EYES IN THE UPDATE? LEARNING WHERE REWARDS LAND IN MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal post-training has moved from broad instruction tuning toward reinforcement learning for image and video reasoning, where outcome rewards can improve final-answer accuracy. Yet these rewards are assigned to complete generated trajectories, while visual evidence is used only at sparse and uneven positions inside the response. A growing line of work has therefore begun to make multimodal RL more token-sensitive through grounding signals or external token-level weighting. We ask a more specific question: **can the allocation of the same trajectory-level reward be inferred from the model's own multimodal computation, rather than from an auxiliary annotator, perception module, or process label?** We propose GAZE-RL, a representation-driven policy optimization method that redistributes completion-level advantage through a latent reward landing field. The field is computed from internal cross-modal geometry and local predictive uncertainty, assigning larger update weight to positions that are both visually anchored and decision-bearing while remaining mean-one, so the total credit of each rollout is preserved and only its positionwise allocation changes. GAZE-RL requires no process supervision, auxiliary perception module, or architectural modification. Trained on a mixed image-video RL dataset, GAZE-RL improves Qwen2.5-VL and InternVL3.5 across broad image and video benchmarks. It also yields stronger evidence sensitivity and shorter, more concentrated reasoning dynamics. These results suggest that multimodal RL benefits from learning how existing rewards should land inside the reasoning trajectory. Our code is open-sourced via https://anonymous.4open.science/r/anonymous-gaze-rl-2ECE/.

1 INTRODUCTION

Multimodal large language models (MLLMs) now combine capable image and video perception with post-training for visual reasoning. Stronger backbones, instruction-tuned assistants, and unified image-video models provide the perceptual substrate (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023; Bai et al., 2023; Li et al., 2024a; Tong et al., 2024a; Bai et al., 2025a; Zhu et al., 2025; Wang et al., 2025b); reinforcement learning and preference optimization increasingly elicit long-form reasoning (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Ahmadian et al., 2024; Shao et al., 2024; Guo et al., 2025; Liu et al., 2025c; Huang et al., 2026b; Meng et al., 2025; Yang et al., 2025b; Zhang et al., 2025a; Deng et al., 2025; Shen et al., 2025). Reasoning quality therefore depends on how post-training turns visual representations into evidence-based decisions along a response.

Yet aggregate accuracy does not always imply faithful visual reasoning. Models remain brittle when shortcut cues are suppressed or tasks require expert visual knowledge, diagrams, spatial reasoning, or temporal aggregation (Yu et al., 2024; Yue et al., 2024; Li et al., 2024b; Guan et al., 2024; Chen et al., 2024; Wang et al., 2025a; 2024; Zhang et al., 2025b; Bi et al., 2025; Xu et al., 2025; Gong et al., 2025; Hu et al., 2025; Lymperaiou et al., 2026; Zhang et al., 2026; Huti et al., 2026; Zhao et al., 2026; Li et al., 2026; Tian et al., 2026). In videos, decisive evidence may occupy only a few frames or events (Maaz et al., 2024; Li et al., 2023b; Mangalam et al., 2023; Li et al., 2024c; Feng et al., 2025; Tang et al., 2025; Qiao et al., 2025; Li et al., 2025; Zeng et al., 2026). A fluent, detailed response can thus contain only a few positions that actually connect visual evidence to the final decision.

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

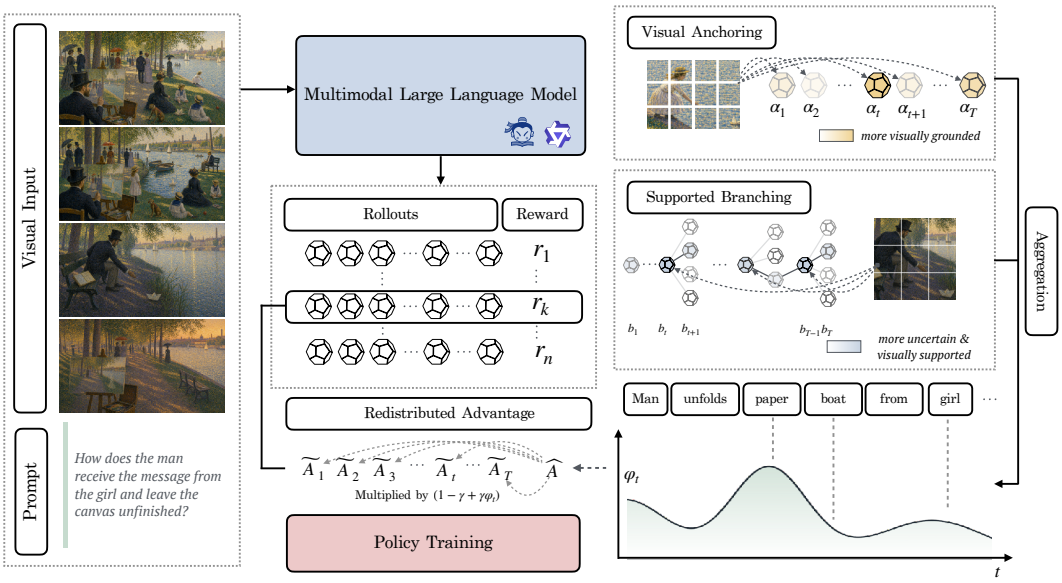

Figure 1: **Overview of GAZE-RL for multimodal policy training.** Two rollout-side signals, *visual anchoring* α_t and *supported branching* b_t (high predictive entropy with visually anchored predecessors), are aggregated into a reward landing field φ_t that redistributes the completion-level advantage $\widehat{A}$ into $\widetilde{A}_t = \big((1-\gamma)+\gamma\varphi_t\big)\widehat{A}$.

Multimodal RL has addressed this problem through curricula, augmentation, cold starts, grounding-aware rewards, and perception-sensitive objectives (Chen et al., 2025; Liu et al., 2025a; Fan et al., 2025; Bai et al., 2025b; Yao et al., 2025; Fan et al., 2026; Wei et al., 2025; Sarch et al., 2025; Xiao et al., 2025; Wang et al., 2025c; Huang et al., 2026a; Ding et al., 2025; Marsili & Gkioxari, 2025; Ou et al., 2026; Jiao et al., 2026). Related token-sensitive methods shape per-token rewards or reweight policy updates using perception, attention, or process scores (Yu et al., 2025a; Xiao et al., 2025; Fan et al., 2025; Sarch et al., 2025; Feng et al., 2026; Huang et al., 2026a; Wang et al., 2025c). These methods recognize that generated positions contribute unequally, but often rely on signals external to the policy's reasoning computation. We ask whether the model's internal multimodal states can instead locate token-level credit while preserving the original trajectory-level reward.

We formulate this as *reward landing*: preserve a rollout's total advantage while assigning more update weight to positions that connect visual evidence to consequential decisions. Such positions should be visually anchored in the model's hidden states and decision-bearing, with locally uncertain continuations supported by visually grounded predecessors. GAZE-RL combines internal cross-modal geometry and predictive uncertainty into a mean-one landing field. It requires no process supervision, auxiliary perception module, or architectural change. Across image and video benchmarks, it improves reasoning-heavy performance and visual evidence sensitivity while concentrating response-side credit. The central question is therefore not only how much reward a trajectory receives, but where that reward lands.

Contributions are as follows:

Takeaways

- We formulate reward landing as allocating fixed trajectory-level credit across response positions that use visual evidence unequally.
- We derive a mean-one, token-level landing field from internal cross-modal geometry and predictive uncertainty, without external labels, modules, or architectural changes.
- GAZE-RL improves Qwen2.5-VL and InternVL3.5 on image and video benchmarks and strengthens evidence sensitivity and response-side credit concentration.

2 Related Work

Post-training multimodal LLMs (MLLMs) with reinforcement learning. RL has progressed from preference alignment to verifier-rewarded reasoning in LLMs (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025). MLLM methods apply outcome-reward algorithms (Shao et al., 2024; Yu et al., 2025b) to visual backbones (Huang et al., 2026b; Liu et al., 2025c; Meng et al., 2025; Zhang et al., 2025a; Feng et al., 2025), often with vision-specific data, perception rewards, or grounding-aware sampling (Yu et al., 2025a; Xiao et al., 2025; Fan et al., 2025; Sarch et al., 2025; Bai et al., 2025b; Yang et al., 2025b; Liu et al., 2025a; Chen et al., 2025). These approaches improve training but generally broadcast one outcome-level advantage across all response positions. GAZE-RL instead redistributes that advantage within each rollout while retaining the verifier, data, and architecture.

Improving visual capability in MLLMs. Stronger open backbones (Bai et al., 2025a; Li et al., 2024a; Wang et al., 2025b) and diagnostic benchmarks probe visual indispensability, professional and logical reasoning, and long-form or temporal understanding (Fu et al., 2025a; Liu et al., 2025b; Tong et al., 2024b; Chen et al., 2024; Singh et al., 2019; Mathew et al., 2021; Masry et al., 2022; Lu et al., 2024; Xiao et al., 2024; Xu et al., 2025; Bi et al., 2025; Xiao et al., 2021; Fu et al., 2025b; Liu et al., 2024; Wu et al., 2024; Yang et al., 2025a; Hu et al., 2025). Yet analyses find that RL-trained MLLMs often use visual evidence sparsely along long responses, with aggregate accuracy masking perception-critical failures (Tang et al., 2025; Ding et al., 2025; Just et al., 2025; Marsili & Gkioxari, 2025; Huti et al., 2026; Zhang et al., 2026; Zhao et al., 2026). Rather than change the backbone or data, GAZE-RL targets the update positions implicated by these diagnostics.

Token-level credit assignment in policy updates. Multimodal RL can add perception or grounding signals to per-token rewards (Yu et al., 2025a; Xiao et al., 2025; Fan et al., 2025; Sarch et al., 2025; Feng et al., 2026) or reweight token-level updates using external perception or attention scores (Huang et al., 2026a; Wang et al., 2025c). GAZE-RL also reweights positions, but derives weights from the policy's internal cross-modal geometry and predictive uncertainty. Its mean-one field changes where rollout credit lands without changing its total.

3 Preliminaries

Multimodal autoregressive policy. The policy π_θ conditions on visual embeddings $\mathcal{V} = (v_1, \ldots, v_M)$ and a prompt $\mathcal{P}$, then generates $\mathcal{Y} = (y_1, \ldots, y_T)$. With state $s_t = (\mathcal{V}, \mathcal{P}, \mathcal{Y}_{<t})$ and action $a_t = y_t$, its log-likelihood factorizes as $\log \pi_\theta(\mathcal{Y} \mid \mathcal{V}, \mathcal{P}) = \sum_{t=1}^{T} \log \pi_\theta(y_t \mid s_t)$. A verifier assigns an outcome reward $r = R(\mathcal{V}, \mathcal{P}, \mathcal{Y})$. After group normalization, advantage-based methods (Schulman et al., 2017; Shao et al., 2024; Yu et al., 2025b; Huang et al., 2026a; Ahmadian et al., 2024) yield a completion-level advantage $\widehat{A}$. Thus, the per-rollout gradient contribution $\widehat{A} \nabla_\theta \log \pi_\theta(\mathcal{Y} \mid \mathcal{V}, \mathcal{P})$ applies the same scalar to every token score $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$.

GRPO. GRPO (Shao et al., 2024) estimates advantages without a learned value function. For each input $(\mathcal{V}, \mathcal{P})$, it samples G responses $\{\mathcal{Y}_j\}_{j=1}^{G}$ from $\pi_{\theta_{\text{old}}}$, collects rewards $r_j = R(\mathcal{V}, \mathcal{P}, \mathcal{Y}_j)$, and computes

$$\widehat{A}_j = \frac{r_j - \text{mean}_{g \in [G]}\, r_g}{\text{std}_{g \in [G]}\, r_g + \delta}, \tag{1}$$

where $\delta > 0$ stabilizes normalization. GAZE-RL uses GRPO, although landing-field reweighting applies to any method with a similar per-rollout advantage. For token y_t, the probability ratio is $\rho_t(\theta) = \pi_\theta(y_t \mid s_t)/\pi_{\theta_{\text{old}}}(y_t \mid s_t)$, and the reference-policy penalty is $\text{KL}_t(\theta) = D_{\text{KL}}(\pi_\theta(\cdot \mid s_t) \| \pi_{\text{ref}}(\cdot \mid s_t))$, where π_{ref} is fixed. The clipped surrogate is $\psi_\epsilon(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}$, with $\epsilon \in (0, 1)$. The GRPO objective is

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\text{rollout}} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\psi_\epsilon(\rho_t(\theta), \widehat{A}) - \beta\, \text{KL}_t(\theta) \right) \right], \tag{2}$$

with $\beta \geq 0$ controlling reference-policy regularization. The method below changes how $\widehat{A}$ is distributed across token-level update terms.

4 METHOD

GRPO (Eq. (2)) weights every response position by the same completion-level advantage $\widehat{A}$. Yet visual evidence and consequential reasoning decisions are not used uniformly along a rollout: some positions ground their next-token choice in $\mathcal{V}$, others drive a branching decision, and many do neither. GAZE-RL therefore replaces this uniform broadcast with a token-level redistribution — a reward landing field that converts $\widehat{A}$ into a landed advantage $\widetilde{A}_t$ before policy improvement.

Definition 4.1 (Reward Landing Field): *Given a rollout $\mathcal{Y}$ of length T with completion-level advantage $\widehat{A}$, a* reward landing field *is a function $\Phi : [T] \to \mathbb{R}_{\geq 0}$ with*

$$\Phi(t) \geq 0, \qquad \frac{1}{T}\sum_{t=1}^{T} \Phi(t) = 1, \qquad \widetilde{A}_t = \Phi(t)\,\widehat{A}. \tag{3}$$

The mean-one constraint preserves the rollout's total scalar advantage, $\sum_t \widetilde{A}_t = T\widehat{A}$, but does not imply identical clipped policy gradients; the uniform field $\Phi \equiv 1$ recovers GRPO. The remainder of this section constructs a rollout-dependent Φ from two signals — visual anchoring (Section 4.1) and visually supported branching (Section 4.2) — and plugs it into the policy update through a landed objective (Section 4.3).

4.1 VISUAL ANCHORING

Visual anchoring provides the first signal for position-wise weighting. In standard advantage-based policy training, the same $\widehat{A}$ is applied to every token-level update term as in Eq. (2), even though visual evidence is usually used unevenly across the generated response. We therefore assign each response position t a rollout-local anchoring score that measures the similarity between its pre-action hidden state and the hidden states associated with the visual embeddings $(v_1, \ldots, v_M)$. This score is not an additional reward and does not require process supervision; it is used only to distribute the already computed advantage $\widehat{A}$ across response tokens.

Let $\mathcal{L}$ denote the internal transformer layers of the multimodal policy π_θ used for multimodal reasoning. For layer $\ell \in \mathcal{L}$, let $h^{\text{txt}}_{\ell,t} \in \mathbb{R}^d$ be the response-side state aligned to s_t before predicting y_t, and let $h^{\text{vis}}_{\ell,m} \in \mathbb{R}^d$ be the hidden state associated with visual embedding v_m. For each response position, we score how close this state is to the visual hidden states, averaged over the selected layers:

$$\alpha_t = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \left\langle \frac{h^{\text{txt}}_{\ell,t}}{\|h^{\text{txt}}_{\ell,t}\|_2}, \frac{1}{M}\sum_{m=1}^{M} \frac{h^{\text{vis}}_{\ell,m}}{\|h^{\text{vis}}_{\ell,m}\|_2} \right\rangle, \qquad \widehat{\alpha}_t = \text{Norm}_{1:T}(\alpha_t). \tag{4}$$

Here $\text{Norm}_{1:T}$ denotes rollout-wise min-max scaling over response-token positions. Thus, $\widehat{\alpha}_t \in [0, 1]$ is the visual anchoring signal passed to the landing-field aggregator, with larger values indicating stronger visual anchoring relative to the rest of the same rollout. If a rollout has little reliable visual contrast, the score range is nearly flat and the resulting allocation remains close to uniform. We select the layers on a held-out visual-dependency diagnostic without benchmark labels and fix them across tasks; for Qwen2.5-VL-7B these are layers 14, 18, 22, and 26 (Section D.3).

4.2 VISUALLY SUPPORTED BRANCHING

Visually supported branching provides the second signal for position-wise weighting. Visual anchoring marks positions that draw on the visual context, but it does not distinguish those that drive consequential reasoning decisions: an anchored position with a near-deterministic next-token distribution is unlikely to be where policy improvement should land. We therefore assign each response token y_t a rollout-local supported-branching score that modulates the local predictive entropy of $\pi_{\theta_{\text{old}}}$ by a coefficient measuring how strongly preceding visually anchored states support the current generation step. Like visual anchoring, this score is not an additional reward and does not require process supervision; it is used only to distribute the already computed advantage $\widehat{A}$ across response tokens. Let $H_t = \text{Ent}(\pi_{\theta_{\text{old}}}(\cdot \mid s_t))$ be the rollout-time predictive entropy (Eq. (14)), so high H_t marks

positions with locally uncertain generation decisions. To keep only uncertainty that is supported by visual evidence, we compute a visual-support coefficient $\zeta_t \in (0,1)$ using a causal similarity kernel over preceding response states. This coefficient is high when the predecessors most similar to position t are more visually anchored than the prefix average, and near $1/2$ when the selection is uninformative; the causal kernel, support statistics, and the full derivation of ζ_t are given in Eqs. (15) to (18). In particular, $K_{\tau t} \propto \exp(\langle z_\tau, z_t\rangle/\kappa - \lambda_{\text{lag}} \log(1+t-\tau))$ weights only predecessors $\tau < t$, and $\zeta_t = \sigma(\chi^{-1} \log[(\varepsilon_{\text{sup}} + u_t)/(\varepsilon_{\text{sup}} + m_t)])$ compares kernel-weighted anchoring u_t with prefix-mean anchoring m_t. The visually supported branching score and its rollout-wise normalized form are

$$b_t = H_t\, \zeta_t, \qquad \widehat{b}_t = \text{Norm}_{1:T}(b_t), \tag{5}$$

with $b_1 = 0$ by convention. The signal $\widehat{b}_t \in [0,1]$ is the supported branching signal passed to the landing-field aggregator, large only when an uncertain position is selectively supported by visually anchored predecessors.

4.3 Aggregation and Landed Policy Optimization

The reward landing field aggregates the raw anchoring α_t, the rollout-normalized anchoring $\widehat{\alpha}_t$, and the supported branching score $\widehat{b}_t$ of Eqs. (4) and (5) into a single mean-one weighting over response positions. We define the combined evidence $e_t = \widehat{\alpha}_t + \widehat{b}_t$ and its rollout-centered version $\tilde{e}_t = e_t - \frac{1}{T}\sum_\tau e_\tau$, and construct the landing field as

$$\varphi_t = T\, \frac{\exp\big((1+\eta \tanh \tilde{e}_t)\, \alpha_t\big)}{\sum_{\tau=1}^{T} \exp\big((1+\eta \tanh \tilde{e}_\tau)\, \alpha_\tau\big)}, \tag{6}$$

where $\eta \geq 0$ is the modulation strength and the prefactor T enforces the mean-one constraint of Eq. (3). The raw anchoring α_t provides the base exponent, while the bounded multiplier $(1+\eta \tanh \tilde{e}_t) \in [1-\eta, 1+\eta]$ sharpens it for above-average positions and flattens it for below-average ones; setting $\eta = 0$ recovers a softmax over raw anchoring alone. Alternative aggregation maps are compared in Section A.6.

Both the uniform field $\Phi \equiv 1$ and the constructed field φ_t of Eq. (6) satisfy the mean-one constraint of Eq. (3), so any convex combination is also a valid landing field. We deploy the one-parameter blend

$$\Phi_\gamma(t) = (1-\gamma) + \gamma\, \varphi_t, \qquad \gamma \in [0,1], \tag{7}$$

in which γ is the *landing strength*: $\gamma = 0$ collapses Φ_γ to the uniform field and recovers GRPO, while $\gamma = 1$ uses the full constructed field. The total rollout credit $\sum_t \widetilde{A}_t = T\widehat{A}$ is preserved for every γ. Our defaults are $\gamma = 0.70$, $\eta = 1.00$, $\kappa = 0.50$, $\lambda_{\text{lag}} = 0.10$, and $\chi = 0.25$, yielding $\Phi_\gamma(t) \geq 0.30$. The full causal kernel costs $O(T^2 d)$ during training; measured scaling is in Section D.6.

GAZE-RL then feeds the landed advantage $\widetilde{A}_t = \Phi_\gamma(t)\, \widehat{A}$ into the clipped policy surrogate ψ_ϵ defined in Section 3, in place of the completion-level $\widehat{A}$ used by GRPO in Eq. (2), while leaving the verifier, group-relative advantage, clipping radius ϵ, and reference-policy penalty $\beta\, \text{KL}_t$ untouched,

$$\mathcal{J}_{\text{GAZE-RL}}(\theta) = \mathbb{E}_{\text{rollout}}\left[\frac{1}{T}\sum_{t=1}^{T}\Big(\psi_\epsilon\big(\rho_t(\theta), \widetilde{A}_t\big) - \beta\, \text{KL}_t(\theta)\Big)\right], \tag{8}$$

with $\rho_t(\theta)$ and $\text{KL}_t(\theta)$ defined in Section 3. The only modification relative to GRPO is the landing field Φ_γ, which redistributes the completion-level advantage across token-level update terms while preserving the total scalar credit assigned to each rollout. We compute it from rollout-time old-policy states and detach it before optimization; the sign of $\widehat{A}$ still determines reinforcement or penalty (Section A).

5 Experiments

5.1 Experimental Setup

We evaluate GAZE-RL on Qwen2.5-VL-3B, Qwen2.5-VL-7B, and InternVL3.5-8B (Bai et al., 2025a; Wang et al., 2025b). Within each backbone, all RL methods share the initial checkpoint, 35%

Table 1: General image-side benchmarks. MME uses the original summed score; all other entries are accuracies or official benchmark scores. Higher is better.

Model and Method	MME	MMVP	MMBench	MMStar	CVBench	POPE	RefCOCO
Qwen2.5-VL-3B	2151.2	65.7	76.4	54.1	70.2	83.2	83.9
+ GRPO	2148.6	67.3	77.0	54.5	71.9	84.1	85.2
+ DAPO	2152.2	68.0	77.4	54.9	**73.2**	**85.3**	85.9
+ VPPO	2149.0	66.3	77.6	53.6	71.5	83.5	85.1
+ GAZE-RL (Ours)	**2165.0**	**68.4**	**78.3**	**56.1**	72.6	85.1	**86.8**
Qwen2.5-VL-7B	2309.5	76.3	82.1	62.6	73.6	85.9	85.1
+ GRPO	2309.7	76.7	82.7	63.2	74.5	87.0	86.7
+ DAPO	2312.7	77.0	83.4	64.1	75.1	87.4	87.3
+ VPPO	2313.2	76.2	83.3	63.8	73.7	86.5	87.2
+ GAZE-RL (Ours)	**2344.1**	**78.7**	**84.0**	**64.4**	**75.9**	**88.0**	**87.9**
InternVL3.5-8B	2364.6	76.0	79.5	67.6	77.0	87.7	88.2
+ GRPO	2360.3	77.3	81.1	68.9	78.4	88.1	88.8
+ DAPO	2367.3	78.0	81.9	**70.5**	79.0	88.5	89.2
+ VPPO	2362.7	76.0	81.6	69.8	77.2	87.8	88.7
+ GAZE-RL (Ours)	**2371.7**	**79.3**	**83.0**	70.1	**80.8**	**89.0**	**90.1**

Table 2: OCR-heavy and reasoning-heavy image benchmarks. All entries follow the official evaluation protocol of each benchmark. Higher is better.

Model and Method	Geometry3K	TextVQA	DocVQA	AI2D	ChartQA	MathVista	LogicVista
Qwen2.5-VL-3B	15.3	78.3	93.5	81.1	80.8	60.9	36.6
+ GRPO	25.0	78.8	93.9	81.7	81.4	62.4	37.5
+ DAPO	26.8	79.4	94.1	82.1	81.8	62.9	38.7
+ VPPO	26.4	79.2	93.8	81.9	81.1	62.0	37.3
+ GAZE-RL (Ours)	**29.1**	**79.9**	**94.2**	**82.8**	**83.0**	**64.1**	**39.1**
Qwen2.5-VL-7B	36.8	84.0	95.1	83.2	83.1	67.2	42.9
+ GRPO	39.6	84.5	94.8	84.2	84.0	68.6	44.7
+ DAPO	41.9	84.9	95.3	**85.3**	**84.7**	69.4	45.8
+ VPPO	44.0	84.7	94.4	84.4	**84.7**	66.7	**46.1**
+ GAZE-RL (Ours)	**47.9**	**85.6**	**95.8**	85.0	84.5	**70.4**	**46.1**
InternVL3.5-8B	40.3	77.6	91.8	83.5	82.7	76.5	57.3
+ GRPO	43.2	78.3	92.1	84.1	84.4	77.3	60.8
+ DAPO	45.0	78.6	92.3	85.0	84.8	78.7	61.6
+ VPPO	44.8	78.0	91.9	84.6	85.3	76.1	60.7
+ GAZE-RL (Ours)	**48.1**	**79.2**	**92.7**	**85.5**	**86.8**	**78.9**	**63.0**

ViRL39K (TIGER-Lab, 2025) / 65% verifier-filtered Video-R1 (Feng et al., 2025) curriculum, verifier, rollout group size, and 200-update budget. Baselines are GRPO (Shao et al., 2024), DAPO (Yu et al., 2025b), VPPO (Huang et al., 2026a), and Video-R1 (Feng et al., 2025). Image benchmarks are MME (Fu et al., 2025a), MMVP (Tong et al., 2024b), MMBench (Liu et al., 2025b), MMStar (Chen et al., 2024), CV-Bench (Tong et al., 2024a), POPE (Li et al., 2023c), RefCOCO (Yu et al., 2016), Geometry3K (Lu et al., 2021), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), MathVista (Lu et al., 2024), and LogicVista (Xiao et al., 2024). Video benchmarks are Video-MME (Fu et al., 2025b), TempCompass (Liu et al., 2024), NExT-QA (Xiao et al., 2021), Video-MMMU (Hu et al., 2025), LongVideoBench (Wu et al., 2024), and VSI-Bench (Yang et al., 2025a). Protocols and implementation details are in Sections B and C.3.

5.2 Main Results

Across the image and video suites in Table 1, Table 2 and Table 3, GAZE-RL improves most consistently on benchmarks that require visually grounded reasoning rather than broad recognition alone. On image-side evaluation, the gains are strongest on MMVP, Geometry3K, MathVista, and LogicVista, where response tokens must bind visual evidence to object, spatial, numerical, or logical decisions. Video-side evaluation suggests that reward landing is especially useful when relevant evidence is temporally sparse and only a few response positions carry the main decision burden.

Table 3: Video benchmarks. Results cover short-video QA, temporal reasoning, knowledge-intensive video QA, long-video understanding, and spatial memory. Higher is better.

Model and Method	Video-MME	TempCompass	NExT-QA	Video-MMMU	LongVideoBench	VSI-Bench
Qwen2.5-VL-3B	58.4	62.9	**67.7**	31.3	52.2	22.3
+ GRPO	58.2	63.1	63.1	32.4	53.1	23.5
+ DAPO	59.0	63.8	65.8	33.3	53.6	24.4
+ Video-R1	58.6	63.4	64.7	32.6	52.4	22.7
+ GAZE-RL (Ours)	**59.5**	**64.8**	67.2	**33.9**	**53.7**	**25.0**
Qwen2.5-VL-7B	60.1	69.0	**72.5**	45.9	55.7	30.8
+ GRPO	60.9	69.4	66.0	47.2	57.1	31.9
+ DAPO	61.4	69.9	68.6	48.0	58.3	32.5
+ Video-R1	61.8	68.5	66.5	47.6	55.4	31.8
+ GAZE-RL (Ours)	**62.2**	**72.0**	71.7	**48.3**	**59.0**	**33.0**
InternVL3.5-8B	64.5	68.2	74.8	48.1	61.3	43.1
+ GRPO	73.9	66.1	75.2	49.4	62.7	44.3
+ DAPO	**75.6**	68.5	76.4	50.8	**63.2**	45.2
+ Video-R1	74.7	68.7	73.1	49.6	61.3	43.8
+ GAZE-RL (Ours)	67.0	**70.1**	**78.9**	**51.2**	62.7	**45.9**

Across the full suite, the results support redistributing outcome-level advantage to visually grounded and decision-bearing tokens without task-specific supervision.

Matched-seed robustness. We reran GRPO, DAPO, and GAZE-RL on Qwen2.5-VL-7B with three paired seeds. Their Image Avg / Video Avg scores (mean±standard deviation) are 57.4±0.2 / 53.2±0.1, 58.5±0.1 / 54.0±0.1, and 60.7±0.1 / 54.8±0.1, respectively; GAZE-RL leads on both averages in each seed. The main tables and Figure 5 retain their representative single-run results; full matched-seed results and uncertainty estimates are in Section D.1.



Figure 2: Credit mass and response length.

Training dynamics. Figure 2 shows that GAZE-RL progressively concentrates more credit on high-weight response positions, while GRPO remains uniform by construction. Top-20 mass measures the landing weight assigned to the highest-weighted 20% of response tokens. At the same time, average response length decreases rather than increases, suggesting that the gains do not come from longer reasoning alone. The trend is consistent with reward landing reducing redundant continuation while preserving visually grounded decision steps. Details are given in Section C.4.

5.3 REWARD LANDING, VISUAL EVIDENCE, AND ABLATIONS

We visualize response-side credit on held-out rollouts from image and video reasoning benchmarks. Figure 3 shows that GAZE-RL concentrates landing mass on visually grounded entities, numerical comparisons, relation or temporal cues, and final decision spans, rather than filler or formatting tokens. Negative-return examples remain interpretable: φ_t identifies where the update lands, while the sign of $\widehat{A}$ determines whether those positions are reinforced or discouraged.

We test whether response-side landing points to visual evidence that affects prediction. Figure 4 visualizes guided and random dropping on the trained checkpoints under matched masking budgets, alongside examples of question-relevant regions. Cross-checkpoint drop magnitudes also depend on each model's masking robustness, so they alone do not isolate ranking quality. A controlled comparison that changes only the mask ranking within each frozen checkpoint is reported in Section D.5.

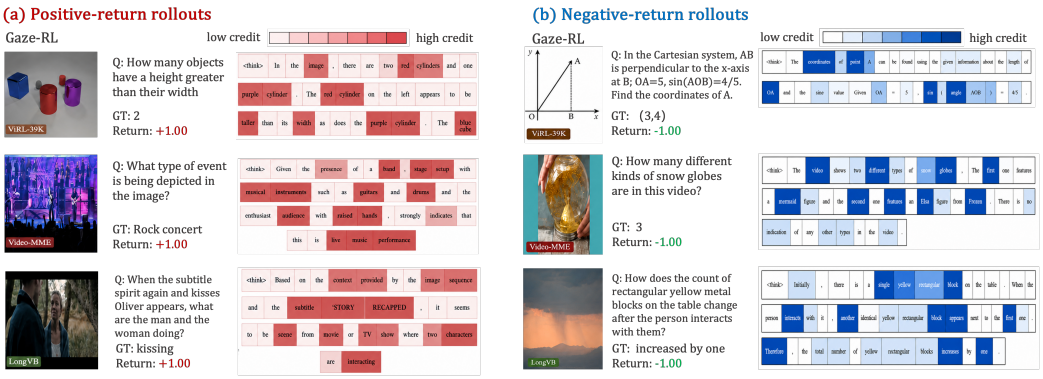

Figure 3: Response-side reward landing on diagnostic rollouts. Red and blue heatmaps show positive- and negative-return examples; darker cells indicate larger landing weight φ_t.

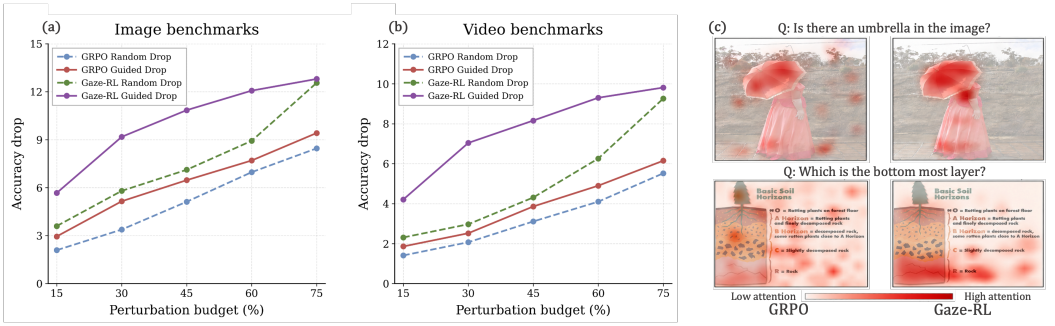

Figure 4: Visual evidence sensitivity. Guided drop masks visual cells ranked by the projected landing field; random drop masks the same budget uniformly. Larger accuracy drops indicate more necessary evidence.

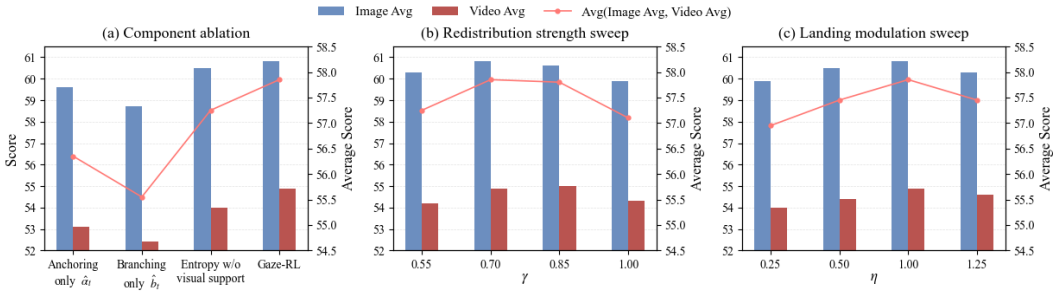

Figure 5: Ablation and sweep on Qwen2.5-VL-7B. Bars report Image Avg and Video Avg; the line reports their mean. Image Avg averages MMVP, Geometry3K, MathVista, and LogicVista; Video Avg averages Video-MME, TempCompass, Video-MMMU, LongVideoBench, and VSI-Bench.

Figure 5 shows that the full landing rule is stronger than using visual anchoring alone, supported branching alone, or entropy without visual support. The sweeps also indicate that moderate landing strength and modulation are preferred: weak redistribution approaches uniform GRPO-style credit, while overly strong concentration can make updates too narrow. The benchmark averages and default-setting consistency check are described in Section C.6.

5.4 Zero-shot Transfer Probe on Image Editing



Figure 6: Image-editing transfer probe. We replace the VLM inside Qwen-Image-Edit with the GAZE-RL-trained Qwen2.5-VL-7B-Instruct while keeping the editing backbone fixed.

Figure 6 provides a qualitative transfer probe beyond benchmark QA. With only the instruction-understanding VLM replaced, the examples suggest improved grounding for local appearance, material, facial attribute, color, weather, and action edits. Since this probe is qualitative and uses no editing-specific reward, we treat it as supporting evidence rather than a primary result. Additional setup details are given in Section C.7.

6 Conclusion and Discussion

Conclusion. This paper revisits multimodal reinforcement learning through a simple but overlooked question: when a verifier rewards an entire response, where should that reward be absorbed inside the generated trajectory? We show that the usual uniform broadcast of completion-level advantage is mismatched with multimodal reasoning, where visual evidence is used sparsely and only a few response positions often mediate the final decision. To address this mismatch, we propose GAZE-RL, a representation-driven policy optimization method that builds a mean-one reward landing field from internal cross-modal geometry and local predictive uncertainty. The method preserves the total credit of each rollout while reallocating update strength toward positions that are visually anchored and decision-bearing. Across image and video benchmarks, GAZE-RL improves both general understanding and reasoning-heavy performance. These results suggest that multimodal RL should not only ask whether a trajectory is rewarded, but also learn where that reward should land.

Discussion. The main implication is that credit localization should be treated as a core design axis in multimodal post-training. Current methods often improve data, rewards, or optimization stability, but still allow the learning signal to diffuse across long textual continuations that do not equally use visual evidence. GAZE-RL points to a complementary direction: use the model's own internal representations to identify where visual evidence and consequential uncertainty meet, without adding process labels, external perception modules, or architectural changes. This suggests that the structure of the policy update itself can be a source of grounded reasoning improvement, not merely the reward model or the training data. This perspective is especially relevant for long-video and complex visual reasoning, where decisive evidence may be temporally sparse, locally expressed, and easily diluted by fluent but weakly grounded reasoning text. Future work could extend reward landing beyond token-level fields toward object-, region-, or event-level credit allocation, and combine internal landing signals with stronger verifiers or memory mechanisms for longer-horizon multimodal reasoning.

Limitations. The method depends on the quality of the backbone's visual-text representations and of the verifier that supplies the rollout-level advantage. It cannot correct an erroneous outcome reward or missing visual evidence; these limitations and further directions are discussed in Appendices E and F.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing the methods, including code implementation and experiment execution during the later stages of the project. We did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; write proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; translate text; clean or reformat datasets; support qualitative or thematic data analysis; or interpret results.

Additionally, we used generative AI tools to improve the wording and readability of the manuscript through discussions with the authors about writing and presentation, correct grammatical errors, and assist in writing and editing code. The experiments were designed by the authors. The authors reviewed and verified the AI-assisted implementations and experimental results, and reviewed and revised all AI-assisted text. The authors take full responsibility for the final content of this work, including its text, code, and experimental results.

REPRODUCIBILITY STATEMENT

Appendix A provides the method details and derivations. Appendices B and C specify the training data, baselines, optimization and reward settings, hyperparameters, compute resources, and evaluation protocols. Appendix D documents additional controlled analyses, including matched-seed comparisons. Our code is available at https://anonymous.4open.science/r/anonymous-gaze-rl-2ECE/ to facilitate reproduction of our results.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966, 2023. URL https://arxiv.org/abs/2308.12966.

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

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning. arXiv preprint arXiv:2505.14231, 2025b. URL https://arxiv.org/abs/2505.14231.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, and Chenliang Xu. VERIFY: A Benchmark of Visual Explanation and

Reasoning for Investigating Multimodal Reasoning Fidelity. arXiv preprint arXiv:2503.11557, 2025. URL https://arxiv.org/abs/2503.11557.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. arXiv preprint arXiv:2504.11468, 2025. URL https://arxiv.org/abs/2504.11468.

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. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-0850. URL https://doi.org/10.52202/079017-0850.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract.html.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex Vision-Language Reasoning via Iterative SFT-RL Cycles. arXiv preprint arXiv:2503.17352, 2025. URL https://arxiv.org/abs/2503.17352.

Yizhuo Ding, Mingkang Chen, Zhibang Feng, Tong Xiao, Wanying Qu, Wenqi Shao, and Yanwei Fu. VTPerception-R1: Enhancing Multimodal Reasoning via Explicit Visual and Textual Perceptual Grounding. arXiv preprint arXiv:2509.24776, 2025. URL https://arxiv.org/abs/2509.24776.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward. In *International Conference on Learning Representations*, pp. 9290–9310, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0fb49ffa040d13a491c7164c5912cf7d-Abstract-Conference.html.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. GRIT: Teaching MLLMs to Think with Images. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129120–129141. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-3884. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a92ee7e49a25412add0883454da6bed9-Abstract-Conference.html.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing Video Reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 109844–109867. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-3310. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8eb3976840e08b80dda9667562574246-Abstract-Conference.html.

Sicheng Feng, Kaiwen TUO, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. RewardMap: Tackling Sparse Rewards in Fine-grained Visual Reasoning via Multi-Stage Reinforcement Learning. In *International Conference on Learning Representations*, pp. 5240–5261, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0967d7c8b171dd81b77c43067c02bebf-Abstract-Conference.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567. Neural Information Processing Systems Foundation, Inc., 2025a. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/

hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025b. doi: 10.1109/cvpr52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Gemma Team. Gemma 3 Technical Report. arXiv preprint arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. arXiv preprint arXiv:2506.07966, 2025. URL https://arxiv.org/abs/2506.07966.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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 (CVPR)*, 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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. arXiv preprint arXiv:2501.13826, 2025. URL https://arxiv.org/abs/2501.13826.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on Token Perception for Multimodal Reinforcement Learning. In *International Conference on Learning Representations*, pp. 103291–103325, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a8e3197f627b9c88f86c0d3eb1ade5d7-Abstract-Conference.html.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. In *International Conference on Learning Representations*, pp. 63794–63812, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/680b2a8135b9c71278a09cafb605869e-Abstract-Conference.html.

Mohamed Huti, Alasdair Mackintosh, Amy Waldock, Dominic Andrews, Maxime Lelièvre, Moritz Boos, Tobias Murray, Paul Atherton, Robin A. A. Ince, and Oliver G. B. Garrod. Visual Reasoning Benchmark: Evaluating Multimodal LLMs on Classroom-Authentic Visual Problems from Primary Education. arXiv preprint arXiv:2602.12196, 2026. URL https://arxiv.org/abs/2602.12196.

Zhengbo Jiao, Shaobo Wang, Zifan Zhang, Wei Wang, Bing Zhao, Hu Wei, and Linfeng Zhang. Credit Where It is Due: Cross-Modality Connectivity Drives Precise Reinforcement Learning for MLLM Reasoning. arXiv preprint arXiv:2602.11455, 2026. URL https://arxiv.org/abs/2602.11455.

Hoang Anh Just, Yifei Fan, Handong Zhao, Jiuxiang Gu, Ruiyi Zhang, Simon Jenni, Kushal Kafle, Ruoxi Jia, and Jing Shi. More Than the Final Answer: Improving Visual Extraction and Logical Consistency in Vision-Language Models. arXiv preprint arXiv:2512.12487, 2025. URL https://arxiv.org/abs/2512.12487.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *Computer Vision – ECCV 2016*, pp. 235–251, Cham, 2016. Springer, Cham. doi: 10.1007/978-3-319-46493-0_15. URL https://link.springer.com/chapter/10.1007/978-3-319-46493-0_15.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. arXiv preprint arXiv:2408.03326, 2024a. URL https://arxiv.org/abs/2408.03326.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Jiaze Li, Hao Yin, Wenhui Tan, Jingyang Chen, Boshen Xu, Yuxun Qu, Yijing Chen, Jianzhong Ju, Zhenbo Luo, and Jian Luan. REVISOR: Beyond Textual Reflection, Towards Multimodal Introspective Reasoning in Long-Form Video Understanding. arXiv preprint arXiv:2511.13026, 2025. URL https://arxiv.org/abs/2511.13026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-Centric Video Understanding. arXiv preprint arXiv:2305.06355, 2023b. URL https://arxiv.org/abs/2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024c. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Xirui Li, Ming Li, and Tianyi Zhou. What does RL improve for Visual Reasoning? A Frankenstein-Style Analysis. arXiv preprint arXiv:2602.12395, 2026. URL https://arxiv.org/abs/2602.12395.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023c. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Shieh. NoisyRollout: Reinforcing Visual Reasoning with Data Augmentation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 3248–3282. Neural Information Processing Systems Foundation, Inc., 2025a. doi: 10.52202/085713-0102. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0440de9623f97491b6de9b53afcfe6f4-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-Around Player? In *Computer Vision – ECCV 2024*, pp. 216–233. Springer, 2025b. doi: 10.1007/978-3-031-72658-3_13. URL https://link.springer.com/chapter/10.1007/978-3-031-72658-3_13.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? arXiv preprint arXiv:2403.00476, 2024. URL https://arxiv.org/abs/2403.00476.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual Reinforcement Fine-Tuning. arXiv preprint arXiv:2503.01785, 2025c. URL https://arxiv.org/abs/2503.01785.

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 (Volume 1: Long Papers)*, pp. 6774–6786. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

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*, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

Maria Lymperaiou, Vasileios Karampinis, Giorgos Filandrianos, Angelos Vlachos, Chrysoula Zerva, and Athanasios Voulodimos. Reasoning or Pattern Matching? Probing Large Vision-Language Models with Visual Puzzles. arXiv preprint arXiv:2601.13705, 2026. URL https://arxiv.org/abs/2601.13705.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Damiano Marsili and Georgia Gkioxari. No Labels, No Problem: Training Visual Reasoners with Multimodal Verifiers. arXiv preprint arXiv:2512.08889, 2025. URL https://arxiv.org/abs/2512.08889.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

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.

Siqu Ou, Tianrui Wan, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Do MLLMs Really See It: Reinforcing Visual Attention in Multimodal LLMs. arXiv preprint arXiv:2602.08241, 2026. URL https://arxiv.org/abs/2602.08241.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, Chong Sun, Chen Li, Jing Lyu, and Honggang Zhang. V-Thinker: Interactive Thinking with Images. arXiv preprint arXiv:2511.04460, 2025. URL https://arxiv.org/abs/2511.04460.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded Reinforcement Learning for Visual Reasoning. arXiv preprint arXiv:2505.23678, 2025. URL https://arxiv.org/abs/2505.23678.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

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.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv preprint arXiv:2504.07615, 2025. URL https://arxiv.org/abs/2504.07615.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Yolo Y. Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, Chao Huang, Zhiyuan Wang, Susan Liang, Xinyi Liu, Yizhi Song, Junhua Huang, Jia-Xing Zhong, Bozheng Li, Daiqing Qi, Ziyun Zeng, Ali Vosoughi, Luchuan Song, Zeliang Zhang, Daiki Shimada, Han Liu, Jiebo Luo, and Chenliang Xu. Video-LMM Post-Training: A Deep Dive into Video Reasoning with Large Multimodal Models. arXiv preprint arXiv:2510.05034, 2025. URL https://arxiv.org/abs/2510.05034.

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. In *International Conference on Learning Representations*, pp. 133928–133965, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8a19c815a8bef25e6094e87f963d28e-Abstract-Conference.html.

TIGER-Lab. ViRL39K. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/TIGER-Lab/ViRL39K.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C. 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. Neural Information Processing Systems Foundation, Inc., 2024a. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. arXiv preprint arXiv:2401.06209, 2024b. URL https://arxiv.org/abs/2401.06209.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. In *International Conference on Learning Representations*, pp. 62624–62650, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9cf6139382f98623d08cc595622f3fb1-Abstract-Conference.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with

MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169. Curran Associates, Inc., 2024. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint arXiv:2508.18265, 2025b. URL https://arxiv.org/abs/2508.18265.

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, 2025c. URL https://arxiv.org/abs/2507.06448.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start. arXiv preprint arXiv:2505.22334, 2025. URL https://arxiv.org/abs/2505.22334.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. arXiv preprint arXiv:2407.15754, 2024. URL https://arxiv.org/abs/2407.15754.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing Multimodal Reasoning Capabilities of MLLMs via Visual Perception Reward. arXiv preprint arXiv:2506.07218, 2025. URL https://arxiv.org/abs/2506.07218.

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.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multi-modal Large Language Models. arXiv preprint arXiv:2504.15279, 2025. URL https://arxiv.org/abs/2504.15279.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025a. doi: 10.1109/cvpr52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-Onevision: Advancing Generalized

Multimodal Reasoning through Cross-Modal Formalization. arXiv preprint arXiv:2503.10615, 2025b. URL https://arxiv.org/abs/2503.10615.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaxing Huang. R1-ShareVL: Incentivizing Reasoning Capability of Multimodal Large Language Models via Share-GRPO. arXiv preprint arXiv:2505.16673, 2025. URL https://arxiv.org/abs/2505.16673.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. arXiv preprint arXiv:2504.07954, 2025a. URL https://arxiv.org/abs/2504.07954.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, pp. 69–85. Springer, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://link.springer.com/chapter/10.1007/978-3-319-46475-6_5.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv preprint arXiv:2503.14476, 2025b. URL https://arxiv.org/abs/2503.14476.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native Interleaved Clue Seeking for Long Video Multi-Hop Reasoning. arXiv preprint arXiv:2601.23224, 2026. URL https://arxiv.org/abs/2601.23224.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. arXiv preprint arXiv:2503.12937, 2025a. URL https://arxiv.org/abs/2503.12937.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: Does Your Multimodal LLM Truly See the Diagrams in Visual Math Problems? In *Computer Vision – ECCV 2024*, pp. 169–186, Cham, 2025b. Springer Nature Switzerland. doi: 10.1007/978-3-031-73242-3_10. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_10. ECCV 2024.

Xintong Zhang, Xiaowen Zhang, Jingrong Wu, Zhi Gao, Shilin Yan, Zhenxin Diao, Kunpeng Gao, Xuanyan Chen, Yuwei Wu, Yunde Jia, and Qing Li. AdaptMMBench: Benchmarking Adaptive Multimodal Reasoning for Mode Selection and Reasoning Process. arXiv preprint arXiv:2602.02676, 2026. URL https://arxiv.org/abs/2602.02676.

Rosie Zhao, Anshul Shah, Xiaoyu Zhu, Xinke Deng, Zhongyu Jiang, Yang Yang, Joerg Liebelt, and Arnab Mondal. On Robustness and Chain-of-Thought Consistency of RL-Finetuned VLMs. arXiv preprint arXiv:2602.12506, 2026. URL https://arxiv.org/abs/2602.12506.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv:2304.10592, 2023. URL https://arxiv.org/abs/2304.10592.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479, 2025. URL https://arxiv.org/abs/2504.10479.

A METHOD DETAILS AND DERIVATIONS

This appendix gives the full construction of the latent reward landing field used by GAZE-RL. We follow the notation of Sections 3 and 4. For a rollout $\mathcal{Y} = (y_1, \ldots, y_T)$, the verifier produces a completion-level advantage $\widehat{A}$, while GAZE-RL constructs a rollout-dependent field over response positions and converts $\widehat{A}$ into landed advantages $\widetilde{A}_t$. The construction is post-hoc with respect to the sampled rollout: it uses the policy's internal hidden states and rollout-time predictive distributions, but introduces no additional reward model, process supervision, auxiliary perception module, or architectural modification.

A.1 NOTATION SUMMARY

Table 4 summarizes the main quantities used in the method and appendix.

Table 4: Notation used in the reward landing construction.

Symbol	Meaning
$\mathcal{V} = (v_1, \ldots, v_M)$	visual embeddings or packed visual memory elements
$\mathcal{P}$	textual prompt
$\mathcal{Y} = (y_1, \ldots, y_T)$	generated response
$s_t = (\mathcal{V}, \mathcal{P}, \mathcal{Y}_{<t})$	autoregressive state before generating y_t
$\widehat{A}$	group-normalized completion-level advantage
$\mathcal{L}$	selected transformer layers used for internal geometry
$h^{\text{txt}}_{\ell,t}$	pre-action response state predicting y_t at layer ℓ
$h^{\text{vis}}_{\ell,m}$	hidden state of visual memory element v_m at layer ℓ
$\alpha_t, \widehat{\alpha}_t$	raw and rollout-normalized visual anchoring scores
H_t	rollout-time predictive entropy at position t
ζ_t	visual-support coefficient for local branching
$b_t, \widehat{b}_t$	raw and rollout-normalized supported-branching scores
$e_t, \widetilde{e}_t$	combined evidence and its centered version
φ_t	constructed mean-one landing field
$\Phi_\gamma(t)$	blended landing field with landing strength γ
$\widetilde{A}_t$	landed token-level advantage

A.2 CREDIT PRESERVATION OF MEAN-ONE LANDING FIELDS

The defining constraint of a reward landing field is

$$\Phi(t) \geq 0, \qquad \frac{1}{T}\sum_{t=1}^{T} \Phi(t) = 1, \qquad \widetilde{A}_t = \Phi(t)\widehat{A}.$$

This constraint is important because GAZE-RL is designed to redistribute the existing completion-level advantage rather than amplify or attenuate it. Indeed,

$$\sum_{t=1}^{T} \widetilde{A}_t = \sum_{t=1}^{T} \Phi(t)\widehat{A} = T\widehat{A}. \tag{9}$$

Thus the average landed advantage over response positions is still $\widehat{A}$. When $\Phi(t) \equiv 1$, the landed advantage reduces to the usual uniform assignment $\widetilde{A}_t = \widehat{A}$, recovering the GRPO-style update.

The constructed field φ_t in Eq. (6) is also mean-one by construction. Let

$$q_t = \exp((1 + \eta \tanh \widetilde{e}_t)\alpha_t), \qquad Z = \sum_{\tau=1}^{T} q_\tau.$$

Then $\varphi_t = Tq_t/Z$, so

$$\frac{1}{T}\sum_{t=1}^{T} \varphi_t = \frac{1}{T}\sum_{t=1}^{T} T\frac{q_t}{Z} = 1. \tag{10}$$

The blended field is $\Phi_\gamma(t) = (1-\gamma) + \gamma\varphi_t$, with $\gamma \in [0,1]$. Because both the uniform field and φ_t are mean-one,

$$\frac{1}{T}\sum_{t=1}^{T}\Phi_\gamma(t) = (1-\gamma) + \gamma\frac{1}{T}\sum_{t=1}^{T}\varphi_t = 1. \tag{11}$$

Therefore, for every γ, the total rollout credit remains $T\widehat{A}$, while γ controls how far the update moves away from uniform token-level credit.

A.3 Rollout-wise scalar normalization

For any rollout-side scalar sequence $q_{1:T}$, the operator $\mathrm{Norm}_{1:T}$ used in Eqs. (4) and (5) applies stabilized min-max scaling over response positions:

$$\mathrm{Norm}_{1:T}(q_t) = \frac{q_t - \min_{\tau\in[T]} q_\tau}{\max_{\tau\in[T]} q_\tau - \min_{\tau\in[T]} q_\tau + \varepsilon_{\mathrm{norm}}}, \qquad \varepsilon_{\mathrm{norm}} > 0. \tag{12}$$

The normalization is local to each generated response, rather than shared across prompts or mini-batches. This is useful because raw anchoring and entropy scales can vary across inputs, modalities, response lengths, and model backbones. Rollout-wise scaling keeps only the within-response contrast needed for positionwise credit allocation. When the score range is nearly flat, the denominator is dominated by $\varepsilon_{\mathrm{norm}}$, the normalized contrast remains small, and the resulting landing field stays close to uniform. This behavior is desirable: if the internal signals do not provide reliable evidence that some positions are more credit-relevant than others, the method should avoid imposing an artificial concentration pattern.

A.4 Stabilized state normalization

The $\mathrm{StateNorm}$ operator used to form the per-position response state applies an ℓ_2 normalization with a small additive denominator:

$$\mathrm{StateNorm}(z) = \frac{z}{\|z\|_2 + \varepsilon_{\mathrm{sup}}}, \qquad \varepsilon_{\mathrm{sup}} > 0. \tag{13}$$

This prevents division by near-zero norms when a layer-averaged hidden state has very small magnitude. We use the same stability constant $\varepsilon_{\mathrm{sup}}$ in the support log-ratio of Eq. (18). The role of this normalization is only to make response-state similarity comparable across positions; it does not change the model's forward pass or add a trainable component.

A.5 Supported-branching coefficient

Visual anchoring identifies positions whose hidden states are close to the visual stream, but not every visually anchored token is a consequential decision point. For example, a token can be visually grounded while its next-token distribution is nearly deterministic. Supported branching therefore adds a second signal: it gives higher weight to positions where the model faces local predictive uncertainty and where the preceding states supporting that uncertainty are themselves visually anchored.

The local predictive entropy at position t is

$$H_t = \mathrm{Ent}\big(\pi_{\theta_{\mathrm{old}}}(\cdot \mid s_t)\big) = -\sum_{c\in\Sigma_{\mathrm{tok}}} \pi_{\theta_{\mathrm{old}}}(c \mid s_t)\log \pi_{\theta_{\mathrm{old}}}(c \mid s_t), \tag{14}$$

where Σ_{tok} is the token vocabulary. The entropy is computed from the rollout-time old policy, matching the policy distribution that generated the response. High H_t indicates that the model has multiple plausible continuations at position t, but entropy alone is not sufficient: uncertainty can arise from harmless wording variation or formatting, rather than from a visually meaningful reasoning branch.

For the support coefficient ζ_t in Eq. (5), we first form a per-position response state by averaging hidden states across the same layer set $\mathcal{L}$ used for visual anchoring and applying $\mathrm{StateNorm}$:

$$z_t = \mathrm{StateNorm}\left(\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} h_{\ell,t}^{\mathrm{txt}}\right). \tag{15}$$

Each predecessor index $\tau < t$ is then weighted by a causal kernel that combines response-state similarity with log-distance decay:

$$\xi_{\tau t} = \frac{\langle z_\tau, z_t \rangle}{\kappa} - \lambda_{\text{lag}} \log(1 + t - \tau), \qquad K_{\tau t} = \frac{\exp(\xi_{\tau t})}{\sum_{j<t} \exp(\xi_{jt})}, \tag{16}$$

where $\kappa > 0$ is a temperature and $\lambda_{\text{lag}} \geq 0$ controls temporal distance decay. The kernel is causal because it only uses predecessor positions $\tau < t$. The similarity term selects earlier response states that are representation-wise close to the current state, while the log-distance penalty prevents very early tokens from dominating purely because of broad semantic similarity. In the implementation, all response states are stacked and the kernel is evaluated with batched matrix multiplication and a strict lower-triangular mask. Cached decoder states are shifted so that the state and old-policy entropy at position t both refer to the pre-action state s_t that predicts y_t. The complete landing field is detached before policy optimization.

For $t > 1$, the kernel-selected anchoring and prefix-average anchoring are

$$u_t = \sum_{\tau<t} K_{\tau t} \widehat{\alpha}_\tau, \qquad m_t = \frac{1}{t-1} \sum_{\tau<t} \widehat{\alpha}_\tau. \tag{17}$$

Here u_t measures how visually anchored the predecessors selected by the causal kernel are, while m_t measures the average visual anchoring of the entire prefix. The comparison between u_t and m_t is what distinguishes visually supported uncertainty from ordinary language uncertainty. If the current position is mainly supported by predecessors that are no more visually anchored than the prefix average, then its entropy should not receive a large visual-branching boost.

The visual-support coefficient compares these two quantities through a stabilized log-ratio:

$$\zeta_t = \sigma\left(\frac{1}{\chi} \log \frac{\varepsilon_{\text{sup}} + u_t}{\varepsilon_{\text{sup}} + m_t}\right), \qquad \chi > 0. \tag{18}$$

Thus $\zeta_t > 1/2$ when the causal kernel selects predecessors that are more visually anchored than the prefix as a whole, $\zeta_t \approx 1/2$ when the selection is uninformative, and $\zeta_t < 1/2$ when the selected predecessors are less visually anchored than the prefix average. The visually supported branching score is then

$$b_t = H_t \zeta_t, \qquad \widehat{b}_t = \text{Norm}_{1:T}(b_t),$$

with $b_1 = 0$ by convention because the first generated token has no predecessor support. This construction makes $\widehat{b}_t$ large only when a position is both locally uncertain and selectively supported by visually anchored earlier states.

A.6 Landing aggregation function

The landing field formula is given in Eq. (6). Here we spell out the design choices behind the aggregation. The raw anchoring α_t serves as the base exponent because it preserves the unnormalized scale of cross-modal similarity: positions with higher absolute anchoring receive larger base weights even before evidence modulation. The normalized anchoring $\widehat{\alpha}_t$ and normalized supported branching $\widehat{b}_t$ are combined as

$$e_t = \widehat{\alpha}_t + \widehat{b}_t, \qquad \widetilde{e}_t = e_t - \frac{1}{T} \sum_{\tau=1}^{T} e_\tau.$$

The centered evidence $\widetilde{e}_t$ indicates whether a position is above or below the rollout's average evidence level. The bounded multiplier $(1 + \eta \tanh \widetilde{e}_t)$ then modulates the raw anchoring exponent:

$$\varphi_t = T \frac{\exp((1 + \eta \tanh \widetilde{e}_t)\alpha_t)}{\sum_{\tau=1}^{T} \exp((1 + \eta \tanh \widetilde{e}_\tau)\alpha_\tau)}. \tag{19}$$

The $\tanh$ nonlinearity prevents unbounded modulation from very large normalized evidence contrasts. For the default setting $\eta \geq 0$, above-average positions sharpen the anchoring-based allocation, while below-average positions flatten it. Setting $\eta = 0$ recovers a mean-one softmax over raw visual anchoring alone:

$$\varphi_t^{\text{anchor}} = T \frac{\exp(\alpha_t)}{\sum_{\tau=1}^{T} \exp(\alpha_\tau)}.$$

The full construction is therefore a controlled interpolation between pure anchoring-based allocation and an evidence-modulated allocation that also accounts for visually supported branching.

The method can also be interpreted as separating two roles. Raw α_t provides an absolute cross-modal compatibility signal, while e_t provides a rollout-local utility signal. Using only normalized evidence would ignore the absolute scale of visual anchoring; using only raw anchoring would ignore whether a visually grounded position is actually decision-bearing. The aggregation in Eq. (19) keeps both.

A.7 Landed policy objective

After constructing φ_t, GAZE-RL uses the blended field

$$\Phi_\gamma(t) = (1-\gamma) + \gamma\varphi_t, \qquad \gamma \in [0,1],$$

and defines the landed advantage

$$\widetilde{A}_t = \Phi_\gamma(t)\widehat{A}.$$

The policy objective is the same clipped surrogate used by GRPO, except that each response position receives $\widetilde{A}_t$ instead of the uniform $\widehat{A}$:

$$\mathcal{J}_{\text{GAZE-RL}}(\theta) = \mathbb{E}_{\text{rollout}}\left[\frac{1}{T}\sum_{t=1}^{T}\Big(\psi_\epsilon(\rho_t(\theta), \widetilde{A}_t) - \beta\,\mathrm{KL}_t(\theta)\Big)\right].$$

All other components are unchanged: the verifier, group-relative advantage computation, clipping radius, and reference-policy penalty are the same as in the corresponding GRPO training run. Thus the method changes where the update lands, not which trajectories are rewarded or how much total advantage they receive.

The sign of the update is still controlled by $\widehat{A}$. If $\widehat{A} > 0$, high-$\Phi_\gamma(t)$ positions receive stronger positive reinforcement. If $\widehat{A} < 0$, the same positions receive stronger discouragement. This is why negative-return examples remain interpretable in the response-side visualizations: the landing field identifies the positions responsible for absorbing the update, while the completion-level advantage determines the update direction.

B Experimental Details

B.1 Training Data and Baselines

The training set is a mixed image-video RL curriculum constructed from ViRL39K (TIGER-Lab, 2025) and a verifier-filtered subset of Video-R1 (Feng et al., 2025). Image data counts for 35% and video data counts for 65%. The image portion covers visual logic, chart reasoning, mathematical diagrams, OCR-heavy question answering, and grounding-oriented question answering. The video portion covers short-video understanding, temporal ordering, event localization, and long-video evidence aggregation. For the three backbones in the main tables, GRPO (Shao et al., 2024), DAPO (Yu et al., 2025b), VPPO (Huang et al., 2026a), Video-R1 (Feng et al., 2025), and GAZE-RL use the same initial checkpoint, curriculum, verifier, 200-update budget, rollout group size, and hardware within each backbone. Their optimization rules differ, but the training data are not a method-specific advantage. The separate Gemma-3-4B comparison in Section D.2 uses a matched image-side curriculum within that backbone.

B.2 Training Configuration

Table 5 summarizes the main training and rollout settings used in our implementation. We train for a fixed number of RL updates, save checkpoints periodically, and use grouped rollouts for advantage normalization. The generation budget is capped at 512 new tokens, which matches the response-length analysis in Figure 2 under the same generation limit. For video inputs, we sample at 0.5 fps and cap the number of sampled frames. Image resolution is controlled by the pixel budget of the corresponding backbone.

Table 5: Main RL training and rollout configuration. These settings are shared by the reported GAZE-RL runs unless otherwise specified.

Category	Setting
Training steps	200
Checkpoint interval	every 5 updates
Hardware	4 NVIDIA H200 GPUs
Gradient accumulation	32 steps
Learning rate	1.0×10^{-6}
Weight decay	0.0
Warmup ratio	0.03
Learning-rate schedule	cosine
Maximum gradient norm	1.0
Rollout group size	$G = 8$
Rollout chunk size	2
Maximum new tokens	512
Minimum generated tokens	1
Maximum sequence length	5120
Maximum generation time	600 seconds

Table 6: Wall-clock comparison for 200 RL training updates. GAZE-RL has nearly the same runtime as vanilla GRPO under this configuration because rollout generation and policy optimization dominate the total training time.

Phase	Vanilla GRPO	GAZE-RL
Wall-clock time	2.1 days	2.1 days

B.3 Optimization and Reward Settings

For advantage processing, we use standard group-normalized advantages with a numerical floor on the group standard deviation. If the reward variance within a group is effectively flat, the group contributes zero normalized advantage, avoiding unstable updates from uninformative groups. We clip normalized advantages before the policy update to reduce the effect of rare outlier groups. The PPO-style update uses clipping with $\epsilon = 0.2$ and a reference-policy KL coefficient of 0.001. The reward combines answer correctness and response-format quality, with the answer component carrying the dominant weight.

The reward settings are intended to keep the verifier signal outcome-level. The format component only encourages parseable responses, while the answer component determines the main correctness signal. GAZE-RL does not add token-level rewards; it only redistributes the resulting completion-level advantage across generated response positions.

C Compute Resources

We report the computational resources used for the representative training configuration in Tables 6 and 7. The numbers are measured for 200 RL updates under the same rollout and optimization configuration used in the main experiments. These measurements are intended to characterize the additional overhead of the landing-field computation relative to vanilla GRPO; they should not be interpreted as hardware-independent complexity estimates.

C.1 Landing-Field Hyperparameters

The landing-field construction follows Section 4. The default landing strength is $\gamma = 0.70$, modulation strength is $\eta = 1.00$, causal-kernel temperature is $\kappa = 0.50$, log-distance decay is $\lambda_{\text{lag}} = 0.10$, and support temperature is $\chi = 0.25$. The sweeps in Figure 5 vary γ and η while keeping other settings fixed.

Table 7: Peak allocated GPU memory for 200 RL training updates on NVIDIA H200 GPUs. The small increase for GAZE-RL comes from storing selected hidden states, rollout-time logits, and intermediate landing-field statistics during training.

Phase	Vanilla GRPO	GAZE-RL
Peak allocated VRAM	77.7 GB	78.1 GB

Table 8: Default landing-field hyperparameters. The sweep in Figure 5 varies the landing strength and modulation strength around this default configuration.

Quantity	Default value
Landing strength γ	0.70
Causal-support temperature κ	0.5
Log-distance decay λ_{lag}	0.1
Landing modulation strength η	1.0
Support temperature χ	0.25
Visual anchoring weight in evidence score	1.0
Supported branching weight in evidence score	1.0
Landing map	exponential mean-one field

C.2 Complexity and Memory Overhead

The additional computation of GAZE-RL comes from constructing the landing field during training. The method does not modify the model architecture and does not require an auxiliary perception network. It uses hidden states from selected transformer layers, rollout-time logits for entropy, and the visual memory already produced by the multimodal policy.

For a rollout of length T, hidden dimension d, selected layer count $|\mathcal{L}|$, and visual memory size M, visual anchoring with pooled visual prototypes costs $O(|\mathcal{L}|Td)$ after the visual prototype of each selected layer is formed. The supported-branching kernel in its full causal form costs $O(T^2d)$, since each response position compares to its predecessors. This cost is incurred only during training-time landing-field computation. In long-response settings, the same causal support computation can be restricted to a finite predecessor window or approximated with a top-k predecessor set without changing the definition of the landing field.

The memory overhead is mainly due to storing selected-layer response hidden states, selected-layer visual hidden states, and rollout-time logits. In our implementation, we control memory through chunked rollout generation and forward computation. The rollout chunk size is 2, and the forward chunk size is 1. Since the landing field is only used for training and analysis, inference with the trained policy has the same architecture and does not require landing-field computation.

C.3 Evaluation Protocol

For image-side evaluation, we report results on MME (Fu et al., 2025a), MMVP (Tong et al., 2024b), MMBench (Liu et al., 2025b), MMStar (Chen et al., 2024), CV-Bench (Tong et al., 2024a), POPE (Li et al., 2023c), RefCOCO (Yu et al., 2016), Geometry3K (Lu et al., 2021), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), MathVista (Lu et al., 2024), and LogicVista (Xiao et al., 2024). For video-side evaluation, we report results on Video-MME (Fu et al., 2025b), TempCompass (Liu et al., 2024), NExT-QA (Xiao et al., 2021), Video-MMMU (Hu et al., 2025), LongVideoBench (Wu et al., 2024), and VSI-Bench (Yang et al., 2025a). We follow the official evaluation protocol of each benchmark whenever available. Accuracy is used unless the benchmark defines a different official score; MME follows the original summed score, and RefCOCO reports the average over the standard referring-expression splits used in our evaluation.

C.4 Training-Dynamics Measurement

For Figure 2, top-20 mass is computed by sorting response tokens by landing weight and summing the mass assigned to the highest-weighted 20% of positions. Because GRPO applies the same completion-level advantage uniformly to all response positions, its top-20 mass is fixed at 20% by construction. Average response length is measured on the same diagnostic rollout set across training checkpoints and counts generated response tokens only. These curves are used only to analyze the evolution of credit concentration and response length; they are not used as training objectives.

C.5 Visual Evidence Perturbation

For Figure 4, we project response-side landing mass onto visual memory elements using the model's internal cross-modal similarity. For images, visual memory elements are pooled into spatial cells; for videos, they are pooled into temporal memory units after frame packing. Guided drop masks the highest-ranked cells under the projected landing score, while random drop masks the same number of cells uniformly. We sweep the perturbation budget and re-run the model with the perturbed input. The reported curve measures the absolute accuracy drop relative to the full visual input.

This perturbation is used only for analysis. It is not used during training and does not provide supervision to the landing field. The comparison between guided drop and random drop tests whether the projected landing field identifies visual evidence that is necessary for the model's prediction. A larger guided-drop accuracy decrease indicates that the selected visual cells are more causally important for the answer.

C.6 Ablation Average Definition

The ablation and hyperparameter sweep in Figure 5 are conducted on Qwen2.5-VL-7B. Image Avg is the arithmetic mean of MMVP, Geometry3K, MathVista, and LogicVista. Video Avg is the arithmetic mean of Video-MME, TempCompass, Video-MMMU, LongVideoBench, and VSI-Bench. These subsets are selected because they emphasize visually grounded reasoning and temporal evidence usage. The full GAZE-RL point uses the default setting and corresponds to the same Qwen2.5-VL-7B result reported in the main-result tables under this averaged benchmark definition.

C.7 Image-Editing Transfer Probe

For Section 5.4, we use the image-editing system only as a transfer probe. The image-editing backbone is kept fixed, and only the instruction-understanding VLM is replaced by the GAZE-RL-trained Qwen2.5-VL-7B-Instruct. Prompts cover local action changes, weather changes, object color changes, facial attributes, and material-specific edits. The purpose of this probe is to test whether reward landing improves instruction grounding in a generation pipeline outside the main QA benchmarks. We do not use editing-specific rewards, human scoring, or AI judging for this analysis.

D Additional Controlled Analyses

D.1 Matched-seed Baseline Comparison

We repeated the Qwen2.5-VL-7B training of GRPO, DAPO, and GAZE-RL with three matched seed IDs. Each method uses the same initial checkpoint, mixed curriculum, verifier, rollout budget, and evaluation protocol. Entries in Tables 9 and 10 are means and standard deviations over the three complete training runs. Image Avg averages MMVP, Geometry3K, MathVista, and LogicVista; Video Avg averages Video-MME, TempCompass, Video-MMMU, LongVideoBench, and VSI-Bench. These multi-run means are distinct from the representative single-run entries in the main benchmark tables and Figure 5.

Table 9: Matched-seed Qwen2.5-VL-7B image results (mean ± standard deviation over three runs).

Method	MMVP	Geometry3K	MathVista	LogicVista	Image Avg
GRPO	76.6 ± 0.2	39.7 ± 0.7	68.5 ± 0.2	44.6 ± 0.3	57.4 ± 0.2
DAPO	77.1 ± 0.1	42.0 ± 0.4	69.3 ± 0.3	45.7 ± 0.1	58.5 ± 0.1
GAZE-RL	78.6 ± 0.1	47.8 ± 0.5	70.3 ± 0.2	46.0 ± 0.1	60.7 ± 0.1

Table 10: Matched-seed Qwen2.5-VL-7B video results (mean ± standard deviation over three runs).

Method	Video-MME	TempCompass	Video-MMMU	LongVideoBench	VSI-Bench	Video Avg
GRPO	60.8 ± 0.2	69.3 ± 0.3	47.1 ± 0.1	57.0 ± 0.2	31.8 ± 0.2	53.2 ± 0.1
DAPO	61.4 ± 0.2	69.8 ± 0.5	48.0 ± 0.1	58.2 ± 0.3	32.4 ± 0.1	54.0 ± 0.1
GAZE-RL	62.1 ± 0.1	71.9 ± 0.3	48.3 ± 0.1	58.9 ± 0.1	33.0 ± 0.2	54.8 ± 0.1

GAZE-RL leads GRPO and DAPO on both averages in each of the three paired seeds. A paired hierarchical bootstrap with 10,000 resamples over seeds and evaluation examples yields an GAZE-RL minus GRPO difference of $+3.3$ Image Avg (95% CI $[2.7, 3.9]$) and $+1.6$ Video Avg ($[1.2, 2.0]$); against DAPO, the differences are $+2.2$ ($[1.7, 2.7]$) and $+0.8$ ($[0.4, 1.2]$), respectively. These intervals concern the two aggregate metrics and do not establish significance for every individual benchmark.

D.2 Independent Backbone Comparison

To examine an independently developed backbone, we trained Gemma-3-4B-IT (Gemma Team, 2025) with a matched image-side curriculum, verifier, rollout budget, and evaluation protocol for all methods in Table 11. MME-P denotes the MME-Perception score. This experiment tests image-side generalization within Gemma; its scores are not directly pooled with the mixed image-video results of the main tables.

Table 11: Gemma-3-4B image-side comparison under shared training conditions.

Method	MMVP	MMBench	MME-P	MMStar	POPE	AI2D	DocVQA	TextVQA	Geo3K
Gemma-3-4B-IT	62.3	68.4	1197	36.1	79.2	69.3	68.7	59.3	30.4
+ GRPO	64.2	69.9	1200	37.0	80.1	71.0	69.5	60.5	32.8
+ VPPO	64.1	69.0	1202	36.4	79.7	70.8	70.1	60.7	33.3
+ PAPO	63.5	70.5	1193	36.6	79.5	70.2	70.6	58.2	32.9
+ GAZE-RL	65.3	71.4	1206	37.8	80.8	72.3	70.8	62.1	35.8

D.3 Anchoring Validity, Layer Selection, and Pooling

Visual anchoring is an internal similarity proxy, not a direct observation of causal visual use. We evaluate it against a held-out behavioral dependency score based on the change in next-token likelihood after replacing the image with a matched irrelevant image or patch-shuffled counterfactual. This diagnostic is not a training reward and uses no benchmark labels for layer selection. Table 12 compares diagnostic signals, including visual states shuffled across samples as a negative control. AUROC treats the top 20% of dependency positions as positives.

Table 12: Held-out agreement with behavioral visual dependency.

Diagnostic signal	Spearman	Top-20% AUROC
Predictive entropy	0.17	0.58
Cross-modal attention mass	0.29	0.65
All-layer visual similarity	0.41	0.75
Selected-layer anchoring	0.46	0.79
Visual states shuffled across samples	0.03	0.51

We select normalized decoder depths once on a held-out training split to maximize correlation with the dependency score and then fix them across tasks. For Qwen2.5-VL-7B these correspond to 1-indexed layers 14, 18, 22, and 26. The layer-set and pooling training ablations use one matched seed and the same 200-update budget; they should not be read as three-seed uncertainty estimates.

Table 13: Layer-set ablation on Qwen2.5-VL-7B.

Layer set	Spearman	AUROC	Image Avg	Video Avg
Early quarter	0.24	0.65	58.9	53.6
Middle quarter	0.39	0.74	60.0	54.2
Late quarter	0.44	0.78	60.5	54.6
All layers	0.41	0.75	60.3	54.4
Selected depths	0.46	0.79	60.8	54.9

Table 14: Visual-prototype pooling alternatives. The default normalizes each visual token before pooling, as in Eq. (4).

Prototype	Spearman	AUROC	MMVP	RefCOCO	TextVQA	Geo3K
Token-normalize then pool	0.46	0.79	78.7	87.9	85.6	47.9
Pool then normalize	0.44	0.77	78.4	87.5	85.3	47.5
Top-16 local pooling	0.48	0.80	78.6	88.2	85.8	47.3

The default is strongest on this mixed subset and Geometry3K, while Top-16 improves localized grounding on RefCOCO and TextVQA. These results support layer selection and the chosen pooling order without equating representation similarity with guaranteed causal attribution.

D.4 Negative-Advantage Landing

Mean-one normalization preserves scalar rollout-credit mass but does not prove that every negative update penalizes the correct cause of an error. The default $\gamma = 0.70$ gives $\Phi_\gamma(t) \geq 0.30$, so each token retains a uniform-update floor before clipping. We compared sign-asymmetric variants with three matched seeds. Perception Avg averages MMVP, POPE, RefCOCO, TextVQA, and DocVQA; Reasoning Avg averages Geometry3K, ChartQA, MathVista, and LogicVista.

Table 15: Sign-asymmetric landing. γ_+ and γ_- are strengths for positive and negative advantages. Entries are mean $\pm$ standard deviation over three seeds.

Variant	γ_+	γ_-	Perception Avg	Reasoning Avg	Video Avg
GRPO	0.00	0.00	85.8 ± 0.1	59.2 ± 0.2	53.2 ± 0.1
Positive-only landing	0.70	0.00	86.6 ± 0.1	61.2 ± 0.2	54.1 ± 0.1
Negative-damped landing	0.70	0.35	87.0 ± 0.1	62.0 ± 0.1	54.5 ± 0.1
Full GAZE-RL	0.70	0.70	87.2 ± 0.1	62.2 ± 0.1	54.8 ± 0.1

We also audited 400 negative-advantage rollouts, 100 each from Geometry3K, ChartQA, TempCompass, and NExT-QA, using two annotators. Error types were 30.5% perceptual, 48.0% reasoning or calculation, 16.0% mixed, and 5.5% extraction or format. On one fixed GAZE-RL checkpoint and the same annotated rollouts, we recomputed each landing rule over identical spans. Table 16 reports percentages of total response-side landing mass.

Table 16: Negative-rollout mass on annotated spans. The first two columns concern reasoning errors; the last concerns perception errors. Lower mass on a correct visual premise and higher mass on an erroneous span are preferable in these respective cases.

Landing rule	Correct visual premise	First erroneous step	Erroneous visual span
Uniform field	23.4	29.8	28.7
Anchoring only	36.2	32.1	49.3
Entropy without visual support	18.9	46.7	35.8
Full GAZE-RL	20.7	53.8	55.1

The full field shifts mass away from correct visual premises toward an erroneous reasoning step while still localizing erroneous visual spans in perception errors. This is an empirical attribution pattern on the audited set, not a guarantee of causal blame.

D.5 FIXED-CHECKPOINT VISUAL MASKING

Figure 4 includes comparisons across trained checkpoints, whose intrinsic masking robustness can differ. To isolate ranking quality, we repeated deletion on each frozen checkpoint with the same examples, budgets, and decoding, changing only the ranking used to select visual cells. The GRPO checkpoint's landing field is computed post hoc from that checkpoint. Drop@30 is the absolute accuracy decrease at 30% masking. Deletion AUC is the normalized trapezoidal area across 15%, 30%, 45%, 60%, and 75% masking; larger values mean the ranking identifies evidence more necessary to prediction.

Table 17: Mask rankings evaluated within each fixed checkpoint.

Checkpoint	Ranking	Image Drop@30	Image AUC	Video Drop@30	Video AUC
GRPO	Random	3.5	5.3	2.3	3.6
GRPO	Attention	4.8	6.6	3.4	4.8
GRPO	Shuffled landing	3.6	5.4	2.4	3.7
GRPO	Landing field	6.4	8.0	5.3	6.7
GAZE-RL	Random	4.4	6.1	3.0	4.4
GAZE-RL	Attention	5.9	7.4	4.3	5.6
GAZE-RL	Shuffled landing	4.5	6.2	3.1	4.5
GAZE-RL	Landing field	8.8	10.5	6.9	8.2

Landing-field ranking produces the largest drops within both checkpoints. The controlled comparison supports ranking quality without attributing cross-checkpoint differences solely to the landing field.

D.6 CAUSAL-KERNEL SCALING

The full predecessor kernel in Eq. (16) costs $O(T^2 d)$ during training-time landing construction. We profiled one rollout on an NVIDIA H200 using the same four selected layers, with 200 latency profiles per length. At lengths up to the submitted 512-token generation cap, end-to-end overhead is measured on complete RL steps. The 1024- and 2048-token rows use cached rollout states, so no end-to-end training overhead is reported for those lengths.

Table 18: Measured landing-kernel cost by response length. Extra peak VRAM is relative to GRPO.

Tokens	Landing latency (ms)	Extra VRAM (GB)	End-to-end step overhead
128	4.7 ± 0.3	0.12	$0.2\% \pm 0.1\%$
256	10.9 ± 0.5	0.21	$0.4\% \pm 0.1\%$
512	27.6 ± 1.2	0.40	$0.8\% \pm 0.2\%$
1024	91.8 ± 3.6	0.74	Cached-state profile only
2048	342.7 ± 11.9	1.46	Cached-state profile only

A finite causal predecessor window reduces the quadratic kernel cost to $O(TWd)$ for window size W. This is an efficiency option for longer generations; the main benchmark results use the full kernel at the 512-token cap. Deployment inference uses the unchanged backbone and performs no landing-field computation.

E LIMITATIONS

Dependence on internal representation quality. GAZE-RL assumes that the model's internal cross-modal geometry contains usable information about where visual evidence is being used. This assumption is reasonable for strong multimodal backbones, but it can fail when the backbone itself has weak visual-text binding, poor OCR grounding, or unstable video memory. In such cases, the landing field may concentrate credit on positions that appear visually anchored in representation space but do not correspond to genuinely relevant evidence. Thus, GAZE-RL should be viewed as a way to better allocate an existing learning signal inside a capable policy, rather than as a replacement for strong multimodal pretraining or reliable visual representation learning.

Outcome reward remains the source of update direction. The landing field determines where the completion-level advantage is absorbed, but it does not determine whether the rollout should be rewarded or penalized. The sign and magnitude of the update still come from the verifier-derived advantage. If the verifier assigns an incorrect or biased reward, GAZE-RL can make the resulting update more concentrated, but it cannot correct the reward itself. This limitation is inherent to methods that redistribute outcome-level rewards without changing the reward model. More reliable verifiers, uncertainty-aware reward models, or calibrated rejection mechanisms remain important for robust multimodal RL.

F FUTURE WORK

From token landing to structured visual credit. A natural next step is to extend reward landing from response-token fields to structured visual credit fields. Instead of only assigning update mass to generated positions, future methods could jointly infer which response tokens, image regions, objects, video frames, and temporal events should absorb the learning signal. This would make the connection between reasoning text and visual evidence more explicit, especially for long-video understanding, spatial reasoning, and multi-object interaction.

Joint modeling of reward quality and landing confidence. In the current formulation, the verifier supplies the trajectory-level advantage and the landing field supplies positionwise allocation. Future work could model these two uncertainties jointly: a rollout with uncertain reward, weak visual anchoring, or flat landing contrast should induce a more conservative update. Such a direction could combine reward calibration, verifier ensembles, and internal landing confidence, allowing the optimizer to decide not only where reward should land, but also how strongly it should be trusted.

More efficient landing computation. The causal support computation can be made more scalable through sparse predecessor selection, windowed kernels, low-rank state summaries, or online updates during generation. These approximations are especially relevant for long-context video reasoning, where both generated responses and visual memories can become large. An efficient implementation

Table 19: Existing assets used in this work. Licenses and terms are taken from the official model, dataset, or project release pages when available.

Asset	Type	Usage in this paper	License / terms
Qwen2.5-VL-3B / Qwen2.5-VL-7B (Bai et al., 2025a)	Model backbone	Main multimodal policy backbones	Apache-2.0 official release
InternVL3.5-8B (Wang et al., 2025b)	Model backbone	Main multimodal policy backbone	Apache-2.0 official release
ViRL39K (TIGER-Lab, 2025)	Training dataset	Image-side RL curriculum	MIT official dataset release
Video-R1 (Feng et al., 2025)	Training dataset / baseline	Video-side RL curriculum and video RL baseline	Apache-2.0 official dataset release
Qwen-Image-Edit	Image-editing model	Qualitative transfer probe	Apache-2.0 official release
MME (Fu et al., 2025a), MMVP (Tong et al., 2024b), MMBench (Liu et al., 2025b), MMStar (Chen et al., 2024)	Evaluation benchmarks	Image-side general and diagnostic evaluation	Public benchmark terms
CV-Bench (Tong et al., 2024a), POPE (Li et al., 2023c), RefCOCO (Yu et al., 2016)	Evaluation benchmarks	Visual perception, hallucination, and grounding evaluation	Public benchmark terms
Geometry3K (Lu et al., 2021), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), AI2D (Kembhavi et al., 2016)	Evaluation benchmarks	OCR-heavy, diagram, and visual reasoning evaluation	Public benchmark terms
ChartQA (Masry et al., 2022), MathVista (Lu et al., 2024), LogicVista (Xiao et al., 2024)	Evaluation benchmarks	Chart, mathematical, and logical reasoning evaluation	Public benchmark terms
Video-MME (Fu et al., 2025b), TempCompass (Liu et al., 2024), NExT-QA (Xiao et al., 2021)	Evaluation benchmarks	Video understanding and temporal reasoning evaluation	Public benchmark terms
Video-MMMU (Hu et al., 2025), LongVideoBench (Wu et al., 2024), VSI-Bench (Yang et al., 2025a)	Evaluation benchmarks	Knowledge-intensive, long-video, and spatial memory evaluation	Public benchmark terms
PyTorch, Transformers, vLLM	Software libraries	Training, inference, and rollout generation	Open-source software licenses

would make reward landing easier to combine with long-video memory, retrieval-augmented visual context, and streaming multimodal policies. Combining these mechanisms with stronger verifiers is another direction for longer-horizon multimodal reasoning.

Beyond verifiable QA. Reward landing is most directly evaluated here under verifiable image and video QA settings, but the underlying idea is not tied to multiple-choice or short-answer evaluation. Future work could test the same principle in instruction-guided image editing, interactive visual dialogue, multimodal planning, robotics, and tool-augmented visual reasoning. These settings may require richer rewards and stronger safety constraints, but they also make the core question more important: when a long multimodal trajectory succeeds or fails, which parts of the trajectory should actually receive the learning signal?

G Existing Assets and Licenses

Table 19 lists the main existing assets used in this work. We use public model releases, public datasets, public benchmarks, and standard open-source software libraries. For each asset, we cite the corresponding paper or project and follow the license or terms specified by the official release. We do not introduce a new dataset in this paper.

For benchmarks and software packages whose licenses are maintained by their official repositories, we follow the corresponding official terms of use. The released code accompanying this paper is intended for research use and should be used together with the licenses of the underlying models, datasets, and software dependencies.

H BROADER IMPACT

This work studies how to improve the allocation of existing reinforcement-learning signals in multimodal reasoning models. The potential positive impact is improved grounding: by concentrating policy updates on response positions that are visually anchored and decision-bearing, the method may reduce weakly grounded reasoning, improve evidence sensitivity, and make image and video reasoning systems less dependent on redundant textual continuation. This can benefit applications that require faithful multimodal understanding, such as visual question answering, educational diagram reasoning, chart analysis, accessibility tools, and long-video retrieval or summarization.

At the same time, stronger multimodal reasoning systems can also be misused. Improvements in visual grounding and image/video understanding may strengthen systems used for surveillance, sensitive visual inference, misinformation analysis, or manipulative image and video editing workflows. The method itself does not introduce a new data source, does not collect private user data, and does not add a new generative capability beyond the underlying models, but it can improve the reliability of existing multimodal policies. Users of released code or checkpoints should therefore follow the safety and usage restrictions of the underlying base models, datasets, and deployment environments.

The evidence-sensitivity analyses in this paper are diagnostic tools rather than safety guarantees. They indicate whether the model's prediction depends on the visual evidence selected by the landing field, but they do not certify factual correctness, fairness, privacy preservation, or robustness under adversarial inputs. Future work should evaluate reward landing under broader safety, fairness, and robustness settings, especially before applying it to high-stakes domains.

I USE OF LARGE LANGUAGE MODELS

This section gives details for the AI use statement at the end of the main text.

Implementation and experiment execution. AI coding assistants helped implement components specified by the authors and run training and evaluation jobs. The authors reviewed and tested all AI-assisted code before using it in reported experiments, and checked the configurations, logs, and outputs of every run behind a reported number.

Writing and LaTeX preparation. AI tools helped organize the LaTeX source for the ICLR 2027 template and helped check and format references; the authors checked every changed reference against its original source. We also used AI tools to polish the grammar, wording, and flow of text written by the authors; no part of the paper was drafted by AI, and the authors verified and edited all polished text. An example editing prompt is:

> *Please check the grammar, spelling, word choice, and sentence flow of the following LaTeX text. Keep all equations, labels, references, and macros unchanged, and avoid changing the scientific content.*