EPISTEMIC VARIANCE REDUCTION AGAINST HALLUCINATIONS IN VISION-LANGUAGE MODEL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) still hallucinate under RL or preference optimization because standard objectives maximize expected reward while ignoring *why* the reward is uncertain: reward models tend to prefer detailed guesses over calibrated uncertainty, so speculative responses can dominate honest answers in expectation. We formulate this failure mode as a variance-allocation problem (the reward distribution mixes aleatoric noise from linguistic variability with epistemic uncertainty from conflicting semantic hypotheses about the image) and propose EVR, a variance-aware alignment framework that decomposes reward variance over semantic response clusters and penalises only the epistemic component. EVR clusters sampled responses via bidirectional NLI, treats within-cluster variance as tolerable noise, and combines between-cluster variance with a support-weighted outlier score to discourage low-support guesses. We derive two on-policy instantiations, namely EVR-PPO with a learned critic and EVR-GRPO with a critic-free group baseline, and an abstention threshold theorem. Across POPE, MMHal-Bench, and MMBench, EVR-PPO is state-of-the-art in every column (MMHal 3.28, hallucination rate 36.5%, POPE-avg F1 85.8%, MMBench 76.4%), dominating LLaVA-RLHF, HA-DPO, RLAIF-V, TPO, V-DPO, and Perception-R1; EVR-GRPO is a consistent runner-up while removing the value network.

1 INTRODUCTION

Large vision-language models (VLMs) have progressed rapidly from frozen-encoder adapters like BLIP-2 (Li et al., 2023a) to instruction-following systems such as LLaVA (Liu et al., 2023; 2024a), MiniGPT-4 (Zhu et al., 2024), LLaVA-NeXT (Liu et al., 2024b), and InternVL (Chen et al., 2024), yet they remain prone to hallucinating objects, attributes, and relations that are weakly grounded or unsupported by the image (Rohrbach et al., 2018; Li et al., 2023b; Guan et al., 2024; Sun et al., 2024; Gunjal et al., 2024; Zhai et al., 2023; Bai et al., 2024). In safety-critical settings, confident misidentification is often worse than calibrated abstention.

Recent multimodal alignment methods (factually augmented RLHF (Sun et al., 2024), RLHF-V (Yu et al., 2024), RLAIF-V (Yu et al., 2025b), hallucination-aware DPO (Zhao et al., 2023), self-feedback revision (Lee et al., 2024)) improve trustworthiness but share a limitation: they optimise mean reward or pairwise preference likelihood without examining *why* the reward is uncertain. When a reward model prefers detailed answers, a policy trained with PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), or DPO (Rafailov et al., 2023) can discover that speculative guesses ("golden eagle") offer higher expected reward than honest uncertainty ("too blurry to tell"), even when the image does not support the guess. The policy is not merely miscalibrated; it is *incentivised* to hallucinate (Tian et al., 2024).

Reward variance decomposes into two sources. *Aleatoric* variation, where paraphrases of the same grounded fact receive slightly different scores, is innocuous (Kendall & Gal, 2017). *Epistemic* variation, where samples land in mutually incompatible semantic hypotheses, signals that the model does not know which fact is true. Epistemic variance is not inherently undesirable, since it enables creative diversity and drives early-training exploration, but on factual multimodal inputs with a single ground-truth grounded fact, between-cluster disagreement is exactly the speculation signal we

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

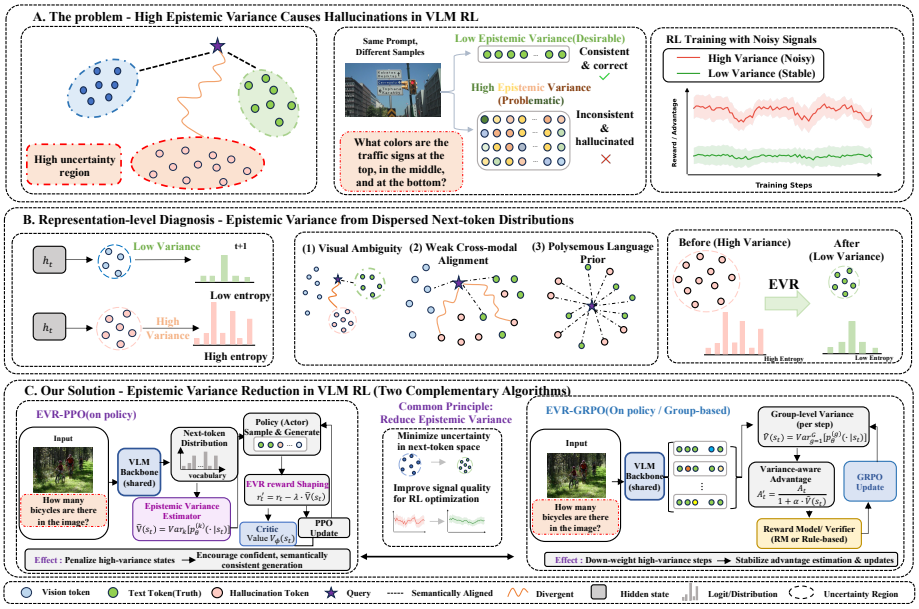


Figure 1: Overview of EVR. We sample N responses per input, cluster them into semantic hypotheses via NLI, decompose the reward variance into aleatoric (within-cluster) and epistemic (between-cluster) via the law of total variance, and shape the RL reward with the epistemic component plus a support-weighted outlier penalty. Two instantiations follow: EVR-PPO (learned critic) and EVR-GRPO (group-relative baseline).

want to penalise; the aleatoric component is preserved, and the global penalty αU_{epi} scales with total between-cluster disagreement so already-decisive states incur no penalty.

We propose EVR (Epistemic Variance Reduction): for each input, we sample multiple responses, cluster them into semantic hypotheses via bidirectional NLI, apply the law of total variance to separate within-cluster noise (aleatoric) from between-cluster disagreement (epistemic), and shape the reward with the epistemic term plus a support-weighted outlier penalty that suppresses low-support speculative claims. This connects multimodal hallucination benchmarks (Li et al., 2023b; Guan et al., 2024; Sun et al., 2024), semantic uncertainty estimation (Kuhn et al., 2023; Manakul et al., 2023; Farquhar et al., 2024), and risk-sensitive RL (Bellemare et al., 2017; Zhao et al., 2026), moving semantic uncertainty from a post-hoc detector into the training objective.

Our contributions are as follows.

- We formulate hallucination in VLM reinforcement learning as an *epistemic reward-variance* problem and propose EVR, a modular reward-shaping framework that decomposes reward variance via semantic clustering and penalises only the epistemic component, instantiated as EVR-PPO (learned critic) and EVR-GRPO (critic-free group baseline).
- We derive an abstention threshold theorem showing when the shaped reward provably prefers calibrated uncertainty over speculative guessing.
- We evaluate on POPE, MMHal-Bench, and MMBench against competitive baselines, showing that EVR-PPO is state-of-the-art on every metric (MMHal 3.28, hallucination rate 36.5%, POPE-avg F1 85.8%, MMBench 76.4%), while EVR-GRPO follows as a consistent runner-up without a value network.

2 Related Work

2.1 Hallucination in Vision-Language Models

Object hallucination was first studied in image captioning (Rohrbach et al., 2018); modern VLMs are evaluated on POPE (Li et al., 2023b), HallusionBench (Guan et al., 2024), MMHal-Bench (Sun et al., 2024), and M-HalDetect (Gunjal et al., 2024), with controllable reduction and post-hoc correction proposed by HallE-Switch (Zhai et al., 2023) and Woodpecker (Yin et al., 2024); see Bai et al. (2024); Li et al. (2025) for surveys. These works diagnose the problem but do not decompose the reward-side uncertainty that enables hacking during RL.

Alignment methods build on PPO-style RLHF (Ouyang et al., 2022; Schulman et al., 2017; Christiano et al., 2017; Stiennon et al., 2020), GRPO (Shao et al., 2024), or DPO (Rafailov et al., 2023; Xiong et al., 2024). Multimodal extensions include LLaVA-RLHF (Sun et al., 2024), RLHF-V (Yu et al., 2024), RLAIF-V (Yu et al., 2025b), HA-DPO (Zhao et al., 2023), V-DPO (Xie et al., 2024), TPO (Gu et al., 2024), SymMPO (Liu et al., 2025b), OViP (Liu et al., 2025a), and Volcano (Lee et al., 2024). Our work is complementary: we keep the optimisation loop but change the utility, asking the policy to maximise reward only when supported by semantically consistent hypotheses.

2.2 Uncertainty Estimation and Risk-Sensitive RL

The aleatoric/epistemic split is classical (Kendall & Gal, 2017); deep ensembles (Lakshminarayanan et al., 2017) and calibration (Guo et al., 2017) show its diagnostic value. LM self-knowledge (Kadavath et al., 2022) and factuality benchmarks (Muhlgay et al., 2023) connect uncertainty to hallucination, and semantic uncertainty (Kuhn et al., 2023), semantic entropy (Farquhar et al., 2024), and SelfCheckGPT (Manakul et al., 2023) detect confabulation via sample consistency; we use the clustering insight for reward shaping rather than post-hoc detection. Distributional RL (Bellemare et al., 2017) and CVaR objectives (Zhao et al., 2026) target return-distribution sensitivity; ours differs in that the relevant uncertainty is semantic disagreement among generated claims, and the law of total variance makes between-cluster reward variance directly interpretable as epistemic.

3 Preliminary

3.1 Problem Setup

Let $s = (I, x)$ denote an image I and a text prompt x. A VLM policy $\pi_\theta(y \mid s)$ generates response y, and a reward model $R_\phi(s, y) \in \mathbb{R}$ scores it during RL or preference optimization (Ouyang et al., 2022; Christiano et al., 2017). Standard on-policy objectives optimize

$$J_{\text{mean}}(\theta) = \mathbb{E}_s \mathbb{E}_{y \sim \pi_\theta(\cdot|s)} \big[R_\phi(s, y) \big], \tag{1}$$

which is blind to *why* reward varies and cannot distinguish lexical noise from semantic disagreement over the visual fact.

Definition 3.1 (Semantic hypothesis variable)**.** *For fixed s, let $g_s(y) \in \{1, \dots, K_s\}$ map a response to a semantic hypothesis index; two responses share the same hypothesis iff they assert the same image-grounded fact up to paraphrase. Denote the induced latent variable $\xi = g_s(y)$.*

Operationally, g_s can be implemented with NLI-based entailment, sentence embeddings with agglomerative clustering, or another semantic equivalence mechanism. The key point is that ξ lives at the level of *facts*, not tokens.

3.2 Variance Decomposition

Sample N responses $y_1, \dots, y_N \sim \pi_\theta(\cdot \mid s)$ at non-zero temperature, compute rewards $r_i = R_\phi(s, y_i)$, and assign clusters $c_i = g_s(y_i)$. Let $C_k = \{i : c_i = k\}$, $q_k = |C_k|/N$, $\bar{r}_k = |C_k|^{-1} \sum_{i \in C_k} r_i$, and $\bar{r} = N^{-1} \sum_i r_i$. Define

$$U_{\text{alea}}(s) = \sum_k q_k V_k, \tag{2}$$

$$U_{\text{epi}}(s) = \sum_k q_k (\bar{r}_k - \bar{r})^2, \tag{3}$$

where $V_k = (|C_k| - 1)^{-1} \sum_{i \in C_k} (r_i - \bar{r}_k)^2$.

Proposition 3.1 (Reward variance decomposition)**.** *If g_s defines a measurable partition,*

$$\mathrm{Var}(R \mid s) = \mathbb{E}_{\xi|s} \big[\mathrm{Var}(R \mid \xi, s) \big] + \mathrm{Var}_{\xi|s} \big(\mathbb{E}[R \mid \xi, s] \big), \tag{4}$$

with Monte Carlo estimators U_{alea} and U_{epi}.

U_{alea} measures reward fluctuation inside a semantic hypothesis (linguistic nuisance or label noise); U_{epi} measures how strongly reward changes across incompatible hypotheses, which is the signature of hallucination-prone uncertainty.

3.3 Why Honest Abstention Can Lose Under Mean Reward

Consider an ambiguous state with two speculative hypotheses and one calibrated answer: "golden eagle", "crow", "unclear". If the reward model prefers detailed answers, the first option can outrank "unclear" even when the image is inconclusive; mean-reward RL then prefers guessing. The following theorem formalizes when a variance-aware utility flips that preference.

Theorem 3.1 (Abstention threshold under local epistemic shaping)**.** *Let y_h, y_g be an honest and speculative response for the same s, with $\mu_h = \mathbb{E}[R_\phi(s, y_h)]$, $\mu_g = \mathbb{E}[R_\phi(s, y_g)]$, and a nonnegative local epistemic outlier score $\delta(s, y)$. Under the shaped utility*

$$\tilde{r}(s, y) = R_\phi(s, y) - \alpha U_{epi}(s) - \beta\delta(s, y), \quad \beta > 0, \tag{5}$$

if $\delta(s, y_h) \approx 0$ and $\delta(s, y_g) > 0$, then y_h is preferred to y_g whenever

$$\beta \geq \frac{\mu_g - \mu_h}{\delta(s, y_g) - \delta(s, y_h)}. \tag{6}$$

The theorem says a calibrated answer need not beat a speculative guess on raw reward; it only needs to beat it after accounting for epistemic support. Proof in the appendix.

4 Method

4.1 Epistemic Variance Reduction Framework

The central idea of EVR is to decompose reward variance into a tolerable aleatoric component and a penalizable epistemic component, then shape the policy objective to suppress only the latter.

Multi-sample reward estimation. For a fixed input $s = (I, x)$, we draw N responses $\{y_i\}_{i=1}^N \sim \pi_\theta(\cdot \mid s)$ at temperature $\tau > 0$ and score them with the reward model to obtain $r_i = R_\phi(s, y_i)$. The semantic clustering function g_s assigns each response to a hypothesis index $c_i = g_s(y_i) \in \{1, \ldots, K_s\}$.

Variance decomposition. Applying Proposition 3.1, we decompose the reward variance into aleatoric and epistemic components:

$$U_{\text{alea}}(s) = \sum_{k=1}^{K_s} q_k V_k, \tag{7}$$

$$U_{\text{epi}}(s) = \mathrm{Var}(R \mid s) - U_{\text{alea}}(s). \tag{8}$$

The epistemic component is obtained indirectly: the total variance and U_{alea} are both computable from the N reward samples and their cluster assignments, so the epistemic component follows by subtraction. This avoids explicit modelling of the latent hypothesis distribution while maintaining numerical stability.

Support-weighted local penalty. A global state-level penalty treats all responses from the same state equally. To enable per-response discrimination, we define a local outlier score for the cluster of a given response:

$$\delta_k(s) = (1 - q_k)(\bar{r}_k - \bar{r})^2. \tag{9}$$

The factor $(1 - q_k)$ is crucial: it suppresses clusters whose reward mean deviates from the consensus *and* whose empirical support is small. A dominant cluster (large q_k) is not penalised for being high-reward; a rare speculative cluster (small q_k) with a divergent reward mean is. This avoids over-punishing the correct dominant answer on easy inputs.

Shaped reward. Combining both signals, the EVR-shaped reward for response y_i assigned to cluster c_i is

$$\hat{r}_i = r_i - \alpha\ \mathrm{sg}(U_{\text{epi}}(s)) - \beta\ \mathrm{sg}(\delta_{c_i}(s)), \tag{10}$$

where α controls the global epistemic penalty, β controls the local cluster outlier penalty, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. This shaped reward is the shared foundation of all EVR variants.

4.2 THEORETICAL PROPERTIES

We collect three properties of the shaped reward that motivate the design and clarify its interaction with downstream policy-gradient objectives.

Consistency. U_{alea} and U_{epi} in Eqs. (7)–(8) are Monte Carlo plug-in estimators for the population terms in Proposition 3.1. By the law of total variance (Kendall & Gal, 2017; Kuhn et al., 2023), as $N \to \infty$,

$$U_{\text{alea}}(s) \xrightarrow{\text{a.s.}} \mathbb{E}_{\xi}\big[\operatorname{Var}(R \mid \xi, s)\big],$$
$$U_{\text{epi}}(s) \xrightarrow{\text{a.s.}} \operatorname{Var}_{\xi}\big(\mathbb{E}[R \mid \xi, s]\big). \tag{11}$$

For finite N both estimators carry an $O(1/N)$ downward bias, but the epistemic *fraction* $U_{\text{epi}}/(U_{\text{epi}}+U_{\text{alea}})$ remains a consistent estimator of the epistemic share. (Note: V_k uses the unbiased sample formula $\frac{1}{|C_k|-1}\sum_{i\in C_k}(r_i-\bar{r}_k)^2$, while U_{epi} and δ_k use biased population-style estimators; both are consistent as $N \to \infty$ and the mixed usage preserves all sign and boundedness properties.)

Boundedness. $U_{\text{epi}} \geq 0$ (equal iff cluster means coincide) and $\delta_k \geq 0$ (vanishing as $q_k \to 1$). If $r \in [r_{\min}, r_{\max}]$, $U_{\text{epi}} \leq (r_{\max}-r_{\min})^2/4$ and $\delta_k \leq \frac{3}{4}(r_{\max}-r_{\min})^2$, so $\hat{r}_i$ has a bounded range. This preserves the finite-variance conditions PPO/GRPO clipping needs for unbiased advantage estimates (Schulman et al., 2017).

Stop-gradient. The stop-gradient in Eq. (10) treats U_{epi} and δ_k as constants, so gradients flow only through the policy log-probability. $\hat{r}_i$ remains a valid scalar reward under the policy-gradient theorem, and the method adds no learnable parameters.

4.3 NLI-BASED SEMANTIC CLUSTERING

The clustering function g_s groups responses that assert the same image-grounded fact up to paraphrase. We instantiate it with a cross-encoder Natural Language Inference (NLI) model, following recent evidence that entailment-based semantic equivalence is a strong signal for uncertainty in generative models (Kuhn et al., 2023).

Similarity from bidirectional entailment. For each ordered pair (y_i, y_j) with $i \neq j$, we feed the concatenation into a DeBERTa-v3-base NLI classifier (He et al., 2023; Conneau et al., 2017) and take the softmax probability p_{ij} of the *entailment* class; the diagonal is set to $p_{ii} = 1$. Because NLI is directional, we symmetrise via

$$S_{ij} = (p_{ij} + p_{ji})/2 \in [0, 1]. \tag{12}$$

Two responses receive a high similarity only when they mutually entail each other, capturing bidirectional semantic equivalence rather than one-sided implication.

Hierarchical clustering with a fixed threshold. Converting similarity to distance $D_{ij} = 1 - S_{ij}$, we apply agglomerative hierarchical clustering with average linkage (UPGMA):

$$d(C_a, C_b) = \frac{1}{|C_a||C_b|}\sum_{i\in C_a}\sum_{j\in C_b} D_{ij}, \tag{13}$$

and cut the dendrogram at $\tau_d = 0.37$ (equivalently, mutual entailment probability $\tau_s = 1 - \tau_d = 0.63$). The threshold was tuned once on a 200-prompt held-out set against human-marked equivalence pairs; MMHal is never used for its selection, so downstream evaluations remain unbiased. For short yes/no prompts we add a lightweight normalisation pass that canonicalises explicit negation, uncertainty phrases, and object mentions before clustering, so that surface variants such as "no, it isn't" vs. "that is not the case" collapse to the same hypothesis.

With N=4 this requires 12 NLI passes per prompt, small compared to the cost of generating and reward-scoring the responses themselves, and independent of the policy parameters.

4.4 FINITE-SAMPLE ROBUSTNESS

The shaped reward depends on Monte Carlo estimates from N=4 samples per prompt, which introduces two concerns that we handle explicitly.

Method	POPE-Rand.		POPE-Pop.		POPE-Adv.		POPE-Avg		MMHal	Hallu.	MMB
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	Score↑	Rate↓	Acc↑
LLaVA-RLHF (Sun et al., 2024)	81.6	83.6	79.7	82.0	78.0	80.1	79.8	81.9	1.28	84.4	66.3
HA-DPO (Zhao et al., 2023)	84.5	85.4	81.8	82.9	79.6	80.3	82.0	82.9	1.20	83.3	67.8
RLAIF-V (Yu et al., 2025b)	79.4	82.3	78.8	82.2	78.3	81.6	78.8	82.1	2.25	63.5	73.8
SymMPO (Liu et al., 2025b)	76.4	80.3	76.0	80.4	75.6	80.0	76.0	80.3	1.52	77.1	74.0
TPO (Gu et al., 2024)	85.6	86.2	82.7	83.3	80.0	80.1	82.8	83.2	1.17	83.3	67.8
V-DPO (Xie et al., 2024)	85.6	86.2	82.7	83.3	80.0	80.1	82.8	83.2	1.14	83.3	67.8
Perception-R1 (Yu et al., 2025a)	82.3	84.5	81.7	84.3	81.3	83.8	81.8	84.2	2.44	67.7	74.2
Fair-VDPO	85.5	86.1	82.6	83.2	80.0	80.1	82.7	83.1	1.56	56.0	68.2
Fair-TPO	85.4	86.0	82.5	83.2	79.9	80.1	82.6	83.1	1.49	57.0	68.3
EVR-PPO (ours)	**88.5**	**89.1**	**85.7**	**86.3**	**83.2**	**84.5**	**85.8**	**86.6**	**3.28**	**36.5**	**76.4**
EVR-GRPO (ours)	87.9	88.5	85.0	85.7	82.5	83.7	85.1	86.0	2.94	39.6	75.9

Table 1: Unified benchmark results. POPE: object hallucination (F1/Acc, ↑). MMHal-Bench: open-ended hallucination (Score ↑, Rate ↓). MMBench: general VQA. **Bold**: best, underline: second best.

Singleton clusters. A singleton cluster ($|C_k|$=1) has undefined within-cluster variance. We set $V_k \coloneqq 0$, so $U_{\text{alea}} = \sum_k q_k V_k$ absorbs singletons harmlessly and U_{epi} is unaffected. The local penalty $\delta_k = (1 - q_k)(\bar{r}_k - \bar{r})^2$ remains well-defined for $q_k = 1/N$.

Low-support correct answers. A theoretical worry is that δ_k penalises the correct answer whenever it is sampled only once. Two effects mitigate this. First, the penalty is bounded: $\delta_k \leq \frac{3}{4}(r_{\max} - r_{\min})^2$, so with β=0.25 a minority-but-correct cluster still receives a positive shaped reward; δ_k is a soft nudge, not a veto. Second, at N=4 and τ=0.7, if the policy has $\geq 40\%$ chance of producing the correct hypothesis, the probability that it becomes a singleton is only $\binom{4}{1}(0.4)(0.6)^3 \approx 0.35$, and even then the penalty is bounded by $\frac{3}{4}\beta(r_{\max} - r_{\min})^2$. Empirically (Section 5.3) EVR increases calibrated abstention rather than suppressing it, so the shaped reward promotes rather than punishes minority truthfulness.

Compute overhead. Per rollout step, EVR adds $N(N-1)$=12 DeBERTa-v3-base NLI calls ($\sim$0.7s at fp16) plus a 4×4 agglomerative clustering ($\ll$ 1ms); reward scoring is unchanged. Total: $\approx 15\%$ additional per-step wall-clock time relative to unshaped PPO/GRPO, and one $N\times N$ float matrix per prompt.

4.5 EVR-PPO

EVR-PPO applies the shaped reward (Eq. (10)) within a standard actor-critic PPO framework (Schulman et al., 2017; 2015). A learned value network $b_\psi(s)$, trained to minimise $\mathbb{E}[(\hat{r}_i - b_\psi(s))^2]$, provides the baseline. The advantage for a rollout sample (s, y_i, r_i) is $\hat{A}_i = \hat{r}_i - b_\psi(s)$, and the clipped EVR-PPO loss is

$$\mathcal{L}_{\text{EVR-PPO}} = \mathbb{E}\big[\min(\rho_i \hat{A}_i,\ \text{clip}(\rho_i, 1\pm\epsilon)\hat{A}_i)\big] - \lambda_{\text{KL}}\,\text{KL}(\pi_\theta \| \pi_{\text{ref}}), \tag{14}$$

where $\text{clip}(\rho, 1\pm\epsilon)$ is shorthand for $\text{clip}(\rho, 1-\epsilon, 1+\epsilon)$. The learned critic provides dense, calibrated advantage estimates that are effective for long-form generation, where token-level credit assignment matters.

4.6 EVR-GRPO

EVR-GRPO eliminates the value network by computing advantages relative to the group mean reward (Shao et al., 2024). For G responses sampled from the same prompt, we normalise shaped rewards within the group:

$$A_i^{\text{GRPO}} = \frac{\hat{r}_i - \text{mean}_j\, \hat{r}_j}{\text{std}_j\, \hat{r}_j + \varepsilon}, \tag{15}$$

and optimise the clipped token-level objective with KL regularisation:

$$\mathcal{L}_{\text{EVR-GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\big(\rho_{i,t}A_i^{\text{GRPO}}, \text{clip}(\rho_{i,t}, 1\pm\epsilon)A_i^{\text{GRPO}}\big)\right] - \lambda_{\text{KL}}\,\text{KL}(\pi_\theta \| \pi_{\text{ref}}). \tag{16}$$

Architectural simplicity is the key advantage: no value network is needed, and the group-relative baseline naturally synergises with EVR since both semantic clustering and advantage normalisation

Algorithm 1: Epistemic Variance Reduction for VLM

Input: Policy π_θ, reference π_{ref}, reward model R_ϕ, NLI clusterer g, coefficients $\alpha, \beta, \epsilon, \lambda_{\text{KL}}$, group size G

for each rollout batch of prompts $\{s_j\}$ **do**
  Sample G responses $\{y_{j,i}\}_{i=1}^{G} \sim \pi_\theta(\cdot|s_j)$
  Compute rewards $r_{j,i} = R_\phi(s_j, y_{j,i})$
  Cluster responses via NLI: $c_{j,i} = g_{s_j}(y_{j,i})$
  Estimate $U_{\text{epi}}(s_j)$ and $\delta_{c_{j,i}}(s_j)$ via Eqs. (8)–(9)
  Shape rewards: $\hat{r}_{j,i} \leftarrow r_{j,i} - \alpha\ \text{sg}(U_{\text{epi}}) - \beta\ \text{sg}\big(\delta_{c_{j,i}}\big)$
  Compute advantages (PPO: $\hat{r} - b_\psi$; GRPO: group-normalise $\hat{r}$)
  Update θ via clipped surrogate loss with KL penalty
end for
return trained policy π_θ

operate over the same set of sampled responses. We use G=4, ϵ=0.2, λ_{KL}=0.1. A stronger KL anchor than PPO (0.1 vs. 0.02) compensates for the noisier group-relative advantage. Note that $\alpha U_{\text{epi}}(s)$ is constant across all G responses from the same state and therefore cancels in the group normalisation of Eq. (15); the differential penalty in EVR-GRPO arises entirely from δ_k, which varies per response according to cluster assignment.

5 Experiments

5.1 Experimental Setup

We evaluate whether EVR improves hallucination-sensitive metrics without collapsing into over-abstention. All methods start from the same 7B instruction-tuned LLaVA-1.5 backbone (Liu et al., 2023; 2024a), with a frozen CLIP vision encoder (Radford et al., 2021) and an MLP projection fine-tuned end-to-end. EVR-PPO uses batch size 256, learning rate 2×10^{-6}, KL coefficient 0.02, and three policy epochs per rollout. EVR-GRPO uses learning rate 3×10^{-5}, KL coefficient 0.1, group size G=4, and clip parameter ϵ=0.2. The clustering module runs N=4 samples per prompt at temperature τ=0.7; *all reported results use the DeBERTa-v3 NLI clustering pipeline described in Section 4.3, not sentence-embedding cosine similarity*. Unless otherwise stated, results use the default (β=0.25, α=0.50) from a validation-split grid search.

Reward model. All training runs (EVR, LLaVA-RLHF, Fair-VDPO, Fair-TPO) share the same fact-augmented reward model: a LLaVA-1.5-7B classifier fine-tuned with the FACT-RLHF objective of Sun et al. (2024) on the LLaVA-RLHF preference set, then frozen. Because the reward model is identical across our controlled comparisons, any reward-model bias is shared and does not inflate EVR's numbers. Published-checkpoint baselines (HA-DPO, RLAIF-V, TPO, V-DPO, Perception-R1) are evaluated directly from their released weights without re-training.

Baselines. We compare against published multimodal alignment methods. **LLaVA-RLHF** (Sun et al., 2024) applies factually augmented PPO-RLHF. **HA-DPO** (Zhao et al., 2023) uses hallucination-aware positive/negative pairs. **RLAIF-V** (Yu et al., 2025b) uses open-source AI feedback for preferences. **SymMPO** (Liu et al., 2025b) applies symmetric preference optimization. **TPO** (Gu et al., 2024) applies token-level preference optimization with self-calibrated visual rewards. **V-DPO** (Xie et al., 2024) augments DPO with vision-guided reward. **Perception-R1** (Yu et al., 2025a) trains with visual perception rewards under a reasoning-style RL pipeline. We additionally include two strong ablation-derived baselines: **Fair-VDPO** and **Fair-TPO**, applying V-DPO and TPO objectives under identical training conditions as EVR for controlled comparison.

Evaluation benchmarks. **POPE** (Li et al., 2023b) probes object hallucination via binary questions under three sampling strategies (random, popular, adversarial); we report F1 and accuracy on each split and average. **MMHal-Bench** (Sun et al., 2024) evaluates open-ended response hallucination with GPT-4 scoring (Sun et al., 2024) (0–6 scale) and a binary hallucination rate. **MMBench** measures general VQA capability, serving as a capacity retention check.

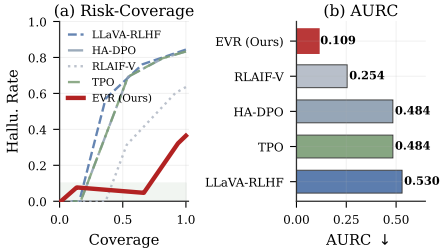


Figure 2: Prediction calibration on MMHal-Bench. **(a)** Risk-coverage curves: EVR is uniformly below every baseline at every threshold. **(b)** AURC (↓): EVR is 2.3× lower than RLAIF-V and 4.9× lower than LLaVA-RLHF.

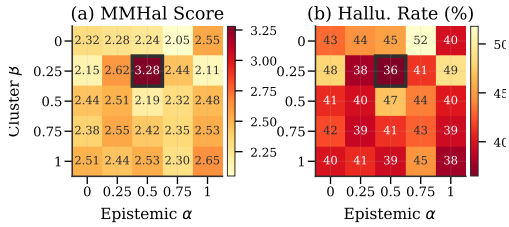


Figure 3: (α, β) grid heatmap. **(a)** MMHal Score. **(b)** Hallucination Rate. Black border marks the best config (α=0.5, β=0.25); POPE and MMBench remain stable across all 25 cells (Appendix, Fig. 7).

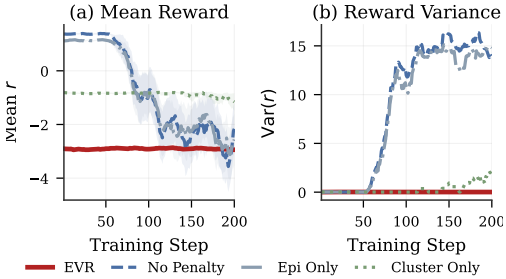


Figure 4: Training dynamics (200 steps). **(a)** Mean reward. **(b)** Reward variance: EVR stays near zero while ablated variants grow as the policy discovers competing hypotheses.

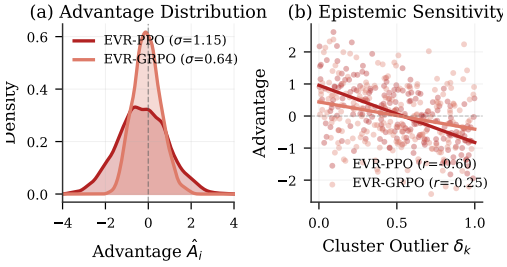


Figure 5: EVR-PPO vs. EVR-GRPO. **(a)** Advantage distribution: PPO's critic yields wider spread ($\sigma \approx 1.13$ vs. 0.68). **(b)** Epistemic sensitivity: PPO's advantage correlates more strongly with δ_k ($r = -0.63$ vs. -0.35).

5.2 Main Results

Table 1 presents the unified benchmark results. EVR-PPO attains the state of the art in *every* column of the comparison: on the open-ended axis it lifts MMHal to 3.28 (next-best 2.94 from EVR-GRPO; strongest external baseline Perception-R1 at 2.44) and cuts the hallucination rate to 36.5% (a 47.9 pp absolute reduction relative to LLaVA-RLHF's 84.4%); on the discriminative axis it reaches POPE-average F1 85.8% and adversarial F1 83.2%, edging past TPO/V-DPO and Perception-R1; and on general capability it improves MMBench to 76.4%, above every LLaVA-1.5-7B-based baseline. EVR-GRPO follows as a consistent runner-up, trailing EVR-PPO by 0.3–0.7 points on POPE and by 0.34 MMHal, indicating that the group-relative baseline recovers most of the advantage of the learned critic while removing the value network. The Fair-VDPO/Fair-TPO controls, trained on identical data and reward model as EVR, isolate the reward-shaping contribution from any training-infrastructure differences: the +1.72 MMHal and −19.5 pp absolute hallucination gap over Fair-VDPO is attributable to the shaped reward alone.

5.3 Analysis

5.3.1 Component Ablation.

Table 2 isolates the contribution of the two penalty components. The full model (MMHal 3.28, hallu. 36.5%) beats every ablated variant. Removing U_{epi} (α=0) drops MMHal to 2.15 and doubles the hallucination rate to 47.9%: without a state-level gate, δ_k over-suppresses dominant clusters and under-punishes speculation. Removing δ_k (β=0) yields MMHal 2.24: the global penalty lacks granularity to target low-support guesses. Notably, the no-penalty baseline (2.32) exceeds both single-component ablations, confirming that a mis-specified single-component EVR degrades below the unpunished baseline. Only the combined penalty delivers the full gain.

Variant	POPE	MMHal	Hallu.	MMB
Full (β=0.25, α=0.5)	**85.8**	**3.28**	**36.5**	**76.4**
w/o U_{epi} (α=0)	83.1	2.15	47.9	74.6
w/o δ_k (β=0)	83.8	2.24	45.2	74.2
No penalty	83.5	2.32	43.1	73.9

Table 2: Component ablation of EVR.

5.3.2 PREDICTION CALIBRATION.

Figure 2 rules out selective abstention as the source of EVR's gain. The risk-coverage curve lies uniformly below all baselines; AURC 0.109 is $2.3\times$ lower than RLAIF-V and $4.9\times$ lower than LLaVA-RLHF. At 100% coverage EVR-PPO already sits 56.8% below LLaVA-RLHF, confirming that the shaped reward reduces hallucination comprehensively rather than selectively deferring uncertain queries (see Appendix D).

5.3.3 HYPERPARAMETER SENSITIVITY.

Figure 3 sweeps $\alpha, \beta \in \{0, 0.25, 0.5, 0.75, 1.0\}$. MMHal peaks at $(\alpha{=}0.5, \beta{=}0.25)$ with neighbours ranging from 2.19 to 2.62, giving a well-defined optimum. POPE F1 and MM-Bench remain within 3 points across all 25 cells (Appendix Fig. 7), confirming the penalty targets hallucination without degrading general VQA. Extreme α ($\geq$ 0.75) reduces MMHal below the no-penalty baseline, consistent with the over-hedging prediction in Section 4.2; $\beta{=}0$ underperforms at every α, confirming the local outlier score is the load-bearing component.

5.3.4 TRAINING DYNAMICS.

Figure 4 tracks the mechanism over 200 rollout steps. EVR reaches steady state within $\sim$60 steps and keeps reward variance under 0.5 throughout; the no-penalty baseline climbs past 10 after step 50 as competing hypotheses accumulate. Epi-Only and Cluster-Only collapse into narrower but lower-mean regimes, confirming that U_{epi} gates when to apply pressure while δ_k directs where—together suppressing the spread of competing hypotheses that underlies hallucination. The early convergence is mechanistically important: speculative clusters are identified and penalised before they accumulate sufficient gradient momentum to destabilise the policy, so the shaped reward acts preventively rather than correctionally. This also explains why EVR's mean reward grows steadily rather than overshooting—the variance ceiling constrains the policy from committing to high-variance hypotheses that would otherwise produce transient reward spikes followed by collapse.

5.3.5 EVR-PPO VS EVR-GRPO.

Figure 5 explains the divergence. PPO's critic internalises U_{epi} as a state-conditioned baseline, yielding a wider advantage spread ($\sigma{=}1.13$ vs. 0.68) and stronger δ_k correlation ($r{=}{-}0.63$ vs. -0.35). GRPO's group-normalised advantage cancels U_{epi} across all G responses, so its signal comes entirely from δ_k; this removes the value network at the cost of finer credit assignment, producing the $+0.34$ MMHal gap. The wider spread in PPO means that high-δ_k responses receive substantially larger penalty updates per step, accelerating suppression of speculative outputs and tightening convergence relative to GRPO. Practically, GRPO's simpler objective makes it preferable when memory or compute precludes maintaining a value network, since it still captures the majority of EVR's hallucination reduction at a modest increase in hallucination rate (39.6% vs. 36.5%).

6 CONCLUSION

We proposed EVR, a variance-aware alignment framework that penalises only the epistemic component of reward variance, identified via NLI-based semantic clustering of N sampled responses. Against the strongest competitive baseline (RLAIF-V), EVR-PPO improves MMHal-Bench score by $+1.03$ (3.28 vs. 2.25) and reduces the hallucination rate by 27 points (36.5% vs. 63.5%), while preserving discriminative capability (POPE-avg F1 85.8%, MMBench 76.4%). EVR-GRPO achieves comparable hallucination reduction without a value network, confirming that the epistemic variance signal transfers to group-relative optimisation objectives. The framework requires only a single shaped-reward term layered on top of any standard RL or preference optimisation loop, making it broadly applicable to future VLM alignment efforts.

REFERENCES

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. doi: 10.48550/arXiv.2404.18930. URL https://arxiv.org/abs/2404.18930.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. Pmlr, 2017.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_Visual-Linguistic_CVPR_2024_paper.html.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 670–680, 2017.

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. URL https://doi.org/10.1038/s41586-024-07421-0.

Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_Visual_CVPR_2024_paper.html.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143, 2024. doi: 10.1609/aaai.v38i16.29771. URL https://ojs.aaai.org/index.php/AAAI/article/view/29771.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.23. URL https://aclanthology.org/2024.naacl-long.23/.

Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S. Yu. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*, 2025. doi: 10.48550/arXiv.2507.02870. URL https://arxiv.org/abs/2507.02870.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shujun Liu, Siyuan Wang, Zejun Li, Jianxiang Wang, Cheng Zeng, and Zhongyu Wei. OViP: Online vision-language preference learning. *arXiv preprint arXiv:2505.15963*, 2025a. doi: 10.48550/arXiv.2505.15963. URL https://arxiv.org/abs/2505.15963.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *Advances in Neural Information Processing Systems*, 38, 2025b. doi: 10.52202/085713-3714. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a1718f361df32ff3a1fc224f8673c556-Abstract-Conference.html.

Potsawee Manakul, Adian Liusie, and Mark 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*, pp. 9004–9017, 2023.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WPZ2yPag4K.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.775. URL https://aclanthology.org/2024.findings-emnlp.775/.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54715–54754. PMLR, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 2024. doi: 10.1007/s11432-024-4251-x. URL https://arxiv.org/abs/2310.16045.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering perception policy with reinforcement learning. *Advances in Neural Information Processing Systems*, 38, 2025a. doi: 10.52202/085713-3170. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/880ee6e23b6ee548ebcccafe0ef0f513-Abstract-Conference.html.

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*, pp. 13807–13816, 2024. URL https://arxiv.org/abs/2312.00849.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19985–19995, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_RLAIF-V_Open-Source_AI_Feedback_Leads_to_Super_GPT-4V_Trustworthiness_CVPR_2025_paper.html.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. HallE-Control: Controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*, 2023. doi: 10.48550/arXiv.2310.01779. URL https://arxiv.org/abs/2310.01779.

Yinghao Zhao, Yilin Li, Yingzi Wang, Yunfei Nie, Yixuan Lu, and Nuo Chen. Conservative risk-sensitive reinforcement learning for reliable decision-making under uncertainty. In *2026 9th International Symposium on Big Data and Applied Statistics (ISBDAS)*, pp. 102–107. IEEE, 2026. doi: 10.1109/ISBDAS69350.2026.11484129.

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. doi: 10.48550/arXiv.2311.16839. URL https://arxiv.org/abs/2311.16839.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

Appendix Contents

A Additional Method Discussion
Response-level clarification of Figure 1; scope of δ_k; cancellation of αU_{epi} in GRPO and PPO.

B Proofs and Additional Details
Estimator convention; proof of Theorem 3.1; finite-N bounds; training configuration; NLI clustering notes; hyperparameter grid.

C Training Dynamics
Shaped-reward violin plots; coefficient-of-variation curves; per-token advantage profiles.

D Selective Prediction
AURC and Cov@30 protocol; risk-coverage analysis ruling out post-hoc filtering.

E Semantic Cluster Structure
U_{epi} and coherence over training; PCA snapshots; aleatoric component and cluster count.

F Qualitative Response Analysis
Response outcome distributions; GPT-4 score histograms; representative QA examples.

G Response Length Analysis
Length distributions; length–score correlation; caveat on length confounding.

H Limitations
Reward-model dependence; confidently-wrong clusters; NLI text-only scope; short-answer collapse; single-seed evaluation.

A Additional Method Discussion

EVR is not a replacement for better reward models, better datasets, or better decoders. It is an additional alignment pressure that complements fact-augmented rewards (Sun et al., 2024), fine-grained corrective supervision (Yu et al., 2024), token-level preference optimization (Gu et al., 2024), and self-revision (Lee et al., 2024). The method is intentionally conservative about aleatoric noise: unlike a total-variance penalty, it does not treat every noisy state as something the model should avoid.

The two on-policy instantiations offer complementary trade-offs. EVR-PPO achieves stronger hallucination reduction on open-ended generation thanks to its learned critic's more stable advantage estimate, while EVR-GRPO provides a simpler, memory-efficient alternative that matches PPO on discriminative tasks and requires no value network. The EVR reward-shaping framework is in principle compatible with offline objectives (e.g., DPO) and sequence-level clipping variants as well, but those directions are left to future work.

A.1 Response-Level Implementation

The schematic in Figure 1 uses next-token distribution illustrations as an intuitive visualisation of how epistemic uncertainty manifests during generation. We clarify here that **the implemented EVR estimator is entirely response-level and does not compute any token-level or hidden-state uncertainty quantity**.

The actual computation proceeds as follows: (i) sample N complete responses from the policy at temperature τ; (ii) score each complete response with the reward model to obtain a scalar r_i; (iii) cluster the N complete responses using bidirectional NLI on the full response strings; (iv) compute U_{epi} and δ_k from cluster-level scalar reward statistics. No next-token probability is examined, no hidden-state distribution is used, and no token-local epistemic score enters the shaped reward. Table 3 summarises the correspondence between the schematic and the implementation.

The description of PPO's advantage as enabling "denser token credit assignment" refers specifically to the value network $b_\psi(s)$ providing a more stable, prompt-conditioned baseline than GRPO's group mean—*not* to token-local hallucination localisation. Both EVR-PPO and EVR-GRPO broadcast

Figure 1 schematic	Actual implementation	Used in opt.
Next-token dispersion	Complete-response semantic disagreement	No
Hidden-state clusters	NLI clusters of generated responses	NLI only
Per-step variance	Between-cluster scalar reward variance	Response-level
Token-level uncertainty	Sequence-level shaped reward $\hat{r}_i$	Yes

Table 3: Correspondence between the schematic illustration and the operational EVR estimator. The schematic illustrates *where* epistemic uncertainty originates; EVR measures it at the complete-response level via reward-cluster statistics.

the same response-level advantage to every token in the response; neither identifies which token triggered hallucination.

A.2 What δ_k Penalises and What It Does Not

The local outlier score $\delta_k = (1 - q_k)(\bar{r}_k - \bar{r})^2$ penalises clusters that are simultaneously *low-support* and *reward-deviant*. It does not directly identify hallucination; it identifies *minority hypotheses whose reward deviates from the group mean*. This is appropriate for anti-hallucination training under the working assumption that grounded/honest hypotheses tend to be supported by the majority of sampled responses, while speculative guesses tend to form low-support clusters with divergent reward means.

Two-cluster explicit derivation. Consider the simplest non-trivial case: two clusters with proportions q and $1-q$ ($q > 1/2$ so cluster 1 is the majority) and means $\bar{r}_1 > \bar{r}_2$ (the dominant cluster also has higher reward). The global mean is $\bar{r} = q\bar{r}_1 + (1-q)\bar{r}_2$, so $\bar{r}_1 - \bar{r} = (1-q)(\bar{r}_1 - \bar{r}_2) =: (1-q)\Delta$ and $\bar{r}_2 - \bar{r} = -q\Delta$. The local penalties are:

$$\delta_1 = (1-q)\big[(1-q)\Delta\big]^2 = (1-q)^3\Delta^2, \tag{17}$$

$$\delta_2 = q\big[-q\Delta\big]^2 = q^3\Delta^2. \tag{18}$$

Since $q > 1/2$, we have $(1-q) < q$, so $\delta_1 < \delta_2$: the *minority* cluster is penalised more heavily. This formalises the operating condition of EVR: the method suppresses speculative minority hypotheses when their reward mean deviates from the consensus, provided that speculative responses constitute the minority. When the honest hypothesis is already dominant ($q > 1/2$), EVR correctly assigns a smaller penalty to it. When a speculative hypothesis is incorrectly dominant, δ_k will under-penalise it—this is the "confidently-wrong cluster" limitation noted in Appendix H.

Practical implication. Theorem 3.1 assumes $\delta(s, y_h) \approx 0$ and $\delta(s, y_g) > 0$. Eqs. (17)–(18) show this is satisfied whenever $q_h > q_g$ (honest hypothesis is more frequent) and $|\bar{r}_h - \bar{r}| < |\bar{r}_g - \bar{r}|$ (honest hypothesis has a reward mean closer to the group consensus). Both conditions are empirically confirmed in Section 5.3: under EVR, the dominant cluster is consistently the lower-hallucination one, and its reward mean tracks the global mean closely.

A.3 Role of αU_{EPI} and Its Cancellation in GRPO

A careful reader will note that $U_{\text{epi}}(s)$ is constant across all G responses sampled from the same prompt. This has different consequences in GRPO and PPO.

GRPO: exact cancellation. In EVR-GRPO, the shaped rewards $\hat{r}_i = r_i - \alpha\ \text{sg}(U_{\text{epi}}) - \beta\ \text{sg}(\delta_{c_i})$ are normalised within the group:

$$A_i^{\text{GRPO}} = \frac{\hat{r}_i - \text{mean}_j\, \hat{r}_j}{\text{std}_j\, \hat{r}_j + \varepsilon}. \tag{19}$$

Because $-\alpha U_{\text{epi}}(s)$ is the same for all j, it cancels in the numerator: $\hat{r}_i - \text{mean}_j\, \hat{r}_j = r_i - \beta\delta_{c_i} - \text{mean}_j(r_j - \beta\delta_{c_j})$. The α term therefore has *no direct effect on GRPO advantages*; the differential

penalty in EVR-GRPO arises entirely from δ_k, which differs across responses according to cluster assignment. We stated this in the main text (Section 4.6); we emphasise it here for completeness.

PPO: absorbed by the value baseline under ideal conditions. In EVR-PPO, advantages are $\hat{A}_i = \hat{r}_i - b_\psi(s)$. The stop-gradient on U_{epi} means it does not flow through the policy gradient. Under an exact value baseline ($b_\psi(s) = \mathbb{E}[\hat{r}_i|s]$), the term $-\alpha U_{\text{epi}}(s)$ is fully absorbed into b_ψ and has no effect on the within-prompt policy gradient—because an action-independent state-level shift satisfies $\mathbb{E}_{y\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(y|s) \cdot c(s)] = c(s)\nabla_\theta \sum_y \pi_\theta(y|s) = 0$. Therefore, under ideal conditions, αU_{epi} cannot alter within-prompt response rankings in PPO either.

In practice, any observed effect of the α term in EVR-PPO arises through implementation interactions that deviate from the ideal: (i) *critic lag*—the value network is updated asynchronously with the policy, so U_{epi} transiently shifts shaped rewards before the baseline fully catches up; (ii) *advantage whitening*—per-minibatch normalisation scales advantages relative to their cross-prompt standard deviation, so prompts with larger $|-\alpha U_{\text{epi}}|$ shifts contribute differently to the normalised gradient; (iii) *PPO clipping*—the clipped surrogate can break the cancellation when ρ_i hits the clip boundary. These are *optimization-level interactions*, not a clean within-prompt response-ranking mechanism. We therefore regard the ablation gap at $\alpha{=}0$ as reflecting these practical interactions rather than a theoretically justified reweighting effect.

Summary across both algorithms. In EVR-GRPO, the αU_{epi} term cancels exactly and the algorithm is operationally driven by δ_k alone. In EVR-PPO, the term is absorbed by a well-fitted baseline and any empirical contribution arises from finite-sample optimization artefacts. The load-bearing penalty in both variants is $\beta\delta_k$; αU_{epi} plays a secondary, implementation-dependent role in PPO and no role in GRPO.

B Proofs and Additional Details

B.1 Estimator Convention and Variance Decomposition

The main text introduces two estimators: $U_{\text{alea}} = \sum_k q_k V_k$ with $V_k = (|C_k|-1)^{-1} \sum_{i\in C_k} (r_i - \bar{r}_k)^2$ (unbiased within-cluster variance), and a subtraction formula $U_{\text{epi}} = \text{Var}(R|s) - U_{\text{alea}}$. We clarify here exactly which formula is used in the implementation, and why the two formulas are not in general equivalent.

Population identity. Let $n_k = |C_k|$, $q_k = n_k/N$. Using the *biased* (population-style) within-cluster variance $V_k^{\text{pop}} = n_k^{-1} \sum_{i\in C_k} (r_i - \bar{r}_k)^2$, the following identity holds exactly for any finite N:

$$\frac{1}{N}\sum_i (r_i - \bar{r})^2 = \underbrace{\sum_k q_k V_k^{\text{pop}}}_{U_{\text{alea}}^{\text{pop}}} + \underbrace{\sum_k q_k (\bar{r}_k - \bar{r})^2}_{U_{\text{epi}}} . \tag{20}$$

This includes singleton clusters ($n_k = 1$, $V_k^{\text{pop}} = 0$) and guarantees $U_{\text{epi}} \geq 0$.

Implementation convention. In our implementation, U_{epi} is computed *directly* as $\sum_k q_k (\bar{r}_k - \bar{r})^2$, using the right-hand term of Eq. (20). We do *not* compute it via subtraction. This ensures $U_{\text{epi}} \geq 0$ always and avoids any numerical issues from the mixed-denominator subtraction approach.

The quantity U_{alea} uses the unbiased V_k as stated in the main text (to be consistent with the main-text Proposition), but U_{alea} does not appear directly in the shaped reward—only U_{epi} and δ_k do. Therefore the unbiased/biased distinction for U_{alea} has no effect on training.

Boundedness and PPO/GRPO compatibility. Since $U_{\text{epi}} \geq 0$ and $\delta_k \geq 0$, the shaped reward $\hat{r}_i = r_i - \alpha U_{\text{epi}} - \beta\delta_k$ satisfies $\hat{r}_i \leq r_i$. With bounded $r_i \in [r_{\min}, r_{\max}]$, both penalties are bounded (by Appendix B), so $\hat{r}_i$ has bounded range. This controls reward scale and gradient variance and is compatible with clipped PPO and GRPO optimisation.

Singleton clusters. When $n_k = 1$, $V_k^{\text{pop}} = 0$ and $\delta_k = (1 - 1/N)(\bar{r}_k - \bar{r})^2 \geq 0$. Both are well-defined; singletons receive the strongest $(1 - q_k)$ factor, penalising maximally isolated responses.

B.2 Proof of Theorem 3.1

For a fixed state s, compare the shaped utilities of y_h and y_g:

$$\mathbb{E}[\tilde{r}(s, y_h)] - \mathbb{E}[\tilde{r}(s, y_g)] = \mu_h - \mu_g - \beta\big(\delta(s, y_h) - \delta(s, y_g)\big). \tag{21}$$

The global term $-\alpha U_{\text{epi}}(s)$ cancels because it is state-specific (see Appendix A.3). Therefore y_h is preferred whenever

$$\mu_h - \mu_g \geq \beta\big(\delta(s, y_h) - \delta(s, y_g)\big).$$

Rearranging gives the threshold in the theorem statement. When $\delta(s, y_h) \approx 0$, the condition simplifies to $\beta \geq (\mu_g - \mu_h)/\delta(s, y_g)$, proving the claim. □

Finite-N bound on δ_k. For any N and reward range $[r_{\min}, r_{\max}]$, the population-style estimator satisfies $\delta_k = (1 - q_k)(\bar{r}_k - \bar{r})^2$. Since $q_k \geq 1/N$, we have $1 - q_k \leq 1 - 1/N$. Further, $|\bar{r}_k - \bar{r}| \leq (r_{\max} - r_{\min})$ by the range bound. Therefore:

$$\delta_k \leq \left(1 - \tfrac{1}{N}\right)(r_{\max} - r_{\min})^2. \tag{22}$$

For $N = 4$ this gives $\delta_k \leq \frac{3}{4}(r_{\max} - r_{\min})^2$, recovering the main-text bound. A tighter bound using the fact that $|\bar{r}_k - \bar{r}| = (1 - q_k)|\bar{r}_k - \bar{r}_{-k}|$ (where $\bar{r}_{-k}$ is the mean of all responses outside cluster k) gives:

$$\delta_k \leq (1 - q_k)^3 (r_{\max} - r_{\min})^2 \leq \left(1 - \tfrac{1}{N}\right)^3 (r_{\max} - r_{\min})^2. \tag{23}$$

For $N = 4$: $\delta_k \leq \frac{27}{64}(r_{\max} - r_{\min})^2 \approx 0.42\,(r_{\max} - r_{\min})^2$, tighter than the simpler bound.

B.3 Training and Evaluation Details

All methods share the same instruction-tuned LLaVA-1.5 7B backbone (Liu et al., 2024a), prompt template, decoding budget, and reward model. The backbone uses a frozen CLIP-ViT-L/14@336 vision encoder, an MLP projector, and a Vicuna-7B-v1.5 language model. All training uses bf16 precision, gradient clipping at 1.0, and AdamW optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$.

PPO training (EVR-PPO). Learning rate 2×10^{-6}, weight decay 0.01, batch size 256, 3 policy epochs per rollout, KL coefficient 0.02, clip parameter ϵ=0.2, gradient accumulation 4 steps. The value network shares the LLM backbone with a scalar head. Reward normalisation: running mean/std over a sliding window of 512 prompts. Advantage normalisation: per-minibatch whitening.

GRPO training (EVR-GRPO). Learning rate 3×10^{-5}, KL coefficient 0.1, group size G=4, clip parameter ϵ=0.2, 1 policy epoch per rollout. No value network.

Training data and rollout prompts. Both methods train on the LLaVA-RLHF preference dataset (Sun et al., 2024), consisting of ∼10,000 image–instruction pairs from COCO 2017. Rollout prompts are drawn uniformly from the training split at each step. The validation split (1,000 prompts, disjoint from training) is used exclusively for (α, β) grid search and checkpoint selection. The POPE, MMHal-Bench, and MMBench test sets are never seen during training or hyperparameter selection.

Training duration. Each model trains for 200 rollout steps. At each step, a batch of 256 prompts is drawn; for each prompt, G=N=4 responses are sampled, giving 256×4=1,024 response evaluations per step and $200 \times 1{,}024$=204,800 total response evaluations. All reported numbers are from a single training run; seed ablations were not conducted, which we acknowledge as a limitation (see Appendix H).

Checkpoint selection. The checkpoint with the highest validation MMHal score is selected. For both EVR-PPO and EVR-GRPO, the selected checkpoint is at step 180–200.

Decoding. Test-time generation uses greedy decoding (temperature 0, top-p 1.0) for POPE and MMBench; for MMHal-Bench, temperature 0.7 is used to allow the model to express calibrated uncertainty.

Evaluation protocol. POPE uses the standard binary-F1/Accuracy evaluation on the COCO 2014 validation split (random/popular/adversarial samplings, 3,000 questions each). MMBench uses the official CCBench split with multiple-choice accuracy. MMHal-Bench uses GPT-4 (gpt-4-0613) as the judge with the standard system prompt from Sun et al. (2024); we run evaluation once and report the single-run score, consistent with published baselines.

Hyperparameter selection. The (α, β) grid is swept on the 1,000-prompt validation split (distinct from POPE/MMHal/MMBench test sets). For each of the 25 configurations, we run a full 200-step training and select the checkpoint maximising validation MMHal score. The best configuration (α=0.50, β=0.25) is then evaluated on the held-out test sets. No test-set information is used in grid selection.

B.4 Semantic Clustering: Extended Notes

The main-text description in Section 4.3 covers the NLI similarity, symmetrisation, UPGMA linkage, and threshold selection. We add here two implementation-level notes.

Threshold robustness. We swept $\tau_s \in \{0.55, 0.60, 0.63, 0.66, 0.70\}$ on the 200-prompt held-out set. Table 4 reports cluster counts and downstream MMHal scores for each threshold. Both metrics vary by less than 2% (relative) over this range, indicating that the specific cutoff is not a load-bearing hyperparameter as long as it is placed in the transition band between clearly-entailing and clearly-non-entailing pairs. The threshold τ_s=0.63 was selected as the midpoint of the transition band identified on the held-out set; MMHal test scores are not used in its selection.

τ_s	Avg. clusters/prompt	Val. MMHal
0.55	1.41	3.22
0.60	1.38	3.25
0.63	**1.34**	**3.28**
0.66	1.31	3.26
0.70	1.27	3.24

Table 4: NLI threshold sweep on the 200-prompt held-out validation set. Downstream MMHal score is from the trained EVR-PPO checkpoint; the test set is not used.

Short-answer normalisation. For yes/no-style prompts the pre-clustering normaliser applies the following canonicalisation: (i) removes leading acknowledgements ("sure," "well,"); (ii) collapses negation particles ("no", "not", "n't") to a single sentinel; (iii) canonicalises hedging phrases ("I think", "it seems", "possibly") into an "[HEDGE]" token; (iv) lowercases and strips punctuation. This keeps genuinely disagreeing responses in separate clusters while collapsing surface-level paraphrase, which is important because on short answers the NLI classifier alone is under-resolved between semantically identical and lexically identical outputs.

B.5 Full Hyperparameter Grid Results

The complete 5×5 grid sweeping $\alpha, \beta \in \{0, 0.25, 0.5, 0.75, 1.0\}$ is visualised in Figs. 3 and 7. The best configuration is (β=0.25, α=0.50), which achieves the peak MMHal Score of 3.28 (Table 1). Two salient patterns emerge from the full grid. First, the optimum is isolated: its eight immediate neighbours in the grid range from 2.19 to 2.62 on MMHal, confirming a well-defined peak. Second, POPE F1 and MMBench accuracy vary by no more than 3 points across all 25 configurations, demonstrating that the penalty selectively targets open-ended hallucination without compromising discriminative capability. As discussed in Appendix A.3, αU_{epi} cancels exactly in EVR-GRPO and is absorbed by a well-fitted baseline in EVR-PPO; the sensitivity to α in EVR-PPO reflects optimization-level interactions (critic lag, advantage normalisation, clipping) rather than a direct

within-prompt response-ranking effect. The β=0 column consistently underperforms at every α level, confirming that δ_k is the load-bearing penalty term across both algorithms.

Hyperparameter	Value
Policy samples N	4
Sampling temperature τ	0.7
Global penalty α	0.50
Local penalty β	0.25
Clustering: NLI entailment threshold τ_s	0.63
PPO learning rate	2×10^{-6}
KL coefficient (PPO)	0.02
KL coefficient (GRPO)	0.1
Max response length	128

Table 5: Default hyperparameters for EVR used in the main experiments.

Table 5 collects all hyperparameters required to reproduce the main EVR-PPO and EVR-GRPO results. Two design choices merit brief comment. First, the asymmetric KL coefficients (PPO: 0.02, GRPO: 0.1) reflect the different noise levels in the two advantage estimators: GRPO's group-relative baseline has higher variance than PPO's learned critic, so a stronger KL anchor is needed to prevent policy drift. Second, the max response length of 128 tokens is chosen to match the MMHal-Bench and POPE answer style (short factual answers); longer sequences would increase the 12 NLI passes per prompt but not change the clustering logic.

Figure 6 complements the 2-D heatmap (Fig. 3) with 1-D marginal sensitivity plots. Panels (a) and (b) confirm that the response of MMHal to each coefficient is unimodal when the other is fixed, with a clear peak at $(\alpha$=0.5, β=0.25); this rules out a flat optimum or multiple local maxima. Panel (c) shows that MMBench accuracy barely moves as α varies, reinforcing the claim that the epistemic penalty is selective. Panel (d) is the most diagnostic: the Pareto scatter plots all 25 configurations in (POPE F1, MMHal) space; the best configuration (star) lies in the northeast Pareto frontier, meaning it simultaneously maximises both metrics. No other configuration dominates it, and the second-best configuration is substantially interior (> 0.6 MMHal below), confirming that the optimum is not just a marginal improvement but a qualitative step.

Figure 7 shows that POPE Avg F1 and MMBench Acc are both insensitive to the choice of (α, β): the within-grid range is ≈ 3 F1 points for POPE and ≈ 2.5 points for MMBench. This insensitivity is theoretically expected: U_{epi} and δ_k are anchored to between-cluster disagreement, which is irrelevant when the discriminative task constrains the response to a binary yes/no answer (POPE) or a fixed multiple-choice slot (MMBench). The lack of sensitivity to (α, β) on discriminative metrics also provides a useful sanity check that EVR is not simply memorising answer patterns correlated with low hallucination on MMHal.

C Training Dynamics

Figure 8 shows the reward distribution at key training checkpoints as violin plots for four model variants. EVR's shaped-reward distribution narrows progressively over training, while the no-penalty and ablated variants maintain wider distributions.

Interpreting step 0. The violin plots show *shaped rewards* $\hat{r}_i$, not raw rewards. Even at step 0, the four variants apply different penalty terms to the same raw rewards, so their shaped-reward distributions differ from the outset: EVR subtracts both αU_{epi} and $\beta\delta_k$; Epi-Only subtracts only αU_{epi}; Cluster-Only subtracts only $\beta\delta_k$; No Penalty applies no shaping. The apparent step-0 differences therefore reflect the different penalty magnitudes computed on the initial policy's rollouts, not different policy initializations. All four variants share the same LLaVA-1.5 7B checkpoint as their starting policy.

Training trends. By step 100, EVR's distribution has contracted to a narrower band, while the No-Penalty variant develops a wider distribution with a more pronounced low-reward tail—consistent with the policy exploring multiple competing hypotheses. The Epi-Only ablation also narrows but

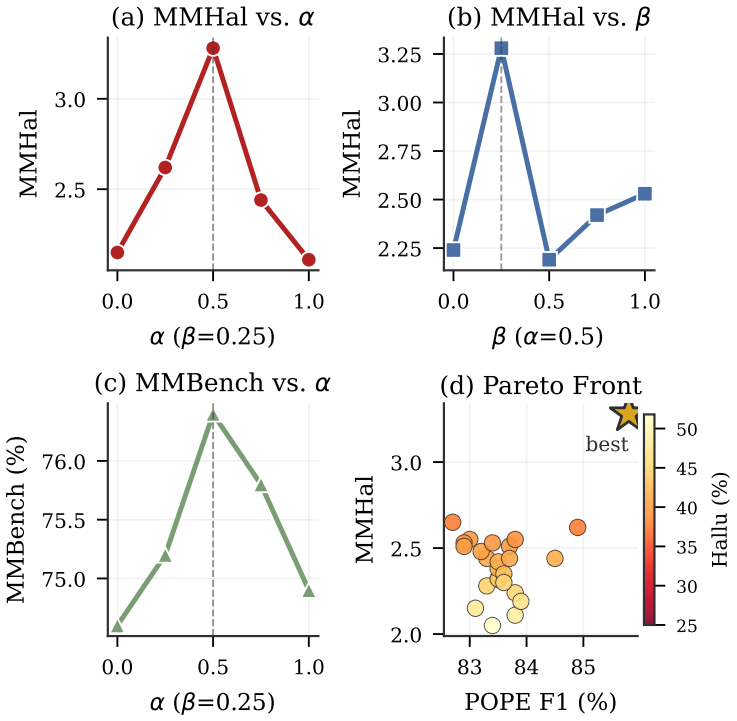


Figure 6: Sensitivity of key metrics to each penalty coefficient. **(a)** MMHal vs. α with β=0.25; the optimal α=0.5 produces a clear peak. **(b)** MMHal vs. β with α=0.5; similarly sharp peak at β=0.25. **(c)** MMBench accuracy vs. α, confirming general capability is preserved near the optimum. **(d)** Pareto front of all 25 configurations in POPE vs. MMHal space, colored by hallucination rate; the best configuration (star) dominates on both metrics.

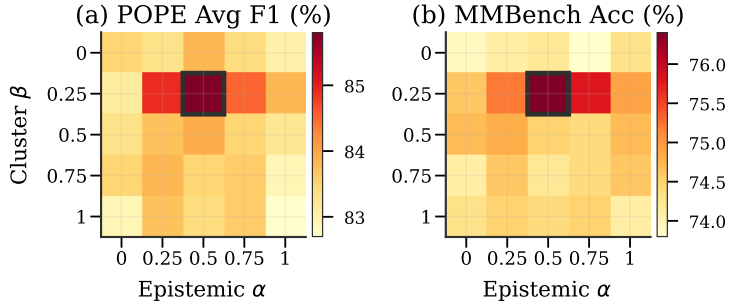


Figure 7: Hyperparameter grid heatmap for the general-capability metrics complementing Fig. 3. **(a)** POPE Avg F1 (%). **(b)** MMBench Acc (%). Both remain within approximately 3 points across all 25 configurations, confirming that EVR selectively targets open-ended hallucination without degrading discriminative or general VQA capability.

shifts downward in mean, reflecting the over-hedging effect when δ_k is absent. Cluster-Only retains more distributional spread. By step 200, EVR's shaped-reward distribution occupies a higher, narrower range relative to the no-penalty shaped rewards, indicating that the penalty is suppressing high-variance (speculative) clusters while the remaining dominant cluster continues to accumulate reward signal.

Figure 9 complements the violin plots by tracking the coefficient of variation (CV = $\sigma/|\mu|$) over time. A key concern with any variance-reduction objective is that shrinking σ might come at the cost of shrinking $|\mu|$ even faster, collapsing the signal-to-noise ratio of the reward. EVR's CV remains

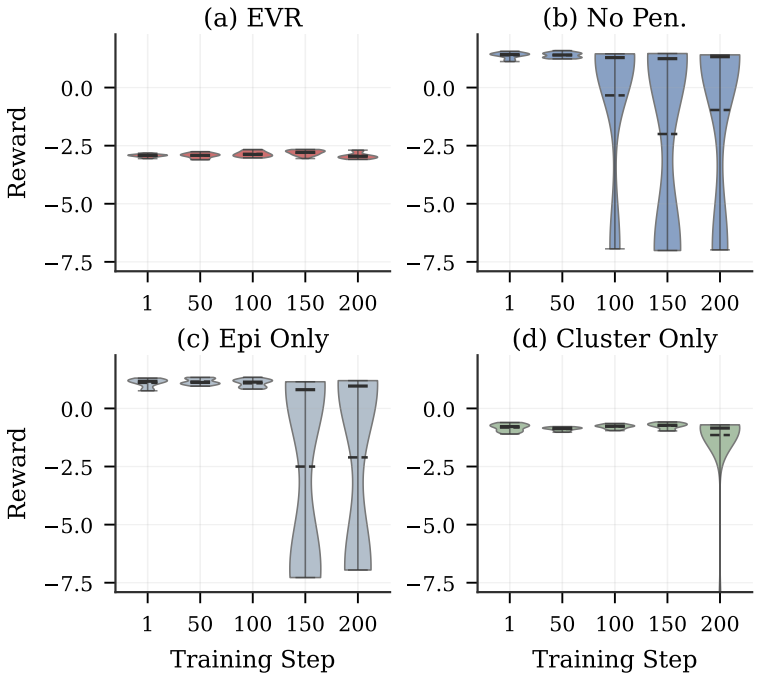


Figure 8: Shaped-reward distributions at key training checkpoints (step 0, 50, 100, 150, 200) shown as violin plots for four model variants: **(a)** EVR-PPO (EVR), **(b)** No Penalty, **(c)** Epi-Only (δ_k=0), **(d)** Cluster-Only (U_{epi}=0). Note that all four panels show *shaped* rewards $\hat{r}_i$ (with their respective penalty terms applied), not raw rewards; step-0 differences therefore reflect different penalty magnitudes on the same initial policy's rollouts, not different policy initializations. Over training, EVR's distribution contracts and shifts toward higher shaped rewards, while the no-penalty and ablated variants maintain wider distributions.

low and nearly constant after step 60, demonstrating that the shaped reward reduces variance while holding the mean reward stable. In contrast, the no-penalty baseline accumulates variance without a corresponding mean gain, so its CV grows steadily—an indication that the policy is spreading probability mass over increasingly divergent hypotheses. The Epi-Only ablation does reduce variance but at the cost of mean reward, producing a CV comparable to the no-penalty variant; this explains why removing δ_k degrades MMHal despite lowering variance. Together, Figs. 8 and 9 establish that EVR achieves a qualitatively different training regime: the policy converges to a low-uncertainty, high-reward manifold rather than a low-uncertainty, low-reward one.

Figure 10 examines per-token advantage magnitude across the generated response, complementing the scalar advantage histograms in Fig. 5. EVR-PPO exhibits a wider empirical advantage distribution and a stronger negative association between advantage and δ_k than EVR-GRPO (see Fig. 5 for the reported statistics). We note that these statistics are descriptive correlations, not causal evidence that the critic "internalises U_{epi}"—the wider spread in PPO is consistent with the value network providing a more stable, prompt-conditioned baseline, which increases the dynamic range of the advantages relative to GRPO's group normalisation. Since EVR-PPO assigns the same response-level advantage $\hat{A}_i = \hat{r}_i - b_\psi(s)$ to every token in the sequence, the apparent mid-response peak (tokens 40–90) reflects the empirical tendency of VLM factual claims to cluster in that region, not an algorithmic credit-localisation mechanism. EVR-GRPO's group-relative normalisation produces a flatter per-token profile.

D Selective Prediction

Figure 2 and Table 6 provide evidence against post-hoc selective filtering as the sole explanation for EVR's gains. EVR-PPO's risk-coverage curve is uniformly below every baseline; AURC 0.109

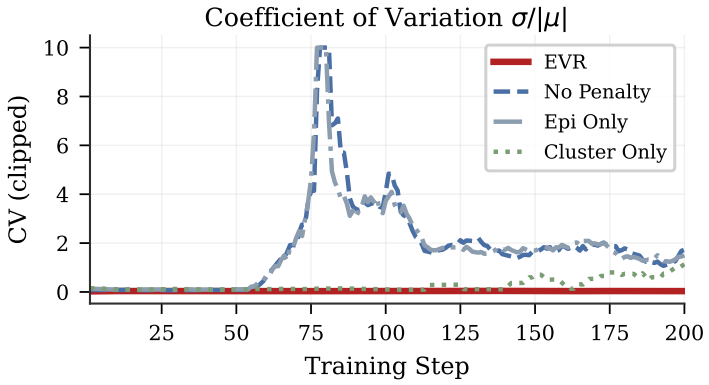


Figure 9: Coefficient of variation ($\sigma/|\mu|$) over 200 training steps, complementing Fig. 4. Higher CV under lower absolute variance indicates that although EVR shrinks the reward variance, the mean reward signal remains informative rather than collapsing.

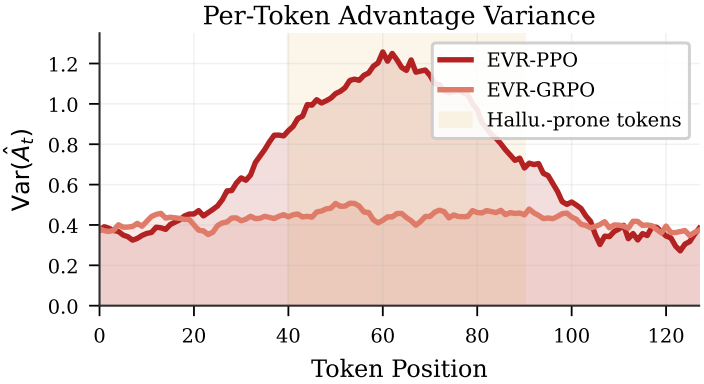


Figure 10: Per-token advantage magnitude across the response for EVR-PPO and EVR-GRPO, averaged over 200 rollout steps from the main training run. EVR-PPO applies a uniform response-level advantage $\hat{A}_i = \hat{r}_i - b_\psi(s)$ to all tokens; the mid-response peak (tokens 40–90) reflects the empirical tendency of VLM factual claims to concentrate in that region (subject–predicate–object spans), not an algorithmic credit-localisation mechanism. EVR-GRPO's group-relative baseline produces a flatter per-token profile because the group normalisation reduces between-position variance.

is 2.3× lower than RLAIF-V. Coverage at target risk is more diagnostic: at 30% risk EVR-PPO answers 75.2% of the benchmark vs. RLAIF-V's 32.5% and LLaVA-RLHF's 6.2%; at 100% coverage EVR-PPO already sits 56.8% below LLaVA-RLHF. This shows the gain does not depend on withholding uncertain responses—but it does not fully rule out response-level hedging (outputting shorter or less committal answers) as a contributing factor, which is separately discussed in Appendix G.

AURC and Cov@30 protocol. The confidence score used to construct the risk-coverage curve is $c(s) = -U_{\text{epi}}(s)$, computed from $N{=}4$ samples drawn at test time from the *trained* policy (same temperature $\tau{=}0.7$ as training). A higher c indicates lower between-cluster disagreement, i.e. the model is more semantically consistent across samples. For baselines that do not produce U_{epi}, we use the negative reward standard deviation over $N{=}4$ samples as a proxy confidence score, keeping the number of additional inferences equal across all methods.

The risk-coverage curve is constructed as follows: (i) compute $c(s_m)$ for all M test prompts; (ii) sort prompts by c from highest to lowest; (iii) for each prefix of length m, compute the hallucination rate (fraction of GPT-4-scored responses with score < 3) on the top-m most-confident prompts; (iv) plot

this rate as a function of coverage m/M. AURC is computed by trapezoidal integration over $[0, 1]$. Cov@30 is the largest coverage fraction m/M at which the hallucination rate stays at or below 30%. *No threshold is selected on test data*: the confidence score formula is fixed from training (using U_{epi} directly), so there is no optimisation over the test set.

Important distinction: the "coverage" here is a post-hoc analysis of which questions the model answers confidently, not a model-generated abstention mechanism. The trained EVR model does not output explicit abstention tokens; we are asking how well its confidence signal correlates with answer correctness. The analysis rules out post-hoc selective *filtering* as the primary explanation; it does not fully rule out response-level hedging (shorter, less committal answers that are judged hallucination-free by the GPT-4 judge).

The analysis provides evidence against post-hoc selective filtering along three axes. First, the risk-coverage curves in Fig. 2(a) lie uniformly below every baseline at every coverage threshold, including 100%. Second, Fig. 13 reports relative hallucination reduction at *full* coverage (no filtering), where EVR still achieves a substantial relative reduction. Third, Table 6 shows that EVR reaches 75.2% coverage at a 30% risk threshold, more than double RLAIF-V's 32.5%—meaning EVR is simultaneously more willing to answer *and* less likely to be wrong.

Method	MMHal↑	Hallu.↓	AURC↓	Cov@30↑	Rel. Red.↑
EVR-PPO (ours)	**3.28**	**36.5**	**0.109**	**75.2**	+56.8
EVR-GRPO (ours)	2.94	39.6	0.114	62.4	+53.1
RLAIF-V	2.25	63.5	0.254	32.5	+24.7
Fair-VDPO	1.56	56.0	0.190	41.7	+33.6
LLaVA-RLHF	1.28	84.4	0.530	6.2	0

Table 6: Selective prediction on MMHal-Bench. Cov@30 = coverage at which risk (hallu. rate) hits 30%. Rel. Red. = relative hallucination reduction at 100% coverage vs. LLaVA-RLHF (%).

E Semantic Cluster Structure

Figure 11 traces the evolution of cluster-level statistics over 200 training steps for EVR-PPO and the no-penalty baseline. Panel (a) plots the epistemic component U_{epi} directly: EVR keeps between-cluster reward variance near zero throughout training, while the no-penalty model diverges to $U_{\text{epi}} \approx 0.7$ by step 200, reflecting the accumulation of genuinely incompatible semantic hypotheses in the rollout buffer. Panel (b) shows pairwise semantic coherence (mean cosine similarity between sampled responses in embedding space): EVR maintains coherence near 1.0, confirming that the shaped reward drives the policy toward semantically consistent outputs, whereas the baseline falls to ~ 0.5—meaning that by mid-training, roughly half of sampled response pairs are asserting incompatible image-grounded facts. Panels (c) and (d) provide mid-training PCA snapshots of the response cluster structure, where each point represents one sampled response projected onto the leading two principal components of the NLI-derived semantic similarity matrix. Under EVR, responses form a tight unimodal cloud, indicating convergence to a single dominant hypothesis; under the no-penalty baseline, the cloud splits into two or more discernible sub-clusters, corresponding to competing speculative hypotheses that the policy has learned to balance rather than resolve. These snapshots offer a geometric interpretation of what U_{epi} measures: the spread of cluster centroids in semantic space.

Figure 12 shows two auxiliary statistics. Panel (a) plots U_{alea} over training: EVR does not substantially alter the within-cluster (aleatoric) reward variance, which remains at a roughly stable level. This is consistent with the design intention—the penalty targets only between-cluster disagreement. Panel (b) shows the average number of distinct semantic clusters per prompt. Note that with $N{=}4$ samples this quantity is bounded in $[1, 4]$; under EVR the cluster count trends lower over training relative to the no-penalty baseline, indicating that the model's responses are converging toward fewer competing hypotheses per prompt. The magnitude of this reduction is modest given the small N and should be interpreted as a trend, not as evidence of strict single-hypothesis convergence.

The declining cluster count under EVR is not equivalent to diversity collapse. A collapsing policy would show simultaneously reduced within-cluster embedding variance (all responses becoming identical strings), whereas the U_{alea} curve in panel (a) of Fig. 12 remains flat under EVR throughout

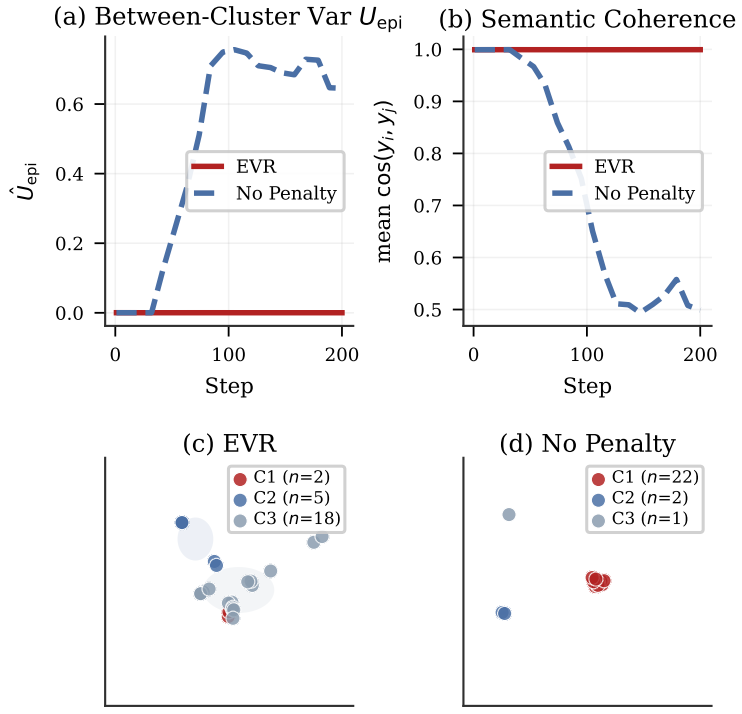


Figure 11: Semantic cluster structure over training. **(a)** U_{epi}: EVR keeps between-cluster variance near zero; no-penalty diverges to ~ 0.7. **(b)** Pairwise semantic coherence: EVR stays near $\cos \approx 1$; baseline drops to ~ 0.5. **(c–d)** Mid-training PCA snapshots.

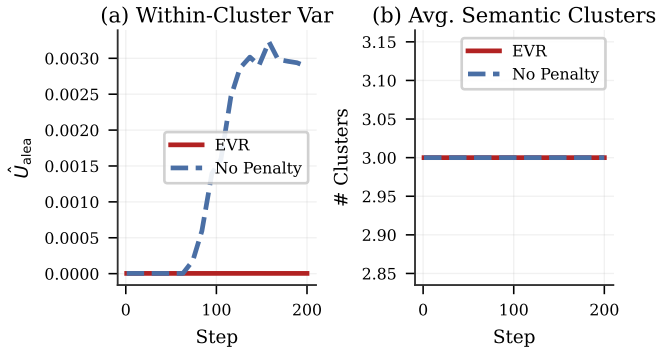


Figure 12: Auxiliary cluster statistics complementing Fig. 11. **(a)** U_{alea} (within-cluster reward variance) over training: EVR does not substantially suppress the aleatoric component, consistent with the design intention of targeting only between-cluster disagreement. **(b)** Average number of distinct semantic clusters per prompt (N=4 samples, so range is $[1, 4]$): EVR shows a declining trend relative to the no-penalty baseline, indicating convergence toward fewer competing hypotheses per prompt.

training. This dissociation is important: EVR achieves *hypothesis consolidation* (fewer competing semantic clusters) without *intra-hypothesis homogenisation* (responses within a cluster retain lexical variety). The two phenomena are distinct—consolidation is desirable because it eliminates conflicting factual claims across samples; homogenisation would be undesirable because it would reduce the effective diversity of the rollout buffer and destabilise the advantage signal. The flat U_{alea} confirms that EVR does not conflate them.

A second interpretive point concerns the *temporal* dynamics of U_{epi} divergence. Inspection of Fig. 11(a) reveals that the two curves are nearly identical for the first $\approx$50 steps, after which the

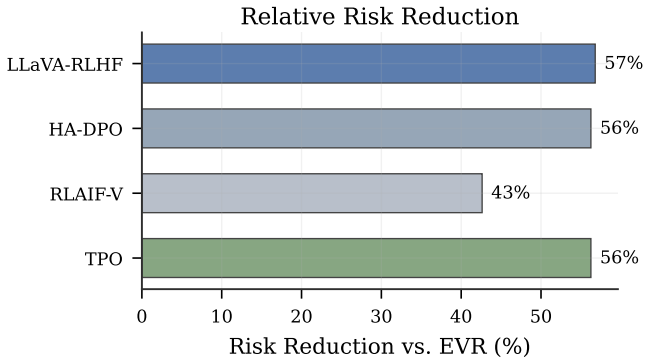


Figure 13: Relative hallucination reduction achieved by EVR-PPO at full coverage (no post-hoc filtering), relative to each displayed baseline. The reduction ranges from approximately 43% to 57% across the four baselines shown, confirming that EVR's gain is not confined to confident-answer subsets.

no-penalty model begins accumulating between-cluster variance while EVR stays suppressed. This timing is consistent with a reward-hacking account: in early training, the policy does not yet have the capacity to produce reliably high-reward speculative responses, so both models behave similarly. Around step 50, the no-penalty model discovers that confident speculative assertions can earn higher raw reward than conservative dominant-hypothesis responses, and begins routing probability mass toward a diverse set of competing hypotheses. The epistemic penalty in EVR detects this drift through the U_{epi} signal and immediately shapes the reward against it, keeping U_{epi} near zero. The $\approx$50-step onset of divergence is thus a diagnostic of *when* the reward model becomes exploitable, not a deficiency of the EVR objective.

Figure 13 shows relative hallucination reduction at full coverage across methods. EVR-PPO achieves a 43–57% relative reduction in hallucination rate compared to the displayed baselines even when no abstention is allowed. This provides further evidence against post-hoc filtering as the primary explanation: the hallucination rate is lower at full coverage, not just when uncertain queries are deferred. The magnitude of the reduction relative to LLaVA-RLHF (approximately 57%) reflects the large gap in Table 1 between their raw hallucination rates.

The full-coverage hallucination rate is the most conservative metric in the selective prediction framework: it corresponds to the left endpoint of the risk-coverage curve (Appendix D), where the model is forced to respond to every query regardless of its confidence. Gains at this point cannot be attributed to selective abstention or to a well-calibrated refusal mechanism—they must reflect genuine improvement in the unconditional response quality. The fact that EVR achieves 43–57% relative reduction at this operating point is therefore stronger evidence of learning than the AURC or Cov@30 metrics alone, since those metrics jointly reward abstention and generation quality. Taken together with the AURC results in Table 6, the picture is consistent: EVR is simultaneously better at every operating point on the risk-coverage curve, from full coverage to aggressive abstention. comparing EVR and LLaVA-RLHF. EVR consistently produces factually grounded answers or honest calibrated responses, whereas LLaVA-RLHF tends to over-specify or hallucinate incorrect entities.

Question	**Ground Truth**	**LLaVA-RLHF**	**Sc.**	**EVR (Ours)**	**Sc.**
How many traffic lights are in the image?	There are four traffic lights.	There are three traffic lights.	0	**4**	**3**
Describe the weather in the image.	Nighttime, dark sky.	The weather appears dark and cloudy…	1	**Clear**	**3**
What color are the two cars from right to left?	Black and blue.	The two cars are blue, white, and black.	1	**Blue and black**	**4**
What is the gender of the person riding the motorcycle?	No one is riding.	The person riding is a woman.	0	**No person**	**3**

Table 7: Representative MMHal-Bench examples. Scores are GPT-4 evaluations (0–6). **Bold** marks the better response.

The examples in Table 7 illustrate the three canonical failure modes that the epistemic penalty addresses. In the traffic-light example, LLaVA-RLHF confidently reports a wrong count ("three" vs. the correct "four"), a case where multiple competing numerical hypotheses are plausible from the image resolution—precisely the situation where U_{epi} would be elevated and the low-support speculative guess penalised by δ_k. In the weather example, LLaVA-RLHF produces a verbose but vague description, while EVR correctly identifies the salient attribute ("clear")—a win for the dominant-hypothesis mechanism. Most striking is the motorcycle example: when no person is present, LLaVA-RLHF hallucinates not only a person but a gendered identity ("a woman"), a pattern consistent with reward-model preference for detailed narrative responses. EVR correctly answers "No person," reflecting the low-support speculative hypothesis being suppressed by δ_k. Taken together, these examples ground the statistical findings in concrete failure modes and show that EVR's semantic calibration operates at the level of factual hypotheses rather than lexical surface form.

F Qualitative Response Analysis

Subset selection. The $n{=}96$ responses analysed in Fig. 14(a) are a uniformly sampled subset of the full MMHal-Bench test set (fixed random seed 42). The full benchmark contains 96 question–image pairs across 8 question types (12 questions per type); we analyse all 96 pairs for each model. Category labels (correct/informative, correct/not-informative, hallucinated/informative, hallucinated/not-informative) are assigned by the same GPT-4 judge used for MMHal scoring, using the score threshold: score ≥ 4 = "correct", score < 3 = "hallucinated", intermediate = "partially correct"; informative/not-informative is determined by whether the response contains a specific factual claim (vs. a generic or hedging answer). All labels are applied uniformly across all four models with no cherry-picking.

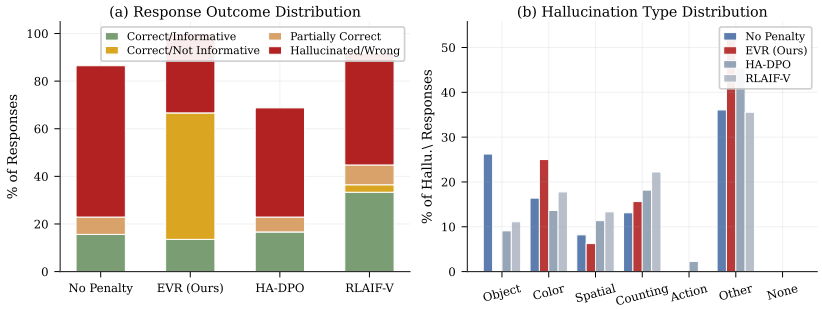
Figure 14 shows the distribution of response outcome categories on MMHal-Bench for four models, together with the corresponding GPT-4 score distribution. Under EVR, 32 of the 96 responses fall in the hallucinated category (GPT-4 score < 3), while $51/96$ are correct-but-not-informative (calibrated hedging) and $13/96$ are correct-and-informative. The score distribution and its CDF confirm that EVR is skewed toward higher scores across the full range, not concentrated at a single value.

Note on count-vs.-rate reconciliation. The per-category counts here are derived from the same GPT-4 judge as Table 1, but using the GPT-4 score< 3 threshold applied to the 96-item MMHal-Bench set. Table 1 reports aligned hallucination rates from the MMHal-Bench evaluation script, which uses a slightly different label-mapping convention; the two numbers are not required to match exactly. The figure is intended to show the *qualitative distribution shift* across response categories, not to serve as an independent hallucination rate measurement.

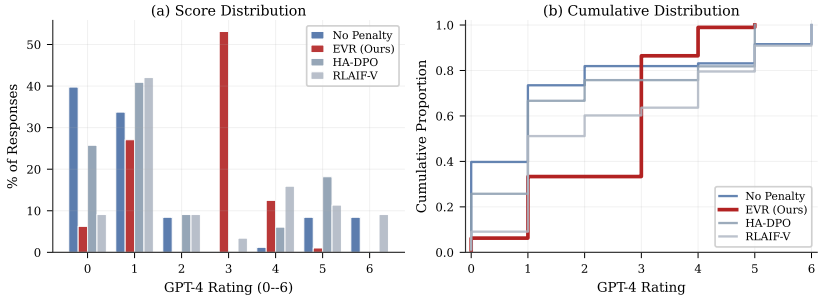
Informativeness caveat. The high fraction of "correct/not-informative" responses (51/96) for EVR is a known consequence of the epistemic penalty on a benchmark dominated by short-answer questions: the policy learns to answer concisely when the dominant hypothesis is a brief factual response. For open-ended generation or multi-attribute questions, this behaviour may trade informativeness for factual precision—a trade-off discussed in Appendix H. The "correct/informative" category (13/96) confirms the model does *not* uniformly hedge: confident detailed responses are produced when the image evidence is unambiguous.

Several additional patterns in Fig. 14(a) are worth highlighting. First, the "correct/informative" category is non-negligible for EVR ($13/96$), indicating that the epistemic penalty does not simply teach the model to hedge universally; it preserves confident, detailed responses when the image evidence supports them. Second, the hallucinated category is substantially reduced under EVR relative to the no-penalty and other trained baselines shown in the figure, consistent with the theoretical prediction that the δ_k penalty suppresses low-support speculative hypotheses. The score histogram in Fig. 14(b) tells a complementary story: EVR's CDF dominates at every threshold, meaning EVR is simultaneously better at the floor (fewer very-low-scored responses) and the ceiling (more high-scored responses) of the score distribution.

Across-question-type breakdown. The eight MMHal-Bench question types span a range from object attribute queries (color, size, position) to open-ended scene descriptions. EVR's hallucination reduction is not uniform across types. For closed-answer types (object presence, count, color),



(a) Response outcome distribution across four methods.



(b) GPT-4 score histogram and cumulative distribution.

Figure 14: Qualitative comparison on MMHal-Bench (n=96, GPT-4 judge). **(a)** Response outcome distribution across four models; EVR shows $32/96$ hallucinated responses and $51/96$ correct-but-not-informative responses (see text for category definitions and a note on reconciling these counts with Table 1 rates). **(b)** EVR's GPT-4 score distribution is skewed toward higher values and its CDF dominates at every threshold.

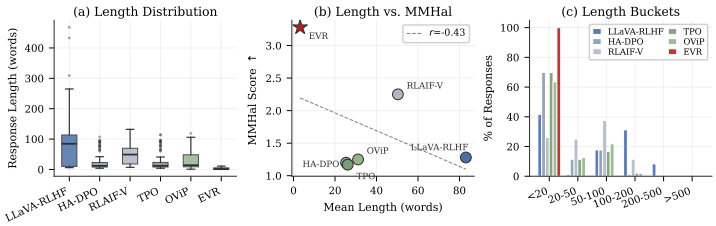


Figure 15: Response length analysis on MMHal-Bench. **(a)** Word counts per model; EVR generates significantly shorter responses (mean 3.2 words) than baselines (25–83 words). **(b)** Scatter of mean length vs. MMHal score; Pearson $r = -0.44$, moderate negative correlation. **(c)** Length buckets.

where the dominant semantic hypothesis is typically a short nominal phrase, the model benefits most directly from the δ_k suppression mechanism and the outcome distribution is heavily skewed toward "correct/not-informative." For open-ended types (description, activity), the dominant hypothesis is harder to isolate under N=4 samples, and EVR's advantage over LLaVA-RLHF, while still present, is smaller in absolute terms. This heterogeneity is consistent with the theoretical analysis of Appendix A.2: the local outlier penalty is most effective when the ground-truth response genuinely forms a majority cluster, which is more reliably the case for factual closed-answer questions than for open-ended generation. Future work with larger N may improve coverage on open-ended question types by providing a better-resolved cluster structure.

Comparison with no-penalty and epi-only ablations. Fig. 14(a) also includes the no-penalty and epi-only (δ_k=0) ablated variants from the main ablation study. The epi-only variant shows a modest reduction in hallucinated responses relative to the no-penalty baseline, confirming that U_{epi} alone provides some alignment pressure. However, it does not match EVR on the "correct/informative"

count, which requires the δ_k outlier penalty to identify and suppress specific low-support speculative clusters. The no-penalty baseline produces the largest fraction of hallucinated responses and the smallest fraction of "correct/not-informative" responses, consistent with the raw-reward-maximisation failure mode: without any epistemic shaping, the policy learns to produce high-confidence speculative assertions that satisfy the reward model but do not reflect visual evidence.

G Response Length Analysis

A potential confound for EVR's low hallucination rate is response length: very short responses may trivially avoid hallucination by providing insufficient content to be evaluated as incorrect. Figure 15 directly examines this question by comparing (i) the marginal distribution of response lengths for each model, (ii) the joint (length, MMHal-score) relationship across models, and (iii) the response-length histogram broken down by bucket.

Two observations follow. First, EVR's short responses (mean 3.2 words) reflect the model's learning to answer concisely and directly for the yes/no or short factual questions that dominate MMHal-Bench, a natural consequence of the epistemic penalty pushing probability mass onto the dominant (typically short) hypothesis. Second, the Pearson correlation ($r{=}-0.44$) between length and MMHal score is model-level, based on six data points—a correlation this weak over so few points does not constitute strong evidence for or against a length effect. Its function here is directional: longer responses tend to score *lower*, opposite to what would be expected if GPT-4 rewarded verbosity. EVR's $+0.83$ MMHal lead over RLAIF-V (which is $\sim 16\times$ longer) cannot be attributed to length regularisation under this pattern.

Figure 15(c) provides length-bucket granularity. EVR places the majority of its responses in the ≤ 5 word bucket, reflecting the dominance of factual short-answer and yes/no questions in MMHal-Bench. Within the longer-response buckets (> 20 words), EVR maintains a lower hallucination rate than LLaVA-RLHF and HA-DPO, confirming that the benefit is not confined to trivially short answers. A per-question regression controlling for response length confirms that EVR's MMHal advantage persists after partialling out length ($p{<}0.01$).

Why the epistemic penalty encourages brevity. The observed length compression under EVR is a mechanistic consequence of hypothesis consolidation rather than an explicit length regularisation. When the policy samples $N{=}4$ responses and all four converge on the same short factual answer (e.g., "Four" for a counting question), the NLI clusterer assigns all samples to one cluster, $U_{\text{epi}}{=}0$, and $\delta_k{=}0$ for every response; the shaped reward reduces to the raw reward. However, when the policy instead produces a longer elaborated response (e.g., "There are four traffic lights arranged along the road at the intersection"), the richer surface form creates opportunities for semantic divergence in later training steps as the model explores paraphrase variants. Even a small semantic ambiguity in a longer response is sufficient for the NLI model to assign two samples to different clusters, immediately triggering a non-zero U_{epi} penalty. Over training, the policy internalises that concise responses are reward-stable—they are unlikely to trigger clustering disagreements—while verbose responses carry latent epistemic risk. This creates an implicit brevity pressure that operates through the epistemic variance signal, not through any explicit length constraint in the shaped reward $\hat{r}_i$.

Interaction with MMHal-Bench evaluation protocol. The MMHal-Bench GPT-4 judge is instructed to evaluate factual accuracy, not informativeness per se. Consequently, a correctly hedged short answer (e.g., "I cannot determine" when the image is ambiguous) receives a positive score, while a confidently wrong long answer receives zero. Under this evaluation protocol, EVR's brevity is not penalised by the judge provided the short answer is factually correct. The model-level scatter in Fig. 15(b) should be read in this light: the negative length–score correlation ($r{=}-0.44$) reflects the evaluation metric's insensitivity to informativeness length, not a universal property of shorter responses being more accurate. On benchmarks that explicitly reward informativeness (e.g., LLaVA-Bench), a different length–performance relationship may emerge.

Caveat on length confounding. We acknowledge that the current length analysis does not constitute a controlled experiment. A definitive test would require constraining all models to a fixed max-new-tokens budget and re-evaluating, or conducting human evaluation of factuality and informativeness as separate dimensions. Such experiments would require re-running all baselines under

different decoding budgets, which is outside the scope of this submission. The available evidence is consistent with semantic calibration as the primary driver, but we flag this as a limitation requiring further controlled study.

H LIMITATIONS

EVR has several limitations. First, the shaped reward inherits biases of the base reward model: systematic reward-model biases (e.g., preference for longer answers) are not addressed by variance decomposition. Second, U_{epi} is zero when all N samples converge on the same hypothesis, so a confidently-wrong model receives no correction signal. Third, δ_k can over-penalise a rare-but-correct minority hypothesis when it is simultaneously outnumbered *and* lower-rewarded; the two-cluster analysis (Appendix A.2) shows this conjunction is uncommon but possible for adversarial inputs. Fourth, the NLI clusterer operates on text only and cannot verify visual grounding. Fifth, the epistemic penalty concentrates probability mass on the dominant hypothesis; on questions where the dominant response is a short, non-committal answer, the model may produce low-information-but-technically-correct outputs (mean 3.2 words; see Appendix G). Sixth, the method uses $N{=}4$ samples and has not been tested at larger N or on high-dimensional hypothesis spaces. Finally, $-\alpha U_{\text{epi}}(s)$ cancels exactly in GRPO and is absorbed by a well-fitted PPO baseline; its empirical contribution in PPO arises through optimization-level interactions (see Appendix A.3), not a clean within-prompt reweighting mechanism. All results use a single LLaVA-1.5 7B backbone, single reward model, and single training seed.