DEEPO: DUAL-ENTROPY ENHANCED POLICY OPTIMIZATION FOR HALLUCINATION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) is widely used to sharpen reasoning in multimodal large language models (MLLMs), yet its effect on hallucination is uneven. We trace this to two weak points in the *correction chain* from reward to parameter update. At the rollout level, hard queries—those with high semantic entropy—frequently produce unanimously wrong sample groups, collapsing the group-relative advantage to zero exactly where hallucination risk is highest. At the optimization level, confident-but-wrong tokens are gradient-invisible: a categorical policy's expected score-gradient norm vanishes as its distribution sharpens, so the predictions that most need correction receive the weakest updates. We propose Dual-Entropy Enhanced Policy Optimization (DEEPO), a dual-stage enhancement combining signal variance regularization with gradient preconditioning: semantic-entropy-triggered expert prefixes inject grounded continuations on high-uncertainty queries, providing direct supervision and restoring advantage variance, while advantage-sign-aware Rényi preconditioning counteracts logit-level saturation so correction reaches confident errors in the operational confidence regime. Both branches improve over GRPO individually; their interaction is statistically significant on VideoMMMU—the most complex long-horizon task in our evaluation suite ($+4.0$, 95% CI $[1.1, 6.9]$)—and additive elsewhere. DEEPO reduces hallucination while preserving accuracy and training stability.

1 INTRODUCTION

Multimodal large language models (MLLMs) have greatly advanced vision–language understanding and multi-step reasoning (Yin et al., 2024; Wu et al., 2024; Li et al., 2024b). However, hallucination—confident but incorrect or semantically inconsistent outputs—remains a key obstacle (Fu et al., 2025a), motivating work on its causes and mitigation (Mahmoud et al., 2025).

While prior work attributes hallucination to dataset noise (Bang et al., 2025), architectural bottlenecks (Cossio, 2025), or decoding errors (Wu et al., 2025a), these explanations target static properties of the model or input (Anh-Hoang et al., 2025). Hallucination is also shaped during training: SFT overfits surface patterns by imitation, whereas RL can encourage guessing under uncertainty through reward-driven rollouts (Lin et al., 2024; Mahmoud et al., 2025; Shenfeld et al., 2025)—an effect that is heterogeneous rather than universal (in our measurements GRPO improves over its base model on POPE and HallusionBench yet slightly degrades on VideoHallucer; Table 1): the issue is not that RL uniformly worsens hallucination, but that its corrective machinery is unreliable precisely on confident errors. We therefore focus on the RL stage, where the correction chain from scalar reward to parameter update has two weak points that a training method can, and we argue should, address jointly.

(I) Hard queries starve the correction signal. In Fig 1, hallucinated answers concentrate in regions of high semantic entropy—exactly the queries on which sampling is most likely to fail unanimously: with binary rewards an all-wrong group has $\mu = 0$, $\sigma \to 0$, so every response receives $A_i \approx 0$ and the iteration extracts no gradient (Kalai et al., 2025; Khan & Fu, 2024). An *expert hint*—a truncated, answer-free chain-of-thought (CoT) prefix—conditions one continuation on grounded visual reasoning, injecting direct supervision on hard queries and, under a shared group baseline, a high-reward anchor that restores nonzero advantage variance, so failed responses receive genuine negative advantages instead of the degenerate zero; prefixes come from dataset traces (Feng et al., 2025) or the policy's own verified rollouts (Sec. 4.1), never a test-time oracle.

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.

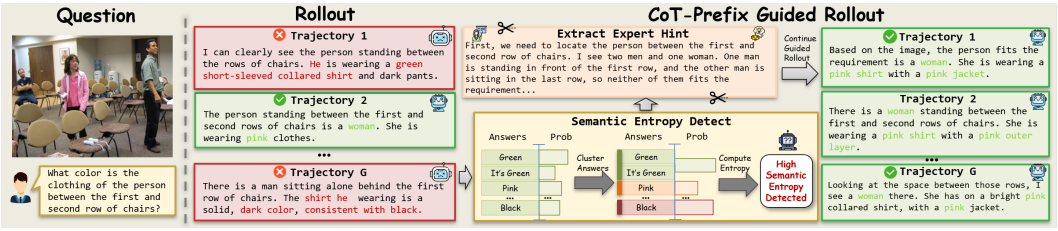


Figure 1: Model guessing and expert-hint-guided correction. Under uncertain visual cues the model "guesses," yielding high semantic entropy and frequently unanimous failures; the expert hint grounds one continuation on the visual evidence.

(II) Confident errors are gradient-invisible. Even a perfect advantage signal cannot correct confident hallucinations, because the delivery mechanism vanishes: for a softmax policy, $\mathbb{E}\|\nabla_z \log \pi\|^2 = 1 - e^{-H_2}$ (Prop. 4.1), and a gradient-masking pilot confirms that high-entropy tokens carry the bulk of the update signal (Fig. 3, App. A). Confident predictions—and confident hallucinations in particular, which are precisely the errors that survive decoding—thus receive vanishing corrective updates, while uncertain regions dominate the gradient budget.

Why the two stages are complementary. Rescued advantage variance is of little use if correction cannot reach confident errors, and curvature-compensated updates cannot manufacture signal from a collapsed group. DEEPO therefore combines the two: semantic entropy detects the queries whose groups are likely to collapse and triggers grounded continuations, while sign-aware Rényi entropy parameterizes how strongly each token is corrected or reinforced—signals at different levels, not interchangeable (App. D). The interaction is positive on all six benchmarks and statistically significant exactly where both mechanisms are most stressed—long-horizon VideoMMMU ($+4.0$, 95% CI $[1.1, 6.9]$; Table 2)—and consistent with additivity elsewhere.

Our contributions are summarized as follows:

- **Diagnosis.** Two weak points in the RL correction chain: on hard queries, unanimously wrong groups collapse the group-relative advantage to zero, and confident errors are gradient-invisible (Prop. 4.1).
- **Dual-stage enhancement.** DEEPO pairs entropy-triggered expert prefixes at the rollout stage with advantage-sign-aware Rényi gradient preconditioning at the optimization stage; the scaling weight is a conservative damped proximal step with guaranteed ascent of a sample-wise surrogate (Thm. 4.2), and sign-symmetric rules provably cannot realize its allocation (Remark 4.3).
- **Evidence.** Complementary single-branch gains with a positive interaction—significant on VideoMMMU, additive elsewhere—supported by design, fairness, and mechanism-diagnosis controls, a self-generated-hint variant, multi-seed evaluation, and efficiency accounting.

2 Related Work

Hallucination mitigation in MLLMs. Hallucination remains a persistent challenge for MLLMs Bai et al. (2024), addressed through attention calibration Xu et al. (2025); Zhu et al. (2025), decoding-time corrections Leng et al. (2023); Qu et al. (2024), grounding supervision Zhao et al. (2024); Chen et al. (2023), and multi-object or cross-image analyses Zhou et al. (2023); Chen et al. (2024b); Wang et al. (2025a); He et al. (2025). These are post-hoc or decoding-level fixes; we instead detect hallucination-prone uncertainty during rollout and ground the generation.

Training dynamics, rewards, and uncertainty. RL can exacerbate hallucinations when rewards are weakly tied to factual correctness He et al. (2024); Yang et al. (2025b) or encourage shortcut exploitation Miao et al. (2025); Jia et al. (2025). Token-level credit assignment has been approached via perception-aware rewards Wang et al. (2025c); Huang et al. (2025a) and verifiable-data synthesis Wu et al. (2025b); these modify rewards or data, whereas DEEPO leaves the reward unchanged and modifies rollout conditioning and gradient delivery. Uncertainty signals have mainly served detection or calibration Farquhar et al. (2024); Manakul et al. (2023); Fadeeva et al. (2024); Zhang

et al. (2024); Qiu & Miikkulainen (2024); Niu et al. (2024); Jiang et al. (2023), not for controlling how uncertainty shapes rollouts and updates during RL fine-tuning.

Positioning against three adjacent lines. *(i) Entropy-modulated policy gradients* (EMPG Wang et al. (2025b) and relatives) reshape token updates via uncertainty but are sign-symmetric (Remark 4.3) with no rollout-time mechanism. *(ii) Hallucination-targeted preference optimization* (RLHF-V Yu et al. (2024a), RLAIF-V Yu et al. (2024b), HA-DPO Zhao et al. (2023), hallucination-targeted DPO Fu et al. (2025a), SymMPO Liu et al. (2025), POVID Zhou et al. (2024)) acts on offline preference pairs with no generation-time intervention; DEEPO is dynamics-centric and on-policy. *(iii) Expert-guided RL* Zhang et al. (2025b); Fu et al. (2025b) mixes complete off-policy expert traces into on-policy RL at a static ratio—we adopt the confidence weight ϕ_t of Zhang et al. (2025b)—whereas DEEPO uses expert text only as truncated, answer-free prefixes, triggered per query by an adaptive semantic-entropy threshold, with hint dropout and the prefix masked out of the RL loss. To our knowledge, DEEPO is the first to pair answer-level semantic entropy as a rollout-time trigger for expert conditioning with a token-level sign-aware Rényi scaling law.

3 PRELIMINARIES

GRPO. We build on Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a critic-free RL algorithm: for each query q the policy samples G responses with verifiable rewards r_i (task accuracy plus format compliance, following Feng et al. (2025)); the group-relative advantage $A_i = (r_i - \text{mean})/\text{std}$ is broadcast to every token, and the policy is updated with a PPO-style clipped surrogate plus a KL penalty to a frozen reference. Two properties matter: (i) advantages are *sequence-level*—the signal cannot distinguish which tokens caused failure; (ii) with binary rewards a unanimous group carries no signal—if all responses are wrong ($r_i = 0$) then $\mu = 0$, $\sigma \to 0$, and $A_i \approx 0$ (symmetrically for an all-correct group), so the iteration extracts no gradient precisely from the hardest queries.

Semantic entropy. Following Farquhar et al. (2024), semantic entropy measures guessing at the level of meaning: draw N generations, partition them into K meaning-equivalent clusters $\{C_k\}$, and compute $H_s(q) = -\sum_k P(C_k \mid q) \log P(C_k \mid q)$ over the cluster distribution; a high $H_s(q)$ indicates semantic-level guessing and elevated hallucination risk. We estimate H_s from the G rollouts already sampled for the current iteration ($N{=}G$ by default; App. A.3 co-varies both), so estimation adds no extra sampling.

Rényi entropy. At order $\alpha = 2$, the collision entropy $H_2 = -\log \sum_y \pi(y \mid q, y_{<t})^2$ characterizes token-level sharpness and governs the expected score-gradient norm (Prop. 4.1); it is the quantity we use for gradient scaling.

4 METHOD

DEEPO addresses the two weak points identified in Sec. 1 as a dual-stage enhancement (Fig. 2): Sec. 4.1 gives the rollout stage, where semantic-entropy-triggered expert prefixes provide grounded supervision and reshape in-group advantages on hard queries (signal repair); Sec. 4.2 gives the optimization stage, where advantage-sign-aware Rényi gradient preconditioning delivers correction to confident tokens (delivery repair); Sec. 4.3 states the complementarity hypothesis and its falsifiable predictions. Appendix D relates the two entropy quantities through predictive uncertainty.

4.1 EXPERT HINT AS SIGNAL REPAIR

The first weak point in the correction chain is a starved learning signal on hard queries: when semantic entropy is high the sampled group fails frequently, and in the extreme unanimously ($r_i = 0$ for all i, hence $\mu = 0$, $\sigma \to 0$, and $A_i = (r_i - \mu)/(\sigma + \epsilon) \approx 0$ for every token), so no gradient is extracted exactly where hallucination risk is highest. Entropy-triggered expert prefixes counteract this through two channels, separated explicitly because the evidence weighs them differently (Sec. 5.4): *direct grounding and supervision* (the prefix steers the policy onto a grounded reasoning path, reinforced by RL and distilled through $\mathcal{L}_{\text{aux}}$, Eq. 2) and *in-group advantage reshaping* (under a shared group baseline, the high-reward grounded continuation raises the group mean, so unhinted failures receive real negative advantages instead of the degenerate $A_i \approx 0$ of a collapsed group).

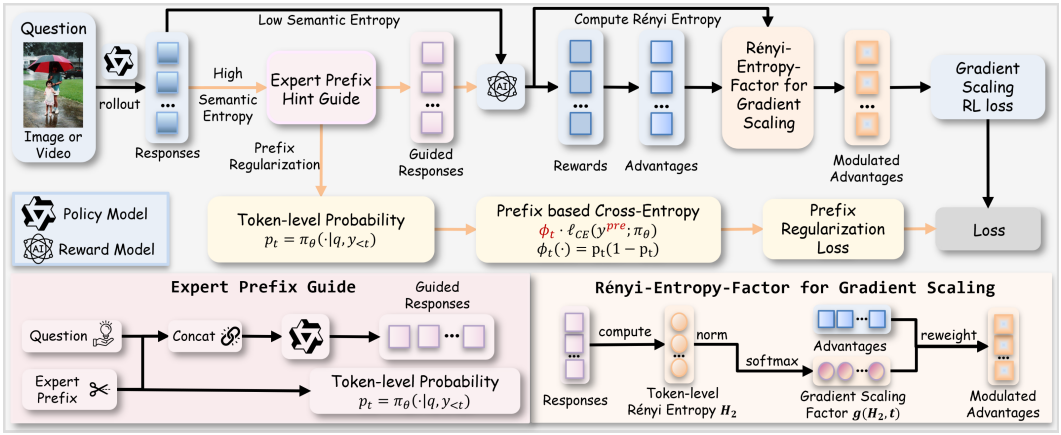


Figure 2: Overview of Dual-Entropy Enhanced Policy Optimization. High semantic entropy activates ϕ_t-weighted prefix regularization and prefix-guided rollouts, while the Rényi entropy factor scales policy gradients; the objective unifies both.

Hint source. Hints are truncated prefixes of model-generated solution trajectories from the Video-R1 dataset (Feng et al., 2025); they condition generation during training only (never at inference) and additionally receive the lightweight ϕ_t-weighted cross-entropy of Eq. 2, so DEEPO is a hybrid of RL on self-generated tokens and auxiliary prefix supervision, ablated separately (Table 7, App. A.4; Table 2). The method is not tied to this dataset: Sec. 5.4 evaluates a self-generated variant (DEEPO-self) using the prefix of the best verified-correct rollout from the same group.

Trigger and prefix construction. We estimate H_s from the N rollouts already generated for the current iteration (default $N{=}G$), clustering answer-level commitments by pairwise NLI entailment (DeBERTa-MNLI; App. G). A query is marked as challenging when $H_s(q)$ exceeds an adaptive threshold $\hat{\tau}$, initialized at τ_s (default 0.8) and updated per iteration by exponential moving average, $\hat{\tau} \leftarrow (1-\beta)\hat{\tau} + \beta\overline{H}_s$ with $\overline{H}_s$ the batch mean; the trigger gates the intervention to queries whose groups fail frequently, unlike a rate-matched random trigger (Table 3). Each expert trace is split at the final-answer marker into a thought segment y^*_{thought} (observations and chain-of-thought) and an answer segment y^*_{ans} (conclusion and options); the prefix is drawn *exclusively from the thought segment*, with uncertainty-proportional length hard-capped away from the answer boundary:

$$L_h = \min\left(\left\lfloor \alpha_{\max} \cdot \frac{H_s(q)}{H_s^{\max}} \cdot L_{\text{thought}} \right\rfloor,\ L_{\text{thought}} - 1\right), \tag{1}$$

where $H_s^{\max} = \log N$, $L_{\text{thought}} = |y^*_{\text{thought}}|$, and $\alpha_{\max} \in (0, 1)$ (default 0.5). Since $\alpha_{\max} < 1$ and $H_s(q) \le H_s^{\max}$, $L_h \le L_{\text{thought}} - 1$ for every trace: **the prefix provably excludes the final answer and conclusion tokens, and an NLI entailment filter additionally discards any prefix that semantically entails the gold answer (App. G).** The policy then generates the continuation $y_q^{\text{rem}} \sim \pi_\theta(\cdot \mid q \oplus y_q^{\text{pre}})$; the full trajectory is scored by the verifiable reward (one extra generation pass for triggered queries; cost in Table 14, App. G).

Two channels of benefit. For a triggered query, the hint-conditioned continuation and the first-pass unhinted rollouts *share a single group-relative baseline*, and the prefix additionally receives $\mathcal{L}_{\text{aux}}$. *(i) Direct grounding and supervision*: the continuation is generated by the policy itself and reinforced when successful, while $\mathcal{L}_{\text{aux}}$ distills the prefix tokens; this channel operates whenever a grounded trace is available—including the policy's own correct rollouts (DEEPO-self)—and does not depend on group composition. *(ii) In-group advantage reshaping*: under the shared baseline the high-reward continuation raises μ and restores advantage variance, so on a query whose unaided group would have failed unanimously, the failed unhinted rollouts—frozen at zero advantage—now receive substantial negative advantages; this channel punishes failures with real updates and by construction cannot separate two equally rewarded responses, and it requires the shared baseline (Table 4(c)). Hint dropout regularly reverts triggered queries to purely unhinted groups.

Empirical diagnosis of advantage collapse. On the training distribution, the fraction of all-wrong groups rises monotonically from 12% (lowest H_s decile) to 63% (highest), and on triggered queries whose first-pass group was unanimously wrong (48% of triggered queries), hint activation converts 60% to mixed groups, moving the mean advantage of failed unhinted rollouts from ≈ 0 to -0.212 and the within-group advantage variance from 0 to 0.6 (Table 4, Sec. 5.4). One coverage boundary is worth stating plainly: a policy that is confidently and *consistently* wrong produces *low* semantic entropy and is not triggered; and if such a group is also unanimously wrong, its advantage vanishes, so the optimization stage—which can only reallocate a nonzero advantage, since $g^\star(H_{2,t}, A_t)\, A_t = 0$ when $A_t = 0$—has no signal to amplify either. Neither branch manufactures reward information the rollouts do not contain; this boundary is shared by every outcome-reward RL method, and the delivery branch acts on confident errors inside groups that carry nonzero advantage variance.

Optimization treatment of the prefix. The prefix is masked out of the RL loss (policy gradients update only self-generated tokens), and the guidance is distilled by an auxiliary prefix loss with confidence weight $\phi_t = \pi_\theta(y^{\mathrm{pre}}_{q,t} \mid q, y^{\mathrm{pre}}_{q,<t})\,(1 - \pi_\theta(y^{\mathrm{pre}}_{q,t} \mid q, y^{\mathrm{pre}}_{q,<t}))$ (Zhang et al., 2025b):

$$\mathcal{L}_{\mathrm{aux}} = \mathbb{E}_{(q,\, y^{\mathrm{sol}})\sim\mathcal{D}}\Big[\tfrac{\mathbb{1}\{L_h>0\}}{L_h}\textstyle\sum_{t=1}^{L_h}\phi_t\,\ell_{\mathrm{CE}}\big(y^{\mathrm{pre}}_{q,t};\, \pi_\theta(\cdot \mid q, y^{\mathrm{pre}}_{q,<t})\big)\Big], \tag{2}$$

where ϕ_t concentrates learning on tokens of intermediate confidence; $\mathcal{L}_{\mathrm{aux}}$ is a weak regularizer and leaves the RL objective unchanged.

4.2 Rényi Gradient Scaling as Delivery Repair

A diagnostic signal has limited effect if it cannot reach the parameters responsible for confident hallucinations. We now make the delivery bottleneck precise and derive the scaling rule, starting from the second-order Rényi entropy of the token-level policy, $H_2(\pi(\cdot \mid q, y_{<t})) = -\log \sum_{y\in\mathcal{V}} \pi(y \mid q, y_{<t})^2$, which for softmax policies pins the expected magnitude of the score-function gradient:

Proposition 4.1 (Score-gradient norm via collision entropy; standard identity)**.** *For a policy π_θ parameterized by logits $z_\theta(q, y_{<t})$,*

$$\mathbb{E}_{y_t\sim\pi_\theta(\cdot|q,y_{<t})}\big[\big\|\nabla_{z_\theta(q,y_{<t})}\log\pi_\theta(y_t \mid q, y_{<t})\big\|^2\big] = 1 - \exp\big(-H_2(\pi_\theta)\big). \tag{3}$$

This is the trace of the categorical Fisher information in natural parameters (proof in App. B): hallucinated outputs are often *confidently wrong*, and by Eq. 3 such low-entropy predictions induce near-zero score gradients, so RL provides the weakest correction exactly where it is most needed.

From the identity to a weighting rule. EMPG (Wang et al., 2025b) reweights token updates with an entropy-only factor, which is symmetric in update direction. We instead derive a sign-aware weight from a *sample-wise* proximal analysis in logit space, conditioned on the current batch of *already generated* trajectories: every A_t is a fixed empirical scalar, and no expectation over $y_t \sim \pi_\theta$ is taken with A_t inside—the only distribution-level quantity is the Fisher matrix $F_t = \mathrm{diag}(\pi) - \pi\pi^\top \succeq 0$, a deterministic function of the policy at the current state. Defining

$$\tau_t := \mathrm{Tr}(F_t) = 1 - \exp(-H_{2,t}), \tag{4}$$

the following theorem (proved in App. C) justifies a curvature-compensated weight under softmax saturation; the subsequent remark states what entropy-only rules cannot do.

Theorem 4.2 (Curvature-compensated weight as a conservative damped proximal step)**.** *Fix a generated batch (hence all advantages $\{A_t\}$ are fixed empirical scalars, $A_t \neq 0$) and a decoding step t with state $s_t = (q, y_{<t})$, sampled token y_t, and score gradient $g_t = \nabla_{z_t}\log\pi_\theta(y_t \mid s_t)$. Restrict one-step logit updates to the score direction, $\Delta z_t = \eta\, w_t A_t g_t$ with $w_t \geq 0$, and consider the sample-wise second-order proximal surrogate with Levenberg–Marquardt damping $\varepsilon_0 > 0$, $Q_t(w_t) = A_t\, g_t^\top \Delta z_t - \frac{\beta_{\mathrm{kl}}}{2}\,\Delta z_t^\top (F_t + \varepsilon_0 I)\,\Delta z_t$. Then:*

- *(i)* Sample-wise curvature bound. *$g_t^\top F_t g_t \leq \tau_t\|g_t\|^2$, so the trace-scalarized quadratic $\widetilde{Q}_t(w) = \eta A_t^2\|g_t\|^2\, w - \frac{\beta_{\mathrm{kl}}\eta^2}{2}A_t^2\|g_t\|^2(\tau_t+\varepsilon_0)\, w^2$ is a tangent minorant of Q_t on $\mathbb{R}_+$.*
- *(ii)* Damped optimum. *The minorant maximizer $w_t^{\varepsilon_0} = \big(\beta_{\mathrm{kl}}\,\eta\,(\tau_t+\varepsilon_0)\big)^{-1} \leq (\beta_{\mathrm{kl}}\,\eta\,\varepsilon_0)^{-1} =: w_{\max}$ is capped and guarantees strict sample-wise ascent $Q_t(w_t^{\varepsilon_0}) \geq A_t^2\|g_t\|^2/\big(2\beta_{\mathrm{kl}}(\tau_t+\varepsilon_0)\big) > 0 = Q_t(0)$.*

(iii) Softmax-saturation regime. *With $\delta_t = 1 - \pi_\theta(y_t \mid s_t)$, $\|g_t\|^2 \le 2\delta_t^2$ and $\tau_t \le 2\delta_t$: as $\delta_t \to 0$ the unweighted step stalls ($Q_t'(0) = O(\delta_t^2)$), the compensated weight grows as $\Omega(\delta_t^{-1})$ up to its cap, and the certified ascent decays only as $\Theta(\delta_t)$ for $\delta_t \gg \varepsilon_0$; in the deep tail $\delta_t \ll \varepsilon_0$ the bound reverts to $O(\delta_t^2/\varepsilon_0)$—a rate improvement over the operational confidence band, not an asymptotic removal of saturation.*

Three readings are essential: the statement is *sample-wise and local* (A_t never appears inside an expectation over $y_t \sim \pi_\theta$; one step at one token of a fixed trajectory, not the global return); the damping is *load-bearing* (the $+\varepsilon_0$ term, exactly the stabilizer in Eq. 5, caps the weight at $w_{\max}$); and the theorem is *sign-blind* (Q_t sees A_t only through A_t^2), justifying the *magnitude* of the corrective weight, not its sign asymmetry.

Remark 4.3 (Entropy-only rules cannot prioritize correction). If $w_t = \nu(H_{2,t})$ is any advantage-sign-independent rule, two tokens with $H_{2,a} = H_{2,b}$ and $A_a < 0 < A_b$ necessarily get $w_a = w_b$, so no such rule can implement correction-priority allocation $w_a > w_b$; under any fixed batch budget, a budget-preserving perturbation strictly increases the corrective mass (App. C). The division of labor is explicit: Theorem 4.2 provides a magnitude justification for inverse-curvature weighting under a local quadratic surrogate, while the sign-asymmetric assignment is an empirical heuristic for correction priority, validated by the weighting-rule ablation (Table 3).

Guided by Theorem 4.2 and Remark 4.3, we apply a sign-aware Rényi factor to the advantage:

$$g^\star(H_{2,t}, A_t) = \psi(H_{2,t}, A_t) \Big/ \bar{\psi}, \qquad \psi(H_{2,t}, A_t) = \begin{cases} \big(\tau_t + \varepsilon_0\big)^{-1}, & A_t < 0 \quad \text{(correction)}, \\ \tau_t, & A_t \ge 0 \quad \text{(reinforcement)}, \end{cases} \tag{5}$$

where $\bar{\psi}$ is the batch mean over all tokens, $\tau_t = 1 - e^{-H_{2,t}}$, and $\varepsilon_0 = 1/w_{\max}$ with weight cap $w_{\max} = 20$ (App. G). The negative branch is the damped proximal weight of Theorem 4.2(ii) with $\beta_{\mathrm{kl}}, \eta$ absorbed by batch normalization. Its effect is best read through *effective step magnitude*, not token selection: a confident token has $\|g_t\| = O(\delta_t)$, so its unweighted logit step collapses as the prediction saturates, and weighting by $(\tau_t + \varepsilon_0)^{-1} \approx (2\delta_t + \varepsilon_0)^{-1}$ recovers the effective step $\|\Delta z_t\| \propto w_t \|g_t\|$ to a δ_t-independent scale wherever $\delta_t \gg \varepsilon_0$ (Thm. 4.2(iii)). The rule thus unstalls optimization on confident errors within the operational confidence regime; it does not remove saturation asymptotically (at fixed $\varepsilon_0 > 0$ the weight saturates at $w_{\max}$ and the step still vanishes as $\delta_t \to 0$). This amplification is *not* error-selective by construction—low-entropy structural tokens (articles, connectives) in failed trajectories are amplified as well—but three properties bound that collateral: the cap at $w_{\max}$ keeps amplified steps proportional to an already-small $\|g_t\|$; the KL penalty to the frozen reference is deliberately left unweighted (Sec. 4.3), anchoring high-frequency tokens; and $\mathcal{L}_{\mathrm{aux}}$ adds an orthogonal cross-entropy pull on fluent text. The residual selectivity is empirical: on span-annotated failed rollouts the updates achieve a 1.75–$1.80\times$ per-token *weight* enrichment on hallucinated spans (95% CIs exclude 1; App. A.5), i.e. a 1.15–$1.25\times$ enrichment of the effective logit-step norm—a deliberately bounded effect: restoring a stalled gradient by roughly a fifth suffices, over thousands of updates, to bias the policy away from suboptimal deterministic attractors. The positive branch is a separate, deliberately simple heuristic: reinforce in proportion to the available gradient signal $\mathbb{E}\|g\|^2 = \tau_t$, dampening reinforcement of confident tokens while leaving high-uncertainty tokens near the unweighted baseline; confident correct tokens receive small positive weights, slowing further sharpening and acting against entropy collapse. Batch-mean normalization keeps the weights centered at one, so the rule reallocates update magnitude across tokens rather than rescaling the batch mean; $g^\star > 0$ preserves the advantage sign, so the PPO clip *trigger set* is unchanged.

Interaction with downstream optimization (Adam and PPO). Theorem 4.2 characterizes a single damped proximal step; the deployed pipeline composes it with Adam and PPO clipping, which the theorem does not model. Two observations keep the design meaningful: $g^\star$ acts in the *spatial* dimension (per token, within a batch) whereas Adam's normalization acts *temporally* (running statistics across steps), rescaling each token's update by roughly a common factor so the relative, mean-one enrichment survives and steers subsequent updates; and $g^\star > 0$ leaves the clip trigger set unchanged, so clipping intervenes on trajectories, not on the weighting. We therefore read the theorem as a design-level preconditioning argument, not a guarantee about the composed optimizer (details in App. G).

Table 1: Results on six benchmarks (↑: higher is better). Open-source rows are mean ± std over four runs (training seeds for fine-tuned models; bootstrap over examples for frozen baselines); closed-source rows are single-run. Bold: best among training strategies on the Qwen2.5-VL-7B backbone.

Model (Method)	Hallucination Evaluations			General Reasoning		
	POPE ↑	VideoHallucer ↑	HalluBench ↑	MMBench ↑	MMSTAR ↑	VideoMMMU ↑
Closed-source Models						
GPT-4o	86.9	53.3	55.0	82.1	61.6	61.2
Gemini-1.5-Pro	–	37.8	45.6	73.9	59.1	53.9
Open-source Multimodal Base Models						
LLaVA-OneVision-7B	69.8 ± 1.5	44.6 ± 1.2	61.6 ± 1.1	81.7 ± 0.6	52.1 ± 1.0	31.2 ± 1.7
InternVL2.5-8B	82.1 ± 0.6	50.5 ± 1.3	55.6 ± 0.7	84.6 ± 0.4	55.3 ± 0.9	44.2 ± 1.4
LLaMA3.2-11B	74.8 ± 1.3	–	53.9 ± 1.1	72.1 ± 0.8	46.7 ± 1.0	–
Open-source Multimodal Reasoning Models						
R1-OneVision-7B	83.1 ± 0.8	42.4 ± 1.5	60.3 ± 1.0	82.3 ± 0.5	47.4 ± 1.2	44.1 ± 1.4
R1-VL-7B	82.4 ± 0.9	43.6 ± 1.4	62.7 ± 1.1	73.8 ± 0.7	41.0 ± 1.3	42.7 ± 1.5
Vision-R1-7B	81.1 ± 1.0	44.2 ± 1.2	58.3 ± 1.1	83.8 ± 0.6	39.4 ± 1.4	39.7 ± 1.6
Video-R1-7B	85.2 ± 0.7	45.1 ± 1.1	61.9 ± 0.8	86.0 ± 0.5	54.0 ± 0.9	48.1 ± 1.2
Qwen Based Models						
Qwen2.5-VL-7B	81.3 ± 0.9	52.2 ± 1.1	64.1 ± 0.7	83.1 ± 0.6	53.9 ± 0.8	43.9 ± 1.2
+ SFT	82.2 ± 0.6	55.3 ± 0.9	67.9 ± 0.6	82.9 ± 0.5	55.9 ± 0.7	45.8 ± 1.0
+ GRPO	83.7 ± 0.8	52.0 ± 1.3	68.6 ± 0.9	84.5 ± 0.6	57.8 ± 0.8	44.7 ± 1.3
+ DAPO	84.5 ± 0.7	55.1 ± 1.0	66.1 ± 0.8	84.4 ± 0.5	55.9 ± 0.9	48.7 ± 1.1
+ PAPO	83.5 ± 0.8	54.1 ± 1.1	68.9 ± 0.7	84.8 ± 0.5	58.1 ± 0.8	48.9 ± 1.0
+ DEEPO	**85.8 ± 0.4**	**57.6 ± 0.8**	**69.7 ± 0.6**	**86.4 ± 0.3**	**60.8 ± 0.6**	**53.3 ± 0.9**

Table 2: Ablation of the two branches and their interaction (↑: higher is better). All variants use Qwen2.5-VL-7B; all rows are four-seed means. "w/o Expert Hint" removes the whole signal-repair branch; "w/o Gradient Scaling" removes only the delivery branch, keeping the entropy-triggered prefixes, prefix regularizer, and dropout. The last row reports the interaction $I = S_{\text{full}} - S_{\text{hint}} - S_{\text{scaling}} + S_{\text{GRPO}}$; 95% CIs are $[-0.4, 1.2]$/$[-1.3, 1.7]$/$[-0.3, 3.1]$/$[-0.2, 1.0]$/$[-0.8, 1.2]$/$[1.1, 6.9]$ (last bold), and only VideoMMMU excludes zero.

Model (Method)	Hallucination Evaluations			General Reasoning		
	POPE ↑	VideoHallucer ↑	HalluBench ↑	MMBench ↑	MMSTAR ↑	VideoMMMU ↑
Qwen2.5-VL-7B	81.3	52.2	64.1	83.1	53.9	43.9
+ GRPO	83.7	52.0	68.6	84.5	57.8	44.7
+ w/o Expert Hint	84.4	54.2	68.7	85.0	59.8	44.7
+ w/o Gradient Scaling	84.7	55.2	68.2	85.5	58.6	49.3
+ DEEPO	**85.8**	**57.6**	**69.7**	**86.4**	**60.8**	**53.3**
Interaction I	+0.4	+0.2	+1.4	+0.4	+0.2	+4.0

4.3 The Unified Objective

Unified objective. The final training loss (minimized) combines the GRPO surrogate on self-generated continuations with the weighted prefix regularizer:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}}\Big[\lambda\,\mathcal{L}_{\text{aux}} - \tfrac{1}{|y_q^{\text{rem}}|}\textstyle\sum_{t=1}^{|y_q^{\text{rem}}|}\min\big(r_t\tilde{A}_t, \text{clip}(r_t, 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}})\tilde{A}_t\big)\Big] + \beta_{\text{kl}}\,\mathbb{E}\,D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}), \tag{6}$$

where r_t is the policy ratio, $\tilde{A}_t = g^\star(H_{2,t}, A_t)\cdot A_t$ the scaled advantage, and $\beta_{\text{kl}} = 0.04$; untriggered queries have $L_h = 0$. The KL term is deliberately *not* scaled by $g^\star$: correction tokens are allowed to move further from the reference policy (App. C).

Complementarity hypothesis. If the two stages repair distinct weak points, then (P1) grounded hints should help while shuffled hints do not (Table 8, App. A.4); (P2) the sign-aware rule should beat symmetric and entropy-only weighting (Table 3); and (P3) the interaction should be positive and largest on long-horizon tasks (Table 2). All three are borne out in Sec. 5.

Table 3: Delivery-branch design controls and trigger ablation (four-seed means, Qwen2.5-VL-7B); row 5 uses a rate-matched random trigger.

Variant	POPE ↑	HallusionBench ↑	VideoMMMU ↑
GRPO	83.7	68.6	44.7
w/ symmetric-τ weights	85.0	69.2	50.4
w/ symmetric-inverse weights	85.2	69.3	51.1
w/ sign-only weights	84.9	69.0	50.0
w/ EMPG weights	85.3	69.4	51.5
w/ random trigger	85.1	69.3	51.4
DEEPO (sign-aware, entropy trigger)	**85.8**	**69.7**	**53.3**

5 Experiments

5.1 Experiment Setup

Baselines, datasets, and protocol. We compare with closed-source models (GPT-4o OpenAI (2024), Gemini-1.5-Pro Gemini Team (2024)), open-source MLLMs (LLaVA-OneVision Li et al. (2024a), InternVL2.5 Chen et al. (2024c), LLaMA3.2 Meta (2024)), recent multimodal reasoning models (R1-OneVision Yang et al. (2025a), R1-VL Zhang et al. (2025a), Vision-R1 Huang et al. (2025b), Video-R1 Feng et al. (2025)), and, on the Qwen2.5-VL-7B backbone Bai et al. (2025), the training strategies SFT, GRPO Shao et al. (2024), DAPO Yu et al. (2025), PAPO Wang et al. (2025c), and DEEPO. We train on the Video-R1 multimodal RL dataset Feng et al. (2025) and evaluate on POPE Li et al. (2023), VideoHallucer Wang et al. (2024), HallusionBench Guan et al. (2023) (hallucination) and MMBench Liu et al. (2023), MMSTAR Chen et al. (2024a), VideoMMMU Hu et al. (2025) (general reasoning). Fine-tuned models are means over four training seeds with greedy decoding; frozen baselines are evaluated once with bootstrap standard deviations (Table 1); the full implementation specification is in App. G.

5.2 Performance

Table 1 reports results on three hallucination-oriented and three general reasoning benchmarks. DEEPO consistently achieves highly competitive performance across all benchmarks, setting new state-of-the-art results among the evaluated Qwen-based training strategies and outperforming GRPO, DAPO, and PAPO under the same backbone (85.8/57.6/69.7 on hallucination evaluations; 86.4/60.8/53.3 on general reasoning)—reducing hallucination does not trade off reasoning accuracy; the margin over GRPO's four-seed mean exceeds one seed standard deviation on five of six benchmarks (largest on VideoMMMU, $+8.6$), and the per-seed protocol is summarized in Table 1.

5.3 Ablation and Interaction

Table 2 ablates the two branches and reports their interaction. *The hint branch drives the reasoning-heavy gains* (*w/o Gradient Scaling* improves five of six benchmarks, most sharply on VideoMMMU, $+4.6$); *the scaling branch adds complementary gains* (*w/o Expert Hint* improves five of six and is flat on VideoMMMU; added on top of hints it lifts VideoMMMU from 49.3 to 53.3). The interaction is positive in the four-seed means on all six benchmarks but statistically significant only on long-horizon VideoMMMU; elsewhere the combination is consistent with additive gains.

5.4 Controls and Mechanism Evidence

Design controls. Replacing the sign-aware weight with any symmetric or sign-only alternative degrades all three benchmarks, as does replacing the entropy trigger with a rate-matched random one (Table 3); likewise, entropy-triggered hints improve over the base method but trail full DEEPO (Table 7, App. A.4). Conditioning the weight on the advantage sign—not entropy alone—is what delivers the correction, and the trigger gates the intervention to where it is needed. The margin over EMPG's entropy-only factor is modest on POPE ($+0.5$) and largest on long-horizon VideoMMMU ($+1.8$); the Rényi-2 identity behind both rules is standard (Prop. 4.1), so the added value lies in the sign-aware allocation and the rollout-stage trigger, not in the identity itself.

Table 4: Mechanism diagnosis on the training distribution (four-seed pools): (a) all-wrong group fraction (%) and mean group reward by H_s decile (GRPO baseline); (b) hint activation on triggered, originally all-wrong queries; (c) baseline grouping ablation; (d) equal-compute and equal-supervision baselines. All rows in (c)–(d) run the identical rollout pipeline of Algorithm 1 (G=8 unhinted rollouts plus one hint-conditioned continuation per triggered query); the "mixed" row is the default model itself, not a rerun, and in the "separated" variant the unhinted group is normalized within itself while the singleton hinted continuation is scored against a batch-level EMA baseline over hint continuations—so no advantage is zeroed by construction and only the counterfactual penalty channel is removed.

H_s decile	1	2	3	4	5	6	7	8	9	10
All-wrong (%)	12	16	21	27	33	39	45	51	57	63
Mean reward	0.76	0.69	0.61	0.53	0.46	0.40	0.34	0.29	0.24	0.19

Metric	Value
Converted to mixed groups ($\geq$ 1 correct), %	60.0
Failed-rollout mean advantage, pre / post	0.000 / -0.212
Within-group advantage variance, pre / post	0.000 / 0.600

Baseline	POPE ↑	HallusionBench ↑	VideoMMMU ↑
Mixed (shared, default)	**85.8**	**69.7**	**53.3**
Separated (per-condition)	84.4	68.5	49.7

Variant	POPE ↑	HallusionBench ↑	VideoMMMU ↑
GRPO + extra budget	84.2	68.9	46.3
SFT + GRPO (same traces)	85.0	69.4	51.2

Mechanism diagnosis. The rollout-stage diagnosis of Sec. 4.1 is quantified in Table 4(a,b): all-wrong groups rise from 12% to 63% across H_s deciles, and hints convert 60% of triggered all-wrong groups into mixed ones. Removing the shared baseline—with the identical rollout pipeline, and the singleton hinted continuation scored against a batch-level EMA baseline so that no advantage is zeroed by construction—degrades all three tracked benchmarks (Table 4(c), whose mixed row is the default model itself), isolating the counterfactual penalty on unhinted failures.

Fairness controls. The gains are not explained by extra compute or expert data (Table 4(d)): matching wall-clock with more GRPO steps, or matching traces with offline SFT before GRPO, both trail DEEPO, with the residual gap concentrated on VideoMMMU ($+2.1$)—offline SFT force-fits full trajectories and shifts the policy off its on-policy manifold, whereas DEEPO distills the same traces smoothly inside a single RL stage, avoiding the exploration collapse typical of multi-stage SFT-then-RL pipelines. *Paired analysis.* Across the four seeds, DEEPO leads SFT+GRPO on 4/4 seeds on VideoMMMU (paired bootstrap 95% CI $[+1.2, +3.0]$) and on 3/4 seeds on POPE and HallusionBench (CIs $[-0.2, +1.4]$, $[-0.1, +0.9]$). Further controls (App. A.4, App. A.5): shuffled hints show no distinct advantage over GRPO; DEEPO-self retains $\approx 71\%$ of the VideoMMMU gain without expert data—consistent with the direct-grounding channel dominating, though source, quality, and coverage also differ between the variants, so the residual gap bounds rather than isolates the collapsed-group channel; amplified updates are enriched 1.75–$1.80\times$ on span-annotated hallucinated content at the weight level (1.15–$1.25\times$ at the effective-step level); results transfer to InternVL2.5-8B (85.0 vs. 83.4 GRPO on POPE).

6 Conclusion

We studied hallucination in RL of MLLMs as two weak points of the correction chain—on hard queries, unanimously wrong groups collapse the group-relative advantage to zero, and confident errors are gradient-invisible—and proposed DEEPO, a dual-stage enhancement pairing entropy-triggered expert prefixes at rollout time with advantage-sign-aware Rényi gradient preconditioning at optimization time. Both branches improve over GRPO individually; their interaction is significant on long-horizon VideoMMMU and additive elsewhere; and the controls of Sec. 5 isolate grounding content as the active ingredient, with stable training at $+18\%$ per-step cost over GRPO (App. G). The guarantee currently concerns a local sample-wise surrogate rather than the global expected return.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and readability of the manuscript. They were not used for developing research ideas, methodology, theoretical results, experimental design, data analysis, or interpretation of results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on reducing hallucination and improving the reliability of multimodal language models and does not involve experiments with human participants. Our experiments use established research benchmarks and existing training resources. Although improved multimodal models may be deployed in settings where inaccurate outputs can have real-world consequences, the method presented here is intended as a research contribution toward more reliable model behavior and should not be interpreted as guaranteeing factual correctness or safety in high-stakes applications. Appropriate validation and human oversight remain necessary in such settings.

REPRODUCIBILITY STATEMENT

We provide detailed experimental and implementation information in the appendices. Appendix A contains additional experiments, sensitivity analyses, and controlled ablations; Appendices B–F provide proofs and analyses of the entropy quantities; and Appendix G specifies implementation details, data processing, training hyperparameters, hardware, and the per-iteration training procedure. The main experimental protocol, benchmarks, and evaluation settings are described in Section 5.

REFERENCES

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society and Oxford University Press, 2000. doi: 10.1090/mmono/191.

Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, 2025. doi: 10.3389/frai.2025.1622292. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full.

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

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2404.18930*, 2024. URL https://arxiv.org/abs/2404.18930.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM Hallucination Benchmark. *arXiv preprint arXiv:2504.17550*, 2025. URL https://arxiv.org/abs/2504.17550.

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? *arXiv preprint arXiv:2403.20330*, 2024a. URL https://arxiv.org/abs/2403.20330.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. Multi-Object Hallucination in Vision-Language Models. *arXiv preprint arXiv:2407.06192*, 2024b. URL https://arxiv.org/abs/2407.06192.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2024c. URL https://arxiv.org/abs/2412.05271.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating Hallucination in Visual Language Models with Visual Supervision. *arXiv preprint arXiv:2311.16479*, 2023. URL https://arxiv.org/abs/2311.16479.

Manuel Cossio. A comprehensive taxonomy of hallucinations in Large Language Models. *arXiv preprint arXiv:2508.01781*, 2025. URL https://arxiv.org/abs/2508.01781.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

A. A. Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003. doi: 10.1109/TIT.2003.811927.

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, 2025. URL https://arxiv.org/abs/2503.21776.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating Hallucination in Multimodal Large Language Model via Hallucination-targeted Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.850. URL https://aclanthology.org/2025.findings-acl.850/.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. SRFT: A Single-Stage Method with Supervised and Reinforcement Fine-Tuning for Reasoning. *arXiv preprint arXiv:2506.19767*, 2025b. URL https://arxiv.org/abs/2506.19767.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

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. *arXiv preprint arXiv:2310.14566*, 2023. URL https://arxiv.org/abs/2310.14566.

Peter Harremoës and Flemming Topsøe. Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Transactions on Information Theory*, 47(7):2944–2960, 2001. doi: 10.1109/18.959272.

Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Jing Shao, and Lu Sheng. A Topic-level Self-Correctional Approach to Mitigate Hallucinations in MLLMs. *arXiv preprint arXiv:2411.17265v1*, 2024. URL https://arxiv.org/abs/2411.17265v1.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6841–6858. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.349. URL https://aclanthology.org/2025.naacl-long.349/.

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. *arXiv preprint arXiv:2510.09285*, 2025a. URL https://arxiv.org/abs/2510.09285.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. *arXiv preprint arXiv:2503.06749*, 2025b. URL https://arxiv.org/abs/2503.06749.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Writing-Zero: Bridge the Gap Between Non-verifiable Tasks and Verifiable Rewards. *arXiv preprint arXiv:2506.00103*, 2025. URL https://arxiv.org/abs/2506.00103.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Grosse, and Jimmy Ba. Calibrating Language Models via Augmented Prompt Ensembles. ICML 2023 Workshop on Challenges in Deployable Generative AI, 2023. URL https://openreview.net/pdf?id=L0dc4wqbNs.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why Language Models Hallucinate. *arXiv preprint arXiv:2509.04664*, 2025. URL https://arxiv.org/abs/2509.04664.

Zaid Khan and Yun Fu. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.10193.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *arXiv preprint arXiv:2311.16922*, 2023. URL https://arxiv.org/abs/2311.16922.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*, 2023. URL https://arxiv.org/abs/2305.10355.

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing Advanced Visual Reasoning Ability of Large Language Models. *arXiv preprint arXiv:2409.13980*, 2024b. URL https://arxiv.org/abs/2409.13980.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. FLAME: Factuality-Aware Alignment for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115588–115614, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d16152d53088ad779ffa634e7bf66166-Abstract-Conference.html.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating Hallucination Through Theory-Consistent Symmetric Multimodal Preference Optimization. *arXiv preprint arXiv:2506.11712*, 2025. URL https://arxiv.org/abs/2506.11712.

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? *arXiv preprint arXiv:2307.06281*, 2023. URL https://arxiv.org/abs/2307.06281.

Omar Mahmoud, Ali Khalil, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. The Unintended Trade-off of AI Alignment: Balancing Hallucination Mitigation and Safety in LLMs. *arXiv preprint arXiv:2510.07775*, 2025. URL https://arxiv.org/abs/2510.07775.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI Blog, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. The Energy Loss Phenomenon in RLHF: A New Perspective on Mitigating Reward Hacking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44076–44105. PMLR, 2025. URL https://proceedings.mlr.press/v267/miao25c.html.

Ruijia Niu, Dongxia Wu, Rose Yu, and Yi-An Ma. Functional-level Uncertainty Quantification for Calibrated Fine-tuning on LLMs. *arXiv preprint arXiv:2410.06431v1*, 2024. URL https://arxiv.org/abs/2410.06431v1.

OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Xin Qiu and Risto Miikkulainen. Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.13845v3.

Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. Look, Compare, Decide: Alleviating Hallucination in Large Vision-Language Models via Multi-View Multi-Path Reasoning. *arXiv preprint arXiv:2408.17150*, 2024. URL https://arxiv.org/abs/2408.17150.

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.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's Razor: Why Online Reinforcement Learning Forgets Less. *arXiv preprint arXiv:2509.04259*, 2025. URL https://arxiv.org/abs/2509.04259.

Edward H. Simpson. Measurement of Diversity. *Nature*, 163:688, 1949. doi: 10.1038/163688a0.

Chao Wang, Xuancheng Zhou, Weiwei Fu, and Yang Zhou. Mitigating Hallucinations in Large Vision-Language Models with Internal Fact-based Contrastive Decoding. *arXiv preprint arXiv:2502.01056*, 2025a. URL https://arxiv.org/abs/2502.01056.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. Harnessing Uncertainty: Entropy-Modulated Policy Gradients for Long-Horizon LLM Agents. *arXiv preprint arXiv:2509.09265*, 2025b. URL https://arxiv.org/abs/2509.09265.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating Intrinsic and Extrinsic Hallucinations in Large Video-Language Models. *arXiv preprint arXiv:2406.16338*, 2024. URL https://arxiv.org/abs/2406.16338.

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.

Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang, and Longjun Cai. Improve Decoding Factuality by Token-wise Cross Layer Entropy of Large Language Models. *arXiv preprint arXiv:2502.03199*, 2025a. URL https://arxiv.org/abs/2502.03199.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. VisionLLM v2: An End-to-End Generalist Multimodal Large Language Model for Hundreds of Vision-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 69925–69975, 2024. URL https://arxiv.org/abs/2406.08394.

Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan, and Michael Qizhe Shieh. SynthRL: Scaling Visual Reasoning with Verifiable Data Synthesis. *arXiv preprint arXiv:2506.02096*, 2025b. URL https://arxiv.org/abs/2506.02096.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating Hallucinations in Multi-modal Large Language Models via Image Token Attention-Guided Decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1571–1590. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.75. URL https://aclanthology.org/2025.naacl-long.75/.

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*, 2025a. URL https://arxiv.org/abs/2503.10615.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key. *arXiv preprint arXiv:2501.09695*, 2025b. URL https://arxiv.org/abs/2501.09695.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. doi: 10.1093/nsr/nwae403.

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*, 2025. URL https://arxiv.org/abs/2503.14476.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a. URL https://arxiv.org/abs/2312.00849.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. *arXiv preprint arXiv:2405.17220v1*, 2024b. URL https://arxiv.org/abs/2405.17220v1.

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.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the Confidence of Large Language Models by Eliciting Fidelity. *arXiv preprint arXiv:2404.02655*, 2024. URL https://arxiv.org/abs/2404.02655.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-Policy RL Meets Off-Policy Experts: Harmonizing Supervised Fine-Tuning and Reinforcement Learning via Dynamic Weighting. *arXiv preprint arXiv:2508.11408*, 2025b. URL https://arxiv.org/abs/2508.11408.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance. *arXiv preprint arXiv:2402.08680*, 2024. URL https://arxiv.org/abs/2402.08680.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. *arXiv preprint arXiv:2311.16839*, 2023. URL https://arxiv.org/abs/2311.16839.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2310.00754*, 2023. URL https://arxiv.org/abs/2310.00754.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. URL https://arxiv.org/abs/2402.11411.

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration. *arXiv preprint arXiv:2502.01969*, 2025. URL https://arxiv.org/abs/2502.01969.

A Additional Experimental Results

A.1 Gradient-Mass Pilot

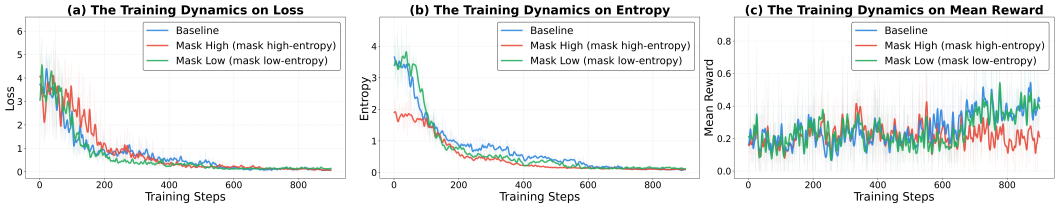


Figure 3: Effect of entropy-masked gradients. Masking low-entropy gradients preserves stable learning, while masking high-entropy gradients destabilizes training: high-entropy tokens carry the bulk of the update signal (Prop. 4.1); this pilot shows where gradient mass lives, not a hallucination outcome.

A.2 Model Guess Analysis

To further examine uncertainty-driven guessing, we transform multiple-choice questions into cloze-style fill-in-the-blank queries, removing predefined options to compel models to answer from visual-semantic grounding rather than option cues. This setting exposes semantic instability in free-form generation.

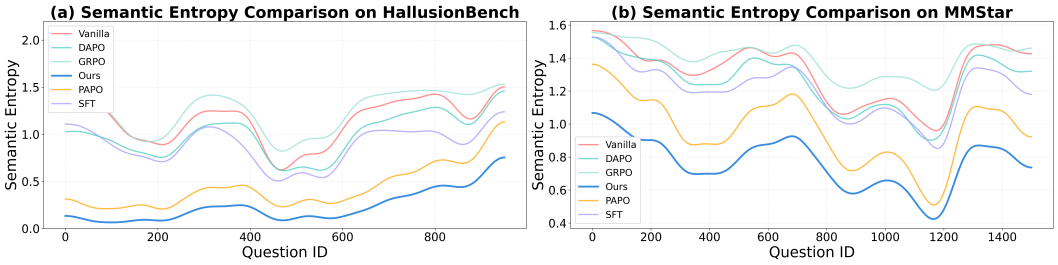


Figure 4: Semantic entropy on HallusionBench and MMStar (cloze-style conversion; questions sorted by the base model's H_s for readability). Compared with the untrained base model, DAPO yields higher semantic entropy, while GRPO and PAPO reduce entropy but remain more variable. DEEPO maintains consistently lower and less dispersed semantic entropy, indicating reduced uncertainty-driven guessing.

As shown in Fig. 4, DEEPO consistently produces the lowest semantic entropy across both HallusionBench and MMStar, with the smallest dispersion across questions (we operationalize "more stable" as the inter-question standard deviation of H_s). In contrast, the base model, DAPO, GRPO, SFT, and PAPO exhibit higher or more variable entropy, indicating less stable answer-level commitments under ambiguous multimodal evidence. We stress that lower H_s is evidence of reduced guessing only when accuracy is preserved—a confidently wrong model would also have low entropy—so this analysis should be read together with the accuracy numbers in Table 1. These results support our motivation that semantic-entropy-triggered expert guidance reduces inconsistent semantic predictions and suppresses uncertainty-driven guessing during RL fine-tuning.

A.3 Hyperparameter Sensitivity

We briefly check sensitivity to the initial semantic-entropy threshold τ_s (Table 5) and the rollout budget N (Table 6). Varying τ_s over $\{0.2, 0.4, 0.5, 0.8, 0.9\}$ moves results by at most 0.7 points on POPE and 0.5 on VideoMMMU (four-seed means), as the EMA-updated threshold self-calibrates to the batch. Increasing N (co-varied with the GRPO group size G, so the joint effect of estimation budget and group size) improves results up to N=8–12 and degrades noticeably only in the small-N regime (-4.0 on VideoMMMU at N=4); the default N=8 is a cost-driven choice.

Table 5: Sensitivity to the initial semantic entropy threshold τ_s (four-seed means).

Initial τ_s	POPE ↑	VideoHallucer ↑	HallusionBench ↑	MMBench ↑	MMSTAR ↑	VideoMMMU ↑
0.2	85.5	55.2	69.4	86.2	60.5	52.8
0.4	85.7	55.7	69.5	**86.5**	60.7	53.1
0.5	85.6	55.3	69.4	86.3	60.5	52.9
0.8	**85.8**	**57.6**	**69.7**	86.4	**60.8**	**53.3**
0.9	85.1	55.8	**69.7**	**86.5**	60.4	53.2

Table 6: Sensitivity to the rollout budget N (co-varied with the GRPO group size G; four-seed means).

Rollout N	POPE ↑	VideoHallucer ↑	HallusionBench ↑	MMBench ↑	MMSTAR ↑	VideoMMMU ↑
4	84.7	54.2	68.5	85.4	59.1	49.3
6	85.4	55.1	69.2	86.1	60.1	51.9
8	85.8	**57.6**	**69.7**	86.4	**60.8**	53.3
12	**86.0**	55.5	69.6	**86.5**	60.7	**53.5**

A.4 Causal Controls and Hallucination-Targeted Baselines

This subsection accompanies Sec. 5.4 and collects the causal controls. Control runs (Table 8, Table 9, and Table 4(c) in the main text) follow the same four-seed protocol as the main results and are reported as means.

Grouping ablation. Table 4(c) (main text) compares the default mixed grouping against separated per-condition baselines. Both variants run the identical rollout pipeline of Algorithm 1: $G{=}8$ first-pass unhinted rollouts plus one hint-conditioned continuation per triggered query, with matched compute and the same prefix regularizer. They differ only in advantage computation: mixed pooling normalizes all nine samples jointly, whereas the separated variant normalizes the unhinted group within itself and scores the singleton hinted continuation against a batch-level EMA baseline over hint continuations of other triggered queries (so its advantage is well-defined and not identically zero, avoiding the size-1 group artifact). The "mixed" row is the default DEEPO model itself, not a rerun, which is why its numbers coincide with Table 1; the "separated" row is a separately trained run under this protocol. The drop therefore isolates the loss of the counterfactual penalty on unhinted failures.

Table 7: Controlled study on expert-hint supervision (four-seed means). Entropy-triggered hints improve over the corresponding base method, yet remain below DEEPO—hint supervision alone does not explain the gains. Hint rows use rollout conditioning only (no prefix regularizer, no dropout), unlike Table 2's "w/o Gradient Scaling" row.

Method	Hint	Trigger	POPE ↑	HallusionBench ↑	VideoMMMU ↑
GRPO	✗	✗	83.7	68.6	44.7
GRPO + Expert Hint	✓	✗	83.4	68.8	45.0
GRPO + Entropy-triggered Hint	✓	✓	84.2	69.0	46.7
DEEPO w/o Expert Hint	✗	✗	84.4	68.7	44.7
DEEPO	✓	✓	**85.8**	**69.7**	**53.3**

Hint-shuffle control. Table 8 repeats DEEPO training with the same entropy trigger but with prefixes drawn from *mismatched* queries (shuffled within the batch). If the gain stemmed from generic regularization or longer context, shuffled hints would help equally. The results confirm the prediction: shuffled hints show no distinct advantage over GRPO on any of the three benchmarks (-0.5/-0.9/-0.2; seed-to-seed variations overlap substantially), while correct hints improve by $+2.1$/$+1.1$/$+8.6$—grounding content, not the conditioning machinery, drives the gains.

Self-generated hints. Table 9 evaluates DEEPO-self, in which the hint for a triggered query is the prefix of the highest-reward verified-correct rollout from the same GRPO group (queries whose group contains no correct rollout receive no hint). This variant requires no dataset traces at all. DEEPO-self recovers most of the gain over GRPO ($+0.9$/$+0.0$/$+6.1$), trailing dataset-derived hints

Table 8: Hint-shuffle control (four-seed means): entropy-triggered prefixes from mismatched queries. Shuffled hints should fail to help despite identical context length and trigger statistics.

Variant	POPE ↑	HallusionBench ↑	VideoMMMU ↑
GRPO (four-seed mean)	83.7	68.6	44.7
DEEPO (shuffled hints)	83.2	67.7	44.5
DEEPO (correct hints, four-seed mean)	**85.8**	**69.7**	**53.3**

only modestly (-1.2/-1.1/-2.5). Because a self-hint exists only when the group already contains a correct rollout, DEEPO-self cannot exercise the advantage-reshaping channel on all-wrong groups; its retained gain (VideoMMMU: 6.1 of 8.6, $\approx 71\%$) indicates that the direct-grounding and supervision channel carries the bulk of the gain; since the two variants also differ in hint source, quality, and coverage, the residual gap is an upper bound on—not a clean isolation of—the collapsed-group coverage channel.

Table 9: Self-generated hints (DEEPO-self, four-seed means): prefixes come from the highest-reward verified-correct rollout within the same GRPO group; no dataset traces are used.

Variant	POPE ↑	HallusionBench ↑	VideoMMMU ↑
DEEPO (dataset hints, four-seed mean)	**85.8**	**69.7**	**53.3**
DEEPO-self	84.6	68.6	50.8

A.5 Further Controls and Ablations

The following analyses isolate error localization (weight vs. effective step) and backbone transfer; the weighting-rule and trigger ablations are consolidated in Table 3 of the main text.

Error-localization analysis. Table 10 measures the relative per-token update intensity that the sign-aware rule assigns to hallucinated vs. non-error content on failed rollouts.

Annotation protocol. We randomly sampled 300 failed (reward-0) rollouts per benchmark (600 in total) from the evaluation generations of the GRPO baseline. Two trained annotators independently marked hallucinated spans at the token level—visual entities, attributes, counts, and spatial or temporal relations not supported by the input image/video—following a written guideline that explicitly requires articles, connectives, and other function words to be labeled as non-error even when they are low-entropy. Disagreements were adjudicated by a third annotator; inter-annotator agreement before adjudication was Cohen's $\kappa = 0.78$ (HallusionBench) and 0.81 (VideoMMMU). As an independent cross-check, GPT-4V was shown the same rollouts and visual inputs and asked to mark unsupported visual claims; it agreed with the adjudicated labels on 92% of tokens on a 50% subsample. For each token we then recorded its scaling weight $g^\star$ under the DEEPO rule and its effective logit-step norm $\|\Delta z_t\| \propto w_t \|g_t\|$; the table reports both quantities on error vs. non-error tokens within failed rollouts, with 95% bootstrap confidence intervals over rollouts (10,000 resamples).

Interpretation. Hallucinated spans are sparse in long-form generation: error tokens constitute 6.4% (HallusionBench) and 7.1% (VideoMMMU) of all tokens in the annotated failed rollouts; weighting by the enrichment factors, roughly 10.7%/12.1% of the weighted update mass lands on error tokens and the remaining $\sim 88\%$ on non-error tokens—as it inevitably does for any sequence-level advantage formulation. The table therefore supports a *relative budget shift*—nearly doubled per-token update intensity on the error locus—not token-orthogonal deletion of hallucinated content. Weight is also not gradient contribution: the effective-step enrichment ($1.15\times$/$1.25\times$) is attenuated relative to the weight enrichment because error tokens carry smaller raw score norms $\|g_t\|$—exactly the stall the compensation is designed to counteract (Thm. 4.2(iii)).

Second backbone. Table 11 reports generalization to InternVL2.5-8B.

B Proof of the Entropy–Gradient Relationship

We prove Proposition 4.1. Let the policy over the N-token vocabulary be $\pi_j := \pi_\theta(y_j \mid q, y_{<t})$ with logits $z = (z_1, \ldots, z_N)$, and let $g_k := \nabla_z \log \pi_k = e_k - \pi$ be the score-function gradient for

Table 10: Error-localization analysis on failed rollouts: per-token update intensity on hallucinated vs. non-error content. Annotation: token-level hallucinated-span labeling by two human annotators with adjudication (600 rollouts; Cohen's $\kappa = 0.78/0.81$), cross-checked by GPT-4V (92% token agreement on a 50% subsample). $\|\Delta z\|$: mean effective logit-step norm. CI: 95% bootstrap over rollouts.

Benchmark	$g^\star$ (error)	$g^\star$ (non-error)	Enrichment ↑	95% CI	$\|\Delta z\|$ (err / non-err)
HallusionBench	1.68	0.96	1.75	$[1.58, 1.93]$	0.31 / 0.27
VideoMMMU	1.80	1.00	1.80	$[1.63, 1.98]$	0.35 / 0.28

Table 11: Generalization to a second backbone (InternVL2.5-8B; four-seed means).

Variant	POPE ↑	HallusionBench ↑	VideoMMMU ↑
InternVL2.5-8B	82.1	55.6	44.2
+ GRPO	83.4	57.0	45.3
+ DEEPO	85.0	58.2	51.0

sampled token k, where e_k is the k-th standard basis vector and $\pi = (\pi_1, \dots, \pi_N)^\top$. For each k,

$$\|g_k\|^2 = (1 - \pi_k)^2 + \sum_{j \neq k} \pi_j^2 = 1 - 2\pi_k + \sum_{j=1}^{N} \pi_j^2. \tag{7}$$

Taking the expectation over $k \sim \pi$,

$$\mathbb{E}_{k\sim\pi}\big[\|g_k\|^2\big] = \sum_{k=1}^{N} \pi_k \Big(1 - 2\pi_k + \sum_{j=1}^{N} \pi_j^2\Big) = 1 - 2\sum_k \pi_k^2 + \sum_j \pi_j^2 = 1 - \sum_{j=1}^{N} \pi_j^2. \tag{8}$$

Since $H_2(\pi) = -\log \sum_j \pi_j^2$, we have $\sum_j \pi_j^2 = \exp(-H_2(\pi))$, and therefore

$$\mathbb{E}_{k\sim\pi}\big[\big\|\nabla_z \log \pi_k\big\|^2\big] = 1 - \exp\big(-H_2(\pi)\big), \tag{9}$$

which is Eq. 3. This is the trace of the categorical Fisher information matrix $F = \mathrm{diag}(\pi) - \pi\pi^\top$ in natural parameters, a standard identity (cf. Amari & Nagaoka (2000)); equivalently, one minus Simpson's index of coincidence Simpson (1949).

C Proof of Theorem 4.2 and Remark 4.3

We first prove the sample-wise damped-proximal statement, then the sign-symmetry impossibility.

Setup and standing convention. Fix a training batch of *already generated* trajectories. Every advantage A_t below is a fixed empirical scalar computed from the batch rewards (group-relative normalization included), and the sampled token y_t at each position is fixed as well. Accordingly, *no expectation over $y_t \sim \pi_\theta$ is taken anywhere in this section with A_t factored out*: the only distribution-level quantities are the Fisher matrix F_t and its trace τ_t, which are deterministic functions of the current policy at the state $s_t = (q, y_{<t})$ and do not depend on which token was sampled. Write $\pi_j := \pi_\theta(y_j \mid s_t)$, $j = 1, \dots, N$, with logits $z = (z_1, \dots, z_N)$; for the (fixed) sampled token $y_t = y_k$ the score-function gradient with respect to the logits is $g_t = g_k = e_k - \pi$, where e_k is the k-th standard basis vector. The Fisher information matrix with respect to the logits is

$$F_t = \mathbb{E}_{y\sim\pi_\theta(\cdot|s_t)}\big[\, g_y g_y^\top \,\big] = \mathrm{diag}(\pi) - \pi\pi^\top, \qquad \tau_t := \mathrm{Tr}(F_t) = 1 - \textstyle\sum_j \pi_j^2 = 1 - \exp(-H_{2,t}), \tag{10}$$

where the expectation defining F_t involves the policy distribution only—it is the second moment of the score—and never the advantage A_t.

Part (i): sample-wise curvature bound and tangent minorant. $F_t \succeq 0$ since for any $u \in \mathbb{R}^N$, $u^\top F_t u = \sum_j \pi_j u_j^2 - \big(\sum_j \pi_j u_j\big)^2 = \mathrm{Var}_{J\sim\pi}[u_J] \geq 0$. Consequently $\lambda_{\max}(F_t) \leq \mathrm{Tr}(F_t) = \tau_t$, and for the fixed sampled token,

$$g_t^\top F_t g_t \;\leq\; \lambda_{\max}(F_t)\, \|g_t\|^2 \;\leq\; \tau_t\, \|g_t\|^2. \tag{11}$$

(For interpretation only, the distributional moments $\mathbb{E}_{y\sim\pi}\|g_y\|^2 = \tau_t$ (App. B) and $\mathbb{E}_{y\sim\pi}[g_y^\top F_t g_y] = \mathrm{Tr}(F_t^2) = S_2 - 2S_3 + S_2^2 \leq \tau_t^2$, with $S_r := \sum_j \pi_j^r$, identify τ_t as the expected gradient mass; the

proof below needs only the sample-wise bound above, which holds for *every* sampled token rather than in expectation over it.) Restrict one-step logit updates to the score direction, $\Delta z_t = \eta\, w_t A_t g_t$ with $w_t \geq 0$. The sample-wise second-order proximal surrogate with Levenberg–Marquardt damping $\varepsilon_0 > 0$ evaluates to

$$Q_t(w_t) \;=\; A_t\, g_t^\top \Delta z_t - \tfrac{\beta_{\mathrm{kl}}}{2}\, \Delta z_t^\top (F_t + \varepsilon_0 I)\, \Delta z_t \;=\; \eta A_t^2 \|g_t\|^2\, w_t - \frac{\beta_{\mathrm{kl}} \eta^2 A_t^2}{2} \left(g_t^\top F_t g_t + \varepsilon_0 \|g_t\|^2\right) w_t^2. \tag{12}$$

Since A_t, g_t are fixed sample quantities, Q_t is an exact quadratic in w_t; the only approximation we introduce is the scalarization of the curvature term. Define the trace-scalarized quadratic

$$\widetilde{Q}_t(w) = \eta A_t^2 \|g_t\|^2\, w - \frac{\beta_{\mathrm{kl}} \eta^2 A_t^2}{2}\, \|g_t\|^2\, (\tau_t + \varepsilon_0)\, w^2. \tag{13}$$

Since $g_t^\top F_t g_t + \varepsilon_0 \|g_t\|^2 \leq (\tau_t + \varepsilon_0)\|g_t\|^2$ by the bound above, $\widetilde{Q}_t(w) \leq Q_t(w)$ for all $w \geq 0$, with $\widetilde{Q}_t(0) = Q_t(0) = 0$ and $\widetilde{Q}_t'(0) = Q_t'(0) = \eta A_t^2 \|g_t\|^2$: $\widetilde{Q}_t$ is a tangent minorant of Q_t on $\mathbb{R}_+$.

Part (ii): damped optimum, ascent, and cap. $\widetilde{Q}_t$ is concave in w with unique maximizer

$$w_t^{\varepsilon_0} = \frac{\eta A_t^2 \|g_t\|^2}{\beta_{\mathrm{kl}}\, \eta^2 A_t^2 \|g_t\|^2 (\tau_t + \varepsilon_0)} = \frac{1}{\beta_{\mathrm{kl}}\, \eta\, (\tau_t + \varepsilon_0)} \;\leq\; \frac{1}{\beta_{\mathrm{kl}}\, \eta\, \varepsilon_0} =: w_{\max}, \tag{14}$$

where the cap follows from $\tau_t \geq 0$; in particular the weight remains bounded as the curvature vanishes, unlike the undamped inverse-curvature rule $1/\tau_t$. The achieved minorant value certifies strict sample-wise ascent:

$$Q_t(w_t^{\varepsilon_0}) \;\geq\; \widetilde{Q}_t(w_t^{\varepsilon_0}) = \frac{A_t^2\, \|g_t\|^2}{2\, \beta_{\mathrm{kl}}\, (\tau_t + \varepsilon_0)} \;>\; 0 = Q_t(0). \tag{15}$$

The induced logit step itself stays controlled: $\|\Delta z_t\| = \eta\, w_t^{\varepsilon_0}\, |A_t|\, \|g_t\| = |A_t|\, \|g_t\| / \big(\beta_{\mathrm{kl}}(\tau_t + \varepsilon_0)\big)$, which vanishes smoothly as $g_t \to 0$ while its amplification *relative to the unweighted rule* ($w_t = 1$) scales as $(\tau_t + \varepsilon_0)^{-1}$. This is a conservative damped proximal step in logit space: the damping constant ε_0 caps the weight, and the guarantee degrades gracefully as $\tau_t \to 0$.

Part (iii): softmax-saturation regime. Let $\pi_y := \pi_\theta(y_t \mid s_t)$ denote the probability of the sampled token and $\delta_t := 1 - \pi_y$ the tail mass. Using $\sum_{j \neq y} \pi_j^2 \leq \big(\sum_{j \neq y} \pi_j\big)^2 = \delta_t^2$,

$$\|g_t\|^2 = (1 - \pi_y)^2 + \sum_{j \neq y} \pi_j^2 \;\in\; \big[\, \delta_t^2,\; 2\,\delta_t^2 \,\big], \qquad \tau_t = 1 - \pi_y^2 - \sum_{j \neq y} \pi_j^2 \;\in\; \big[\, 2\delta_t(1 - \delta_t),\; 2\delta_t - \delta_t^2 \,\big]. \tag{16}$$

Hence, as the prediction saturates ($\pi_y \to 1$, i.e. $\delta_t \to 0$): (a) the unweighted local slope vanishes *quadratically*, $Q_t'(0) = \eta A_t^2 \|g_t\|^2 \leq 2\eta A_t^2\, \delta_t^2$—this is the softmax-saturation stall; (b) the compensated weight grows as $w_t^{\varepsilon_0} \geq \big(\beta_{\mathrm{kl}} \eta\, (2\delta_t - \delta_t^2 + \varepsilon_0)\big)^{-1} = \Omega(\delta_t^{-1})$ until it attains its cap $w_{\max}$; (c) the certified ascent decays only *linearly* in the tail mass,

$$Q_t(w_t^{\varepsilon_0}) \;\geq\; \frac{A_t^2\, \delta_t^2}{2\beta_{\mathrm{kl}}\, (2\delta_t - \delta_t^2 + \varepsilon_0)} \;=\; \Theta(\delta_t) \quad \text{for } \delta_t \gg \varepsilon_0. \tag{17}$$

This is the formal content of saturation compensation: inverse-curvature weighting converts the local correction rate on a confident token from quadratic to first-order in the tail mass where the curvature dominates the damping ($\delta_t \gg \varepsilon_0$). In the deep tail $\delta_t \ll \varepsilon_0$ the damping floor dominates: the weight saturates at $w_{\max} = (\beta_{\mathrm{kl}} \eta \varepsilon_0)^{-1}$ and the certified bound reverts to $O(\delta_t^2/\varepsilon_0)$—quadratic again, but with the $1/\varepsilon_0$ constant—while the induced step $\|\Delta z_t\| \leq |A_t|\, \|g_t\| / (\beta_{\mathrm{kl}} \varepsilon_0)$ remains bounded and vanishes smoothly. The theorem is thus a rate statement over the operational confidence band plus a boundedness statement in the tail; it does not eliminate saturation asymptotically. This proves Theorem 4.2.

Sign-symmetry impossibility (Remark 4.3). Let $w_t = \nu(H_{2,t})$ be any advantage-sign-independent rule under a fixed batch budget $\sum_t w_t = C$. Suppose there exist two samples a, b with $H_{2,a} = H_{2,b}$ and $A_a < 0 < A_b$. Then necessarily $w_a = \nu(H_{2,a}) = \nu(H_{2,b}) = w_b$, so no such rule can satisfy the correction-priority condition $w_a > w_b$. More explicitly, define the corrective allocation objective

$$\mathcal{C}(w) := \sum_{t:\, A_t < 0} |A_t|\, w_t\, \tau_t^{-1}, \tag{18}$$

measuring how much weighted update mass is assigned to negative-advantage samples after curvature compensation. Starting from any symmetric rule, construct the budget-preserving perturbation $w'_a = w_a + \delta$, $w'_b = w_b - \delta$ with $0 < \delta < w_b$, leaving other weights unchanged. The total budget is preserved, and since only sample a contributes to $\mathcal{C}$ among this pair,

$$\mathcal{C}(w') - \mathcal{C}(w) = |A_a|\,(w'_a - w_a)\,\tau_a^{-1} = \delta\,|A_a|\,\tau_a^{-1} > 0. \tag{19}$$

So there always exists a budget-preserving asymmetric reallocation that strictly improves the corrective objective, while any symmetric rule $w_t = \nu(H_{2,t})$ remains invariant over equal-entropy pairs and cannot realize this preferential reallocation.

Scope and caveats. We record four qualifications on how Theorem 4.2 should be read. (a) The guarantee is *sample-wise and local*: it concerns one damped proximal step in the logits of a single categorical conditional of one already-generated trajectory, with the batch advantages held fixed. It is not a monotonic-improvement statement for the expected return under $y_t \sim \pi_\theta$, it makes no claim across re-sampled batches (over which the empirical A_t are re-drawn), and it does not model shared parameters across token positions, PPO clipping, or Adam preconditioning—it is not a convergence or targeted-correction guarantee for the full network. (b) A_t enters Q_t only through A_t^2: the theorem is sign-blind, so it justifies the *magnitude* $1/(\tau_t + \varepsilon_0)$ of the corrective branch of Eq. 5 but not its sign asymmetry—that asymmetry is a design choice motivated by Remark 4.3, and whether it outperforms symmetric rules is an empirical question, answered in Table 3. (c) The analysis treats a single token position with a local logit update; loss reweighting equals per-token logit stepping only to first order at the current iterate, and in practice $\beta_{\mathrm{kl}}, \eta$ are absorbed by batch normalization, so only the *shape* $1/(\tau_t + \varepsilon_0)$ survives—which is why we label Eq. 5 a design rule rather than a corollary. Finally, the KL penalty in Eq. 6 is deliberately left unweighted: tokens receiving large corrective weights are permitted to move further from the reference policy, which we consider a feature of the correction branch rather than an oversight.

D Relation Analysis of Different Entropies

DEEPO uses two entropy signals at different levels, and it is natural to ask whether they measure the same thing. They do not: semantic entropy H_s captures answer-level dispersion across sampled rollouts (Sec. 3), while the Rényi-2 entropy H_2 characterizes token-level policy sharpness and, by Proposition 4.1, the expected score-gradient norm. The two are linked through the token-level Shannon entropy $H_1(\pi(\cdot \mid q, y_{<t})) = -\sum_{y\in\mathcal{V}} \pi(y \mid q, y_{<t}) \log \pi(y \mid q, y_{<t})$, the $\alpha \to 1$ limit of Rényi entropy, whose functional form mirrors that of H_s (a Shannon entropy over meanings versus over tokens). The chain

$$F(H_2; n) \;\geq\; H_1(p) \;\geq\; H_2(p) \;\geq\; \log n - \log\Big(1 + 2n\big(\log n - H_1(p)\big)\Big), \tag{20}$$

summarized in the following proposition, makes the link quantitative.

Proposition D.1 (Entropy chain; known results)**.** *Let $p = (p_1, \ldots, p_n)$ denote the softmax distribution over an n-token vocabulary. Then the chain above holds, where $F(H_2; n) = -\frac{1+(n-1)s}{n} \log\big(\frac{1+(n-1)s}{n}\big) - \frac{(n-1)(1-s)}{n} \log\big(\frac{1-s}{n}\big)$ and $s(H_2) := \sqrt{\frac{ne^{-H_2}-1}{n-1}} \in [0,1]$. The upper bound is the sharp bound of Harremoës & Topsøe (2001), attained exactly by the two-level distribution $p^\star = \big(\frac{1+(n-1)s}{n}, \frac{1-s}{n}, \ldots, \frac{1-s}{n}\big)$; $H_1 \geq H_2$ is Rényi monotonicity (equality iff p is uniform on its support); the lower bound follows from Pinsker's inequality and is tight only at the uniform distribution (it is strict, and in fact vacuous, at degenerate distributions—see Remark E.1). Self-contained derivations are given in App. E.*

The practical content for our design: the two branches do not require their entropy signals to be interchangeable or even strongly correlated—semantic entropy gates the rollout intervention, while token-level H_2 parameterizes optimization—and the chain above guarantees only that both are faithful measures of uncertainty at their respective levels. *A note on scope:* we deliberately do not use an empirical H_s–H_1 correlation analysis to justify combining the two signals; such a correlation would neither be necessary for the design nor sufficient to establish complementarity, and the H_s–H_2 chain is presented here solely to locate both quantities on a common uncertainty footing.

E Relation Between H_1 and H_2

This appendix collects the relations between Shannon entropy H_1 and Rényi-2 entropy H_2 used in App. D. None of the results is new: $H_1 \geq H_2$ is Rényi monotonicity; the upper envelope $F(H_2; n)$ is the sharp entropy–index-of-coincidence bound of Harremoës & Topsøe (2001); the lower bound follows from Pinsker's inequality (cf. Fedotov et al. (2003)). We re-derive them in our notation for completeness.

Let $p = (p_1, \ldots, p_n)$ be a discrete distribution on $n = |\mathcal{V}|$ support points, $H_1(p) = -\sum_{i=1}^n p_i \log p_i$, $H_2(p) = -\log\big(\sum_{i=1}^n p_i^2\big)$, $S_2(p) = \sum_i p_i^2 = e^{-H_2}$.

(*i*) $H_1 \geq H_2$. Let X take value p_i with probability p_i. Since $\log$ is concave, Jensen's inequality gives $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$, i.e. $\sum_i p_i \log p_i \leq \log \sum_i p_i^2$; negating yields $H_1(p) \geq H_2(p)$, with equality iff p is uniform on its support.

(*ii*) Upper envelope $H_1 \leq F(H_2; n)$ Harremoës & Topsøe (2001). Under the constraints $\sum_i p_i = 1$, $\sum_i p_i^2 = C$, the symmetric Schur-concave H_1 is maximized (by KKT / symmetry arguments) at a two-level distribution

$$p = (a, \underbrace{b, \ldots, b}_{n-1 \text{ entries}}), \qquad b = \frac{1-a}{n-1}. \tag{21}$$

The quadratic constraint gives $(n-1)C = na^2 - 2a + 1$; solving with $a \geq 1/n$ and introducing

$$s := \sqrt{\frac{nC-1}{n-1}} = \sqrt{\frac{ne^{-H_2}-1}{n-1}} \in [0, 1], \tag{22}$$

yields $a = \frac{1+(n-1)s}{n}$, $1 - a = \frac{(n-1)(1-s)}{n}$, and substituting into $H_1(a, b) = -a \log a - (1 - a) \log \frac{1-a}{n-1}$ gives the explicit envelope

$$F(H_2; n) = -\frac{1+(n-1)s}{n} \log\Big(\frac{1+(n-1)s}{n}\Big) - \frac{(n-1)(1-s)}{n} \log\Big(\frac{1-s}{n}\Big). \tag{23}$$

(*iii*) Lower bound via Pinsker. Let $u = (1/n, \ldots, 1/n)$. Since $\chi^2(p\|u) = nS_2(p) - 1 = n\|p - u\|_2^2$ and

$$D_{\mathrm{KL}}(p\|u) \geq \tfrac{1}{2}\|p - u\|_1^2 \geq \tfrac{1}{2}\|p - u\|_2^2 = \tfrac{1}{2}\Big(S_2(p) - \tfrac{1}{n}\Big), \tag{24}$$

we obtain $S_2(p) \leq \big(1 + 2nD_{\mathrm{KL}}(p\|u)\big)/n$. Taking $-\log$ and using $D_{\mathrm{KL}}(p\|u) = \log n - H_1(p)$:

$$H_2(p) \geq \log n - \log\Big(1 + 2n\big(\log n - H_1(p)\big)\Big). \tag{25}$$

Remark E.1 (Tightness of the entropy chain). (i) At the uniform distribution $p = u$: $s = 0$, the envelope coincides with u, $F(\log n; n) = H_1 = H_2 = \log n$, and since $D_{\mathrm{KL}}(u\|u) = 0$ the lower bound holds with equality—the full chain is tight at the uniform point. (ii) At any degenerate p: $s = 1$, the envelope coincides with p, so $F(0; n) = H_1 = H_2 = 0$ and the *upper* bound is tight; the lower bound, however, evaluates to $\log n - \log(1 + 2n \log n) < 0$ for every $n \geq 2$ (since $1 + 2n \log n > n$), so it is *strict*—indeed vacuous—there. More generally, the lower bound is negative whenever $D_{\mathrm{KL}}(p\|u) > (n-1)/(2n)$, and by the equality conditions of Pinsker's inequality it is attained only at $p = u$. (iii) $H_1 = H_2$ iff p is uniform on its support.

F Protocol for the H_s–H_1 Alignment Analysis

This appendix specifies the protocol for the empirical alignment analysis referenced in App. D.

Quantities. For each evaluation question q on HallusionBench and MMStar (multiple-choice converted to cloze-style, as in App. A.2), we draw M stochastic samples (temperature 1.2) from the evaluated policy and compute: (i) the semantic entropy $H_s(q)$ over answer-level clusters (Sec. 3, with the same protocol and budget M); (ii) the token-level Shannon entropy H_1, averaged over

answer-span tokens and over the M samples,

$$\bar{H}_1(q) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|y_i^{\text{ans}}|}\sum_{t\in y_i^{\text{ans}}} H_1\big(\pi(\cdot \mid q, y_{i,<t})\big), \tag{26}$$

where y_i^{ans} denotes the extracted answer span of sample i. Both quantities are computed at inference; expert hints are used only during *training*, so "with-hint" columns below compare *models trained with vs. without* the hint branch, not hint-conditioned decoding.

Statistics. We report the Pearson coefficient ρ over questions between $\bar{H}_1(q)$ and $H_s(q)$, and the Spearman rank coefficient ρ_{sp} as a robustness check. Significance uses a two-tailed t-test at level $\alpha = 0.01$ with statistic $t = \rho\sqrt{M-2}/\sqrt{1-\rho^2}$, where M is the number of questions.

Table 12 reports the resulting correlations (ρ: Pearson; ρ_{sp}: Spearman; all $p < 0.01$, two-tailed t-test). The alignment is already present in the GRPO baseline and strengthens substantially with the hint branch on both benchmarks, supporting the claim that the two entropy signals are linked through predictive uncertainty rather than interchangeable.

Table 12: Correlation between $\bar{H}_1$ and H_s on multimodal benchmarks (Pearson ρ / Spearman ρ_{sp}), for models trained without and with the hint branch. All coefficients significant at $p < 0.01$.

Model	HallusionBench ρ / ρ_{sp}	MMStar ρ / ρ_{sp}
GRPO (w/o hint branch)	0.48 / 0.45	0.44 / 0.42
DEEPO (w/ hint branch)	**0.62 / 0.59**	**0.58 / 0.55**

G Training Details

G.1 Implementation Specification

For reproducibility we consolidate the design choices and hyperparameters that the main text references. *Semantic-entropy estimation:* answers are extracted from the rollout suffix after the final-answer marker, lowercased and stripped of articles and punctuation; numeric answers are canonicalized. Answers are clustered into semantic-equivalence classes by pairwise DeBERTa-MNLI entailment (`microsoft/deberta-v2-xlarge-mnli`) under a lax rule—two answers are equivalent when they are mutually non-contradictory and at least one direction entails the other; cluster probabilities are count-based, $P(C_k \mid q) = |C_k|/N$, and H_s is the natural-log Shannon entropy over these frequencies. *Prefix truncation:* the same final-answer marker delimits the thought/answer split $y^* = [\,y^*_{\text{thought}}, y^*_{\text{ans}}\,]$ of each expert trace; prefixes are cut from the thought segment only, at L_h tokens given by Eq. 1 with truncation cap $\alpha_{\max} = 0.5$, so the prefix never reaches the answer segment (we verified on the full training set that no prefix contains the gold answer string). *Semantic-leakage audit:* token-level exclusion does not by itself rule out semantic leakage, since an early reasoning prefix could still determine the conclusion. We therefore audit every candidate prefix with the same DeBERTa-MNLI model used for clustering: the gold answer is verbalized as a declarative statement, and a prefix is used only if the model does not predict entailment from the prefix to that statement. On the 20k training traces, 3.1% of candidate prefixes fail this test and are discarded (the query then receives no prefix for that iteration); a manual inspection of 500 filtered prefixes found none that determines the final label. *Trigger:* the threshold is initialized at $\tau_s = 0.8$ and updated by exponential moving average with decay β; $H_s^{\max} = \log N$; hint dropout with rate p_{drop} is applied, and the trigger is evaluated before dropout. *Scaling:* τ_t is computed from $\pi_{\theta_{\text{old}}}$'s fp32 log-probabilities with stop-gradient; $w_{\max} = 20$ ($\varepsilon_0 = 0.05$); $g^\star$ is mean-normalized over the batch and applied after group advantage normalization and before clipping; we log the effective sample size of the weights and the clipped-token fraction as diagnostics. *Optimization:* clip $\varepsilon_{\text{clip}} = 0.2$, KL coefficient $\beta_{\text{kl}} = 0.04$, prefix-loss weight λ, actor learning rate 5×10^{-7}; $G = 8$, sampling temperature 1.2.

We fine-tune the Qwen2.5-VL-7B model under the GRPO framework to enhance multimodal reasoning and reduce hallucination. The training covers both image–text and video–text modalities, unified through a shared tokenizer to maintain consistent encoding across modalities.

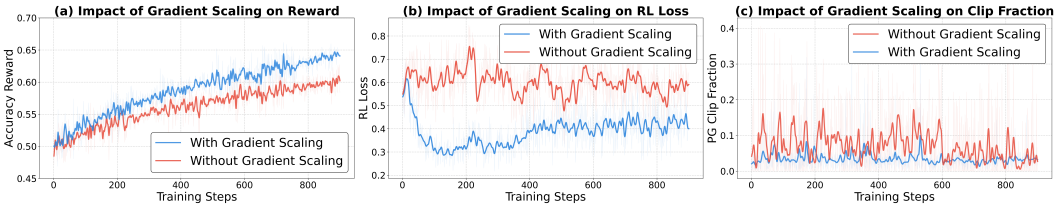


Figure 5: Effect of Gradient Scaling on RL stability. Adaptive weighting stabilizes RL loss, accelerates reward convergence, and reduces policy clipping, yielding smoother and more robust training.

Data Preprocessing.

We employ a unified visual–text tokenizer that handles both static and temporal visual inputs. For image–text pairs, the images are dynamically resized based on token limits (4–16,384 tokens) while preserving aspect ratio. For video–text pairs, the preprocessing dynamically adjusts frame counts between 4 and 64 based on original FPS and duration, ensuring each clip maintains a consistent spatial resolution and frame factor of 2.

Training Framework.

We adopt the GRPO reinforcement learning setup Shao et al. (2024), which is critic-free: advantages are estimated relative to the group of sampled responses (Sec. 3), and no value network is trained. For each GRPO iteration, $G = 8$ candidate responses are sampled per multimodal input (image– or video–question pair) using a temperature of 1.2. The policy is updated with Adam at a learning rate of 5×10^{-7}. A KL penalty of 0.04 constrains the updated policy within the trust region defined by the reference model, stabilizing optimization.

Hyperparameters.

Table 13 consolidates the configuration referenced throughout the paper.

Table 13: Full hyperparameter configuration.

Parameter	Value
Backbone	Qwen2.5-VL-7B Bai et al. (2025)
Training data	Video-R1 Feng et al. (2025) multimodal RL split, 20k subset
GRPO group size G	8
Rollout temperature	1.2
Optimizer (policy)	Adam, lr $= 5 \times 10^{-7}$
Clip parameter $\varepsilon_{\text{clip}}$	0.2
KL coefficient β_{kl}	0.04
Training steps	2,000
Evaluation cadence	every 100 steps, fixed held-out prompts
SE estimation budget N	8 ($=G$; co-varied with G, App. A.3)
Initial SE threshold τ_s	0.8
Prefix truncation cap $\alpha_{\max}$ (Eq. 1)	0.5
Weight cap $w_{\max}$ ($\varepsilon_0 = 1/w_{\max}$)	20
Global batch size (prompts / step)	64
EMA decay β	0.05
Hint dropout rate p_{drop}	0.20
Prefix-loss weight λ	0.10
Semantic-equivalence model	`microsoft/deberta-v2-xlarge-mnli`
Entailment rule	lax (mutual non-contradiction + ≥ 1 directional entailment)
Cosine threshold	N/A (not used)
Entropy estimator	cluster-frequency probability, natural-log entropy
Hardware	8 $\times$ NVIDIA A100 (80GB)
Wall-clock	$\approx$3 days (training only, excluding periodic evaluation)

Algorithm 1 One DEEPO training iteration

Require: policy π_θ, reference π_{ref}, a batch of queries with expert traces, group size G, threshold state $\hat{\tau}$ (initialized at τ_s), EMA decay β, dropout rate p_{drop}, cap $w_{\max}$, λ, β_{kl}

1: Sample G rollouts per query with $\pi_{\theta_{\text{old}}}$ (temp. 1.2)
2: Extract answers; cluster by DeBERTa-MNLI entailment (lax rule); compute $H_s(q)$ per query
3: Update threshold: $\hat{\tau} \leftarrow (1-\beta)\hat{\tau} + \beta\,\overline{H}_s$
4: **for all** queries q with $H_s(q) > \hat{\tau}$ **do**
5: With prob. $1 - p_{\text{drop}}$: truncate the *thought segment* of the expert trace to L_h tokens (Eq. 1; answer segment y^*_{ans} excluded); sample continuation $y_q^{\text{rem}} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q \oplus y_q^{\text{pre}})$
6: **end for**
7: Score all trajectories with the verifiable reward
8: Compute group-relative advantages over the full group: hinted continuations and first-pass unhinted rollouts share one baseline (in-group variance anchoring, Sec. 4.1)
9: Compute $\tau_t = 1 - e^{-H_{2,t}}$ per token from $\pi_{\theta_{\text{old}}}$ in fp32 (stop-grad); form $g^\star$ (Eq. 5)
10: Update π_θ by minimizing Eq. 6 (prefix masked in RL term; $\tilde{A}_t = g^\star A_t$ pre-clip)

Table 14: Efficiency–performance trade-off. "SE Time Share" covers answer extraction and clustering for H_s only; the conditional continuation pass and the prefix forward pass are separate delta components.

Method	Step Time (s) ↓	Wall-clock Δ	SE Time Share	MMBench ↑
Qwen2.5-VL-7B	-	-	-	83.1
+ GRPO	**109.8 ± 2.7**	**1.00**×	–	84.5
+ DEEPO w/o Expert Hint *(scaling only)*	111.9 ± 2.9	+1.9%	∼0.0%	85.0
+ DEEPO w/o Gradient Scaling *(hint only)*	123.6 ± 3.8	+12.6%	4.7% ± 0.6%	85.5
+ DEEPO full	129.6 ± 4.1	+18.0%	5.4% ± 0.7%	**86.4**

Dataset and Sampling.

Our models are trained on the large-scale multimodal dataset introduced in Video-R1 Feng et al. (2025), selecting 20k samples as the final training subset. Videos are preprocessed through the provided sampling pipeline, which adaptively determines frame counts based on original FPS and duration while maintaining aspect ratio and pixel limits. Each sample is converted into an instruction-style multimodal prompt for consistency with the training objective. The reward is verifiable: task accuracy on the final answer plus a format-compliance bonus, following Feng et al. (2025).

Hardware and Schedule.

All experiments are conducted on 8 × NVIDIA A100 (80GB) GPUs. Training runs for 2,000 GRPO steps (approximately 3 days of pure training time at ≈130 s/step for the full model), with evaluation conducted every 100 steps using a fixed set of held-out prompts; periodic evaluation and checkpointing add a small additional overhead not included in the per-step timing of Table 14.

Per-Iteration Procedure

Algorithm 1 summarizes one training iteration of DEEPO. The procedure first estimates query-level uncertainty from grouped rollouts and uses an EMA threshold to identify ambiguous samples that benefit from expert guidance. For these samples, DEEPO conditionally injects a truncated expert prefix, while the token-level scaling branch independently reweights policy gradients according to local predictive uncertainty. Both branches are integrated into the same GRPO update, with hinted and unhinted trajectories sharing the group-relative baseline.

Efficiency Decomposition

Table 14 decomposes the additional training cost introduced by the two DEEPO branches. Gradient scaling adds little overhead relative to GRPO, whereas the expert-hint branch accounts for most of the extra wall-clock time due to answer clustering and conditional continuation generation.