VISUAL EVIDENCE PROTOTYPE-GUIDED POLICY OPTIMIZATION FOR VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards has improved visual reasoning in large vision-language models. Existing methods either rely on visual dependence, while strong visual dependence does not imply correct use of visual information, or require external supervision over task-relevant image regions. To address these issues, we introduce **Visual Evidence Prototypes (VEPs)**, which derive fine-grained visual supervision from on-policy rollout outcomes without region annotations. Specifically, for each question, VEPs aggregate the attention distributions of correct and incorrect rollouts over visual tokens into Success and Failure prototypes, and contrast them to capture the key differences. Empirical analyses show that VEPs identify key visual and response tokens for visual reasoning. Intervening on only the top 10% visual tokens selected by VEPs reduces answer accuracy by 20.15–28.45 percentage points. Based on these findings, we propose **VEP-RL**, which uses VEP to reweight trajectory-level updates and redistribute token-level credit to selected tokens. Across fourteen visual reasoning and perception benchmarks, VEP-RL improves average accuracy over the strongest baseline for each backbone by 2.44, 1.93, and 3.24 absolute percentage points on three models, respectively, with its effectiveness consistently validated in out-of-distribution benchmarks.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has significantly improved the reasoning capabilities of large vision-language models (LVLMs) (Shao et al., 2024; DeepSeek-AI et al., 2025; Bai et al., 2025; Wang et al., 2025b; Huang et al., 2025). By rewarding correct answers, RLVR enables LVLMs to learn from on-policy rollouts without annotated reasoning trajectories (Wang et al., 2025a; Liu et al., 2025). However, answer-level rewards alone do not ensure that LVLMs effectively use visual information during reasoning. Prior studies have shown that models tend to answer questions using textual cues or prior knowledge alone, without access to the image (Chen et al., 2024; Zhang et al., 2024).

To address this limitation, recent work has incorporated additional visual supervision into RLVR Xiao et al. (2026a); Miao et al. (2026). One line of work estimates visual dependence through image perturbations or internal representations and incorporates these estimates into policy optimization. For example, PAPO (Wang et al., 2025b) and VPPO (Huang et al., 2025) perturb the image and use the resulting changes in token probabilities to measure how strongly the response depends on visual input, while VGPO (Wang et al., 2026) measures the similarity between response and visual representations. Such visual dependence guides policy optimization through auxiliary objectives, selective token updates, or advantage reweighting. However, as Liu et al. (2026b) observe, strong visual dependence does not indicate that the model uses visual information correctly.

Another growing line of work leverages explicit visual supervision constructed by external tools, human annotators, or stronger models. Saliency-R1 (Gong et al., 2026) aligns model saliency with human-annotated bounding boxes, while Vision-OPD (Yuan et al., 2026) uses an evidence-centered crop as the teacher input. OPD-V (Aniri et al., 2026) further combines external visual supervision with visual dependence, using a zoomed-in image and a mask image for token-level visual self-distillation. Despite their effectiveness, these methods require fine-grained annotations of task-relevant image regions, which we refer to as *visual evidence* in this work.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, we propose automatically constructing Visual Evidence Prototypes (VEPs) by aggregating attention patterns from on-policy rollouts according to answer correctness. For each image–question pair, we partition the rollouts into success and failure groups according to answer correctness. Within the group, we summarize each rollout's attention distribution over visual tokens and aggregate these distributions to form a Success VEP and a Failure VEP. We then contrast the two prototypes to emphasize visual tokens that differ between them. As shown in Figure 1, compared with methods that rely on external visual supervision (e.g., Vision-OPD) and visual dependence estimation (e.g., VPPO), VEPs automatically derive supervision from answer-level rewards to localize task-relevant visual evidence, without requiring region or process annotations.

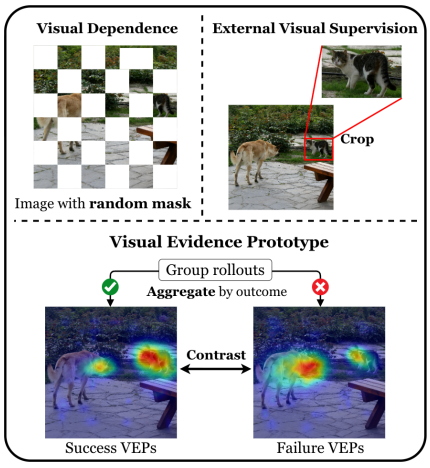


Figure 1: Comparison of visual dependence, external visual supervision, and VEPs derived from rollout outcomes.

VEPs have demonstrated several characteristics for effective visual supervision across three models: (i) VEPs can effectively distinguish independently sampled correct and incorrect rollouts, achieving a mean AUROC of 0.7127; (ii) Intervening only on the visual tokens (top 10%) selected by VEPs reduces answer accuracy by 20.15–28.45 percentage points; (iii) VEPs can be used to localize supporting tokens in correct rollouts and harmful ones in incorrect rollouts.

Based on these characteristics, we propose VEP-RL, which uses VEPs to refine credit assignment in policy optimization at both the trajectory and token levels. At the trajectory level, we use each rollout's relative similarity to the Success and Failure VEPs to determine how strongly to reinforce correct rollouts or penalize incorrect ones. At the token level, we intervene on visual tokens selected by the corresponding VEP and use the resulting probability changes to select key response tokens. We then redistribute the advantage toward these tokens, strengthening positive updates on supporting tokens and negative updates on harmful tokens.

Experiments on fourteen visual reasoning and perception benchmarks demonstrate consistent improvements across three models. Compared with the strongest baseline for each backbone, VEP-RL improves average accuracy by 2.44, 1.93, and 3.24 absolute percentage points on Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-9B, respectively. On three unseen benchmarks, VEP-RL improves average accuracy over the base models by 4.29 and 1.27 percentage points on Qwen3-VL-8B and Qwen3.5-9B, respectively. These improvements demonstrate the effectiveness of VEP-RL.

2 Visual Evidence Prototypes

Problem setup. For an image I and a question q, let $Q = (I, q)$. A vision-language policy π samples a group of rollouts $\mathcal{R}_Q = \{r_g\}_{g=1}^{G}$, where $r_g = (y_{g,t})_{t=1}^{|r_g|}$. Based on answer correctness, the verifier partitions them into correct rollouts $\mathcal{R}_Q^+$ and incorrect rollouts $\mathcal{R}_Q^-$.

Instead of relying on localizing evidence with human annotations, we construct VEPs from correct and incorrect rollouts for the same question. We first derive a visual attribution over image tokens for each rollout (§2.1). We then aggregate these attributions by outcome to construct success and failure VEPs, and contrast them to characterize their differences (§2.2).

2.1 Visual Attribution

To construct VEPs, we first transform each rollout's attention over visual tokens into a ***visual attribution***. Direct attention from the answer to visual tokens does not account for the attention from intermediate response tokens, whereas uniformly averaging their visual attention ignores their different relevance to the answer. We therefore weight each response token by its connection to the final answer when aggregating its attention to the visual tokens.

For rollout r, let $\mathcal{I}_Q$, $\mathcal{T}_r$, and $\mathcal{A}_r$ denote the positions of visual tokens, response tokens preceding the final answer, and tokens in the generated final answer, respectively. Let $A_r(u, v)$ denote the attention to position v when predicting token u, averaged over heads and a selected set of decoder layers $\mathcal{L}$. We compute

$$w_{r,t} = \frac{\sum_{a\in\mathcal{A}_r} A_r(a,t)}{\sum_{u\in\mathcal{T}_r}\sum_{a\in\mathcal{A}_r} A_r(a,u)}, \qquad v_{r,i} = \frac{\sum_{t\in\mathcal{T}_r} w_{r,t} A_r(t,i)}{\sum_{j\in\mathcal{I}_Q}\sum_{t\in\mathcal{T}_r} w_{r,t} A_r(t,j)}. \tag{1}$$

Here, $w_{r,t}$ measures the relative attention assigned to response position t when generating the final answer. We use these weights to aggregate the response tokens' visual attention and normalize it over $\mathcal{I}_Q$, yielding the rollout-level visual attribution $\mathbf{v}_r = (v_{r,i})_{i\in\mathcal{I}_Q}$.

2.2 Visual Evidence Prototypes

For the same question, visual attributions can vary across rollouts with the same outcome. To aggregate the visual patterns for each outcome, we average the attributions separately over correct and incorrect rollouts:

$$\boldsymbol{\mu}_Q^+ = \frac{1}{|\mathcal{R}_Q^+|}\sum_{r\in\mathcal{R}_Q^+} \mathbf{v}_r, \qquad \boldsymbol{\mu}_Q^- = \frac{1}{|\mathcal{R}_Q^-|}\sum_{r\in\mathcal{R}_Q^-} \mathbf{v}_r. \tag{2}$$

We call $\boldsymbol{\mu}_Q^+$ and $\boldsymbol{\mu}_Q^-$ the ***Success VEP*** and ***Failure VEP***, respectively. However, visual tokens with high attribution in correct rollouts may also receive high attribution in incorrect ones. To distinguish the visual evidence emphasized by correct and incorrect rollouts, we construct contrastive VEPs by taking the differences between the two prototypes:

$$\mathbf{D}_Q^+ = \boldsymbol{\mu}_Q^+ - \boldsymbol{\mu}_Q^-, \qquad \mathbf{D}_Q^- = \boldsymbol{\mu}_Q^- - \boldsymbol{\mu}_Q^+. \tag{3}$$

where subtraction is performed element-wise between attribution values at corresponding visual-token positions. We call $\mathbf{D}_Q^+$ and $\mathbf{D}_Q^-$ the ***Success-Specific Contrastive VEP*** and ***Failure-Specific Contrastive VEP***, respectively. These contrastive VEPs highlight where correct and incorrect rollouts differ in their visual attributions.

3 Characterizing Visual Evidence Prototypes

To better understand what VEPs capture, we conduct three pilot experiments that examine whether VEPs distinguish rollout outcomes (§ 3.1), identify important visual tokens (§ 3.2), and localize key response tokens (§ 3.3). Detailed pilot settings are provided in Appendix C.

3.1 VEPs Distinguish Correct and Incorrect Rollouts

To investigate whether VEPs provide an informative signal about a rollout's visual reasoning quality, we use answer correctness as a verifiable proxy and test whether VEPs can distinguish correct from incorrect rollouts. Specifically, we construct VEPs from one rollout set and evaluate them on an independently sampled set for each question. If VEPs capture visual patterns reflecting successful and failed reasoning, correct rollouts should align more closely with the Success VEP relative to the Failure VEP. We thus define the ***VEP Similarity Score*** as:

$$s_r = \cos\left(\mathbf{v}_r, \boldsymbol{\mu}_Q^+\right) - \cos\left(\mathbf{v}_r, \boldsymbol{\mu}_Q^-\right). \tag{4}$$

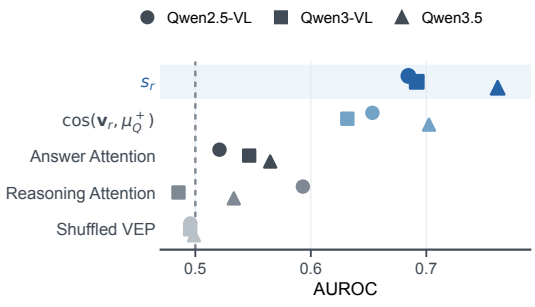


Figure 2: AUROC for distinguishing independently sampled correct and incorrect rollouts.

where cos denotes cosine similarity. Higher scores indicate stronger alignment with visual patterns related to successful reasoning. As shown in Figure 2, VEP Similarity Score achieves a mean AUROC of 0.7127 across the three models, compared with 0.6624 using similarity to the Success VEP alone.

Answer Attention, the total attention to visual tokens averaged over answer tokens achieves 0.5441. Reasoning Attention, the total attention to visual tokens averaged over response tokens, reaches 0.5373. Shuffling the outcome labels used to construct VEPs reduces AUROC to 0.4967. Together, these results suggest that VEPs may signal a rollout's visual reasoning quality.

Finding 1: *Success and Failure VEPs capture visual patterns that distinguish correct and incorrect rollouts across independent samples.*

3.2 Contrastive VEPs Identify Supporting Visual Tokens

Pilot 3.1 shows that VEPs capture visual patterns that distinguish correct and incorrect rollouts. We next examine whether the visual tokens highlighted by these patterns support correct answers. Specifically, we select 10% of the visual tokens under each method and replace their hidden states with the mean visual state of the same image and resample a complete response. Under the same intervention setting, we compare $\mathbf{D}_Q^+$ with the Success VEP $\boldsymbol{\mu}_Q^+$, Saliency-R1 (Gong et al., 2026), answer attention, TAM (Li et al., 2025), random visual tokens, and $\mathbf{D}_Q^-$.

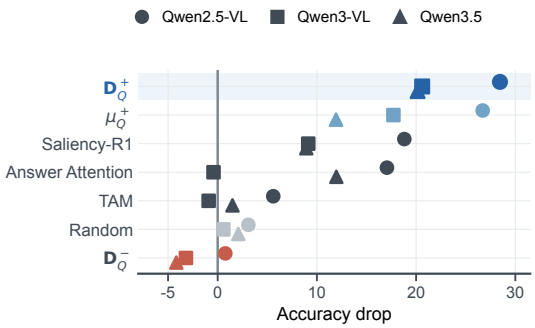


Figure 3: Accuracy drop after replacing 10% of the visual-token hidden states selected by each method. All methods use the same intervention setting across three models.

As shown in Figure 3, intervening on only 10% visual tokens selected by $\mathbf{D}_Q^+$ causes the largest accuracy drop across all three models, with decreases of 28.45, 20.61, and 20.15 percentage points. The drops are consistently larger than those caused by random selection and saliency methods, showing that $\mathbf{D}_Q^+$ identifies visual tokens important for correct answers. The Success VEP $\boldsymbol{\mu}_Q^+$ shows a similar effect, indicating that the base prototype also captures supporting visual tokens. In contrast, intervening on $\mathbf{D}_Q^-$-selected tokens might cause slight improvement in accuracy, suggesting that $\mathbf{D}_Q^-$ may capture visual tokens that hinder reasoning.

Finding 2: *Contrastive VEPs identify supporting visual tokens, and intervening on them significantly reduces answer accuracy across three models.*

3.3 Contrastive VEPs Locate Supporting and Harmful Response Tokens

Pilot 3.2 shows that VEPs identify visual tokens that are important for correct answers. We further ask whether VEPs can also identify key response tokens in correct and incorrect rollouts. For each rollout, we intervene on the visual tokens selected by $\mathbf{D}_Q^+$ for correct rollouts and $\mathbf{D}_Q^-$ for incorrect rollouts, and measure how the probability of each observed response token changes under the same prefix.

Specifically, let $p_{r,t}$ and $\widetilde{p}_{r,t}(\mathbf{D})$ denote the probabilities assigned to the observed token at position t in rollout r under the original and intervened forward passes, respectively. We define

$$\Delta_{r,t}(\mathbf{D}) = \log p_{r,t} - \log \widetilde{p}_{r,t}(\mathbf{D}). \quad (5)$$

A larger positive $\Delta_{r,t}$ indicates that the selected visual tokens have a stronger influence on $y_{r,t}$. For each rollout, we select the response token with

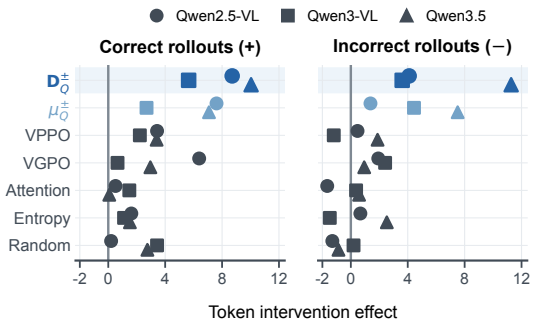


Figure 4: Effect on answer accuracy when the selected response token is retained or suppressed. $\mathbf{D}_Q^\pm$ and $\boldsymbol{\mu}_Q^\pm$ denote contrastive and base VEPs.

the largest positive $\Delta_{r,t}$, excluding final-answer tokens. We then compare two generations from the same prefix, where the selected token is retained in one and suppressed in the other. For correct rollouts, we compute $\mathrm{Acc}(\mathrm{retain}) - \mathrm{Acc}(\mathrm{suppress})$, where a positive value indicates a supporting token. For incorrect rollouts, we compute the reverse difference, $\mathrm{Acc}(\mathrm{suppress}) - \mathrm{Acc}(\mathrm{retain})$, where a positive value indicates a harmful token. We apply the same intervention to tokens selected by VPPO (Huang et al., 2025), VGPO (Wang et al., 2026), reasoning attention, entropy, and random selection.

Figure 4 shows intervention effects for the two outcomes across all three models. ***Suppressing the $\mathbf{D}_Q^+$-selected token decreases answer accuracy for correct rollouts, whereas suppressing $\mathbf{D}_Q^-$-selected token raises accuracy for incorrect rollouts***. Compared with baseline selectors, contrastive VEPs produce larger accuracy changes in each model and outcome, showing that $\mathbf{D}_Q^+$ and $\mathbf{D}_Q^-$ locate supporting and harmful response tokens, respectively.

Finding 3: *Contrastive VEPs locate supporting tokens in correct rollouts and harmful tokens in incorrect rollouts, even when final-answer tokens are excluded.*

Together, the three pilots show that VEPs distinguish correct and incorrect rollouts, identify supporting visual tokens, and locate supporting or harmful response tokens. These findings motivate VEP-guided policy optimization at both the trajectory and token levels.

4 Visual Evidence Prototype-Guided Policy Optimization

Based on the pilot findings, we propose VEP-RL, as illustrated in Figure 5. VEP-RL uses the VEP Similarity Score to reweight trajectory-level updates and VEP-selected response tokens to redistribute credit within each rollout. We build VEP-RL on Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which estimates each rollout's advantage relative to other rollouts sampled for the same input. The advantage determines the direction and strength of the policy update. For rollout $r_i \in \mathcal{R}_Q$ with verifiable reward R_i, its normalized advantage is $\widehat{A}_i = (R_i - \mathrm{mean}(\{R_j\}_{j=1}^G))/\,\mathrm{std}(\{R_j\}_{j=1}^G)$, which is applied uniformly to tokens within a rollout.

In particular, to construct each outcome prototype from multiple rollouts, VEP-RL applies VEP guidance only when a rollout group contains at least two correct and two incorrect rollouts, i.e., $|\mathcal{R}_Q^+| \geq 2$ and $|\mathcal{R}_Q^-| \geq 2$. For all other groups, we use the original GRPO update objective.

Trajectory-level guidance. Finding 1 shows that the VEP Similarity Score distinguishes correct and incorrect rollouts and may signal a rollout's visual reasoning quality. We therefore use this score to refine GRPO's trajectory-level update strength from a visual supervision perspective. Specifically, we standardize VEP Similarity Score s_i separately within $\mathcal{R}_Q^+$ and $\mathcal{R}_Q^-$ to obtain $\widehat{s}_i$, and define:

$$\widehat{A}_i^{\mathrm{traj}} = \begin{cases} [1 + \lambda_{\mathrm{traj}}\,\mathrm{clip}(\widehat{s}_i, -1, 1)]\,\widehat{A}_i, & r_i \in \mathcal{R}_Q^+, \\ [1 - \lambda_{\mathrm{traj}}\,\mathrm{clip}(\widehat{s}_i, -1, 1)]\,\widehat{A}_i, & r_i \in \mathcal{R}_Q^-. \end{cases} \tag{6}$$

Here, λ_{traj} controls the trajectory-level weighting strength, and we set $\lambda_{\mathrm{traj}} = 0.1$ in all experiments. Since the multiplier changes only the update magnitude, the GRPO advantage determines its direction. Thus, correct rollouts more similar to the Success VEP are updated more strongly, while incorrect rollouts more similar to the Failure VEP are also updated more strongly.

Token-level guidance. Finding 3 shows that contrastive VEPs locate supporting response tokens in correct rollouts and harmful ones in incorrect rollouts. We therefore use $\mathbf{D}_Q^+$ for correct rollouts and $\mathbf{D}_Q^-$ for incorrect rollouts to intervene on the selected visual tokens, and rank response tokens by the resulting probability change $\Delta_{r_i,t}$ defined in Eq. (5). We select the highest-scoring response tokens and redistribute the trajectory-level advantage toward them. Specifically, let $\mathcal{U}_i$ be all response-token positions. We rank these positions by $\Delta_{r_i,t}$ and select up to the top 5% with positive scores as $\mathcal{K}_i \subseteq \mathcal{U}_i$. We then reweight the trajectory-level advantage at each response position:

$$\widehat{A}_{i,t}^{\mathrm{VEP}} = \left[1 + \lambda_{\mathrm{tok}} \left(\mathbf{1}[t \in \mathcal{K}_i] - \frac{|\mathcal{K}_i|}{|\mathcal{U}_i|}\right)\right] \widehat{A}_i^{\mathrm{traj}}. \tag{7}$$

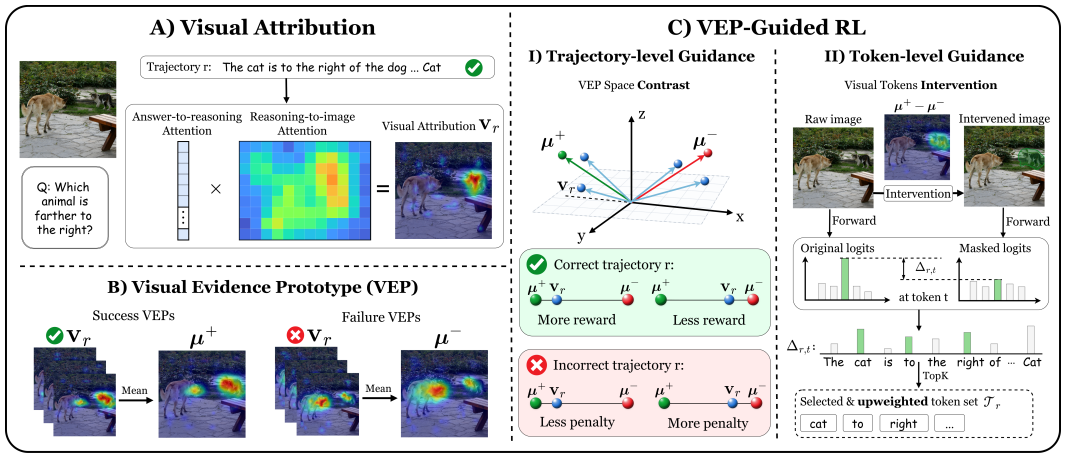


Figure 5: Overview of VEP-RL. A) Visual attribution for each rollout derived from answer-reasoning-image attention. B) Correct and incorrect rollouts are aggregated into Success and Failure VEPs. C) VEP-RL uses VEP Similarity Score for trajectory-level reweighting and contrastive VEPs for token-level credit redistribution.

Here, λ_{tok} controls the strength of token-level redistribution, and we set $\lambda_{\text{tok}} = 0.1$. The multiplier averages to one over $\mathcal{U}_i$, selected tokens receive more weight, and the remaining tokens receive less.

VEP-RL objective. Let $y_{i,t}$ denote the t-th token of rollout r_i, and let $\rho_{i,t} = \pi_\theta(y_{i,t} \mid Q, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid Q, y_{i,<t})$ be its importance ratio. Building on the GRPO objective given in Appendix B.1, VEP-RL replaces the trajectory-level advantage with the token-specific advantage $\hat{A}^{\text{VEP}}_{i,t}$ defined above. The resulting objective is

$$\mathcal{J}_{\text{VEP-RL}}(\theta) = \mathbb{E}_{Q,\{r_i\}} \left[\sum_{i=1}^{G} \sum_{t=1}^{|r_i|} \frac{\min\left\{ \rho_{i,t} \widehat{A}^{\text{VEP}}_{i,t},\ \text{clip}(\rho_{i,t}, 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}) \widehat{A}^{\text{VEP}}_{i,t} \right\} - \beta D^{i,t}_{\text{KL}}}{G|r_i|} \right]. \tag{8}$$

Here, $\varepsilon_{\text{low}} = 0.20$ and $\varepsilon_{\text{high}} = 0.28$ set the clipping ranges, while $\beta D^{i,t}_{\text{KL}}$ regularizes the policy toward the reference model. Thus, VEP-RL changes how the verified trajectory reward is assigned across rollouts and tokens while retaining the GRPO surrogate and KL regularization.

5 Experiments

5.1 Experimental Setup

Training Setup We train Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3.5-9B with VEP-RL on ViRL-39K (Wang et al., 2025b). Our comparisons include standard RL methods, GRPO and DAPO (Shao et al., 2024; Yu et al., 2025); visual RLVR methods, including PAPO, VPPO, VGPO, VAPO, NoisyRollout, VL-Rethinker, Octopus, PRCO, and ContextRL (Wang et al., 2025b; Huang et al., 2025; Wang et al., 2026; Tian et al., 2025; Liu et al., 2025; Wang et al., 2025a; Ding et al., 2026; Miao et al., 2026; Lu et al., 2026), and visual OPD methods, Vision-OPD, VAD, and OPD-V (Yuan et al., 2026; Zhang et al., 2026; Aniri et al., 2026). The training settings for our GRPO, DAPO, and VEP-RL runs are detailed in Appendix B.1.

Evaluation Setup We evaluate on visual reasoning benchmarks including MathVerse, MathVista, DynaMath, MATH-Vision, MMK12, ScienceQA, We-Math, Geometry3K, and LogicVista (Zhang et al., 2024; Lu et al., 2024; Zou et al., 2025; Wang et al., 2024; Meng et al., 2025; Lu et al., 2022; Qiao et al., 2025; Lu et al., 2021; Xiao et al., 2024). We further evaluate vision-centric reasoning on the vision-dependent subset of MathVerse, CLEVR Counting, HallusionBench, VLMsAreBlind, and V*Bench (Zhang et al., 2024; Johnson et al., 2017; Guan et al., 2024; Rahmanzadehgervi et al.,

Table 1: Main results on 14 benchmarks (accuracy, %). Best and second-best results within each base model are shown in **bold** and underlined, respectively.

Method	*Visual Reasoning*									*Vision-Centric*					*Avg.*
	MathVerse	MathVista	DynaMath	MathVision	MMK12	ScienceQA	We-Math	Geo3K	LogicVista	MathVerse-V	CLEVR	HalluBench	VLMsAreBlind	V*	
Base model: Qwen2.5-VL-7B-Instruct															
NoisyRollout	57.52	70.80	64.35	20.16	47.70	86.35	63.16	**52.91**	48.77	50.38	83.50	71.82	46.56	79.06	60.22
VL-Rethinker	70.09	61.50	62.36	32.03	64.90	77.48	68.85	42.26	45.86	67.05	71.00	69.61	44.49	70.68	60.58
PAPO	67.25	70.50	63.04	27.97	72.75	89.58	67.47	41.26	48.32	63.46	89.50	72.24	48.14	79.06	64.32
GRPO	70.23	68.70	65.41	28.79	74.50	89.39	71.09	42.43	48.10	67.35	86.50	69.72	45.42	79.58	64.80
DAPO	68.39	69.20	65.41	30.51	79.90	89.11	69.60	43.26	48.55	64.37	89.00	66.98	49.63	80.10	65.29
VPPO	71.42	65.50	66.28	30.13	81.35	91.65	71.90	44.43	49.22	68.88	84.50	70.03	49.65	82.20	66.22
VAPO	**72.66**	69.40	65.93	30.58	74.25	89.86	**73.39**	45.09	**49.66**	69.65	88.00	70.35	46.10	82.72	66.26
VGPO	69.91	70.40	65.44	29.17	81.80	90.90	69.89	45.42	48.32	64.76	93.00	70.45	48.40	83.25	66.51
Ours	**72.66**	**72.30**	**67.24**	**32.16**	**83.20**	**91.77**	72.53	49.25	**49.66**	**70.03**	**93.50**	**74.13**	**52.03**	**84.82**	**68.95**
Base model: Qwen3-VL-8B-Instruct															
OPD-V	59.22	69.60	51.55	32.44	60.40	91.20	62.99	28.62	40.27	55.89	92.00	72.24	**67.73**	85.34	62.11
ContextRL	72.71	71.00	74.80	48.19	67.60	95.14	78.45	64.06	56.15	70.80	88.50	72.34	52.37	87.43	71.40
PRCO	79.13	75.40	76.79	44.44	72.25	93.35	78.62	65.89	57.49	75.54	82.50	**76.45**	61.24	82.72	72.99
DAPO	81.74	70.70	76.21	53.22	77.55	95.64	82.01	68.39	60.85	78.36	89.50	75.08	52.78	**89.01**	75.07
Octopus	82.16	73.90	77.55	53.59	77.40	96.51	84.02	68.39	63.53	**80.66**	89.00	75.92	57.65	85.34	76.12
VPPO	83.81	69.60	78.48	55.21	**80.60**	97.05	83.97	67.89	63.53	80.20	90.50	75.29	58.20	85.34	76.41
Ours	**84.13**	**75.60**	**80.01**	**55.76**	79.45	**97.45**	**84.77**	**74.04**	**64.43**	80.58	**95.50**	76.03	63.60	85.34	**78.34**
Base model: Qwen3.5-9B															
OPD-V	80.00	70.60	60.50	55.62	83.60	92.45	83.33	43.09	55.03	72.78	77.50	70.24	80.80	93.19	72.77
Vision-OPD	83.58	71.60	63.99	58.55	85.80	92.41	88.05	46.26	67.79	75.99	87.50	71.92	85.08	91.62	76.44
VAD	88.94	75.30	74.41	63.67	86.85	95.05	**91.15**	67.89	**77.40**	86.47	86.00	76.34	84.84	**94.76**	82.08
GRPO	89.27	76.10	86.85	61.30	86.65	98.44	87.93	84.86	70.92	85.93	95.50	**77.92**	88.49	88.48	84.19
DAPO	89.82	75.50	86.58	63.54	92.25	97.12	88.51	87.02	72.71	87.69	95.00	75.81	87.57	85.86	84.64
Ours	**92.57**	**76.80**	**89.91**	**72.65**	**94.10**	**98.89**	**91.15**	**90.85**	75.17	**89.98**	**97.00**	77.29	**92.90**	91.10	**87.88**

2024; Wu & Xie, 2024). We report accuracy on each benchmark and their macro-average. Evaluation settings and benchmark details are provided in Appendix B.2.

5.2 Main Results

Performance across Model Backbones. VEP-RL achieves the highest average accuracy on all three model backbones (Table 1). Compared with the strongest baseline, it improves average accuracy by 2.44, 1.93, and 3.24 percentage points on Qwen2.5-VL, Qwen3-VL, and Qwen3.5, respectively. On Qwen3.5, VEP-RL further exceeds VAD, the strongest visual OPD baseline, by 5.80 points.

Visual Reasoning and Vision-Centric Performance. VEP-RL consistently improves both benchmark groups across all three backbones, with gains of $2.10/2.93$, $1.72/2.30$, and $3.23/2.39$ points in visual reasoning and vision-centric accuracy, respectively. The gains span perception-focused benchmarks such as CLEVR and VLMsAreBlind and visual mathematical reasoning benchmarks such as Geo3K and MathVision. We report the statistical analysis of Qwen3.5-9B results in Appendix B.2.2.

5.3 Ablation Studies

We conduct ablations on both VEP construction and VEP-RL guidance, as shown in Table 2. **First, each component of VEP construction contributes to performance.** Replacing the full answer–response–visual attribution with answer attention or reasoning attention reduces performance, while using a single rollout instead of outcome-level aggregation and removing the contrast during token selection also leads to consistent drops. To examine whether the improvement depends on a rollout contributing to its own outcome prototype, we further construct a leave-one-out (LOO) variant that excludes each rollout from the prototype of its corresponding outcome when computing its VEP guidance. LOO VEP achieves 86.31 on VR and 89.26 on VC, close to the 86.90 and 89.65 of the default VEP-RL, showing that the gain is largely preserved without self-inclusion. **Second, both forms of VEP-RL guidance are effective.** Trajectory-only and token-only guidance both improve

Table 2: Ablation of VEP construction and VEP-RL guidance on Qwen3.5-9B. Category-average accuracy (%). VR and VC denote Visual Reasoning and Vision-Centric, respectively. In the attribution column, A, R, and V denote final-answer, reasoning (response), and visual tokens. We report more ablations of VEP-RL guidance on Qwen2.5-VL-7B in Appendix B.3.

	VEP Construction			VEP-RL Guidance		Accuracy (%)	
Variant	Attribution	Aggregation	Contrast	Trajectory	Token	VR	VC
GRPO	–	–	–	–	–	82.48	87.26
DAPO	–	–	–	–	–	83.67	86.39
VEP construction ablations							
Attribution							
Answer attention	$A \to V$	Outcome mean	✓	✓	✓	85.18	87.96
Reasoning attention	$R \to V$	Outcome mean	✓	✓	✓	85.70	88.47
Aggregation							
Single rollout	$A \to R \to V$	Single rollout	✓	✓	✓	85.47	88.29
Leave-one-out VEP	$A \to R \to V$	LOO mean	✓	✓	✓	86.31	89.26
Contrast							
w/o prototype contrast	$A \to R \to V$	Outcome mean	–	✓	✓	86.14	88.98
VEP-RL guidance ablations							
Trajectory only	$A \to R \to V$	Outcome mean	✓	✓	–	85.52	88.63
Token only	$A \to R \to V$	Outcome mean	✓	–	✓	85.18	88.14
VEP-RL	$A \to R \to V$	Outcome mean	✓	✓	✓	**86.90**	**89.65**

over DAPO, while combining them achieves the best performance, indicating that trajectory- and token-level guidance provide complementary benefits.

5.4 Analyses

Analysis of VEP Guidance. To investigate the effect of VEP guidance, we measure the VEP Similarity Score of both correct and incorrect rollouts before and after VEP-RL training. As shown in Figure 6, both distributions shift toward higher scores. Probability mass in lower similarity score regions is reduced and squeezed into higher similarity score regions within each outcome category. We call this empirical redistribution the **VEP Squeeze Effect**. This pattern is consistent with our trajectory-level guidance that correct rollouts with higher s_r receive stronger reinforcement, whereas incorrect rollouts with lower s_r receive stronger penalties. This effect, together with the following error analysis in Appendix E.1, suggests that our method likely optimizes the model's visual perception capability.

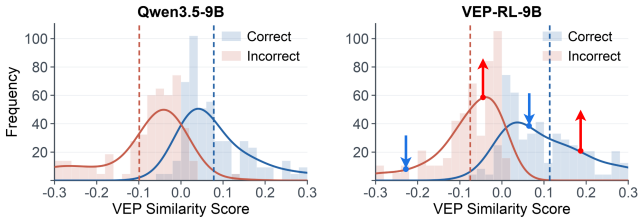


Figure 6: Distribution of the VEP Similarity Score s_r.

Analysis of Intervention Ratio. To examine the choice of the default 10% intervention ratio, we measure accuracy drops under intervention ratios from 5% to 20%. Beyond accuracy, we also measure the mean token log-likelihood of the original correct response, which reflects the model's confidence in that response. As shown in Figure 7, intervening on only 5–10% of the selected visual tokens with D_Q^+ yields significant accuracy drops and a lower likelihood. We therefore adopt 10% as the intervention ratio in VEP-RL. The effect of VEP intervention becomes marginal beyond 10%, echoing observations (Yang et al., 2025) that visual tokens are highly redundant and ***useful visual evidence is sparse***. In contrast, random masking is not evidence-aware and

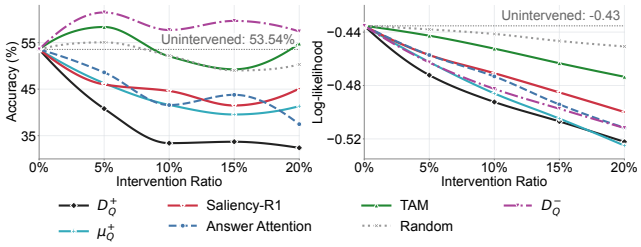


Figure 7: Answer accuracy and log-likelihood.

removes visual tokens uniformly, often requiring ratios as high as 60% or 80% (Wang et al., 2025b) to sufficiently intervene on critical visual evidence.

Generalization to Out-of-Domain Tasks. To evaluate whether VEP-RL generalizes beyond the training domains, we further evaluate the trained models on MMMU-Pro (Yue et al., 2025), MM-Star (Chen et al., 2024), and CV-Bench (Tong et al., 2024). As shown in Table 3, VEP-RL consistently improves both evaluated backbones over their corresponding base models. On Qwen3-VL-8B, VEP-RL improves the average accuracy from 69.19% to 73.48%, and also outperforms VPPO on all three benchmarks. On Qwen3.5-9B, VEP-RL reaches 78.74% average accuracy, improving the base model by 1.27 percentage points. It also exceeds Vision-OPD-9B by 10.39, 6.94, and 3.75 points, respectively. These results show that the improvements of VEP-RL extend to unseen visual reasoning tasks. More error analysis and case studies of VEP and VEP-RL are provided in Appendix E.

Table 3: Performance on out-of-domain visual reasoning benchmarks.

Model	MMMU-Pro	MM-Star	CV-Bench	Avg.
Qwen3-VL-8B	55.90	70.90	80.78	69.19
GRPO	51.97	72.13	84.95	69.68
VPPO	53.52	73.27	86.13	70.97
VEP-RL-8B	**58.11**	**74.87**	**87.45**	**73.48**
Qwen3.5-9B	67.40	77.13	87.87	77.47
GRPO	62.89	76.60	87.76	75.75
Vision-OPD	57.12	72.73	85.29	71.71
VEP-RL-9B	**67.51**	**79.67**	**89.04**	**78.74**

6 Related Work

RLVR for Visual Reasoning. Existing RLVR methods improve visual reasoning through better rollout construction, reward design, and optimization objective. Some methods enrich training rollouts through data augmentation, replay, recombination, or resampling (Liu et al., 2025; Wang et al., 2025a; Ding et al., 2026; Lu et al., 2026; Miao et al., 2026). Some methods complement answer correctness with perception rewards (Xiao et al., 2026a; Tian et al., 2025), saliency-alignment rewards (Gong et al., 2026), or contrastive visual dependency rewards (Huang et al., 2026). Closer to our work, some methods incorporate perceptual signals into the optimization objective (Wang et al., 2025b; Yu et al., 2026), or leverage visual dependency and visual focus to provide fine-grained optimization signals (Huang et al., 2025; Wang et al., 2026), operating at trajectory or token levels.

Visual Supervision Signals. Existing methods construct visual supervision from different sources. Visual dependency measures sensitivity to random perturbations (Wang et al., 2025b; Huang et al., 2025), or similarity between response and visual representations (Wang et al., 2026), but strong dependence does not necessarily imply correct visual reasoning (Liu et al., 2026b; Xiao et al., 2026b). Explicit guidance through privileged annotated visual regions (Yuan et al., 2026; Liu et al., 2026a; Sun et al., 2026; Zhang et al., 2026; Aniri et al., 2026) makes it dependent on the availability and quality of external evidence. Saliency-based signals localize relevant visual regions through attention or activation maps (Li et al., 2025; Gong et al., 2026), but saliency alone does not explicitly distinguish between the visual evidence contributing to successful or failed reasoning processes.

7 Conclusion

We introduced Visual Evidence Prototypes (VEPs), which derive fine-grained visual supervision from on-policy rollout outcomes without region annotations. VEPs aggregate visual attributions from correct and incorrect rollouts into Success and Failure prototypes and contrast them to capture differences in the visual evidence used across outcomes. Empirical analyses show that VEPs identify supporting visual and response tokens, and that intervening on only the top 10% selected visual tokens reduces answer accuracy by 20.15–28.45 percentage points. Building on these findings, VEP-RL uses VEP to reweight trajectory-level updates and contrastive VEPs to redistribute token-level credit. Across fourteen visual reasoning and perception benchmarks, VEP-RL consistently improves three model backbones over their strongest baselines, with gains also extending to unseen benchmarks.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, code development and debugging, and proving mathematical claims. All AI-assisted text, code, analyses, and artifacts were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

Our research is conducted entirely within the scope of LVLM reasoning and relies exclusively on publicly available academic benchmarks. These datasets contain no personal, sensitive, or private user information. We collected no new data for this study, and no human subjects were involved. Our primary objective is to improve AI models' visual reasoning capabilities. Given the focus on abstract problem-solving and the use of established public benchmarks, we are not aware of any direct or foreseeable ethical concerns or negative societal impacts arising from our methodology or findings.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility of our work, we describe the experimental setup, including training data, optimization, and rollout configuration, in Appendix B.1, and the main training hyperparameters are summarized in Table 4. Appendix B.2 describes the evaluation benchmarks, inference settings, metrics, and scoring procedures. For public baselines, we follow the inference settings in their official repositories. Appendix C describes the pilot-study settings.

REFERENCES

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

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*, 2025. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024. doi: 10.52202/079017-0850.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Runxin Xu, Qihao Zhu, Zhihong Shao, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Yi Ding, Ziliang Qiu, Bolian Li, and Ruqi Zhang. Learning self-correction in vision-language models via rollout augmentation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08503.

Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. Saliency-R1: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2604.04500.

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. 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*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/

Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Haojie Huang, Xinlei Yu, Chengming Xu, Zhangquan Chen, Cheng Yang, Qingdong He, Yu Yang, Jiangning Zhang, and Xiaobin Hu. Evidence-RL: Towards evidence-intensive visual reasoning. *arXiv preprint arXiv:2608.08021*, 2026. URL https://arxiv.org/abs/2608.08021.

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.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 48–58, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_Token_Activation_Map_to_Visually_Explain_Multimodal_LLMs_ICCV_2025_paper.html.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026a. URL https://arxiv.org/abs/2605.21924.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. NoisyRollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025. URL https://arxiv.org/abs/2504.13055.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=JAI7afWA9e.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6774–6786, 2021. URL https://aclanthology.org/2021.acl-long.528/.

Pan Lu, Swaroop Mishra, Tony 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. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.09513.

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. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02255.

Xingyu Lu, Jinpeng Wang, Yifan Zhang, Shijie Ma, Xiao Hu, Tianke Zhang, Haonan Fan, Kaiyu Jiang, Changyi Liu, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, and Chun Yuan. ContextRL: Enhancing MLLM's knowledge discovery efficiency with context-augmented RL. *arXiv preprint arXiv:2602.22623*, 2026. URL https://arxiv.org/abs/2602.22623.

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. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. URL https://arxiv.org/abs/2503.07365.

Ziqi Miao, Haonan Jia, Lijun Li, Chen Qian, Yuan Xiong, Wenting Yan, and Jing Shao. Seeing with you: Perception-reasoning coevolution for multimodal reasoning. *arXiv preprint arXiv:2603.28618*, 2026. URL https://arxiv.org/abs/2603.28618.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 20023–20070, 2025. URL https://aclanthology.org/2025.acl-long.983/.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024. URL https://openaccess.thecvf.com/content/ACCV2024/html/Rahmanzadehgervi_Vision_language_models_are_blind_ACCV_2024_paper.html.

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

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-Zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026. URL https://arxiv.org/abs/2606.25319.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *arXiv preprint arXiv:2509.25848*, 2025. URL https://arxiv.org/abs/2509.25848.

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. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024. doi: 10.52202/079017-2771.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a. URL https://arxiv.org/abs/2504.08837.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.14804.

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

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2506.07218.

Wenyi Xiao, Xinchi XU, and Leilei Gan. VL-calibration: Decoupled confidence calibration for large vision-language models reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44791–44815, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2074. URL https://aclanthology.org/2026.acl-long.2074/.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. URL https://arxiv.org/abs/2407.04973.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual policy optimization for multimodal reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.23206.

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, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025. doi: 10.18653/v1/2025.acl-long.736.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026. URL https://arxiv.org/abs/2607.28590.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.14624.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

APPENDIX CONTENTS

A Limitations 14

B Experimental Setup 15

B.1 Training Details 15

B.2 Evaluation Details 16

B.3 Ablation Studies 19

C Pilot Study Details 20

C.1 VEPs Distinguish Correct and Incorrect Rollouts 22

C.2 Contrastive VEPs Identify Supporting Visual Tokens 22

C.3 Contrastive VEPs Locate Supporting and Harmful Response Tokens 23

D Theoretical Analysis 24

D.1 Contrastive Prototype Recovery 24

D.2 Local Policy Improvement 25

E Additional Analysis 27

E.1 Error Analysis 27

E.2 Selected-token Frequencies 27

E.3 Case Study 28

A LIMITATIONS

While our results demonstrate the effectiveness of VEP and VEP-RL, it is important to acknowledge their current limitations and future directions.

Computational Overhead. Our method introduces a limited computational overhead. To construct VEPs and compute token-level guidance, VEP-RL requires an additional intervened forward pass for rollouts in eligible groups. Empirically, we found this increases total training time by approximately 3–4% relative to GRPO (e.g., from 58.3 to 60.2 hours on Qwen2.5-VL-7B).

Scope of Generalization. Our experiments have demonstrated the effectiveness of VEP-RL across three models. We have not yet verified the efficacy of our method on extremely large-scale models (e.g., 32B+ parameters) due to computational resource constraints.

B EXPERIMENTAL SETUP

B.1 TRAINING DETAILS

Table 4: Training hyperparameters used for GRPO, DAPO, and VEP-RL.

Hyperparameter	**Value**
Training data	ViRL-39K
Training epochs	2
Optimizer	AdamW
Learning rate	1×10^{-6}
Actor global batch size	128
Rollout batch size	256
Group size G	8
Rollout temperature / top-p	1.0 / 1.0
Training data order	Shuffled with seed 42
Rollout sampling seed	42
Maximum prompt / response length	4096 / 4096
Precision	BF16
Minimum / maximum image pixels	262,144 / 4,000,000
KL coefficient	1×10^{-4}
GRPO/DAPO/VEP-RL clip-low / clip-high	0.20 / 0.28
Verifier	MathRuler boxed-answer grader
Accuracy / format reward weight	0.9 / 0.1
VEP visual-token selection ratio	10%
VEP response-token selection	Top 5% positive scores
VEP $\lambda_{\mathrm{traj}}/\lambda_{\mathrm{tok}}$	0.1 / 0.1

To ensure a fair comparison with VPPO (Huang et al., 2025) and PAPO (Wang et al., 2025b), following the same training settings as VPPO (Huang et al., 2025), we train GRPO, DAPO, and VEP-RL on ViRL-39K for two epochs from Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3.5-9B. Table 4 lists the hyperparameters.

GRPO objective. Using the importance ratio defined in the main text, our GRPO objective is

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{Q,\{r_i\}}\left[\sum_{i=1}^{G}\sum_{t=1}^{|r_i|}\frac{\min\left\{\rho_{i,t}\hat{A}_i, \mathrm{clip}(\rho_{i,t}, 1-\varepsilon_{\mathrm{low}}, 1+\varepsilon_{\mathrm{high}})\,\hat{A}_i\right\} - \beta D_{\mathrm{KL}}^{i,t}}{G|r_i|}\right]. \quad (9)$$

Here, $\hat{A}_i$ is the group-normalized trajectory advantage computed from the rollout reward. The clipping bounds and KL regularization are shared with VEP-RL.

General Training Hyperparameters. We use full-parameter training for two epochs, including the vision tower, with AdamW, a learning rate of 10^{-6}, and BF16 precision. Each question produces eight rollouts sampled with temperature 1.0 and top-$p = 1.0$. The rollout batch size is 256 and the actor global batch size is 128. Both the prompt and response lengths are capped at 4096 tokens. Images contain between 262,144 and 4,000,000 pixels after preprocessing. The GRPO, DAPO, and VEP-RL objectives use the VeRL default clip-low 0.20 and clip-high 0.28, together with low-variance KL regularization with coefficient 10^{-4}. We shuffle the training data with seed 42 and use the same seed for rollout sampling.

Each prompt requires the reasoning process inside `<think>...</think>` and the final answer inside `\boxed{...}`. We use MathRuler to extract the final boxed content and compare it with the reference answer. The reward assigns weights 0.9 and 0.1 to answer correctness and format compliance, respectively; other answer-like spans are not considered.

VEP-RL-specific Hyperparameters. We average attention over heads and over the decoder layers at 45%, 65%, and 85% of model depth, rounding each index upward. The resulting one-based layer indices are listed in Table 5. For each question, the Success and Failure VEPs are the mean attributions of its correct and incorrect rollouts. We rank visual tokens directly by the signed difference between these prototypes, without taking the positive part or renormalizing the scores. We apply VEP guidance

only when a rollout group contains at least two correct and two incorrect rollouts. Otherwise, we use the original update in Eq. (9).

For token-level guidance, correct rollouts use $\mathbf{D}_Q^+$ and incorrect rollouts use $\mathbf{D}_Q^-$. We replace the top 10% selected visual states with the mean visual state of the same image before the first language-decoder layer. For Qwen3-VL, we apply the same replacement to each DeepStack visual stream. We record the original-image sampled-token log probabilities during rollout generation. For groups eligible for VEP guidance, we teacher-force the sampled response once with the selected visual states replaced to obtain the intervened log probabilities. Candidate positions are all response tokens. We rank them by $\Delta_{r,t}$ from Eq. (5) and select up to the top 5%, retaining only positions with positive scores. Final-answer tokens remain eligible during training.

Prototype construction and token selection use stop-gradient. If a rollout has no eligible positive score, every response token keeps the trajectory-level advantage. We set $\lambda_{\mathrm{traj}} = \lambda_{\mathrm{tok}} = 0.1$.

Table 5: Decoder layers used to construct VEPs. Layer indices are one-based and are obtained by applying $\lceil rL \rceil$ at depth ratios $r \in \{0.45, 0.65, 0.85\}$. For Qwen3.5, we use the nearest full-attention layers at these depths.

Backbone	**Depth L**	**Decoder layers**
Qwen2.5-VL-7B	28	13, 19, 24
Qwen3-VL-8B	36	17, 24, 31
Qwen3.5-9B	32	16, 20, 28

Computation Environment. Training uses one node with eight NVIDIA A800-SXM4 80GB GPUs. The implementation extends VeRL and uses vLLM for rollout generation and evaluation. We will release the training and evaluation code, configurations, and model checkpoints.

Training Time. VPPO uses a masked-image forward pass for every rollout (Huang et al., 2025). VEP-RL uses an intervened pass only in groups with at least two correct and two incorrect rollouts. Table 6 compares training time and throughput. The VEP-RL measurements include attention extraction and VEP construction.

Table 6: Training time and throughput. VPPO rows are estimates obtained by applying its reported 9.7% overhead (Huang et al., 2025) to the GRPO rows. Throughput is measured in questions per second, and overhead is relative to GRPO. Active (%) is the fraction of training rollout groups with at least two correct and two incorrect responses.

Backbone	**Method**	**Time (h)**	**Q/s**	**Overhead (%)**	**Active (%)**
Qwen2.5-VL-7B	GRPO	58.3	0.370	0.00	–
	VEP-RL	60.2	0.359	+3.26	27.8
	VPPO (est.)	64.0	0.337	+9.70	–
Qwen3-VL-8B	GRPO	75.8	0.285	0.00	–
	VEP-RL	78.6	0.275	+3.69	20.6
	VPPO (est.)	83.2	0.260	+9.70	–
Qwen3.5-9B	GRPO	86.1	0.251	0.00	–
	VEP-RL	89.2	0.242	+3.60	17.9
	VPPO (est.)	94.5	0.229	+9.70	–

VEP-RL adds about 3.5% training time over GRPO, compared with the 9.7% VPPO estimate according to its paper. Its intervened pass is used in 17.9–27.8% of groups.

B.2 Evaluation Details

B.2.1 Models and Evaluation Settings

To ensure a fair comparison with the public baselines, we evaluate all models on the complete benchmark files with the same scoring rules. For the public baselines in Table 1, we follow the inference settings in their official repositories. The training data and inference settings for each model

are listed below. For our trained GRPO, DAPO, and VEP-RL models on all three backbones, we report average accuracy@8: we score all eight responses to each question separately and average their correctness scores.

Qwen2.5-VL-7B.

- **NoisyRollout** (Liu et al., 2025): Qwen2.5-VL-7B-Instruct, trained on Geo3K. Its evaluation script uses temperature 0, top-$p = 0.95$, one response, and a 2048-token response limit.
- **VL-Rethinker** (Wang et al., 2025a): Qwen2.5-VL-7B-Instruct, trained on ViRL-39K. Its evaluation script uses temperature 1.0, top-$p = 0.95$, one response, and an 8192-token response limit.
- **PAPO** (Wang et al., 2025b): Qwen2.5-VL-7B-Instruct, trained on ViRL-39K. PAPO-Eval uses temperature 1.0, top-$p = 1.0$, eight responses per question, and a 2048-token response limit.
- **VPPO** (Huang et al., 2025): Qwen2.5-VL-7B-Instruct, trained on ViRL-39K. Its PAPO-Eval setting uses temperature 1.0, top-$p = 1.0$, eight responses per question, and a 2048-token response limit.
- **GRPO, DAPO, and VEP-RL**: Qwen2.5-VL-7B-Instruct, trained on ViRL-39K with the settings in Table 4. Inference matches VPPO: temperature 1.0, top-$p = 1.0$, eight responses per question, and a 2048-token response limit.
- **VAPO** (Tian et al., 2025): Qwen2.5-VL-7B-Instruct, trained on the ViRL-39K-derived VAPO-Thinker-36K set with visual anchors. Its VLMEvalKit/LMDeploy evaluation uses temperature 0, one response, and a 1024-token default response limit.
- **VGPO** (Wang et al., 2026): Qwen2.5-VL-7B-Instruct, trained on ViRL-39K. Its Look-Back-based evaluation uses temperature 0 and one response per question.

Qwen3-VL-8B.

- **OPD-V** (Aniri et al., 2026): Qwen3-VL-8B-Instruct, trained on Vision-OPD-6K. Its evaluation script uses temperature 0, a 32,768-token response limit, and seed 42.
- **ContextRL** (Lu et al., 2026): Qwen3-VL-8B-Instruct, trained on a 29K-example set of VQA and multimodal math questions selected from FineVision. The paper uses VLMEvalKit for evaluation but does not report its decoding parameters; the 16K response limit is a training setting.
- **PRCO** (Miao et al., 2026): Qwen3-VL-8B-Instruct, trained on ViRL-39K. Its two-stage VLMEvalKit inference uses temperature 0 and top-$p = 1.0$ for both stages, with 1024 tokens for the Observer and 2048 for the Solver.
- **Octopus** (Ding et al., 2026): Qwen3-VL-8B-Instruct, first trained on a 10K-example self-correction set and then on ViRL-39K. Its vLLM inference uses the self-correction prompt, temperature 1.0, top-$p = 0.95$, and a 16,384-token response limit.
- **VPPO** (Huang et al., 2025): Qwen3-VL-8B-Instruct, trained on ViRL-39K. Its PAPO-Eval setting uses temperature 1.0, top-$p = 1.0$, eight responses per question, and a 2048-token response limit.
- **GRPO, DAPO, and VEP-RL**: Qwen3-VL-8B-Instruct, trained on ViRL-39K with the settings in Table 4. Inference matches VPPO: temperature 1.0, top-$p = 1.0$, eight responses per question, and a 2048-token response limit.

Qwen3.5-9B.

- **OPD-V** (Aniri et al., 2026): Qwen3.5-9B, trained on Vision-OPD-6K. Its evaluation script uses temperature 0, a 32,768-token response limit, and seed 42.
- **Vision-OPD** (Yuan et al., 2026): Qwen3.5-9B, trained on Vision-OPD-6K. Its evaluation script uses temperature 0, a 32,768-token response limit, and seed 42.

- **VAD** (Zhang et al., 2026): Qwen3.5-9B, trained on three-view examples built from Vision-OPD-6K. Its evaluation script uses temperature 0, an 8192-token response limit, and seed 42, with thinking disabled.
- **GRPO, DAPO, and VEP-RL**: Qwen3.5-9B, trained on ViRL-39K with the settings in Table 4. Inference uses temperature 0, eight responses per question, and an 8192-token response limit.

We report accuracy for each benchmark. Before scoring HallusionBench, we map *Yes* to 1 and *No* to 0. For V*Bench, we report micro accuracy over its 191 examples. For VLMsAreBlind, we use the official eight-group macro average, keeping Olympic rings and pentagons separate and combining the word and blank variants of the grid task. For CV-Bench, we average the ADE20K/COCO 2D score and the Omni3D 3D score. The category and overall scores in Table 1 are unweighted macro-averages of the reported benchmark scores.

B.2.2 Stability across Training Seeds

We repeat DAPO and VEP-RL on Qwen3.5-9B with seeds 42, 43, and 44. The runs use the same base checkpoint, training data, number of updates, batch sizes, and sampling hyperparameters; the seed controls data order and rollout sampling. Table 7 reports the 14-benchmark macro accuracy for every run.

Table 7: Qwen3.5-9B results across three training seeds (14-benchmark macro accuracy, %).

Method	Seed 42	Seed 43	Seed 44	Mean $\pm$ std
DAPO	84.64	84.18	84.79	84.54 ± 0.32
VEP-RL	87.88	87.88	87.29	87.68 ± 0.34
VEP-RL $-$ DAPO	+3.24	+3.70	+2.50	$+3.15 \pm 0.61$

VEP-RL improves DAPO by 3.24, 3.70, and 2.50 points across the three training seeds, for a mean paired gain of 3.15 ± 0.61 points. Treating each seed pair as one observation on the fixed 14-benchmark suite, the gain is significant under a two-sided paired t-test ($t(2) = 9.00$, $p = 0.012$), with a 95% confidence interval of $[1.64, 4.65]$ points for the mean gain.

B.2.3 Evaluation Benchmarks

We evaluate fourteen benchmarks covering mathematical reasoning, scientific and logical reasoning, and vision-centric tasks.

Mathematical Reasoning. MathVerse varies how information is divided between the question and the diagram, while MathVista places mathematical problems in charts, diagrams, and synthetic scenes (Zhang et al., 2024; Lu et al., 2024). MATH-Vision adds competition-level problems across a broad range of subjects and difficulty levels (Wang et al., 2024). DynaMath measures robustness to visual and numerical variations of the same problem (Zou et al., 2025). We-Math organizes problems by knowledge concept and reasoning step, and Geometry3K focuses on geometry problems that require the diagram and text to be interpreted together (Qiao et al., 2025; Lu et al., 2021).

Scientific and Logical Reasoning. MMK12 contains K–12 mathematics and science problems with human-verified answers (Meng et al., 2025). ScienceQA covers questions from natural, social, and language science (Lu et al., 2022), while LogicVista examines inductive, deductive, numerical, spatial, and mechanical reasoning in visual contexts (Xiao et al., 2024).

Vision-centric Reasoning. MathVerse-V contains the vision-dependent MathVerse problems whose solutions require information from the diagram (Zhang et al., 2024). CLEVR Counting isolates object counting in controlled synthetic scenes (Johnson et al., 2017). HallusionBench studies visual illusion and language hallucination (Guan et al., 2024), and VLMsAreBlind probes basic visual operations such as counting, recognizing intersections, and judging spatial relations (Rahmanzadehgervi et al., 2024). V*Bench tests fine-grained visual search in high-resolution and crowded images (Wu & Xie, 2024).

Out-of-domain Evaluation. For out-of-domain evaluation, we test Qwen3-VL and Qwen3.5 models on MMMU-Pro, MM-Star, and CV-Bench. These benchmarks are not used for training. They test multi-disciplinary reasoning, fine-grained perception, and 2D/3D spatial understanding. We report the unweighted mean of the three benchmark scores.

B.3 Ablation Studies

To give a further validation of VEP-RL, we additionally ablate trajectory- and token-level guidance on Qwen2.5-VL-7B. Table 8 reports the two category averages used in the main ablation study. Beyond ablation of components, we also test whether the gains could arise from random reweighting rather than VEP-based selection.

Table 8: VEP-RL guidance ablations on Qwen2.5-VL-7B. VR and VC denote Visual Reasoning and Vision-Centric category-average accuracy (%).

Variant	Trajectory	Token	VR	VC
DAPO	–	–	62.66	70.02
Trajectory only	✓	–	64.58	73.64
Token Shuffle	✓	Shuffle	64.63	73.58
Token only	–	✓	64.13	72.87
Trajectory Shuffle	Shuffle	✓	64.17	72.95
Excluded Final-Answer Token	✓	No Final-Answer	65.38	74.56
VEP-RL	✓	✓	**65.64**	**74.90**

First, shuffling trajectory weights while retaining token-level guidance yields results close to the token-only variant (64.17/72.95 versus 64.13/72.87 on VR/VC). Similarly, shuffling token weights while retaining trajectory-level guidance yields results close to the trajectory-only variant (64.63/73.58 versus 64.58/73.64). Both are below full VEP-RL (65.64/74.90), suggesting that where the weights are assigned matters. Second, excluding final-answer tokens from token-level guidance retains 65.38/74.56, indicating that non-final-answer response tokens provide the main contribution to the improvement. Third, trajectory- and token-level guidance each improve over DAPO in both categories. Combining them gives the highest VR and VC accuracy.

C Pilot Study Details

Setting. All three pilots use questions from the ViRL-39K training set and evaluate Qwen2.5-VL, Qwen3-VL, and Qwen3.5. Within each pilot, the compared methods use the same questions, prompts, decoding settings, and sampling seeds. We sample eight rollouts or continuations per condition and compute confidence intervals over questions rather than individual responses. Selector comparisons use 20,000 paired question-level sign flips with Holm correction.

Baselines. We compare rollout-level scores in Pilot 1, visual-token selectors in Pilot 2, and response-token selectors in Pilot 3. The following descriptions specify the scoring or selection signals used in each pilot.

Pilot 1: Distinguishing rollout outcomes. Pilot 1 compares the VEP Similarity Score with the Success VEP Similarity, Answer Attention, Reasoning Attention, and Shuffled VEP. Prototypes are all constructed from a support set and evaluated on independently sampled rollouts.

- **VEP Similarity Score.** We construct the Success and Failure VEPs, $\boldsymbol{\mu}_Q^+$ and $\boldsymbol{\mu}_Q^-$. For an evaluation rollout r, we compute

$$s_r = \cos(\mathbf{v}_r, \boldsymbol{\mu}_Q^+) - \cos(\mathbf{v}_r, \boldsymbol{\mu}_Q^-),$$

  where $\mathbf{v}_r$ is its visual attribution. A larger score indicates greater similarity to the success prototype relative to the failure prototype.
- **Success VEP Similarity.** We use only similarity to the success prototype, $\cos(\mathbf{v}_r, \boldsymbol{\mu}_Q^+)$.
- **Answer Attention.** We sum the direct attention assigned to visual tokens at each final-answer prediction position and average over these positions:

$$s_r^{\text{ans}} = \frac{1}{|\mathcal{A}_r|} \sum_{a \in \mathcal{A}_r} \sum_{i \in I_Q} A_r(a, i),$$

  where $\mathcal{A}_r$ and I_Q denote final-answer token positions and visual positions, respectively. This score measures overall visual attention during final-answer prediction without outcome grouping.
- **Reasoning Attention.** We sum the direct attention assigned to visual tokens at each response prediction position and average over all response positions, including final-answer tokens:

$$s_r^{\text{attn}} = \frac{1}{|\mathcal{P}_r|} \sum_{t \in \mathcal{P}_r} \sum_{i \in I_Q} A_r(t, i),$$

  where $\mathcal{P}_r$ denotes all response token positions, including final-answer tokens, and I_Q denotes visual positions. This score measures overall visual attention across the full response without outcome grouping.
- **Shuffled VEP.** We permute the support outcomes before constructing the Success and Failure VEPs, then compute the same two-prototype similarity score.

Pilot 2: Selecting visual tokens. Pilot 2 compares seven visual-token selectors. Each selector chooses 10% of the visual states, which are replaced with the mean visual state of the same image. We then sample eight complete responses per question and report the accuracy decrease relative to unintervened generation. All selectors use the same intervention position and budget.

- **Success-Specific Contrastive VEP.** We rank visual tokens by $\mathbf{D}_Q^+ = \boldsymbol{\mu}_Q^+ - \boldsymbol{\mu}_Q^-$ and select the highest-scoring positions. This emphasizes visual evidence attributed more strongly to successful rollouts than to failed rollouts.
- **Success VEP.** We rank visual tokens directly by $\boldsymbol{\mu}_Q^+$, without subtracting the failure prototype. This retains visual evidence shared by successful and failed rollouts.
- **Saliency-R1.** We use the saliency extraction method of Saliency-R1 (Gong et al., 2026). It composes answer-to-reasoning and reasoning-to-visual attention weights to trace visual information through intermediate reasoning tokens. These weights are combined with

visual value vectors, attention output projections, and target-token unembedding weights to estimate visual contributions to answer logits. The contributions are summed across layers and attention heads, followed by ReLU to retain positive contributions. The per-answer-token maps are then summed, without additional column normalization of the attention-rollout matrices, to obtain the final saliency map for ranking visual tokens.

- **Answer Attention.** We rank visual tokens by the direct attention assigned to them when predicting final-answer tokens. Unlike visual attribution, this signal does not trace attention through intermediate reasoning tokens.
- **TAM.** Token Activation Map (Li et al., 2025) first projects visual hidden states onto the output classifier weight vector of each target text token and retains positive activations. It then estimates an interference map by aggregating the activation maps of preceding text tokens, weighted by their relevance to the target token, while excluding tokens identical to the target. This interference map is subtracted from the target-token map with a scale fitted by least squares. The resulting map is processed with a rank Gaussian filter and positive clipping to produce the final token-specific visual activation map. We use the resulting visual map to rank visual tokens.
- **Random.** We sample visual tokens uniformly without replacement, using the same 10% selection budget.
- **Failure-Specific Contrastive VEP.** We rank visual tokens by $\mathbf{D}_Q^- = \boldsymbol{\mu}_Q^- - \boldsymbol{\mu}_Q^+$ and select the highest-scoring positions. This emphasizes visual evidence attributed more strongly to failed rollouts than to successful rollouts.

Pilot 3: Selecting response tokens. Pilot 3 compares Success-Specific and Failure-Specific Contrastive VEP and Success and Failure VEP selectors with VPPO, VGPO, Reasoning Attention, Token Entropy, and Random Selection. All selectors operate on the same eligible response positions, excluding final-answer tokens and special tokens.

- **Success-Specific and Failure-Specific Contrastive VEP.** We use $\mathbf{D}_Q^+$ for correct rollouts and $\mathbf{D}_Q^-$ for incorrect rollouts to select visual states for replacement. Under the same response prefix, we measure the change in the log-probability of each observed response token:
$$\Delta_{r,t}(\mathbf{D}) = \log p_{r,t} - \log \widetilde{p}_{r,t}(\mathbf{D}),$$
where $p_{r,t}$ and $\widetilde{p}_{r,t}(\mathbf{D})$ are the probabilities before and after visual intervention, respectively. We select the response token with the largest positive score.
- **Success and Failure VEP.** We use $\boldsymbol{\mu}_Q^+$ for correct rollouts and $\boldsymbol{\mu}_Q^-$ for incorrect rollouts, replacing the contrastive prototypes while retaining the same intervention and response-token scoring procedure.
- **VPPO.** Following VPPO (Huang et al., 2025), we rank candidate response tokens by the change in their predictive distributions induced by random image masking, with the response prefix held fixed. This measures sensitivity to image perturbation without outcome-specific visual selection.
- **VGPO.** Following VGPO (Wang et al., 2026), we rank candidate response tokens using hidden-state similarity to an aggregated visual representation. This measures visual alignment in representation space.
- **Reasoning Attention.** We rank candidate response tokens by their direct attention to visual tokens, without using outcome prototypes or visual interventions for scoring.
- **Token Entropy.** We rank candidate positions by the entropy of the next-token predictive distribution under the original image and response prefix, selecting the position with the highest entropy. This measures predictive uncertainty rather than visual dependency.
- **Random.** We select one response-token position uniformly from the same eligible candidate set.

Every response-token selector is evaluated with the same retain-or-suppress continuation experiment. Starting from the original prefix, the retain condition forces the selected token for one decoding step, whereas the suppress condition forbids that token at the same step. For correct rollouts, we report $\mathrm{Acc}(\mathrm{retain}) - \mathrm{Acc}(\mathrm{suppress})$; for incorrect rollouts, we report $\mathrm{Acc}(\mathrm{suppress}) - \mathrm{Acc}(\mathrm{retain})$.

C.1 Success and Failure VEPs Distinguish Rollout Outcomes

For each question, Pilot 1 samples two sets of eight rollouts with independent seeds. One set constructs the Success and Failure VEPs; the other is scored by their relative cosine similarity. We then swap the two sets. We retain an assignment if its construction set has at least two correct and two incorrect rollouts and its evaluation set has both outcomes. Qwen2.5-VL contributes 750 questions and 9,456 evaluation responses. The same procedure applies to Qwen3-VL and Qwen3.5. Sampling uses temperature 1.0, top-$p = 1.0$, and a 4096-token response cap.

Evaluation outcomes enter only the AUROC calculation. For Reasoning Attention, we average total attention to visual tokens over all response positions. The 95% CIs use 10,000 question-level bootstrap samples.

Table 9: Pilot 1 AUROC with 95% confidence intervals.

Model	Signal	AUROC	95% CI
Qwen2.5-VL	Success–Failure prototype contrast	**0.6845**	[0.6253, 0.7401]
	Success VEP only	0.6533	[0.5968, 0.7086]
	Answer Attention	0.5209	[0.4566, 0.5828]
	Reasoning Attention	0.5931	[0.5356, 0.6481]
	Shuffled support outcomes	0.4957	[0.4883, 0.5028]
Qwen3-VL	Success–Failure prototype contrast	**0.6917**	[0.6371, 0.7437]
	Success VEP only	0.6316	[0.5782, 0.6835]
	Answer Attention	0.5466	[0.4875, 0.6040]
	Reasoning Attention	0.4854	[0.4298, 0.5397]
	Shuffled support outcomes	0.4955	[0.4871, 0.5033]
Qwen3.5	Success–Failure prototype contrast	**0.7618**	[0.7157, 0.8044]
	Success VEP only	0.7023	[0.6567, 0.7461]
	Answer Attention	0.5648	[0.5135, 0.6127]
	Reasoning Attention	0.5333	[0.4891, 0.5783]
	Shuffled support outcomes	0.4988	[0.4911, 0.5065]

The two-prototype score gives the highest AUROC on each model, while shuffled outcomes remain near chance.

C.2 Contrastive VEPs Identify Supporting Visual Tokens

Pilot 2 uses 950, 880, and 560 questions for Qwen2.5-VL, Qwen3-VL, and Qwen3.5, respectively. Each selector ranks visual tokens, and we replace the top 10% of their hidden states with the mean visual state of the same image before the first decoder layer. For Answer Attention, the ranking uses direct attention from final-answer predictors to visual tokens. We then generate eight complete responses per question at temperature 1.0 and top-$p = 1.0$. The response cap is 4096 tokens for Qwen2.5-VL and 2048 for Qwen3-VL and Qwen3.5.

Accuracy drop is measured against clean generation. We average the eight response accuracies for each question and use the same aggregation for every selector. The 95% CIs use 20,000 question-level, benchmark-stratified bootstrap samples.

Table 10: Pilot 2 accuracy drop in percentage points. † marks a paired 95% CI above zero.

Selector	Qwen2.5-VL	Qwen3-VL	Qwen3.5
$\mathbf{D}_Q^+$	**28.45**	**20.61**	**20.15**
95% CI	[22.34, 34.54]	[14.02, 27.68]	[10.57, 29.08]
$\boldsymbol{\mu}_Q^+$	26.71	17.71	11.93
Saliency-R1	18.81	9.13	8.93
Answer Attention	17.06	-0.42	11.96
TAM	5.60	-0.91	1.49
Random	3.09	0.57	2.08
95% CI	[-1.21, 7.34]	[-4.17, 5.02]	[-2.86, 6.99]
$\mathbf{D}_Q^-$	0.77	-3.20	-4.17
Extra drop vs. Random	25.36†	20.04†	18.07†
95% CI	[18.72, 32.07]	[12.83, 27.21]	[8.27, 27.44]

$\mathbf{D}_Q^+$ produces the largest drop on all three models. Its extra drop over Random ranges from 18.07 to 25.36 points.

C.3 Contrastive VEPs Locate Supporting and Harmful Response Tokens

Pilot 3 uses groups of eight with at least three correct and three incorrect rollouts, selecting one of each for intervention. The VEPs are computed from the full group. After replacing the $\mathbf{D}_Q^+$- or $\mathbf{D}_Q^-$-selected visual states as in Pilot 2, we rank non-special response tokens with nonempty decoded text by $\Delta_{r,t}$ in Eq. (5). We test all eligible positions and then exclude final-answer tokens.

We retain the token with the largest positive score at its original prefix or suppress it for one step. Both conditions sample eight continuations with the same question-level seed, temperature 1.0, top-$p = 1.0$, and a 4096-token cap. The effect is $\mathrm{Acc}(\mathrm{retain}) - \mathrm{Acc}(\mathrm{suppress})$ for correct rollouts and the reverse for incorrect rollouts. The 95% CIs use 20,000 question-level, benchmark-stratified bootstrap samples.

Table 11: Pilot 3 intervention effects in accuracy points, with 95% CIs.

			Correct rollouts		Incorrect rollouts	
Model / checkpoint	Questions	Response tokens	Effect	95% CI	Effect	95% CI
Qwen2.5-VL / PAPO	3,170	All	+16.17	[12.07, 20.43]	+12.85	[8.95, 16.88]
		Excl. final answer	+8.71	[5.60, 11.95]	+4.10	[1.46, 6.86]
Qwen3-VL / VPPO-8B	1,350	All	+5.65	[1.48, 10.00]	+3.61	[0.47, 7.78]
		Excl. final answer	+5.65	[1.39, 10.00]	+3.61	[0.59, 7.87]
Qwen3.5-9B / official	1,730	All	+11.56	[7.37, 15.97]	+11.13	[6.94, 15.39]
		Excl. final answer	+10.04	[6.07, 14.16]	+11.27	[7.08, 15.53]

After excluding final-answer tokens, all six effects have positive 95% CIs.

D THEORETICAL ANALYSIS

D.1 CONTRASTIVE PROTOTYPE RECOVERY

Attribution model. Fix an image–question pair Q and the sampling policy, and omit Q from the notation. For outcome $y \in \{+, -\}$, assume the normalized attribution of rollout r admits the decomposition

$$\mathbf{v}_r = \mathbf{b} + \mathbf{u}^y + \boldsymbol{\epsilon}_r, \qquad \|\mathbf{b}\|_1 = 1 - \alpha, \qquad \|\mathbf{u}^y\|_1 = \alpha, \tag{10}$$

where $0 < \alpha < 1$, $\mathbf{b}, \mathbf{u}^+, \mathbf{u}^-$ are nonnegative, and the supports of $\mathbf{u}^+$ and $\mathbf{u}^-$ are disjoint. The common component $\mathbf{b}$ can overlap either support. The noise has conditional mean zero and zero coordinate sum, so $\mathbf{v}_r$ remains in the probability simplex. We apply this model after the normalization in Eq. (1). The components $\mathbf{b}, \mathbf{u}^+, \mathbf{u}^-$ belong to the assumed data-generating model and are not defined from $\mathbf{D}^{\pm}$.

Let $n_y > 0$ be the number of rollouts with outcome y, and assume independent samples within each outcome. Each noise coordinate is sub-Gaussian with parameter σ_y^2, meaning $\mathbb{E}[\exp(t\epsilon_{r,j}) \mid y] \le \exp(t^2\sigma_y^2/2)$ for all real t. For $d = |\mathcal{I}_Q|$ visual positions and $\delta \in (0, 1)$, define

$$e_y = \sigma_y \sqrt{\frac{2\log(4d/\delta)}{n_y}}, \qquad e = e_+ + e_-. \tag{11}$$

The sample means satisfy $\boldsymbol{\mu}^y = \mathbf{b} + \mathbf{u}^y + \overline{\boldsymbol{\epsilon}}^{\,y}$. Concentration of each sample mean and a union bound over the two groups and d positions give

$$\Pr\big(\|\overline{\boldsymbol{\epsilon}}^{\,+}\|_\infty \le e_+,\ \|\overline{\boldsymbol{\epsilon}}^{\,-}\|_\infty \le e_-\big) \ge 1 - \delta. \tag{12}$$

When an independent rollout group is partitioned by outcome, this probability statement is conditional on the observed counts n_+ and n_-.

Proposition 1 (Contrastive prototype recovery)**.** *Under the attribution model above, let* $S^y = \mathrm{supp}(\mathbf{u}^y)$*, with* e *defined in Eq.* (11)*. With probability at least* $1 - \delta$*,*

$$\min_{j \in S^y} u_j^y > 2e \quad \Longrightarrow \quad \mathrm{Top}_{|S^y|}(\mathbf{D}^y) = S^y. \tag{13}$$

Proof of Proposition 1. We prove the positive case; the negative case follows by exchanging outcomes. Let $S^+ = \mathrm{supp}(\mathbf{u}^+)$, $m = |S^+|$, and $u_{\min} = \min_{j \in S^+} u_j^+ > 2e$. On the event in Eq. (12), subtracting the prototypes eliminates the common component:

$$\boldsymbol{\mu}^+ - \boldsymbol{\mu}^- = \mathbf{u}^+ - \mathbf{u}^- + \overline{\boldsymbol{\epsilon}}^{\,+} - \overline{\boldsymbol{\epsilon}}^{\,-}. \tag{14}$$

For $j \in S^+$, disjoint support gives $u_j^- = 0$, so $D_j^+ \ge u_{\min} - e > e$. For $j \notin S^+$, the population difference is nonpositive, so $D_j^+ \le e$. Every target position therefore ranks above every other position, establishing exact Top-m recovery with probability at least $1 - \delta$. The comparison does not depend on the shape or peak value of $\mathbf{b}$.

The raw Success VEP can instead rank only common positions. To see this, suppose $d \ge 3m$, take three disjoint sets B, S^+, S^-, each of size m, and choose

$$b_j = \frac{1 - \alpha}{m}\mathbf{1}[j \in B], \qquad u_j^{\pm} = \frac{\alpha}{m}\mathbf{1}[j \in S^{\pm}], \qquad 0 < \alpha < \frac{1}{2}. \tag{15}$$

Without noise, every common position has raw attribution $(1 - \alpha)/m > \alpha/m$. Thus, the Top-$m$ positions of $\boldsymbol{\mu}^+$ are exactly B, whereas those of $\mathbf{D}^+$ are exactly S^+. The two rankings remain different on the finite-sample event whenever

$$\frac{\alpha}{m} > 2(e_+ + e_-), \qquad \frac{1 - 2\alpha}{m} > 2e_+. \tag{16}$$

For any fixed $\alpha \in (0, 1/2)$, both inequalities hold for sufficiently small nonzero noise scales. The counterexample therefore does not rely on a noiseless limit. □

Interpretation. The proposition guarantees recovery of the outcome component in Eq. (10); it does not identify annotated or causal evidence. Common attention may still be useful for reasoning, but its magnitude need not identify the contrastive target. Subtraction also carries estimation error from both outcome groups, so recovery depends on both the remaining signal and within-outcome variation. Here m counts visual positions.

D.2 Local Policy Improvement

Update model. Let $J(\theta)$ be expected return and fix the current parameter θ_0. Decompose a sampled policy-gradient update direction into selected and remaining response-token contributions:

$$H = H_{\mathcal{K}} + H_{\mathcal{N}}, \qquad H_w = aH_{\mathcal{K}} + bH_{\mathcal{N}}. \tag{17}$$

The contributions include the signed advantages and the original averaging factors. We hold the response and selection sizes fixed, so $q = |\mathcal{K}_i|/|\mathcal{U}_i|$ is constant. The weights $a = 1 + \lambda_{\text{tok}}(1-q)$ and $b = 1 - \lambda_{\text{tok}}q$ are exactly those in Eq. (7). We assume $0 < q < 1$ and $0 < \lambda_{\text{tok}} < 1/q$, ensuring $0 < b < a$. Both directions use the same trajectory-level guidance. We compare the local policy-gradient component before clipping and omit optimizer preconditioning and the full KL-to-reference dynamics.

Sparse ascent and gradient separation. Assume the selected contribution has a mean ascent direction $\mathbf{h}$, while the remaining contribution has zero mean:

$$\mathbb{E}H_{\mathcal{K}} = \mathbf{h}, \qquad \mathbb{E}H_{\mathcal{N}} = 0, \qquad \nabla J(\theta_0)^\top \mathbf{h} > 0. \tag{18}$$

Let $F \succeq 0$ be the Fisher matrix at θ_0, under a fixed old-policy state distribution. Assume

$$\mathbb{E}[H_{\mathcal{K}}^\top F H_{\mathcal{N}}] = 0, \qquad V_{\mathcal{K}} = \mathbb{E}[H_{\mathcal{K}}^\top F H_{\mathcal{K}}] > 0, \qquad V_{\mathcal{N}} = \mathbb{E}[H_{\mathcal{N}}^\top F H_{\mathcal{N}}] > 0, \tag{19}$$

with finite moments. VPPO uses a related gradient decomposition to motivate visual-token filtering (Huang et al., 2025). Here we use the split only to compare selected and remaining response-token contributions under the same quadratic KL budget.

Proposition 2 (Local policy improvement)**.** *Under the assumptions above, the leading-order return gains at the same expected quadratic KL budget satisfy*

$$\frac{\Delta J^{(1)}_{\text{VEP}}}{\Delta J^{(1)}_{\text{uniform}}} = \sqrt{\frac{V_{\mathcal{K}} + V_{\mathcal{N}}}{V_{\mathcal{K}} + (b/a)^2 V_{\mathcal{N}}}} > 1. \tag{20}$$

Proof of Proposition 2. For a parameter displacement u, averaging over the fixed state distribution d_{old} gives

$$\mathbb{E}_{s\sim d_{\text{old}}} D_{\text{KL}}(\pi_{\theta_0}(\cdot \mid s) \,\|\, \pi_{\theta_0+u}(\cdot \mid s)) = \tfrac{1}{2} u^\top F u + o(\|u\|^2). \tag{21}$$

Given a budget $\kappa > 0$, choose comparison step sizes

$$\eta_0 = \sqrt{\frac{2\kappa}{V_{\mathcal{K}} + V_{\mathcal{N}}}}, \qquad \eta_w = \sqrt{\frac{2\kappa}{a^2 V_{\mathcal{K}} + b^2 V_{\mathcal{N}}}}. \tag{22}$$

Equation (19) implies

$$\tfrac{1}{2}\mathbb{E}[(\eta_0 H)^\top F(\eta_0 H)] = \tfrac{1}{2}\mathbb{E}[(\eta_w H_w)^\top F(\eta_w H_w)] = \kappa. \tag{23}$$

The two directions therefore have the same expected quadratic KL cost. We use this normalization only for the theoretical comparison; VEP-RL does not add a learning-rate controller.

Writing $c = \nabla J(\theta_0)^\top \mathbf{h} > 0$, Eq. (18) gives

$$\Delta J^{(1)}_{\text{uniform}} = \eta_0 c, \qquad \Delta J^{(1)}_{\text{VEP}} = a\eta_w c. \tag{24}$$

Substituting Eq. (22) yields Eq. (20). The inequality is strict because $b/a < 1$ and $V_{\mathcal{N}} > 0$. □

Finite local updates. Suppose the gradient of J is L-Lipschitz along all realized update segments, and both directions have finite Euclidean second moments. Define $M_0 = \mathbb{E}\|H\|^2/(V_{\mathcal{K}} + V_{\mathcal{N}})$ and $M_w = \mathbb{E}\|H_w\|^2/(a^2 V_{\mathcal{K}} + b^2 V_{\mathcal{N}})$. Taylor bounds for the two updates give

$$\mathbb{E}[J(\theta_0 + \eta_w H_w) - J(\theta_0 + \eta_0 H)] \geq C\sqrt{\kappa} - L\kappa(M_0 + M_w), \tag{25}$$

where

$$C = \sqrt{2}\, c \left[\frac{a}{\sqrt{a^2 V_{\mathcal{K}} + b^2 V_{\mathcal{N}}}} - \frac{1}{\sqrt{V_{\mathcal{K}} + V_{\mathcal{N}}}}\right] > 0. \tag{26}$$

The bound is positive for sufficiently small κ. The comparison matches quadratic KL costs, not exact finite-step KL divergences.

Scope. The sparse-ascent condition in Eq. (18) is not implied by Proposition 1. Spatial recovery and response-token selection are different steps. Pilot 3 tests response-token selection through answer accuracy. If the selector misses a contribution with nonzero mean, or if cross terms provide important noise cancellation, the improvement need not hold. The proposition gives a local comparison under the stated assumptions; it does not establish convergence of the full shared-parameter training process.

E Additional Analysis

E.1 Error Analysis

Error analysis. To understand when VEP-RL fails, we analyze 3,000 sampled responses from Qwen3.5-9B, Vision-OPD, and VEP-RL-9B on evaluation benchmarks and classify each incorrect response. We observe three failure patterns as shown in Figure 8: (i) Visual Perception Errors, where the model misperceives visual details; (ii) Language Dominance Error, where correctly perceived visual evidence is subjected to language priors; and (iii) Logic Reasoning Errors, where the model correctly extracts the visual information but reasons incorrectly. We first used GPT-5.6 Sol to assign a preliminary category to each incorrect response based on its image, question, ground-truth answer, and model rollout, and then manually reviewed the assignments to determine the final labels. The error breakdown indicates that VEP-RL first improves visual perception compared with Vision-OPD and the base model. Second, VEP-RL reduces language-dominance errors, potentially because VEP-RL explicitly reinforces visually supported tokens.

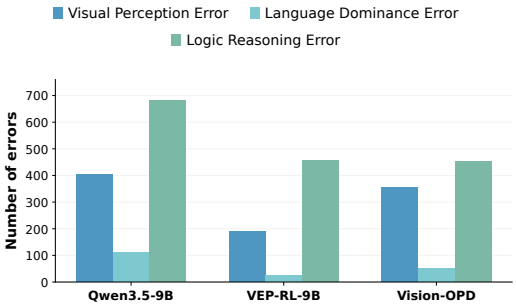


Figure 8: Error breakdown by category.

E.2 Selected-token Frequencies

We showcase what VEP selects in Pilot 3.3. Figure 9 combines the tokens selected by $\mathbf{D}_Q^+$ and $\mathbf{D}_Q^-$ across the three model-specific sets. We remove leading and trailing whitespace. We also omit whitespace-only strings, Unicode replacement and control characters, tag or LaTeX command fragments identified in context, and punctuation-only strings except %.

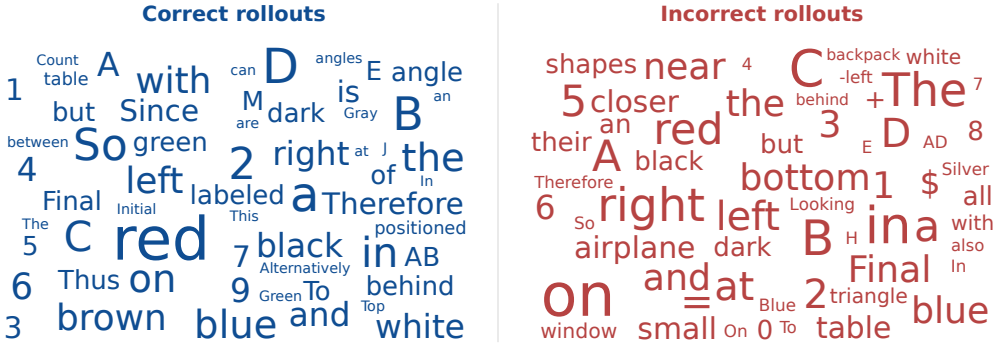


Figure 9: Word clouds of the Top-1 response tokens selected by contrastive VEPs. Text size indicates frequency within each panel, not intervention effect, and is not comparable across panels.

E.3 CASE STUDY



Question: What kind of animal is on the blue sail?
(A) spider (B) dog
(C) fish (D) bird

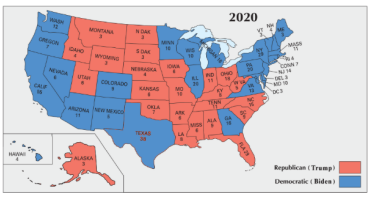
Ground Truth : A Model Answer : D

Model Rollout :

```
Looking at the blue sail in the image
(second from the left), there is a small
black silhouette near the top of the
sail. The shape resembles a bird — likely
a seagull or similar coastal bird — which
is commonly used as a logo or emblem on
sailing vessels, especially those
associated with marine environments or
racing clubs.

None of the other options (spider, dog,
fish) match the visual shape or context.
Birds are frequently depicted on sails
due to their association with flight and
the sea.

\boxed{D}
```



Question: Based on the map, did the Democratic Party win Texas in the 2020 elections? For scoring, encode Yes as 1 and No as 0 in the final answer.

Ground Truth : 1 Model Answer : 0

Model Rollout :

```
Looking at the map, Texas is colored
red(orange-red), which corresponds to the
Republican (Trump) candidate according to
the legend.

The number 38 in Texas indicates the
number of electoral votes won by Trump in
that state. Therefore, the Democratic
Party did not win Texas in the 2020
election.

\boxed{0}
```



Question: What is the color of the telephone?
(A) green. (B) white
(C) orange. (D) blue

Ground Truth : C Model Answer : B

Model Rollout :

```
Looking at the image, there is a telephone
booth or payphone visible on the right
side of the street, near the stone wall
and tree.

The structure appears to be white or off-
white in color, consistent with many
public telephones found in Israel (which
this scene resembles due to Hebrew signage
and license plates). There's no clear
green, orange, or blue telephone visible.

The most plausible answer based on visual
evidence is **white**.

\boxed{B}
```



Question: According to the image, is Key West the southernmost point of Florida?For scoring, encode Yes as 1 and No as 0 in the final answer.

Ground Truth : 0 Model Answer : 1

Model Rollout :

```
Based on the map provided, Key West is
depicted at the very southern tip of
Florida, with no other labeled locations
or illustrations extending further
south.

This aligns with geographical fact — Key
West is indeed the southernmost point of
the contiguous United States and of
Florida.

Therefore, according to the image, yes,
Key West is shown as the southernmost
point.

\boxed{1}
```

Figure 10: **Visual Perception Errors and Language Dominance Errors. Top left:** The model mistakes the small spider emblem on the blue sail for a bird. **Top right:** It reads the blue fill of Texas as red and consequently identifies the wrong winning party. **Bottom left:** An assumption about the typical appearance of public telephones overrides the target object's visual evidence. **Bottom right:** The model relies on its knowledge of Key West rather than the spatial arrangement depicted in the illustration.



Question: What is the direction of this push?
A. toward his finger
B. away from his finger

Ground Truth : B Model Answer : A

Model Rollout :

The child is pressing the elevator button with his finger. When you press a button, you apply force in the direction of the button's surface — that is, **toward** the button.
...
In other words, the force applied by the finger is directed into the button, so from the perspective of the finger, the push is toward the object being pressed — which is aligned with the direction of the finger's motion. Thus, the correct answer is:
\boxed{A}



Question: What is the color of the broom?
(A) black (B) yellow
(C) gray (D) red

Ground Truth : A Model Answer : D

Model Rollout :

Looking at the image, there is a broom visible on the right side, partially hidden under the tree branches. The broom has a red handle and a gray or black bristle head. However, the most prominent color associated with the broom's structure — particularly its handle — is red.

The question asks for "the color of the broom," which typically refers to its main or most noticeable part. In this case, the red handle stands out.

Thus, the correct answer is:

\boxed{D}



Question: What is the color of the man's helmet?
(A) white (B) red
(C) yellow (D) blue

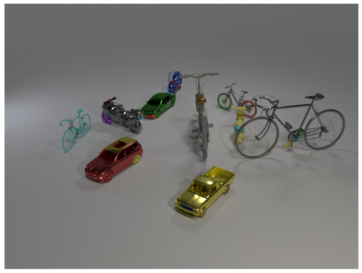
Ground Truth : C Model Answer : C

Model Rollout :

Looking at the image, in the background near the center-right, there is a construction worker visible. He is wearing a hard helmet, and its color is clearly yellow — a standard safety color for construction helmets.

Thus, the correct answer is:

\boxed{C}



Question: How many different items are there in the image?

Ground Truth : 10 Model Answer : 10

Model Rollout :

Let's count the distinct items in the image:

1、Cyan road bike (left) 2、Red SUV 3、Silver motorcycle 4、Green sedan 5、Blue chopper-style motorcycle (back) 6、Gray bicycle (center, with basket) 7、Yellow pickup truck (front) 8、Yellow scooter or small vehicle (near gray bike) 9、Silver mountain bike (back right) 10、Large gray road bike (right)

That's 10 distinct items.

\boxed{10}

Figure 11: **Logic Reasoning Errors and Successful Cases. Top left:** The model recognizes that the finger pushes into the elevator button but reverses the direction relative to the finger. **Top right:** It identifies the broom's red handle and dark bristles, yet incorrectly selects the handle as the broom's defining color. **Bottom left:** The model locates the construction worker and correctly identifies the yellow helmet. **Bottom right:** It enumerates the objects in the scene and correctly counts ten distinct items.

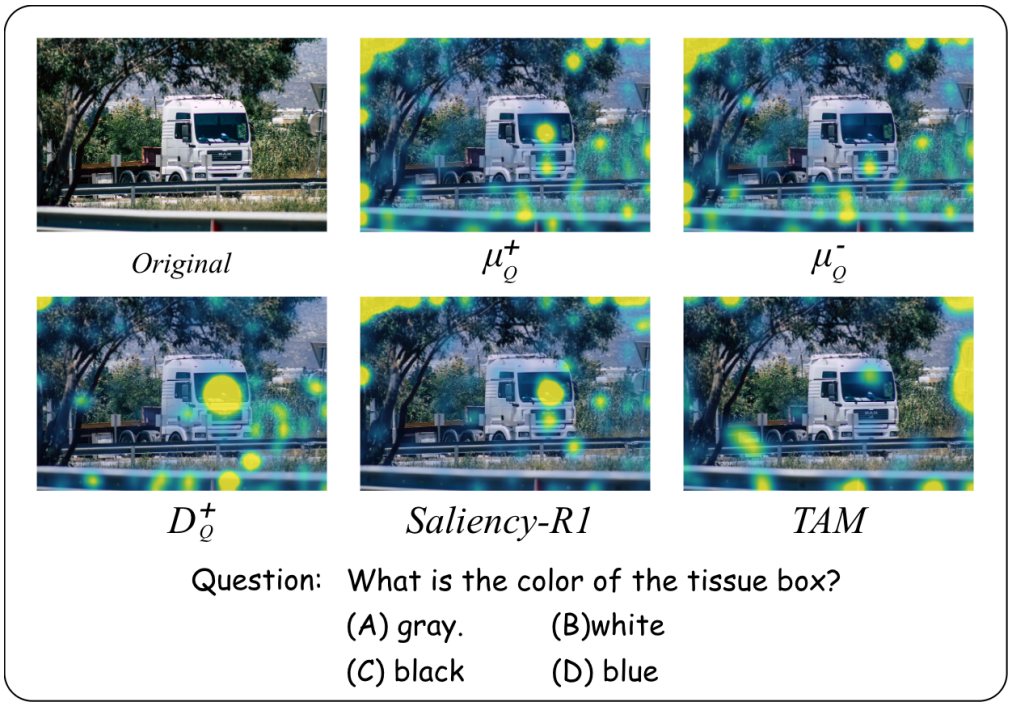


Figure 12: The tissue box is a small object inside the truck windshield. Both our contrastive map D_Q^+ and Saliency-R1 highlight the windshield region, while TAM's most prominent responses appear elsewhere in the scene.

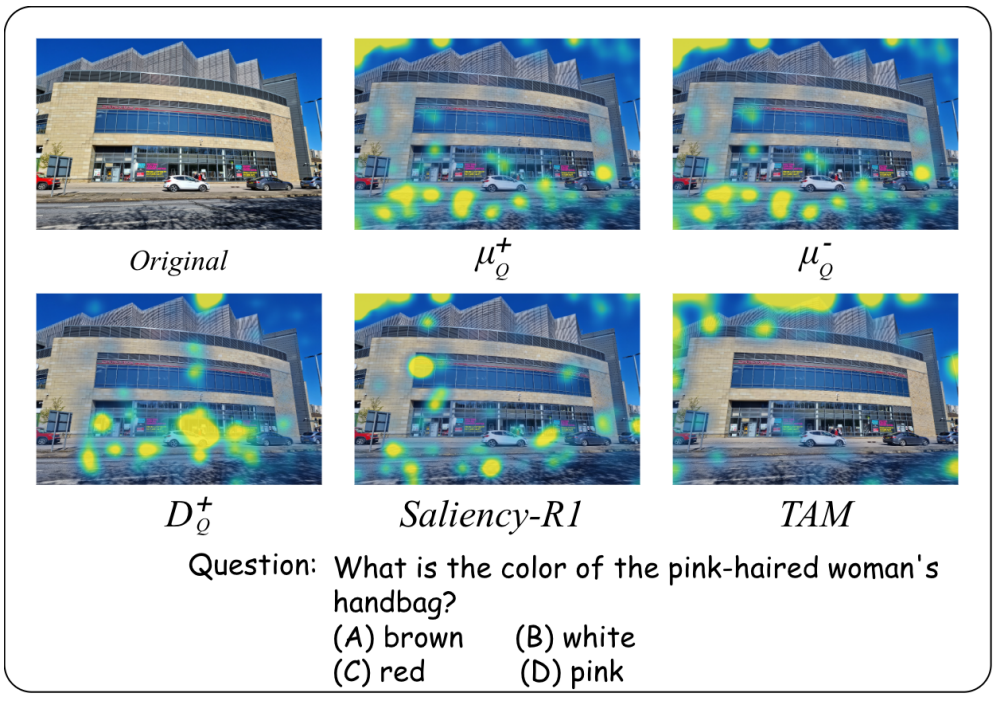


Figure 13: The pink-haired woman is located near the white car in front of the building. Only our contrastive map D_Q^+ clearly highlights her location; Saliency-R1 and TAM focus mainly on other regions.