Figure 3(b) compares the KL distributions of hallucinated responses and responses that do not hallucinate, showing whether hallucinated samples have

higher mismatch overall. Evidence boxes are used only for this offline diagnostic and are not provided to the model input.

B Data Construction Details and Statistics

This appendix documents the evidence extraction prompts, box-metadata construction procedure, filtering rules, and dataset-specific preprocessing. Figure 4 shows the high-level pipeline, and Table 5 reports the source-level statistics of the annotated pools before final training sampling. The important implementation convention is that evidence boxes are never rendered into the image. Each annotated example stores the original image, question, answer, box coordinates, and a single-/multi-evidence tag. EASE consumes the boxes only when constructing the target attention distribution.

B.1 Step 1. Evidence Phrase Extraction

For each source example (I, q, a), GPT-4.1-mini is prompted to enumerate the minimal set of visible entities that must be inspected to verify the answer. The output provides temporary text queries for localization rather than final supervision labels.

Evidence phrase extraction prompt

Given the question q and reference answer a, identify the smallest set of visible and localizable objects or regions that a human would inspect to verify the answer. Include only evidence that directly supports the answer. Return a JSON list of concise noun phrases. Do not include explanations, full sentences, the whole image, generic background, scene-level descriptions, or non-localizable concepts. Return `[]` if no localizable visual evidence is required.

The extractor returns a structured list of concise noun phrases. It is asked for the smallest set of objects needed to verify the answer, rather than all objects mentioned in the question. We remove phrases referring to the whole scene, the background, the image itself, or non-localizable abstractions, and merge near duplicates by string overlap and head-noun matching.

B.2 Step 2. Box Localization

Each intermediate evidence phrase is localized independently, and the resulting coordinates are kept as supervision metadata. We use Grounding DINO (Liu et al., 2024) as the primary phrase-grounding model and keep its top prediction when the confidence exceeds $\delta_{\text{det}} = 0.35$. Candidates below this confidence threshold are discarded.

As a complementary localization signal, we also query a locally deployed Qwen3.5-27B (Qwen Team, 2026) with the original image and the same intermediate phrase to obtain normalized box coordinates. The two localization outputs are merged query by query. If their IoU is at least 0.7, their union is used as a conservative evidence box B_k. If they disagree, the Grounding DINO box is used as the primary localization, and the Qwen box is retained only if it passes Step 3 validation for the same evidence entity. Queries for which neither model returns a valid box are dropped.

B.3 Step 3. Quality Validation

The annotation pipeline validates each proposed box independently. Gemini 2.5 Flash-Lite (Comanici et al., 2025) receives the original image, question, reference answer, intermediate evidence phrase, and proposed box coordinates, and judges whether the boxed region contains answer-relevant visual content. We remove boxes rejected by the verifier. Using a verifier from a different model family reduces the chance that a single model's extraction bias determines both the intermediate phrase queries and the final boxes.

Independent human audit. Three master's students who were not involved in this work independently audit 1,000 sampled instances containing 1,586 evidence boxes. They assess box validity, namely whether a proposed box contains answer-relevant visual evidence, required-evidence coverage, namely whether the proposed set omits evidence needed to verify the answer, and sample-level evidence sufficiency, namely whether the complete box set suffices to verify the answer. The audit yields 95.3% box precision, 98.6% required-evidence recall, and 97.8% sample-level evidence sufficiency, indicating that the automatically constructed evidence boxes provide reliable training-only supervision.

Annotation cost. The three-stage pipeline produces approximately 127k evidence-annotated samples. Evidence phrase extraction with GPT-4.1-mini costs approximately \$18, box localization with Grounding DINO and a locally deployed Qwen3.5-27B incurs no external API cost, and

box validation with Gemini 2.5 Flash-Lite costs approximately $30. The total external API cost is therefore approximately $48.

Quality validation prompt

Given the original image, question q, reference answer a, evidence query e, and proposed box coordinates $b = [x_1, y_1, x_2, y_2]$, decide whether the region specified by b contains visible evidence matching e and directly helps verify a. Use the full image for spatial context, but base the final decision on the content inside the proposed box. Return VALID if the box encloses answer-relevant visual evidence. Return INVALID if the box is empty, localizes the wrong object or region, is too vague, or does not help verify the answer. Return only VALID or INVALID.

Boxes are clipped to image boundaries, converted to the coordinate convention used by the visual-token grid, and stored as $\mathcal{B}$. The final output is a metadata-augmented training tuple $(I, q, a, \mathcal{B})$ in which I remains the original unmarked image. During RL rollouts, the policy sees only (I, q), and during the actor update, EASE uses $\mathcal{B}$ to construct the Gaussian-mixture attention target.

B.4 Training Pool Construction

After validation, examples with one validated evidence box enter the single-evidence pool, and examples with at least two validated evidence boxes enter the multi-evidence pool. The single-evidence pool provides local grounding supervision, while the multi-evidence pool provides cross-region supervision. Table 5 summarizes the available annotated examples in each pool. The main EASE training distribution is then sampled from these pools in a balanced 1:1 ratio.

Table 5 reports 20,257 multi-evidence examples from CLEVR (Johnson et al., 2017), SuperCLEVR (Li et al., 2023b), SpaCE10 (Gong et al., 2025), and ViRL39K (Wang et al., 2025a). It also includes 107,714 single-evidence examples from ZwZ74K (Wei et al., 2026) and the same source corpora. The full annotated pool contains 127,971 examples before final training sampling.

C Experimental Settings

C.1 Training Datasets

All controlled runs draw training data from the evidence-annotated VQA pools described in Appendix B. The pools are built from ZwZ74K, CLEVR, SuperCLEVR, SpaCE10, and ViRL39K. Each training record keeps the original image, question, reference answer, validated box coordinates, and a tag indicating whether one or multiple evidence boxes are available. The intermediate evidence phrases used for localization are not stored as training fields. Unless otherwise noted, training uses a balanced mixture with equal numbers sampled from the single-evidence and multi-evidence pools. All controlled RL variants use the same sampled image-question-answer data and final-answer verifier. EASE additionally uses the stored box metadata during actor updates to construct its auxiliary attention target.

C.2 Baselines

- **ThinkLite-VL** improves data-efficient visual reasoning through MCTS-guided sample selection (Wang et al., 2025c).
- **VL-Rethinker** trains VLMs to reconsider their answers through RL-driven self-reflection (Wang et al., 2025a).
- **MM-Eureka** studies rule-based RL for multimodal reasoning and releases strong 7B-scale models (Meng et al., 2025).
- **NoisyRollout** reinforces visual reasoning with augmented rollouts that perturb the visual input (Liu et al., 2025a).
- **PAPO$_D$** introduces perception-aware policy optimization to improve reasoning under visual uncertainty (Wang et al., 2025d).
- **VPPO-RL** estimates token-level visual dependence from perturbed images and uses it to modulate policy optimization (Huang et al., 2025).
- **VGPO** measures visual focus from internal model states and uses visually guided advantages (Wang et al., 2026).

C.3 Evaluation Benchmarks

- **HR-Bench** evaluates high-resolution image perception and fine-grained visual recognition (Wang et al., 2025b).
- **V*** focuses on guided visual search and localized object understanding (Wu and Xie, 2024).
- **CV-Bench** evaluates vision-centric perception skills in two-dimensional and three-

Table 5: **Evidence-annotated data pools.** The table reports source-level annotated examples available after filtering. Counts are measured before final training sampling.

Source corpus	Multi-evidence	Single-evidence	Total available	Primary supervision role
ZwZ74K	0	74,000	74,000	Fine-grained local region grounding
CLEVR	4,528	1,531	6,059	Controlled compositional reasoning
SuperCLEVR	3,992	926	4,918	Harder synthetic multi-object reasoning
SpaCE10	1,759	2,371	4,130	Spatial relations and cross-region evidence
ViRL39K	9,978	28,886	38,864	General visual reasoning with localized evidence
Total	**20,257**	**107,714**	**127,971**	**Evidence-annotated pool**

dimensional visual contexts (Tong et al., 2024).

- **POPE** probes object hallucination through binary visual questions with controlled object presence (Li et al., 2023a).
- **HallusionBench-Image** diagnoses hallucination and visual illusion in image-grounded question answering (Guan et al., 2023).
- **MathVista** evaluates mathematical reasoning in visual contexts across diverse problem types (Lu et al., 2024).
- **MathVerse**$_V$ denotes the vision-centric subset of MathVerse, where solving the problem requires information from the image (Zhang et al., 2024).
- **WeMath** provides a diagnostic evaluation of multimodal mathematical reasoning (Qiao et al., 2024).
- **MMK12** covers K–12 multimodal reasoning problems and is used to assess general visual problem solving (Meng et al., 2025).
- **LogicVista** evaluates logical reasoning in visual contexts, complementing the perception and mathematical reasoning suites with diagram-grounded reasoning tasks (Xiao et al., 2024).

C.4 Implementation Details

Training and evaluation setup. Unless otherwise noted, models are trained for 2 epochs with learning rate 1×10^{-6}, rollout batch size 512, and maximum response length 1,024. Evaluation uses greedy decoding with temperature 0.0 for all experiments.

EASE configuration. Unless otherwise noted, EASE uses the middle-layer attention index $\ell^\star = \lfloor 2L/3 \rfloor$, samples at most 64 valid response tokens for the auxiliary loss, and applies model-to-target KL only to reward-positive trajectories. We set $\lambda_{\text{attn}} = 0.001$ and background smoothing $\alpha = 0.1$.

Algorithm 1: EASE training procedure.

Input: Evidence data $\mathcal{D}$, policy π_θ, verifier R, layer $\ell^\star$, reward threshold τ, attention weight λ_{attn}
Output: Fine-tuned policy π_θ
for each training step **do**
  Sample evidence tuples $(I, q, a, \mathcal{B})$ from $\mathcal{D}$;
  Roll out grouped responses $y^{(g)} \sim \pi_\theta(\cdot \mid I, q)$ and compute rewards $r^{(g)} = R(y^{(g)}, a)$;
  Map $\mathcal{B}$ to the visual-token grid and build the smoothed Gaussian-mixture target P_{target} by Eq. 5;
  Compute the base RL actor loss from the grouped rewards;
  foreach trajectory with $r^{(g)} \geq \tau$ **do**
    Sample valid response tokens $\mathcal{S}^{(g)}$ and extract layer-$\ell^\star$ response-to-vision attention;
    Average attention heads to obtain $P^{g,t}_{\text{model}}$ and compute the KL loss in Eq. 8;
  end
  Add the reward-gated attention loss in Eq. 9 to the actor loss;
  Update π_θ;
end

Algorithm 1 summarizes the training procedure used by EASE.

D Reward-Gate Coverage and Prompt Difficulty

We estimate initial solvability using eight rollouts from the frozen Qwen3-VL-4B policy and group prompts by their number of successful rollouts. During EASE training, 47.4% of sampled rollouts are reward-positive and 92.3% of training groups contain at least one gate-passing rollout. Among prompts initially gated off, 48.6% later produce at least one positive-reward rollout. Table 6 shows that, although group-level gate coverage is lower for harder prompts, EASE improves over DAPO in

Table 6: **Reward-gate coverage and EASE gains by initial prompt difficulty.** Initial solvability is measured with eight frozen-policy rollouts.

Initial solvability	Prompt count	Group gate coverage	Δ(EASE – DAPO)
0/8 (Hard)	1,200	78.5%	3.4
1–3/8	2,600	88.7%	3.8
4–6/8	3,400	95.1%	2.7
7–8/8 (Easy)	2,800	98.6%	1.6

every difficulty group.

E Evidence-Attention Alignment Metrics

This appendix defines the diagnostics reported in Figure 5. They measure whether response-to-vision attention is spatially aligned with answer-relevant evidence regions. Evidence boxes are used only for offline analysis and are never provided to the model input.

Response-to-vision attention. For each generated response, we extract response-to-vision attention from the same layer used by EASE, $\ell^\star = \lfloor 2L/3 \rfloor$, where L denotes the number of transformer layers in the language model. For a response token t, attention head h, and visual token $v \in \mathcal{V}$, we compute

$$p_h^t(v) = \frac{\exp((q_{h,t}^{\ell^\star})^\top k_{h,v}^{\ell^\star}/\sqrt{d})}{\sum_{v' \in \mathcal{V}} \exp((q_{h,t}^{\ell^\star})^\top k_{h,v'}^{\ell^\star}/\sqrt{d})}. \quad (10)$$

We average over heads and valid response tokens to obtain a single visual-token distribution:

$$P_{\text{model}}(v) = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \frac{1}{H} \sum_{h=1}^{H} p_h^t(v), \quad (11)$$

where $\mathcal{S}$ is the set of valid response tokens used for evaluation. P_{model} is normalized over the visual-token set $\mathcal{V}$.

Mapping evidence boxes to visual tokens. Let the input image have width W_{img} and height H_{img}, and let the visual-token grid have width W_v and height H_v. Each visual token $v_{i,j}$ is assigned an image-plane center

$$c_{i,j} = \left(\frac{j+0.5}{W_v} W_{\text{img}}, \frac{i+0.5}{H_v} H_{\text{img}} \right). \quad (12)$$

For an evidence box $B_k = (x_{1,k}, y_{1,k}, x_{2,k}, y_{2,k})$, we define a binary token mask

$$M_k(v_{i,j}) = \mathbb{1}[x_{1,k} \le c_{i,j}^x \le x_{2,k} \wedge y_{1,k} \le c_{i,j}^y \le y_{2,k}]. \quad (13)$$

The union evidence mask is

$$M_{\text{evid}}(v) = \max_{k=1}^{K} M_k(v). \quad (14)$$

Evidence Attention Mass. Evidence Attention Mass measures the total response-to-vision attention assigned to any annotated evidence region:

$$\text{EAM} = \sum_{v \in \mathcal{V}} P_{\text{model}}(v) M_{\text{evid}}(v). \quad (15)$$

Higher values indicate that a larger fraction of visual attention falls inside answer-relevant evidence boxes.

Pointing Accuracy. Pointing Accuracy measures whether the strongest attention peak falls inside an evidence region. We first identify the most-attended visual token:

$$v^\star = \arg\max_{v \in \mathcal{V}} P_{\text{model}}(v). \quad (16)$$

A sample is counted as correct for pointing if $M_{\text{evid}}(v^\star) = 1$. The reported score is the average over evaluation examples:

$$\text{PointingAcc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[M_{\text{evid}}^{(i)}(v_i^\star) = 1]. \quad (17)$$

Multi-evidence Coverage. Multi-evidence Coverage is computed only on examples with $K > 1$ evidence boxes. For each box B_k, we compute its attention mass

$$m_k = \sum_{v \in \mathcal{V}} P_{\text{model}}(v) M_k(v). \quad (18)$$

Because larger boxes naturally cover more visual tokens, we compare m_k with a size-normalized uniform-attention baseline

$$u_k = \frac{\sum_{v \in \mathcal{V}} M_k(v)}{|\mathcal{V}|}. \quad (19)$$

A box is counted as covered if

$$m_k > \beta u_k, \quad (20)$$

where we use $\beta = 2$ by default. The coverage score for one multi-evidence example is

$$\text{Coverage} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}[m_k > \beta u_k]. \quad (21)$$

The reported Multi-evidence Coverage is the average of this quantity over all multi-evidence examples.

Table 7: **Blind pairwise evaluation of reasoning quality.** GPT-5.4-mini compares anonymized EASE and DAPO responses on the answer-correct subset.

Comparison	EASE win	DAPO win	Tie
EASE vs DAPO	83.3%	5.2%	11.5%

Interpretation. These metrics operationalize attention-based diagnostics of the evidence-acquisition process targeted by EASE. Evidence Attention Mass measures total evidence alignment, Pointing Accuracy measures whether the strongest attention peak lands on evidence, and Multi-evidence Coverage measures whether attention is distributed across multiple supporting regions rather than concentrated on a single salient region. These diagnostics do not establish causal necessity of the attended regions. Instead, they provide quantitative evidence that EASE increases the spatial alignment between response-to-vision attention and annotated visual evidence.

F Quantitative Reasoning Quality Evaluation

We conduct a blind pairwise comparison of EASE and DAPO on 1,000 held-out evidence-annotated instances, retaining only examples for which both final answers are correct under the task verifier. GPT-5.4-mini receives the original image, question, reference answer, annotated evidence, and anonymized responses in randomized order, and assesses evidence-groundedness, visual-claim correctness, hallucination avoidance, and reasoning validity. EASE wins substantially more comparisons than DAPO, providing an output-level assessment of evidence-grounded reasoning quality.

G Responsible Research Details

Artifact use, licenses, and intended use. This work uses existing research artifacts, including VLM checkpoints, VQA training corpora, evaluation benchmarks, grounding models, LLM backends, PyTorch, CUDA, and multimodal RL training code, for research evaluation of evidence-anchored supervision in multimodal RLVR. We cite the original creators of the models, datasets, benchmarks, and baseline systems in the main paper and appendix. We use these artifacts according to the licenses, access conditions, and intended-use requirements specified by their providers. The artifacts produced by our work, including code, prompts, configurations, box metadata, and evaluation scripts, are intended for research on grounded multimodal reasoning, RLVR, and reproducibility analysis, rather than direct safety-critical deployment.

Model-assisted annotation and validation. LLMs and VLMs are used as research tools in the evidence annotation pipeline. As described in Appendix B, GPT-4.1-mini extracts intermediate evidence phrases, a locally deployed Qwen3.5-27B provides complementary box localization, and Gemini 2.5 Flash-Lite validates proposed box coordinates against the original image, question, answer, and phrase. These model outputs are used only to construct and filter training metadata.

Result reporting. We report benchmark accuracy or score, average score across benchmark groups, ablation results, and attention-alignment diagnostics. For controlled comparisons, Base, GRPO, DAPO, and EASE use the same training data, backbone family, verifier, decoding protocol, and evaluation suite within each backbone.

Use of AI assistants. The authors used AI assistants for language polishing and submission-compliance drafting support. All method design, code implementation, experimental design, experimental execution, result analysis, scientific claims, and final writing decisions were conducted, verified, and approved by the authors.

H Failure Case Analysis

To provide deeper insights into the boundaries of our proposed EASE, we present two representative failure cases in Figure 8. EASE improves evidence acquisition by encouraging response-to-vision attention to align with annotated supporting regions, but correct attention to visual evidence does not guarantee correct recognition or reasoning. We therefore focus on cases where the model attends to the relevant evidence yet still produces an incorrect answer.

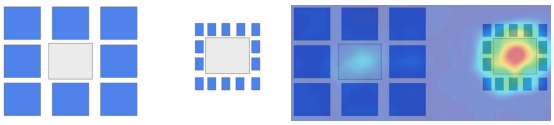
Failure Case 1. Visual perception error. The first example is a contextual size-illusion case from HallusionBench-Image. The question asks whether the right grey box is smaller than the left grey box, while the correct answer is no. The model attends to the relevant grey boxes but is misled by the surrounding blue squares. It treats the smaller contextual squares around the right box as evidence

that the right grey box is also smaller, rather than directly comparing the two grey boxes. This failure suggests that EASE improves where the model looks, but it cannot fully correct a visual representation that is distorted by misleading context.

Failure Case 2. Reasoning after evidence acquisition. The second example is a geometry problem from MathVerse-V. The model correctly identifies the tangent-radius relation and derives $\angle AOB = 54°$. However, it then misapplies a circle-angle relation by setting $\angle C = \angle AOB$, which leads to the incorrect choice. The relevant diagram elements are identified, but the final answer is wrong because the generated solution applies an invalid geometric inference. This failure shows that evidence acquisition is a necessary component of grounded multimodal reasoning, but it is not sufficient when the downstream reasoning policy remains flawed.

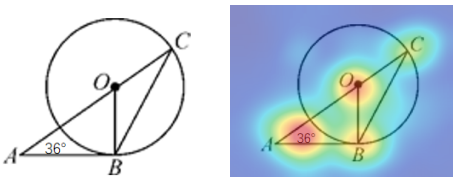
Together, these cases complement the positive examples in Figure 7. They indicate that EASE primarily addresses spatial evidence alignment within response-to-vision attention during standard RL, while residual errors can still arise from contextual perception failures or from reasoning mistakes after the relevant evidence has been attended.

Failure Case 1: Visual Perception Error (HallusionBench-Image, filename: ./VD/illusion/1_0.png)



Question: <image>. Is the right grey box smaller than the left grey box?
Ground Truth: 0
Model Prediction: It seems that yes, the grey box on the right is smaller than the one on the left, as it is shorter in both width and height.
Additionally, the area of the grey box on the right is also noticeably smaller than that of the left grey box. This is further confirmed by the fact that the small blue squares surrounding the right grey box are also much smaller than the large blue squares around the left grey box.
Thus, the answer is: \boxed{1}.
Failure Analysis: The model looks at the relevant grey boxes but is misled by the surrounding blue squares. It infers that the right grey box is smaller from the contextual pattern instead of directly comparing the two grey boxes, showing a visual perception error under size illusion.

Failure Case 2: Reasoning Error (MathVerse-V, id: 20)



Question: <image>. As shown in the figure, AB is tangent to circle O at point B, then angle C is equal to ()
Choices:\nA:36°\nB:54°\nC:60°\nD:27°
Ground Truth: D
Model Prediction: Since AB is tangent to circle O at B, the radius is perpendicular to the tangent. So OB⊥AB, which means ∠OBA = 90°.
In right triangle ABO, ∠A = 36°. I can find the third angle:

∠AOB = 90° - 36° = 54°

Now, I need to find ∠C. Cause angles related to the same arc are equal. Since both ∠AOB and ∠C are "associated with" arc OB, so

∠C = ∠AOB = 54°.

Thus, the correct answer is: \boxed{A}.
Failure Analysis: The model correctly uses the tangent-radius relation to obtain ∠AOB=54°, but then misapplies the circle-angle relation by setting ∠C=∠AOB. The relevant diagram elements are identified, yet the final answer is wrong because of an incorrect geometric inference.

Figure 8: **Representative failure cases.** The left case shows a contextual size illusion from HallusionBench-Image, where the model is misled by surrounding blue squares when comparing the two grey boxes. The right case shows a MathVerse-V geometry example, where the model identifies relevant diagram relations but applies an incorrect circle-angle inference.