RISK IS NOT REWARD: ACTION-RELATIVE PROCESS REWARDS FOR VLM HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can assert visual facts unsupported by an image. Decoding-time corrections may repair such errors, but can also damage correct answers. Detecting that the original answer is unreliable is therefore not the same as knowing whether a *specified* correction will help: queries with identical error risk can have different repair and damage rates. Existing selection strategies based on error risk or visual neglect do not, by themselves, estimate the action-relative quantity needed for intervention. We formulate hallucination mitigation as a one-decision comparison between the corrected continuation and leaving the decoder unchanged. Action-Relative Process Reward (ARPR) gating learns this difference from offline paired outcomes. A process-reward assessor combines static visual evidence with decoder features at the pre-generation boundary, then predicts the fixed action's net gain without generating either continuation online. A reward gate applies the correction only when its predicted gain warrants it. We derive the optimal decision rule, bound regret from reward-estimation and gating errors, and characterize when additional visual evidence can change the decision. On COCO POPE, accuracy improves by 3.59 percentage points on LLaVA-1.5-7B and 1.01 points on Qwen3.5-9B; applying the correction universally instead reduces LLaVA accuracy to 70.86%.

1 INTRODUCTION

Vision-language models (VLMs) can generate fluent claims unsupported by an image, including nonexistent objects and incorrect attributes or relations (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023). Decoding-time interventions attempt to restore visual fidelity by changing input conditions, layer contrasts, context representations, or attention (Leng et al., 2024; Wang et al., 2024b; Chuang et al., 2024; Wang et al., 2025; Fazli et al., 2026; Sun et al., 2026). Yet the intervention itself can introduce an error. On LLaVA POPE, applying our fixed decoder correction to every query repairs 877 wrong answers but damages 2,171 correct ones. Detecting a likely hallucination therefore does not establish that *this particular correction* is worth applying.

Original-answer error risk is the probability that base decoding is wrong. The reward relevant to a fixed correction is its expected change in correctness *relative to leaving base decoding intact*: error risk times repair rate, minus correct-answer probability times damage rate. Two queries can have equal error risk but opposite correction values because their repair and damage rates differ. A perfectly calibrated error detector is therefore not, in general, a sufficient correction policy (Section 3). The missing distinction concerns the consequence of an action, not whether prior methods use visual evidence or ever select interventions.

Selective and reward-guided interventions already exist (Kogilathota et al., 2026; Duan et al., 2025; Yang et al., 2026; Sun et al., 2025; Zhang et al., 2026b). Our target is narrower: at one pre-generation branch point, should a *specified* decoder correction replace the original continuation? Action-Relative Process Reward (ARPR) treats this keep-or-correct choice as a process decision. A static visual evidence state supplies image-derived context; a process reward model (PRM) interface predicts the fixed action's gain over no action from pre-action features; and a reward gate selects the continuation (Figure 1). The reward is action-relative because the training target compares paired base and

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.

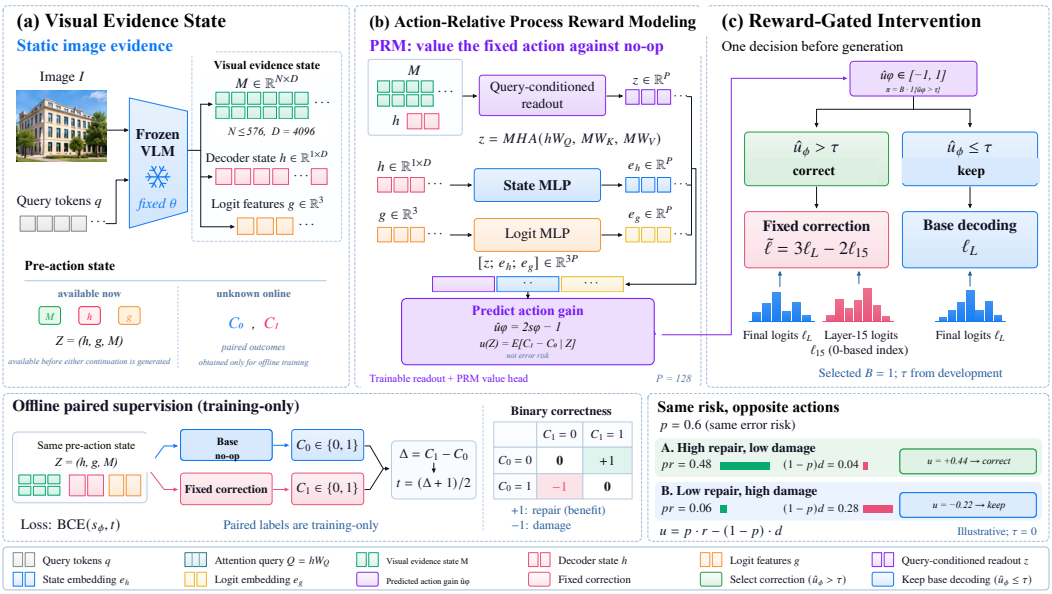


Figure 1: **Error risk versus action-relative process reward.** **(a)** A frozen VLM supplies static visual evidence and decoder state. **(b)** The one-decision PRM estimates the value of a specified correction relative to base decoding, using paired outcomes only during training. **(c)** A reward gate keeps the base decoder or applies that correction. The lower example holds error risk fixed while repair and damage imply opposite optimal actions; its rates are illustrative, not measurements. The evaluated policy uses $B = 1$ and development-selected thresholds.

corrected continuations from the same prefix. Their outcomes are available only offline: inference does not generate both answers or choose among correction operators.

Overall, our contributions are summarized as follows:

- We distinguish original-answer error risk from the net reward of a specified correction, derive an optimal one-decision rule and regret bound, and characterize when additional visual evidence can change that decision.
- We introduce ARPR gating: a one-decision PRM trained on paired base and corrected outcomes that gates a fixed decoder action before generation, without evaluating both continuations online.
- Across LLaVA-1.5-7B and Qwen3.5-9B, ARPR improves pooled POPE accuracy over greedy by 3.59 and 1.01 percentage points, respectively, and AMBER-D attribute/relation balanced accuracy by 4.07 and 1.82 points. MMHal score and hallucination rate also improve on both backbones, supporting transfer beyond binary object questions.

2 Related Work

Hallucination mechanisms and decoding interventions. CHAIR, POPE and AMBER test object and broader visual hallucination (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023). VCD contrasts visual inputs, ICD perturbs instructions, and DoLa contrasts layer predictions (Leng et al., 2024; Wang et al., 2024b; Chuang et al., 2024). DeCo selects preceding-layer evidence, CEI re-injects context according to a commitment signal, and AIR rectifies attention imbalance (Wang et al., 2025; Fazli et al., 2026; Sun et al., 2026). OPERA and SECOND also modify or select decoding interventions (Huang et al., 2024; Park et al., 2025). NoLan suppresses a text-only language-prior branch, ILVAD boosts visual evidence identified through inter-layer attention, and PTI steers modality-specific KV states once at prefill time (Ren et al., 2026; Xie et al., 2026; Zhang et al., 2026a). These methods specify different ways to alter decoding; the present decision compares one fixed operator with leaving the decoder unchanged under a matched reward objective. SeeMe instead restructures visual tokens during inference (Tang et al., 2026); its evidence-editing mechanism is complementary

to our decision about whether a fixed decoder correction should be applied. CausalLens targets visually sensitive attention heads (Ji et al., 2026), providing another distinct correction operator. Output-polarity shifts and plausibility truncation can confound apparent POPE gains (Yin et al., 2025); we therefore retain class-conditional and cross-benchmark diagnostics. We ask when an operator's repair outweighs its damage. Analyses of prompt-copying heads and dual grounding/hallucination pathways caution against a single-margin causal explanation (Rudman et al., 2026; Liu et al., 2026).

Detection and selective correction. HALP predicts hallucination before generation; TruthPrInt intervenes using internal truthfulness representations; HaloProbe studies Bayesian detection and mitigation (Kogilathota et al., 2026; Duan et al., 2025; Zohrabi et al., 2026). VGD learns an internal grounding verifier and selectively corrects objects (Yang et al., 2026). V-ITI explicitly addresses when to intervene by detecting visual neglect before applying visual recall (Sun et al., 2025). Selection is therefore established. Our question is whether a *specified* correction earns more reward than keeping the base decoder, tested against matched original-answer risk and multimodal-hidden controls. We do not reduce every earlier selection rule to an error-risk threshold.

Visual verification and process rewards. Process supervision evaluates intermediate decisions (Lightman et al., 2023). Process advantage verifiers define a step's reward by the progress it makes toward a correct final answer (Setlur et al., 2024). Perceval studies perception-centric PRMs, PaLMR aligns visual reasoning processes, and CAVE assigns credit to belief updates and evidence acquisition (Min et al., 2026; Li et al., 2026; Guo et al., 2026). Phase-wise self-reward and programmatic visual verification offer related guidance (Zhang et al., 2026b; Sahu et al., 2024). Paired revision outcomes have also been used to learn whether revising a retrieval-augmented answer helps (Motlagh et al., 2026). Our PRM interface evaluates one pre-generation VLM decoder action against no action, using paired final outcomes rather than claim-level or token-level process labels. Factually augmented RLHF and multimodal preference optimization are complementary backbone-alignment approaches (Sun et al., 2023; Wang et al., 2024a).

3 Preliminaries: From Error Risk to Action-Relative Reward

3.1 The keep-or-correct decision

A frozen VLM defines $p_\theta(y_t \mid I, q, y_{<t})$ over a vocabulary $\mathcal{Y}$, given image I, query q, and generated prefix $y_{<t}$. We analyze one decision at a fixed prefix. Let $Y^{(0)}$ be the continuation under the unmodified decoder and $Y^{(1)}$ the continuation under a specified correction action. Their binary correctness values are $C_0, C_1 \in \{0, 1\}$, and the correction's reward *relative to* leaving the decoder unchanged is $\Delta = C_1 - C_0$. Thus $\Delta = 1$ denotes repair, $\Delta = -1$ denotes damage, and $\Delta = 0$ means that correctness is unchanged, whether both answers are right or both are wrong. This binary reward counts false presence and false absence as errors; it is a broader surrogate than unsupported positive claims. Both branches use the same initial state, stopping rule, and evaluation target. For stochastic decoding, the joint sampling scheme must also be specified; the expectation below includes that randomness. No outcome-dependent feature is available to the policy at decision time.

Let Z denote the available pre-action information. Define

$$p(Z) = \Pr(C_0 = 0 \mid Z), \quad r(Z) = \Pr(C_1 = 1 \mid C_0 = 0, Z), \quad d(Z) = \Pr(C_1 = 0 \mid C_0 = 1, Z). \tag{1}$$

Conditional rates on zero-probability events may be assigned arbitrarily; their contributions below are zero. The action-relative process reward is

$$u(Z) = \mathbb{E}[\Delta \mid Z] = p(Z)r(Z) - (1 - p(Z))d(Z). \tag{2}$$

For example, at $p = 0.6$, rates $(r, d) = (0.8, 0.1)$ give $u = 0.44$, whereas $(0.1, 0.7)$ give $u = -0.22$. These are illustrative distributions, not experimental measurements. The same error risk can justify opposite actions. Predicting p alone is sufficient only under additional restrictions on r and d; a more accurate detector need not be a better intervention policy. The reward is action-relative, not a generic score for whether the original answer is correct: changing the specified correction changes C_1 and its target u.

3.2 Optimal decisions and the cost of gating

Let $\pi(Z) \in \{0, 1\}$ select the correction and let $c \geq 0$ be a cost in the same units as correctness reward. The expected improvement over preserving the original decoder, net of this cost, is $J(\pi) = \mathbb{E}[\pi(Z)(\Delta - c)]$. Ties are resolved by keeping the original decoder.

Proposition 1 (One-decision rule and regret)**.** *The optimal measurable policy is* $\pi^*(Z) = \mathbf{1}\{u(Z) > c\}$*. For a candidate gate* $B(Z) \in \{0, 1\}$ *and estimate* $\hat{u}(Z)$*, let* $\hat{\pi}(Z) = B(Z)\mathbf{1}\{\hat{u}(Z) > c\}$*. Then*

$$J(\pi^*) - J(\hat{\pi}) \leq \mathbb{E}[(1 - B)(u - c)_+] + \mathbb{E}[B\,|\hat{u} - u|], \tag{3}$$

where $(x)_+ = \max(x, 0)$.

The terms separate opportunities lost to gating from value-estimation error. A cheap gate trades computation for potentially irrecoverable loss. This population result is not a finite-sample neural-scorer guarantee; proofs appear in Appendix B.

3.3 When can additional visual evidence help?

Let $S = (h, g)$ contain the decoder hidden state and logit features; h is already multimodal. Let $Z = (S, M)$ additionally expose patch memory M. For information X, define the oracle value $V(X) = \mathbb{E}[(\mathbb{E}[\Delta \mid X] - c)_+]$.

Proposition 2 (Decision-relevant visual information)**.** *For the same action, reward, and cost,* $V(Z) \geq V(S)$*. Strict improvement occurs if and only if, on a set of* S *with positive probability,* $u(Z) - c$ *has both strictly positive and strictly negative conditional values with positive probability.*

Conditional Jensen's inequality proves the result. Redundant memory gives no advantage, and a finite-data compressed readout need not realize the oracle gain. Hence we test hidden-only and wrong-image controls and equal budgets; score correlation alone cannot establish complementary value.

4 Method: Estimating and Acting on Relative Reward

From the decision rule to functional modules. Equation 2 shows why an error-risk score alone cannot choose a correction: repair and damage rates can differ at the same risk (Appendix B.1). We therefore train the PRM interface on paired corrected and unmodified outcomes, as in Equation 8. The instantiated correction is the fixed layer-contrast decoder action defined in Equation 9; the assessor is trained for that action rather than receiving arbitrary action identifiers as inputs. Proposition 2 gives static image-derived evidence a narrower role: it can improve an oracle decision only when it changes which side of the action threshold the conditional reward lies on, given the already multimodal decoder state. This motivates the visual readout and its hidden-only control. The regret decomposition in Proposition 1 motivates an optional gate without treating it as free. Appendix B.3 links the PRM's paired loss to this decision regret. These population results specify information, reward and decision roles; whether the finite learned readout realizes them is an empirical question. The guarantees compare policies over the same correction operator; performance against a different decoder such as NoLan or ILVAD is an empirical question.

4.1 Visual Evidence State

The frozen vision encoder E_θ and multimodal projector P_θ produce

$$M = P_\theta(E_\theta(I)) \in \mathbb{R}^{N \times D}. \tag{4}$$

N is the number of projected image tokens and D is their width. The assessor masks padding when N varies; backbone-specific dimensions are reported in Appendix C. Captured from the same multimodal prefill, this image-derived memory stays fixed throughout generation and requires no second vision encoding. It is a static, potentially lossy input to the reward assessor, not a guarantee of new information beyond the decoder state. Query-conditioned readout cannot overwrite it, although the scorer may still use textual shortcuts.

4.2 Beyond Error Risk: Action-Relative Process Reward Modeling

Let $h \in \mathbb{R}^{1\times D}$ be the last prompt-position state of the final decoder block, recorded from the *multimodal* prefill. For binary object-presence queries, define text-only and visual yes/no margins

$$m_T = \ell_T(\texttt{Yes}) - \ell_T(\texttt{No}), \quad m_V = \ell_V(\texttt{Yes}) - \ell_V(\texttt{No}), \quad g = [m_T, m_V, m_V - m_T]. \tag{5}$$

The token IDs are fixed by the tokenizer and prompt configuration; these are not probabilities of a complete multi-token answer. The difference $m_V - m_T$ is a diagnostic contrast, not an identified causal visual contribution. A large m_T is neither necessary nor sufficient for hallucination.

With readout width P, external projections form

$$\begin{aligned} Q &= hW_Q \in \mathbb{R}^{1\times P}, \qquad K = MW_K \in \mathbb{R}^{N\times P}, \\ V &= MW_V \in \mathbb{R}^{N\times P}, \qquad z = \mathrm{MHA}(Q, K, V) \in \mathbb{R}^P. \end{aligned} \tag{6}$$

We suppress the singleton query axis when using Q and z in subsequent residuals and concatenations. Here $W_Q, W_K, W_V \in \mathbb{R}^{D\times P}$. Multi-head attention also contains its own input and output projections and uses scaled dot-product attention (Vaswani et al., 2017); its head count is in Appendix C. Separate MLPs $f_g : \mathbb{R}^3 \to \mathbb{R}^P$ and $f_h : \mathbb{R}^{1\times D} \to \mathbb{R}^P$ form $e_g = f_g(g)$ and $e_h = f_h(h)$. The direct state branch retains information that can be lost by attending to patch memory alone. A fusion MLP $f_s : \mathbb{R}^{3P} \to \mathbb{R}$ gives the hybrid assessor

$$s_\phi = \sigma(f_s(\mathrm{LN}([z; e_h; e_g]))), \qquad \hat{u}_\phi = 2s_\phi - 1. \tag{7}$$

The state branch does not assume that h lacks visual evidence; it explicitly preserves that evidence while allowing memory to contribute when useful. We also test a residual readout $[z + Q; e_g]$, a memory-only readout $[z; e_g]$, and hidden-only $[e_h; e_g]$ controls. Hidden widths, head count, selected readout width and parameter-matched controls are specified in Appendix C. GELU, dropout and LayerNorm follow the implementation (Ba et al., 2016). None of these readouts certifies that a claim is supported by a particular patch.

Supervision determines the meaning of the score. For each training state, execute the fixed base and correction branches offline and label their complete short answers. Use the soft target

$$t_i = \frac{C_{1,i} - C_{0,i} + 1}{2} \in \{0, \tfrac{1}{2}, 1\}, \qquad \mathcal{L}(\phi) = -\frac{1}{n}\sum_i [t_i \log s_{\phi,i} + (1 - t_i)\log(1 - s_{\phi,i})]. \tag{8}$$

For unweighted population cross-entropy, its minimizer satisfies $2s^*(Z) - 1 = \mathbb{E}[\Delta \mid Z]$. Paired outcomes are training labels only; inference evaluates neither continuation in advance. The action is fixed per scorer, so changing its layer, strength, or plausibility mask changes the target and requires new validation. The paired-data and matched risk/benefit training paths are present in the revision implementation. This is a process-reward interface for one correction decision at a specified generation boundary, not token-level or multi-step process supervision.

For a matched risk control, the same architecture instead predicts C_0 with unweighted BCE, and interventions are ranked by $1 - s_\phi$. The optimizer, data, action and early-stopping budget are unchanged. Neither head is guaranteed to be calibrated after finite-data training or distribution shift (Guo et al., 2017); both use development thresholds. Image-disjoint partitions precede training and selection. The population identity above does not license selecting thresholds from benchmark labels.

4.3 Reward-Gated Intervention

The instantiated action contrasts final logits ℓ_L with a fixed intermediate-layer readout ℓ_k from the same visual forward pass:

$$\widetilde{\ell} = (1 + \alpha)\ell_L - \alpha\ell_k, \qquad \alpha \geq 0. \tag{9}$$

The action extrapolates *away from* the intermediate prediction, without assuming that this layer is intrinsically more visual. Its fixed experimental strength and layer are specified in Appendix E. A separately labelled masked variant restricts candidates to

$$\mathcal{A}_\beta = \{v \in \mathcal{Y} : \ell_L(v) \geq \max_{w\in\mathcal{Y}} \ell_L(w) + \log\beta\}, \quad 0 < \beta \leq 1, \tag{10}$$

setting logits outside the set to $-\infty$. This relative-logit mask is shift-invariant. Masked and unmasked actions are trained and evaluated separately; backbone-specific development selection chooses the action without access to POPE outcomes. Neither variant is identified with DeCo, whose operator uses a different layer-selection and additive-correction rule.

For the optional candidate heuristic, $B = \mathbf{1}\{m_T > \theta\}$. The value policy uses $\pi = B\mathbf{1}\{\hat{u}_\phi > c\}$; setting $B = 1$ evaluates the assessor everywhere. A tuned threshold τ may replace c as a practical policy choice, but the regret bound then includes an additional $\Pr(B = 1)|\tau - c|$ term. The no-intervention policy must be available during development selection. A high positive text margin targets one particular yes-bias regime and need not cover other errors.

Open-response rewards and a transfer boundary. For an open answer, binary correctness is replaced by a bounded reward $R(Y) \in [0, 1]$. The same target $t = (R(Y^{(1)}) - R(Y^{(0)}) + 1)/2$ estimates the expected reward difference; the decision and information results still apply. The binary repair–damage decomposition is specific to binary rewards and is not reinterpreted as an MMHal judge score.

We instantiate a separate transfer study using 500 COCO training images and 150 development images with 128-token descriptions. CHAIR's canonical object extraction (Rohrbach et al., 2018) supplies a predicted object set $O(Y)$ and an annotation/reference object set $A(I)$. We define

$$R_{\text{obj}}(Y, I) = \frac{2|O(Y) \cap A(I)|}{|O(Y)| + |A(I)|}, \tag{11}$$

with zero reward when both sets are empty. This rewards object coverage and penalizes fabricated objects, but does not score attributes, relations or reasoning. The action is the masked variant at $\beta = 0.1$; $B = 1$ and the three features become normalized next-token entropy, the top-two logit gap, and top-one probability. These inputs and a separate assessor match the long-answer training horizon. Transfer to MMHal is evaluated without using its reference answers to train or select the policy. The object-set training reward is not an MMHal judge score; its transfer must be evaluated separately. Appendix B.4 isolates the mismatch term that the object-set scorer cannot remove.

5 Experiments

Evaluation Setup. We freeze LLaVA-1.5-7B (Liu et al., 2023) and Qwen3.5-9B (Qwen Team, 2026), training backbone-specific assessors on COCO images disjoint from official POPE. Comparisons are within backbone. POPE has three 3,000-query splits (Li et al., 2023); AMBER-D has 14,216 queries (Wang et al., 2023); and MMHal has 96 open-response cases (Sun et al., 2023). POPE-trained policies transfer unchanged to AMBER. ARPR averages three frozen seeds; invalid answers count as errors. Table 1 abbreviates POPE Random/Popular/Adversarial as R/P/A and reports accuracy (Acc) and F1 in percent. MMHal reports score (0–6) and hallucination rate (HR, 0–1). AMBER-D reports accuracy (%) by existence (E), attribute (A) and relation (R). Original is greedy decoding. Preprocessing, data splits, selected actions, judging, and image-cluster intervals are detailed in Appendix E.

Baselines. Under matched prompts and 3/128-token limits, we rerun VCD, ICD, DoLa, DeCo and dynamic CEI on both backbones; AIR, ILVAD and PTI run on LLaVA (Leng et al., 2024; Wang et al., 2024b; Chuang et al., 2024; Wang et al., 2025; Fazli et al., 2026; Sun et al., 2026; Xie et al., 2026; Zhang et al., 2026a). NoLan-Base/Plus runs on both (Ren et al., 2026). The released NoLan code samples; our evaluation uses greedy decoding. ILVAD retains its published POPE settings and PTI its released 100-image directions and coefficients. Author equations and native development probes check the adapters. Qwen AIR/ILVAD/PTI lack validated architecture-specific rules or vectors (Appendix E). Literature scores with different judges or budgets remain separate in Table 6; shared-GPU runtimes do not support latency claims.

5.1 Action-relative correction across backbones

Table 1 groups literature methods by family across three benchmarks. Internal controls appear in Table 3, with their open-response variants retained in Table 5. On LLaVA POPE, ARPR improves

Table 1: **Three-benchmark comparison with recent public baselines. Bold**/underline mark best/second displayed values within each backbone, not significance.

Family	Method	POPE ↑						MMHal		AMBER-D ↑		
		R-Acc	R-F1	P-Acc	P-F1	A-Acc	A-F1	Score↑	HR↓	E	A	R
LLaVA-1.5-7B												
General	Original	89.67	89.78	86.23	86.83	79.80	81.79	2.021	0.667	70.3	72.1	73.7
Contrast	VCD	87.50	87.73	84.27	85.03	78.30	80.47	1.740	0.719	67.0	73.2	71.7
	ICD	89.87	89.71	87.30	87.43	81.27	82.50	2.010	0.667	73.1	72.7	73.3
	DoLa	89.67	89.76	86.20	86.77	79.77	81.73	2.010	0.667	70.4	72.2	73.9
	NoLan+	90.77	90.70	88.03	88.28	82.53	83.81	1.396	0.854	77.8	71.6	69.2
Fusion	DeCo	88.83	89.29	84.67	85.86	78.37	81.15	1.698	0.719	62.8	68.7	66.0
Injection	Dynamic CEI	90.57	90.28	87.70	87.69	82.20	83.11	2.177	0.635	77.3	76.6	65.2
Attention	AIR	89.03	89.52	85.83	86.86	77.60	80.70	2.167	0.625	65.9	66.9	69.2
	ILVAD	90.03	90.14	87.20	87.68	80.23	82.18	1.688	0.729	72.6	71.8	71.3
Steering	PTI	89.67	89.67	86.73	87.12	80.27	81.97	2.062	0.635	74.2	72.6	71.8
Selective	**ARPR (ours)**	**91.12**	**90.74**	**90.52**	**89.96**	**84.84**	**85.17**	**2.312**	**0.602**	**78.2**	**77.8**	**75.1**
Qwen3.5-9B												
General	Original	92.63	92.20	89.50	89.29	87.53	87.52	3.406	0.573	90.2	79.5	90.1
Contrast	VCD	92.33	91.99	88.17	88.82	86.70	87.40	3.292	0.583	86.0	80.9	89.1
	ICD	93.20	93.02	89.00	89.23	86.03	86.70	3.365	0.552	87.6	79.4	86.3
	DoLa	92.83	92.43	89.43	89.37	87.40	87.54	3.427	0.542	89.6	79.3	89.8
	NoLan+	90.77	91.60	85.50	89.75	82.37	87.79	3.219	0.573	82.3	71.2	37.5
Fusion	DeCo	92.87	93.14	88.73	89.71	86.10	87.51	3.760	0.490	83.8	71.5	84.0
Injection	Dynamic CEI	93.07	92.75	89.47	89.40	86.97	87.25	1.833	0.573	91.3	80.6	90.4
Selective	**ARPR (ours)**	**93.41**	**93.88**	**90.73**	**90.42**	**88.56**	**88.52**	**3.924**	**0.429**	**92.3**	**81.6**	**91.6**

greedy accuracy by 3.59 percentage points (pp), and exceeds dynamic CEI, an earlier strong decoding baseline, by 2.00 pp. Applying the action everywhere instead reduces accuracy to 70.86%: 877 repairs are outweighed by 2,171 damaged correct answers. ARPR estimates this fixed action's reward relative to greedy decoding and applies a development-selected gate.

On Qwen, ARPR gains 1.01 pp over greedy. Unrestricted action reaches 89.51%, with 131 repairs and 165 damaged answers. However, hidden-only value controls remain competitive: ARPR minus hidden is $+0.34$ pp on LLaVA and $+0.40$ pp on Qwen. The gain supports selective correction, without establishing an incremental patch-memory advantage.

Recent public interventions. Against NoLan-Plus, pooled POPE gains are $+1.72$ pp on LLaVA and $+4.69$ pp on Qwen. LLaVA gains are $+3.01$ pp over ILVAD and $+3.27$ pp over PTI. Figure 3a instead measures image-level breadth: among 500 images with 18 POPE questions each, ARPR improves 250 and harms 63 after averaging its three fixed policies (NoLan-Plus: 164/39; PTI: 49/15). Panel b conditions net corrections on Original errors per image: ARPR gains 0.74/1.09/1.60 correct answers per image in the 2–3/4–5/$\geq$6-error strata, but loses 0.22 in the 0–1-error stratum (162 images; image 95% interval $[-0.41, -0.05]$). These label-defined strata are descriptive, not used for policy selection. Panel c separates repairs from damage; Appendix Table 7 reports polarity and invalid outputs.

5.2 Risk–Reward Distinction, Sensitivity and Robustness

What aggregate accuracy hides. Figures 2 and 3 separate three questions that a pooled score conflates: whether the original answer is wrong, whether the fixed action changes it, and whether that change helps. The positive benefit–risk gaps across representations in Figure 2a favor learning the action's value directly. At equal budgets, risk ranking can find repairs but incurs substantially more damage; value ranking withholds many such changes as its budget grows (Figure 2b–c). Figure 3a–c adds the image-level view: improvements extend beyond a few images, yet nearly-correct images remain a failure stratum, and pooled scores can conceal different repair–damage tradeoffs. Panel d checks open responses: the higher MMHal score coincides with a lower hallucination rate. These diagnostics support gating on predicted action reward rather than error suspicion alone, without implying uniform benefit on every image.

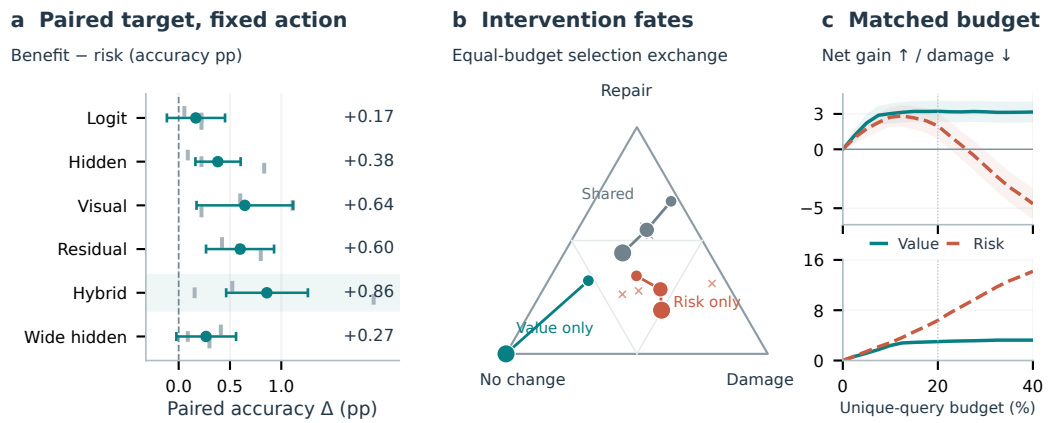


Figure 2: **Risk versus action reward on LLaVA POPE. (a)** Benefit-minus-risk accuracy. **(b)** Outcomes at equal budgets. **(c)** Gain and damage versus budget; bands cluster by image.

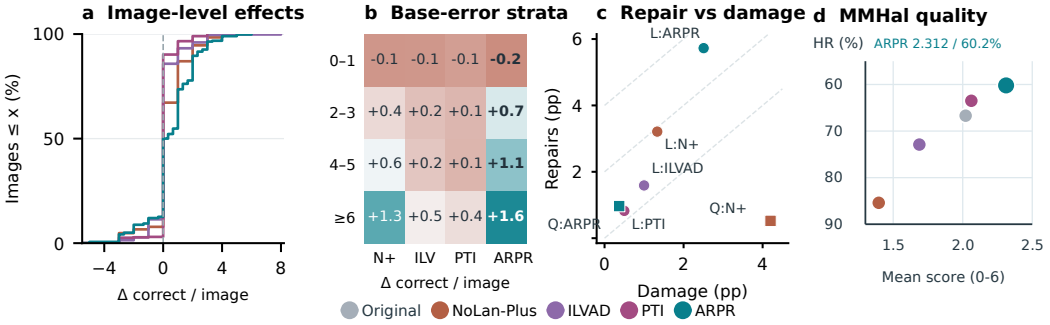


Figure 3: **POPE and MMHal comparisons with recent baselines. (a)** Image-level net-correction ECDF. **(b)** Gain by Original-error stratum. **(c)** POPE repairs versus damage. **(d)** LLaVA MMHal mean score versus hallucination rate. N+: NoLan-Plus; ILV: ILVAD.

Error risk is not action reward. Paired-benefit supervision exceeds error-risk supervision by 0.17–0.86 pp across six matched width-256 assessors (Figure 2a; image-cluster 95% intervals and seed ticks). Panel b tests what changes when benefit ranking replaces risk ranking at the same budget. At 10%, benefit-only calls make 175 repairs and 83 damages, versus 216/205 for risk-only calls. At 20%, benefit-only calls yield 3 repairs, no damages and 1,862 unchanged answers; risk-only calls yield 578 repairs, 909 damages and 544 unchanged answers (counts over three seed–case replays). Thus the gain here chiefly avoids harmful risk-ranked actions, rather than finding additional repairs. At 30%, every benefit-only call leaves the answer unchanged, a saturation limit. In the repair–damage–unchanged simplex, larger dots denote the 10/20/30% unique-query budgets and crosses the three seeds at 20%. Panel c ranks the same fixed action across equal budgets on all 9,000 cases; its gain shading is an image-cluster 95% interval. At 20% the paired accuracy difference is 1.24 pp $[0.66, 1.82]$. Budget allocation uses scores, never test correctness; these replays do not establish an incremental memory effect. Figure 4a selects the ten most frequent POPE object categories by frequency alone. ARPR improves *cup* most, but slightly hurts *backpack*; an orthogonal polarity audit finds No-question gains (5.80–15.13 pp) and Yes-question losses (4.96–5.04 pp). Score-decile bin MAE is 0.032 (b), while repair/damage overlap differs across baselines (c). These are descriptive heterogeneity checks, not a new selection criterion.

Table 2: **Ablation and sensitivity on LLaVA POPE.** Gain versus Original; CI: paired 95% interval; Act: coverage.

Configuration	Gain	CI	Act%
Benefit-head representation			
Hidden only, 256	+3.54	[+2.78,+4.30]	26.8
Visual memory, 256	+2.45	[+1.65,+3.25]	16.9
Hybrid, 256	+3.47	[+2.65,+4.29]	50.6
ARPR selected, 128	**+3.59**	[+2.83,+4.34]	**38.5**
Fixed-score threshold replay			
No gate, $\tau = -0.1$	+3.51	[+2.71,+4.31]	70.5
No gate, $\tau = 0$	+3.54	[+2.77,+4.31]	47.3
No gate, $\tau = +0.1$	+3.56	[+2.83,+4.29]	8.6

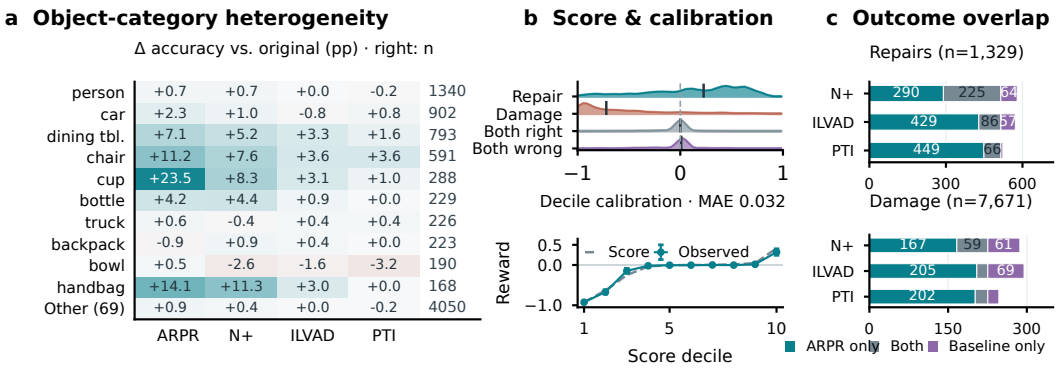


Figure 4: **Outcome heterogeneity on LLaVA POPE. (a)** Gain by object category. **(b)** Fixed-action scores and calibration. **(c)** Repair/damage overlap with recent baselines.

Memory and transfer boundaries. Memory swap/zero changes accuracy by at most 0.03 pp (intervals include zero), with hidden states and action fixed. This tests patch readout, not backbone image use, and does not realize Proposition 2's oracle gain. The stability is consistent with visual evidence already encoded in the frozen hidden state, but cannot distinguish such redundancy from an ineffective memory readout. Appendix E and Appendix F provide robustness checks and replayed cases.

Ablation and threshold stability. Matched width-256 hybrid gains 3.47 pp versus 3.54 pp for hidden-only, leaving memory's incremental effect unproven (Table 2). Frozen width-128 scores at $\tau = -0.1, 0, +0.1$ gain 3.51–3.56 pp while coverage spans 70.5–8.6%. Coverage thus shifts by 61.9 pp while gain varies by only 0.05 pp, so intervention frequency alone cannot explain the improvement. The gate sweep remains in Appendix Figure 5; no test result selects a configuration.

5.3 Frozen transfer beyond object existence

AMBER reuses POPE policies. Table 3 reports attribute/relation balanced accuracy (BA), since existence is all-negative. ARPR's E/A/R accuracies are 78.2/77.8/75.1% on LLaVA and 92.3/81.6/91.6% on Qwen. Qwen's 54.1-pp relation gap to NoLan-Plus coincides with 62.1% unparsed answers (Appendix E); Appendix Figure 7 gives the full error accounting.

Table 3: **Frozen controls (%).** P: POPE accuracy; BA: AMBER balanced accuracy.

	LLaVA		Qwen	
Control	P	BA	P	BA
Original	85.23	71.77	89.89	81.38
Always act	70.86	55.10	89.51	79.19
Risk selector	87.51	72.17	90.31	81.48
Hidden value	88.49	72.43	90.50	81.20
ARPR	**88.83**	**75.85**	**90.90**	**83.20**
Δ *vs. Original*	+3.59	+4.07	+1.01	+1.82

MMHal transfer. Caption-reward assessors use separate COCO data, never MMHal labels (Sun et al., 2023). ARPR scores 2.312 on LLaVA and 3.924 on Qwen; PTI scores 2.062 on LLaVA and DeCo scores 3.760 on Qwen. Figure 3d and Appendix F.2 show the open-response outcomes. Relative to Original, both backbones improve their mean score and reduce hallucination rate, so transfer does not trade one reported axis for the other.

6 Conclusion

Action-relative process reward addresses a decision that error risk alone cannot resolve: whether a specified correction improves the answer over leaving the decoder unchanged. ARPR learns this value from paired outcomes and gates one fixed decoder action before generation. It improves POPE on two VLM backbones, while matched controls isolate the value of reward supervision from error-risk supervision. Incremental patch-memory benefit remains unproven. The current one-decision score, static visual features and fixed correction action also limit the setting; richer action choices require their own paired outcome estimates and evaluation.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. arXiv preprint arXiv:1607.06450, 2016. URL https://arxiv.org/abs/1607.06450.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edc36117f795ca52a0cbf6a7b3882859-Abstract-Conference.html.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. TruthPrInt: Mitigating Large Vision-Language Models Object Hallucination Via Latent Truthful-Guided Pre-Intervention. arXiv preprint arXiv:2503.10602, 2025. URL https://arxiv.org/abs/2503.10602.

Mehrdad Fazli, Bowen Wei, and Ziwei Zhu. Inject to heal: Alleviating hallucination in LVLMs via context embedding injection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41177–41193, 2026. doi: 10.18653/v1/2026.findings-acl.2048. URL https://aclanthology.org/2026.findings-acl.2048/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. arXiv preprint arXiv:1706.04599, 2017. URL https://arxiv.org/abs/1706.04599.

Tengda Guo, Jie Leng, Hanlei Li, Yaoyuan Liang, Qingyue Zhang, Dian Yang, Mingyu Zhang, Yuhua Fu, and Shao-Lun Huang. CAVE: A Structured Credit Assignment Approach for Fragmented Visual Evidence Reasoning. arXiv preprint arXiv:2605.16416, 2026. URL https://arxiv.org/abs/2605.16416.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ji_CausalLens_Sensitivity-Guided_Multi-Head_Causal_Intervention_for_Hallucination_Mitigation_in_Large_CVPR_2026_paper.html.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting Hallucinations in Vision-Language Models without Generating a Single Token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, 2026. doi: 10.18653/v1/2026.eacl-long.287. URL https://aclanthology.org/2026.eacl-long.287/.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Yantao Li, Qiang Hui, Chenyang Yan, Kanzhi Cheng, Fang Zhao, Chao Tan, Huanling Gao, Jianbing Zhang, Kai Wang, Xinyu Dai, and Shiguo Lian. PaLMR: Towards Faithful Visual Reasoning via Multimodal Process Alignment. arXiv preprint arXiv:2603.06652, 2026. URL https://arxiv.org/abs/2603.06652.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. arXiv preprint arXiv:2305.20050, 2023. URL https://arxiv.org/abs/2305.20050.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. arXiv preprint arXiv:2310.03744, 2023. URL https://arxiv.org/abs/2310.03744.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. Dual-Pathway Circuits of Object Hallucination in Vision-Language Models. arXiv preprint arXiv:2605.13156, 2026. URL https://arxiv.org/abs/2605.13156.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101, 2017. URL https://arxiv.org/abs/1711.05101.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving Vision-language Models with Perception-centric Process Reward Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33099–33109, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Min_Improving_Vision-language_Models_with_Perception-centric_Process_Reward_Models_CVPR_2026_paper.html.

Nicholas Kashani Motlagh, Tim Anderson, Jeremy Gwinnup, and Grant Erdmann. Return or Revise? Learning When Revision Helps Retrieval-Augmented QA. arXiv preprint arXiv:2609.30087, 2026. URL https://arxiv.org/abs/2609.30087.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. SECOND: Mitigating Perceptual Hallucination in Vision-Language Models via Selective and Contrastive Decoding. arXiv preprint arXiv:2506.08391, 2025. URL https://arxiv.org/abs/2506.08391.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Official model snapshot c202236235762e1c871ad0ccb60c8ee5ba337b9a; accessed September 25, 2026.

Lingfeng Ren, Weihao Yu, Runpeng Yu, and Xinchao Wang. NoLan: Mitigating object hallucinations in large vision-language models via dynamic suppression of language priors. arXiv preprint arXiv:2602.22144, 2026. URL https://arxiv.org/abs/2602.22144.

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. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Carsten Eickhoff, Ritambhara Singh, and Kyle Mahowald. Mechanisms of Prompt-Induced Hallucination in Vision–Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41894–41912, 2026. doi: 10.18653/v1/2026.acl-long.1941. URL https://aclanthology.org/2026.acl-long.1941/.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting Hallucination in Vision-LLMs via Claim Decomposition and Program of Thought Verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024. doi: 10.18653/v1/2024.emnlp-main.470. URL https://aclanthology.org/2024.emnlp-main.470/.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. arXiv preprint arXiv:2410.08146, 2024. URL https://arxiv.org/abs/2410.08146.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in LVLMs via attention imbalance rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8930–8940, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Sun_Mitigating_Object_Hallucinations_in_LVLMs_via_Attention_Imbalance_Rectification_CVPRF_2026_paper.html.

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, and Yanan Cao. V-ITI: Mitigating Hallucinations in Multimodal Large Language Models via Visual Inference-Time Intervention. arXiv preprint arXiv:2512.03542, 2025. URL https://arxiv.org/abs/2512.03542.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv preprint arXiv:2309.14525, 2023. URL https://arxiv.org/abs/2309.14525.

Kai Tang, Jinhao You, Bohua Zhang, Yichen Guo, Yiding Sun, Dongxu Zhang, Chenxi Li, Xiande Huang, and Shanghang Zhang. SeeMe: Mitigating hallucinations in large vision-language models through effective visual token engineering. arXiv preprint arXiv:2607.04163, 2026. URL https://arxiv.org/abs/2607.04163.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. arXiv preprint arXiv:1706.03762, 2017. URL https://arxiv.org/abs/1706.03762.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/24079b91da7257cb78805262996152b8-Abstract-Conference.html.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. arXiv preprint arXiv:2406.11839, 2024a. URL https://arxiv.org/abs/2406.11839.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. arXiv preprint arXiv:2311.07397, 2023. URL https://arxiv.org/abs/2311.07397.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024b. doi: 10.18653/v1/2024.findings-acl.937. URL https://aclanthology.org/2024.findings-acl.937/.

Yutong Xie, Zhenglin Hua, Ran Wang, Wing W. Y. Ng, Xizhao Wang, and Yuheng Jia. Finding the correct visual evidence without forgetting: Mitigating hallucination in LVLMs via inter-layer visual attention discrepancy. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.20965.

Lei Yang, Xinze Liu, Dayan Wu, Ding Wang, Hengjie Zhu, Zihao Zhang, Tianzhu Hu, Hanqi Wu, Peng Fu, and Zheng Lin. Hallucinations Leave a Grounding Signature:Verifier-Guided Decoding for Selective Object Correction. arXiv preprint arXiv:2607.27823, 2026. URL https://arxiv.org/abs/2607.27823.

Hao Yin, Guangzong Si, and Zilei Wang. The Mirage of Performance Gains: Why Contrastive Decoding Fails to Mitigate Object Hallucinations in MLLMs? arXiv preprint arXiv:2504.10020, 2025. URL https://arxiv.org/abs/2504.10020v4. Version 4, revised June 2026.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25293–25303, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Prefill-Time_Intervention_for_Mitigating_Hallucination_in_Large_Vision-Language_Models_CVPR_2026_paper.html.

Yu Zhang, Chuyang Sun, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. Mitigating Multimodal Hallucination via Phase-wise Self-reward. arXiv preprint arXiv:2604.17982, 2026b. URL https://arxiv.org/abs/2604.17982.

Reihaneh Zohrabi, Hosein Hasani, Akshita Gupta, Mahdieh Soleymani Baghshah, Anna Rohrbach, and Marcus Rohrbach. HaloProbe: Bayesian Detection and Mitigation of Object Hallucinations in Vision-Language Models. arXiv preprint arXiv:2604.06165, 2026. URL https://arxiv.org/abs/2604.06165.

Appendix Contents

A. AI Use Statement. Discloses the scope of AI assistance and author responsibility.

B. Proofs and Boundary Cases. Derives the reward identities and regret bounds, and clarifies when error risk and visual evidence do or do not justify intervention.

C. Assessor Architecture and Training Details. Records assessor dimensions, training, and model-selection details.

D. Inference Contract and Process Extension. Specifies the online decoding procedure and the scope of a possible multi-boundary extension.

E. Benchmark Details and Additional Diagnostics. Documents benchmark evaluation settings, baseline audits, and transfer analyses.

F. Qualitative Case Studies. Shows POPE and MMHal examples with original images, answers, policy scores, and decisions.

A AI Use Statement

Large Language Models (LLMs) were used during the preparation of this manuscript exclusively for language-related assistance. In particular, they were employed to improve grammatical correctness, wording, sentence fluency, and the overall clarity of the writing. The use of LLMs did not extend to any scientific aspects of the study, such as developing research ideas, designing the methodology, conducting experiments, analyzing data, or drawing conclusions from the results. All research content and scientific contributions originated from the authors. The authors have thoroughly reviewed the manuscript and bear full responsibility for its accuracy and content.

B Proofs and Boundary Cases

Action-value decomposition. Since $C_1 - C_0$ is 1 on $(C_0, C_1) = (0, 1)$, -1 on $(1, 0)$, and zero otherwise, conditioning on Z gives

$$\mathbb{E}[C_1 - C_0 \mid Z] = \Pr(C_0 = 0, C_1 = 1 \mid Z) - \Pr(C_0 = 1, C_1 = 0 \mid Z) = pr - (1 - p)d.$$

This identity requires paired actions with a common initial state and common evaluation target. It makes no assumption that logits separate additively into independent causal visual and linguistic mechanisms.

Proof of Proposition 1. Conditioning on Z gives $J(\pi) = \mathbb{E}[\pi(Z)(u(Z) - c)]$. Pointwise maximization over $\{0, 1\}$ yields $\pi^* = \mathbf{1}\{u > c\}$. When $B = 0$, the regret is exactly $(u - c)_+$. When $B = 1$, regret is zero if the decisions agree. If $u > c$ but $\hat{u} \le c$, regret is $u - c \le u - \hat{u} = |u - \hat{u}|$. If $u \le c$ but $\hat{u} > c$, regret is $c - u \le \hat{u} - u = |\hat{u} - u|$. Integrating the cases proves Equation 3. For threshold τ, apply the same argument to the effective estimate $\hat{u} + c - \tau$, and use $|(\hat{u} + c - \tau) - u| \le |\hat{u} - u| + |\tau - c|$. □

Proof of Proposition 2. Let $U = \mathbb{E}[\Delta \mid Z] - c$. Since S is a component of Z, the tower property gives $\mathbb{E}[U \mid S] = \mathbb{E}[\Delta \mid S] - c$. Convexity of $x \mapsto x_+$ implies

$$\mathbb{E}[U_+ \mid S] \ge (\mathbb{E}[U \mid S])_+.$$

Expectation over S gives $V(Z) \ge V(S)$. If both $U > 0$ and $U < 0$ have positive conditional probability, the conditional inequality is strict: when $\mathbb{E}[U \mid S] \ge 0$, the gap is $\mathbb{E}[(-U)_+ \mid S] > 0$; when $\mathbb{E}[U \mid S] < 0$, the gap is $\mathbb{E}[U_+ \mid S] > 0$. A positive-probability set of such S yields a strictly positive total gap. If all conditional values lie on the same side of zero, that conditional gap is zero, even when the memory changes the magnitude of the value. □

Detection precision is a different quantity. For error event $H = \{C_0 = 0\}$ inside a first-stage set $B = 1$, let $\rho = \Pr(H \mid B = 1)$, $a = \Pr(G = 1 \mid H, B = 1)$, and $b = \Pr(G = 1 \mid \neg H, B = 1)$ for a second-stage error flag G. When the denominator is nonzero,

$$\Pr(H \mid B = 1, G = 1) = \frac{\rho a}{\rho a + (1 - \rho) b}. \tag{12}$$

For $0 < \rho < 1$, precision improves over ρ exactly when $a > b$. Filtering also multiplies the first-stage error-detection recall by a; it cannot increase that recall. AUROC above 0.5 does not imply $a > b$ at every threshold. These detector quantities must not be confused with the precision and recall of affirmative answers on POPE.

Precision, base rates, and an AUROC counterexample. Bayes' rule gives Equation 12. Subtracting ρ yields

$$\frac{\rho(1-\rho)(a-b)}{\rho a + (1-\rho)b},$$

whose sign is the sign of $a - b$ under the stated nondegeneracy conditions. The error-detection recall is $\Pr(B = 1, G = 1 \mid H) = \Pr(B = 1 \mid H)a$. As a counterexample to a threshold-wise AUROC guarantee, take error scores $\{0.9, 0.8, 0.1\}$ and correct-answer scores $\{0.95, 0.2, 0\}$, with larger scores predicting error. Five of nine positive–negative pairs are ranked correctly, giving AUROC $5/9 > 1/2$. Thresholding strictly above 0.9 selects only a correct answer, giving error-detection precision zero.

What the two cross-entropy targets estimate. For target $T = (\Delta + 1)/2$, the conditional unweighted binary cross-entropy is minimized by $s^*(Z) = \mathbb{E}[T \mid Z]$. This follows by differentiating $-\bar{t} \log s - (1 - \bar{t}) \log(1 - s)$; endpoints follow by continuity. Thus $2s^* - 1 = u$. A zero-difference pair has target $1/2$, not a label that claims each action is factually correct. It can represent either two correct answers or two incorrect answers.

For correctness label C_0 and positive class weight $w > 0$, the conditional loss is $-w\eta \log s - (1 - \eta) \log(1 - s)$, with minimizer

$$s^* = \frac{w\eta}{w\eta + 1 - \eta}, \qquad \eta = \frac{s^*}{w + (1-w)s^*}.$$

This population inverse is not a guarantee of calibration for a fitted model. In particular, $2s - 1$ from this weighted correctness objective is not an estimate of paired action value. These objectives need distinct checkpoints and metadata even when the network architecture is identical.

Binary margins and their limited interpretation. For labels $y^* \in \{-1, +1\}$, a two-token classifier with visual margin m_V is correct when $y^* m_V > 0$, with equality determined by a declared tie rule. Writing $\delta = m_V - m_T$ gives $m_V = m_T + \delta$ by definition. For a negative label and a strict affirmative prediction, hallucination occurs exactly when $m_T + \delta > 0$, or $\delta > -m_T$. This algebra does not identify causal components. A generated multi-token answer and its parsing rule may also differ from the two-token classifier, so the margin is a feature rather than an exact correctness oracle.

B.1 Paired action rewards and the limits of risk detection

The two potential continuations $Y^{(0)}$ and $Y^{(1)}$ in Section 3 refer to the *same* image, query, prefix, generation horizon, and evaluation rule. The action, including its layer, coefficient, and optional vocabulary mask, is fixed before the scorer is trained. For a stochastic decoder, its random draws and any coupling of the two branches belong to the declared evaluation setting. Both branches are executed offline to obtain a paired target; neither outcome is revealed to the online gate. This is an observed paired-rollout setting, not an assumption that unobserved outcomes can be recovered from observational data. The expectation $\mathbb{E}[C_1 - C_0 \mid Z]$ depends on the two conditional marginal means, whereas separate repair and damage rates also depend on how stochastic branch outcomes are coupled.

Write $A = \{C_0 = 0, C_1 = 1\}$ for repair and $D_{\text{evt}} = \{C_0 = 1, C_1 = 0\}$ for damage. For any pre-action-measurable policy π,

$$\mathbb{E}[\pi\Delta] = \Pr(\pi = 1, A) - \Pr(\pi = 1, D_{\text{evt}}), \tag{13}$$

$$J(\pi) = \Pr(\pi = 1, A) - \Pr(\pi = 1, D_{\text{evt}}) - c \Pr(\pi = 1). \tag{14}$$

Neutral pairs, including two correct or two incorrect continuations, contribute zero reward. Positive mean value does not imply that every selected example is repaired or that the probability of damage is zero.

The comparison with applying this *same* action to every example is also exact. Let $C_\pi = (1-\pi)C_0 + \pi C_1$. Then

$$\mathbb{E}[C_\pi - C_1] = \Pr(\pi = 0, D_{\text{evt}}) - \Pr(\pi = 0, A). \tag{15}$$

Selective correction beats always correcting precisely when the correct answers it preserves by withholding the action outnumber the repairs it misses. This criterion concerns the *fixed* correction operator; it does not compare a policy with a distinct decoding method such as NoLan.

For our pre-action policy π_a and a distinct decoder b with correctness C_b, set $C_a = C_1$ and define $u_a(Z) = \mathbb{E}[C_a - C_0 \mid Z]$ for our fixed action and $u_b(Z) = \mathbb{E}[C_b - C_0 \mid Z]$. On the same cases and evaluation target, the comparison instead decomposes as

$$\mathbb{E}[C_{\pi_a} - C_b] = \mathbb{E}[\pi_a u_a - u_b] = \mathbb{E}[(\pi_a - 1)u_a] + \mathbb{E}[u_a - u_b]. \tag{16}$$

The first term is the effect of withholding action a relative to using it everywhere; the second is an operator difference. Neither the regret bound nor Equation 15 signs that second term. The NoLan, ILVAD, and PTI comparisons therefore measure performance, not the causal contribution of our selector alone.

Proposition 3 (When error risk is decision-sufficient)**.** *Let $P = p(Z)$ and fix $c \geq 0$. A gate that observes only P can implement the full-information Bayes value if and only if*

$$\Pr(u > c \mid P) \Pr(u < c \mid P) = 0 \quad \textit{almost surely}.$$

Equivalently, the strict Bayes action can be modified on indifferent states $u = c$ to be measurable with respect to P. Without this condition, calibrating P perfectly cannot recover the optimal value.

Proof. For a P-measurable decision $g(P)$, conditional regret is $\mathbb{E}[(u-c)_+ \mid P]$ when $g = 0$, and $\mathbb{E}[(c-u)_+ \mid P]$ when $g = 1$. One of these can be zero if and only if, conditional on that P, the corresponding strictly positive or strictly negative value event has probability zero. Such a choice exists almost surely exactly under the stated product condition. Values $u = c$ impose no constraint, so the action there can be chosen to make the gate measurable. □

For a concrete counterexample at $c = 0$, give two equally likely states the same $p = 0.6$ but respective $(r, d) = (0.8, 0.1)$ and $(0.1, 0.7)$. Their values are 0.44 and -0.22. Both are valid joint distributions of (C_0, C_1), yet any risk-only gate must take the same action in both. This is a distributional counterexample, not a benchmark measurement. A risk score can still be useful when repair and damage rates are restricted so that the sign of $pr - (1-p)d - c$ is determined by p; accurate error detection alone does not supply that restriction.

B.2 What a static visual evidence state can and cannot add

The static visual evidence readout supplies an image-derived observation $M = f(I)$ to an assessor that already receives multimodal state S. It does not model transitions. The relevant theoretical question is whether M changes the conditional value of a *fixed* correction after S is known. Let $u_S = \mathbb{E}[\Delta \mid S]$ and $u_{SM} = \mathbb{E}[\Delta \mid S, M]$, with all conditional expectations referring to the same reward, action, and data distribution.

Proposition 4 (Exact oracle gain and an information ceiling)**.** *Set $U = u_{SM} - c$. Then*

$$V(S, M) - V(S) = \mathbb{E}[\min\{\mathbb{E}[U_+ \mid S],\ \mathbb{E}[(-U)_+ \mid S]\}], \tag{17}$$

$$0 \leq V(S, M) - V(S) \leq \mathbb{E}|u_{SM} - u_S| \leq \sqrt{2\,\mathrm{I}(\Delta; M \mid S)}. \tag{18}$$

The last inequality uses natural logarithms; it is vacuous if the conditional mutual information is infinite. Because $M = f(I)$, $V(S, M) \leq V(S, I)$. Equality with this full-image oracle follows, for example, if $\mathbb{E}[\Delta \mid S, I] = \mathbb{E}[\Delta \mid S, M]$ almost surely.

Proof. Condition on S. Write $a = \mathbb{E}[U_+ \mid S]$ and $b = \mathbb{E}[(-U)_+ \mid S]$, so $u_S - c = \mathbb{E}[U \mid S] = a - b$. The conditional value of observing M is a, whereas the value without it is $(a-b)_+$. Their difference is $\min\{a, b\}$, proving Equation 17. The nonnegative-part function is one-Lipschitz; replacing u_{SM} by u_S and taking expectations gives the first upper bound in Equation 18. For conditional laws $P_{\Delta|S,M}$ and $P_{\Delta|S}$, $|\Delta| \leq 1$ implies

$$|u_{SM} - u_S| \leq 2\,\mathrm{TV}(P_{\Delta|S,M}, P_{\Delta|S}) \leq \sqrt{2 D_{\mathrm{KL}}(P_{\Delta|S,M} \| P_{\Delta|S})},$$

where the second inequality is Pinsker's inequality. Jensen's inequality and the definition of conditional mutual information yield the final bound. Finally, (S, M) is a function of (S, I); applying the same conditional Jensen argument as in Proposition 2 establishes $V(S, M) \leq V(S, I)$. The stated equality condition makes their conditional values identical. □

Equation 17 is sharper than a generic statement that "more visual information helps." The memory has decision value only where it can move the conditional action value across the intervention threshold. If M is conditionally redundant given S, then $u_{SM} = u_S$ and the gain is zero. Positive conditional mutual information is not sufficient for a positive gain: memory may alter the magnitude of value while leaving its sign unchanged. It can also encode image details unrelated to this action's reward. These statements concern an oracle with access to exact conditional expectations, not the trained assessor.

The actual readout is a deterministic compression $F = \Phi(S, M) = [z; e_h; e_g]$ followed by a learned score. Data processing gives $V(F) \leq V(S, M)$, and finite-data estimation can reduce realized value further. Because projections and attention need not preserve all of S, this information ordering does not guarantee $V(F) \geq V(S)$. Thus the architectural presence of patch memory cannot certify grounding. Nor does a memory swap that leaves the backbone state unchanged test whether the backbone used the image: it isolates the readout's incremental reliance on M. The reported near-zero accuracy changes under swaps and zeroing bound the present implementation claim, even though the oracle proposition continues to hold.

B.3 Why the action-relative PRM target differs from error risk

The assessor is a process-reward model in the restricted sense that it scores a *specified correction decision at a specified generation boundary*. Its reward is relative to leaving the decoder unchanged: the paired target measures the outcome of the fixed action against the unmodified continuation, whereas error risk measures only the latter. It is trained on paired downstream outcomes, not on a generic claim that a partial answer looks plausible. The current implementation performs one pre-generation decision; a multi-boundary process model would require new state-action targets and a downstream policy definition.

For $T = (\Delta + 1)/2$ and any measurable prediction $s(Z) \in (0, 1)$, define the population soft-label cross-entropy

$$\mathcal{L}(s) = \mathbb{E}[-T \log s(Z) - (1 - T) \log(1 - s(Z))].$$

Since $\mathbb{E}[T \mid Z] = (u(Z) + 1)/2 =: s^*(Z)$, conditional minimization gives s^* and the exact excess-risk identity

$$\mathcal{L}(s) - \mathcal{L}(s^*) = \mathbb{E}[D_{\mathrm{KL}}(\mathrm{Bern}(s^*(Z)) \| \mathrm{Bern}(s(Z)))] \,. \tag{19}$$

This identity holds with fractional labels because cross-entropy is affine in T. It requires no assumption that Δ is Bernoulli. At population optimum, $\hat{u} = 2s - 1$ is conditional paired value; finite-sample optimization, model misspecification, and shift need not attain that optimum.

Proposition 5 (Excess loss, gate regret, and selected harm)**.** *Let* $\epsilon_{\mathrm{CE}} = \mathcal{L}(s) - \mathcal{L}(s^*)$, $\hat{u} = 2s - 1$, *and* $\hat{\pi} = B\mathbf{1}\{\hat{u} > \tau\}$ *for a pre-action gate* B. *For fixed cost* c,

$$J(\pi^*) - J(\hat{\pi}) \leq \mathbb{E}[(1 - B)(u - c)_+] + \sqrt{2\epsilon_{\mathrm{CE}}} + \Pr(B = 1)|\tau - c|. \tag{20}$$

If $R_{\hat{\pi}} = \Pr(\hat{\pi} = 1, A)$ *and* $D_{\hat{\pi}} = \Pr(\hat{\pi} = 1, D_{\mathrm{evt}})$, *then*

$$D_{\hat{\pi}} \leq R_{\hat{\pi}} - \tau \Pr(\hat{\pi} = 1) + \mathbb{E}[\hat{\pi}|u - \hat{u}|]. \tag{21}$$

Proof. Pinsker's inequality for Bernoulli distributions and Equation 19 imply

$$\mathbb{E}|u - \hat{u}| = 2\mathbb{E}|s^* - s| \leq \mathbb{E}\sqrt{2D_{\mathrm{KL}}(\mathrm{Bern}(s^*) \| \mathrm{Bern}(s))} \leq \sqrt{2\epsilon_{\mathrm{CE}}}.$$

Apply Proposition 1 to the effective estimate $\hat{u} + c - \tau$. The triangle inequality bounds its error by $|\hat{u} - u| + |\tau - c|$ on $B = 1$, yielding Equation 20. On selected states $\hat{u} > \tau$, hence $\mathbb{E}[\hat{\pi}u] \geq \tau \Pr(\hat{\pi} = 1) - \mathbb{E}[\hat{\pi}|u - \hat{u}|]$. Equation 13 says $\mathbb{E}[\hat{\pi}u] = R_{\hat{\pi}} - D_{\hat{\pi}}$; rearranging proves Equation 21. □

The loss-to-regret statement is a population implication, not a guarantee that the fitted network has small ϵ_{CE}. The harm inequality controls net repair minus damage; it does not bound the probability of damaging an individual answer without information about repair mass. Likewise, a threshold τ chosen for finite-sample validation need not equal the decision-theoretic cost c. The first term in Equation 20 is lost opportunity from the cheap candidate gate, even if the scorer is perfect.

Why the correction operator needs its own value model. For any two vocabulary tokens i, j, the fixed logit action in Equation 9 changes their first-step margin by

$$\widetilde{m}_{ij} = (1+\alpha)m^L_{ij} - \alpha m^k_{ij} = m^L_{ij} + \alpha(m^L_{ij} - m^k_{ij}). \tag{22}$$

If both tokens survive the optional mask, this is exactly their log odds under the corrected softmax. If $m^L_{ij} > 0$ and $\alpha > 0$, token j overtakes i only when $m^k_{ij} > (1+\alpha)m^L_{ij}/\alpha$; if j is excluded by the mask, the flip is impossible. The algebra predicts a *directional change* in generation, not whether the change is correct. In particular, an action can repair a false claim and damage a true one under the same margin pattern. The paired reward therefore has to be measured for this specific operator. For another action a', define both potential rewards under a declared joint coupling of any decoding randomness. Even with unchanged inputs Z,

$$|\mathbb{E}[\Delta_a \mid Z] - \mathbb{E}[\Delta_{a'} \mid Z]| \le \mathbb{E}[|\Delta_a - \Delta_{a'}| \mid Z].$$

Training on action a alone places no bound on the right-hand side for a different layer, strength, or mask. This is why those choices are fixed for each assessor and validated separately.

A coverage constraint and computation cost. When interventions have a maximum expected coverage b, maximize $\mathbb{E}[\pi(u-c)]$ subject to $\mathbb{E}[\pi] \le b$. A Lagrange multiplier $\lambda \ge 0$ yields the threshold rule $\pi = \mathbf{1}\{u > c + \lambda\}$, with possible randomization where $u = c + \lambda$ to meet a binding budget. This explains score-ranked budget curves when scores estimate value; a risk ranking has no such guarantee. If computational cost depends on state, the same pointwise derivation uses $u(Z) > c(Z)$ for a pre-action-known incremental action cost. The selected $B = 1$ policy pays for text-only prefill and assessor evaluation on every example before choosing the continuation; a cheap gate could skip some assessor calls but still pays for its own features. These costs are separate terms in end-to-end latency relative to greedy decoding. Consequently, a positive reward advantage or sparse action rate alone cannot establish a wall-clock speed advantage.

B.4 Open-response reward transfer and its boundary

For this analysis, extend the object-set training functional R_P in Equation 11 conceptually to MMHal images using their true object sets. Those sets and R_P values are not observed or used for MMHal selection. Write R_J for the evaluator rating divided by six, so both rewards lie in $[0,1]$. The hallucination-rate indicator derived from a rating below three is a different target. For $a \in \{0,1\}$ let $e_a = R_J(Y^{(a)}) - R_P(Y^{(a)})$ and define conditional action values $u_P = \mathbb{E}[R_P(Y^{(1)}) - R_P(Y^{(0)}) \mid Z]$ and $u_J = \mathbb{E}[R_J(Y^{(1)}) - R_J(Y^{(0)}) \mid Z]$. The exact discrepancy and a bound are

$$u_J - u_P = \mathbb{E}[e_1 - e_0 \mid Z], \tag{23}$$

$$|u_J - u_P| \le \mathbb{E}[|e_1| + |e_0| \mid Z]. \tag{24}$$

No independence between the two errors or between the two generated answers is used. Small absolute proxy error for both branches is a sufficient condition for transferring the decision sign away from the threshold; it is not established by using the same images for training. For example, if at a particular Z the proxy mismatch is bounded by $|u_J - u_P| \le \varepsilon_P(Z)$ and scorer error by $|\hat{u}_P - u_P| \le \varepsilon_S(Z)$, then $\hat{u}_P > c + \varepsilon_P(Z) + \varepsilon_S(Z)$ is sufficient for positive judge value. Neither bound is known for the fitted MMHal policy, so this condition is interpretive rather than a certified decision rule.

Proposition 6 (Judge regret of a proxy-value policy)**.** *Let $\pi^*_J = \mathbf{1}\{u_J > c\}$ and $\hat{\pi}_P = B\mathbf{1}\{\hat{u}_P > \tau\}$, where $\hat{u}_P$ estimates u_P from pre-action information. Then, in normalized judge-reward units,*

$$\begin{aligned} J_J(\pi^*_J) - J_J(\hat{\pi}_P) \le{}& \mathbb{E}[(1-B)(u_J - c)_+] + \mathbb{E}[B|u_J - u_P|] \\ &+ \mathbb{E}[B|u_P - \hat{u}_P|] + \Pr(B=1)|\tau - c|. \end{aligned} \tag{25}$$

Here $J_J(\pi) = \mathbb{E}[\pi(R_J(Y^{(1)}) - R_J(Y^{(0)}) - c)]$.

Proof. Apply Proposition 1 with judge value u_J and effective estimate $\hat{u}_P + c - \tau$. On $B = 1$, insert and subtract u_P, then use the triangle inequality. □

The middle term is an *objective mismatch*, which better optimization of the proxy scorer cannot remove. The next term is measured against the *target-domain* conditional proxy value; because the assessor is trained on COCO descriptions, it also absorbs any conditional transport shift to MMHal, beyond fitting and calibration error. Neither term is measured by the reported transfer experiment. Object extraction can give identical R_P to two captions with the same named objects but different attributes or relations, while a judge can score them differently. Conversely, a caption may omit a visible object without hallucinating. Thus paired object reward is a training proxy for a subset of open-answer errors, but it does not theoretically entail a positive MMHal rating gain or improved hallucination rate. The observed 96-case transfer estimates in Section 5 must carry that empirical burden.

C Assessor Architecture and Training Details

The hybrid assessor uses three feed-forward maps. The gap branch has input, hidden and output widths 3, 64 and P, respectively; the decoder-state branch uses D, 768 and P; and the fusion head uses $3P$, 128 and 1. Both backbones have $D = 4096$. The development-selected hybrid sets $P = 128$; matched LLaVA architecture and objective controls use $P = 256$, while the Qwen risk and hidden controls use $P = 128$.

Table 4: Actual assessor sizes. All use $D = 4096$, three scalar features, and four attention heads where applicable.

Assessor	Readout width	Trainable parameters
Gap-only	256	50,433
Hidden-only	256	3,427,073
Memory-only / residual	256	3,492,609
Hybrid	256	6,869,249
Wide hidden	256	6,770,177
Selected hybrid	128	4,942,593

Training uses AdamW (Loshchilov & Hutter, 2017), gradient clipping at one, and early stopping on the first development half. A fixed threshold grid includes never intervening. The second development half selects the architecture and action. These are both development subsets, not additional test sets. Evaluation retains invalid predictions and fixed sample counts.

D Inference Contract and Process Extension

Scope, state handling, and cost. Algorithm 1 fixes the decision for the continuation. It does not rescore every generated token. Running text-only prefill first, then capturing the multimodal state, preserves the image-conditioned decision state. The once-computed text margin is not a prefix-updated text-only reference. Caching projected memory permits reuse, but the method still pays for text prefill, assessor operations, and intermediate vocabulary projection. Memory projection costs $O(NDP)$ and its attention readout costs $O(NP)$; "frozen backbone" does not mean training-free scoring or zero overhead.

Beyond one decision boundary. A multi-boundary extension would define a claim or action boundary j, query the observed state with a representation of that specific claim, and estimate $\mathbb{E}[R(Y^{(a)}) - R(Y^{(0)}) \mid Z_j]$ under a fixed downstream policy. It would need labels for intermediate interventions, preservation of already supported claims, and controls for errors propagated by the generated prefix. An object's supported existence cannot certify an arbitrary attribute or relation involving that object. These are requirements for the extension, not capabilities established by the current one-decision implementation. Summing one-step values without controlling future decisions would not give a justified sequence-level objective.

Algorithm 1 Binary one-boundary action-relative reward gating (fixed action)

Require: Image I, query q, frozen VLM, action-relative reward assessor s_ϕ; fixed action (k, α, β), optional candidate gate B, development threshold τ

Run text-only prefill; retain m_T.
Run multimodal prefill; retain M, h, m_V, visual cache, and layer readout.
Set $\pi \leftarrow 0$.
if $B = 1$ (all states for the selected policy) **then**
    Set $g \leftarrow [m_T, m_V, m_V - m_T]$ and $\hat{u} \leftarrow 2s_\phi(h, M, g) - 1$.
    Set $\pi \leftarrow \mathbf{1}\{\hat{u} > \tau\}$.
end if
for each token in the declared continuation **do**
    Use base logits if $\pi = 0$; otherwise use Equation 9 and the declared optional mask in Equation 10.
    Decode under the fixed generation rule and advance the visual cache.
end for

E Benchmark Details and Additional Diagnostics

Evaluation data and decoding. LLaVA uses FP16, its standard conversation template, padded images and $N = 576$ projected visual tokens. Qwen uses BF16, its native template with thinking disabled, and aspect-ratio images with variable $N \leq 576$; assessor attention masks padding. Both backbones have $D = 4096$. Binary training/development use 12,000/3,000 queries over 2,000/500 image-disjoint COCO images. Two disjoint development halves select epochs/thresholds and architecture/action; all official POPE images are excluded. Each POPE split has 3,000 queries over the same 500 test images, and 5,171 unique image–question pairs map to the 9,000 cases. Binary/open continuations are capped at 3/128 tokens. AMBER-D has 14,216 discriminative queries on 1,004 images and reuses POPE checkpoints and thresholds. MMHal uses all 96 cases, judged with `gpt-4o-2024-11-20`. ARPR reports seeds 17, 42 and 123, never a test-best seed.

For binary POPE, development selects hybrid readout width $P = 128$, the unmasked action at zero-based decoder-layer index $k = 15$ (the sixteenth block) with $\alpha = 2$, and $B = 1$ on both backbones. Assessor dimensions and parameter counts are in Appendix C. Paired 95% intervals resample images with all their questions and average the three fixed policies: 5,000/10,000 draws for primary LLaVA/Qwen comparisons and 5,000 for additional diagnostics. Intervals condition on fitted models; secondary comparisons are unadjusted for multiplicity.

Optimization and frozen thresholds. Assessors use AdamW with learning rate and weight decay 10^{-4}, up to 30 epochs, patience five and binary batch size 64. In seed order 17/42/123, the selected binary thresholds are $(0.01, 0.14, -0.06)$ on LLaVA and $(0.04, -0.09, 0.01)$ on Qwen.

Candidate models and data isolation. For LLaVA, the original three families (gap, hidden, visual), two objectives (risk, benefit), two actions (unmasked, masked) and three seeds give 36 fixed checkpoints. The extension adds residual, hybrid and wide-hidden width-256 models for both objectives and actions, plus hybrid-benefit widths 128 and 512: 48 further checkpoints. Residual preserves the projected query by adding it to the attention readout. Wide-hidden uses a $4096 \to 1536 \to 256$ MLP as a capacity control; its parameter count is reported rather than claimed equal to hybrid. The binary selection searches memory-bearing benefit configurations, with gated and ungated profiles, by mean development accuracy. Hidden-only controls are evaluated without attributing their gain to the static visual evidence readout.

All official POPE image IDs are excluded from the 2,000/500 training and development image sets. A separate custom COCO set is excluded from configuration selection and is not counted as an independent validation set. The open training/development images are subsets of the same isolated COCO partitions. MMHal images and answers never supply training labels.

Qwen evaluation details. Qwen3.5-9B is frozen at its official checkpoint. It is a vision–language model with 32 decoder blocks and hidden width 4,096, not a text-only Qwen checkpoint. The template disables thinking. The image processor preserves aspect ratio with at most 576 merged

tokens; attention in the assessor masks padding. The native block-15 state is normalized and projected by the same output head to define the intermediate readout.

Qwen uses the same 12,000/3,000 binary cases and the same image-disjoint development halves. Hybrid-benefit, hidden-benefit and hybrid-risk heads at width 128, two actions and three seeds yield 18 new binary checkpoints. The eligible set is fixed to $B = 1$. Only the memory-bearing benefit action is selected on the second development half; the hidden and risk rows use that same action. The open study recollects 500/150 descriptions and trains hybrid and hidden width-128 benefit heads, with batch size 64 and the same optimizer/epoch budget. The primary open policy is hybrid; its three thresholds use only the first development half. No MMHal rating selects its architecture or threshold. Qwen's paired intervals resample test images 10,000 times and average the three fixed seed policies within each draw.

Baseline evaluation details. Table 1 uses common-prompt greedy adaptations under the 3/128-token binary/open limits; published decoding defaults may differ. NoLan-Plus adaptively suppresses the language prior. The ARPR policies transfer from POPE to AMBER without retuning, while separate caption-reward assessors handle MMHal. VCD uses diffusion step 500, contrast strength one and relative plausibility threshold 0.1. Its image perturbation is deterministic and shared across questions about the same image. DeCo uses layers 20–28, top-$k = 20$, top-$p = 0.9$ and $\alpha = 0.6$. Dynamic CEI uses injection layer index 10, $\alpha_{\max} = 0.25$ and cutoff 0.55. The released immediate-repeat handling and early-layer projection are retained. AIR applies the authors' language-attention layer/head rules and variance projection without modifying the vision encoder.

All LLaVA reruns use FP16 backbone weights and the common 3/128-token limits. In particular, the CEI paper uses 512-token open answers and its released configuration defaults to 8-bit loading; our numbers must not be presented as exact reproduction of that environment. No test-score-driven baseline hyperparameter tuning is performed. For Qwen DeCo, the same layers 20–28, top-k, top-p and α operate on Qwen hidden states. Qwen dynamic CEI retains the layer-10 injection, cosine/mass coefficient and immediate-repeat rule. These are matched-setting adaptations, not author-released Qwen implementations. On full Qwen POPE, DeCo/CEI obtain pooled accuracy 89.23/89.83%, versus 90.90% for ARPR. On AMBER-D, their E/A/R accuracies are 83.8/71.5/84.0% and 91.3/80.6/90.4%; invalid-output rates are 7.9/21.7/11.2% and 1.2/9.7/3.6%, respectively. Invalid outputs count as errors. DeCo's Qwen MMHal mean is 3.760 versus ARPR's 3.924. AIR's published LLaVA head list does not fit Qwen's 16-query-head, three-linear/one-full-attention pattern. PTI's released 32×128 key/value head vectors do not match Qwen's 4×256 heads, and ILVAD's LLaVA-specific fixed-token attention rule does not describe Qwen's variable image positions. We report no scores for these unsupported Qwen ports.

NoLan-Base uses unit contrast and NoLan-Plus uses symmetric-KL adaptive contrast with $\beta = 0.8$. Both use synchronized image/text-only branches. The released method samples, so our greedy, fixed-horizon results are matched-setting adaptations rather than published-number reproductions. ILVAD uses the authors' POPE values $T = 10$, $\alpha = 5$, $\tau = 5$, $\beta = 1$, layers 0–30 on all three tasks without transfer tuning. Both reruns keep the common prompt, padding and 3/128-token caps. Incomplete or numerically nonfinite outputs are not replaced with guessed answers; parser-invalid answers count as incorrect. SeeMe has no verified author inference release, while CausalLens requires a separate validation of its prompt/image-position assumptions before it can be compared under our shared inputs (Tang et al., 2026; Ji et al., 2026). Other screened recent methods lack a verified inference specification or an independent trained checkpoint; no published score is substituted for a matched rerun.

Open-response development. The 150 image-disjoint development descriptions are divided into two 75-image selection halves. No MMHal image, reference or rating is used for training or selection. On LLaVA development descriptions, always acting lowers mean object-set reward from 0.7920 to 0.6877 (31 improved, 98 degraded descriptions). All seed policies enter the reported MMHal mean. On Qwen's 96 MMHal cases, that mean changes from 3.406 to 3.924; hallucination rate changes from 0.573 to 0.429.

AMBER transfer and additional decoding baselines. The complete AMBER discriminative release contains 14,216 questions over 1,004 images. We use its official images, queries and annotations, retaining all questions and mapping parsed binary answers to the evaluator's Yes/No

categories; invalid answers count as errors. Existence has 4,924 negative cases; attributes have 3,814 positive and 3,814 negative cases; relations have 975 positive and 689 negative cases. Table 1 reports dimension accuracy, not AMBER's generative Cover/Hal/Cog metrics. Balanced accuracy in Table 3 (BA) is the mean of positive and negative recall on the pooled 9,292 attribute/relation queries. Existence is excluded there because it has no positive examples. All three POPE-selected policies, actions and thresholds transfer unchanged; AMBER supplies no training or selection signal. No byte-identical images occur between its 1,004 images and the 2,500 binary training/development images; this check does not detect re-encoded copies. This is transfer of our assessor, not a claim about backbone pretraining exclusion.

NoLan answer-format diagnostic. As a post-hoc check, we rerun NoLan-Plus on all 1,664 Qwen AMBER relation queries with 16 rather than three output tokens, changing no model, prompt, or contrast parameter. Accuracy rises from 37.50% to 38.10%, an image-paired difference of $+0.60$ pp $[-0.06, 1.31]$; the invalid-answer rate falls from 62.08% to 60.04%, or -2.04 pp $[-2.94, -1.14]$. Longer generations often start descriptive image summaries without an explicit yes/no answer. Thus the poor relation score cannot be attributed solely to three-token truncation. This exploratory horizon comparison does not replace the frozen three-token main-table result or select its setting.

DoLa projects hidden states 0,2,...,14 without an extra final normalization, uses the released divergence criterion to select a premature layer, contrasts log probabilities with final layer 32, and applies relative plausibility 0.1. Layer selection is independent per example in Qwen batches. ICD uses contrast strength one, plausibility 0.1, and the fixed released "confused objects detector" system instruction. The unmodified and disturbed branches share the image and receive the same chosen continuation token. Both use each backbone's native template, greedy decoding, repetition penalty one and common 3/128-token limits. These controlled adaptations differ from the original releases' default sampling, repetition and length settings; they are not exact reproductions of published numbers. No test result selects a layer or disturbance prompt.

MMHal scoring contract. We use the complete official MMHal prompt, including all five rubric examples, with image-content annotations, questions, reference answers and model responses. Ratings range from 0 to 6, with hallucination defined as a rating below 3. An ambiguous rating or failed request is recorded as a failure rather than converted to zero or dropped. Complete-set means require all 96 cases. The fixed judge is `gpt-4o-2024-11-20`, with temperature zero and a 512-token completion limit. These are newly scored results, not scores from the original `gpt-4-0314` evaluator. Requests without an unambiguous rating are first retried with the unchanged prompt and judge configuration; the first valid rating is retained, irrespective of its value. Two Qwen DeCo responses remained conditional after this retry. For these, the same judge and official rubric received a format-only instruction to choose one final rating from the supplied annotations; both returned a unique rating of 2. No score is imputed.

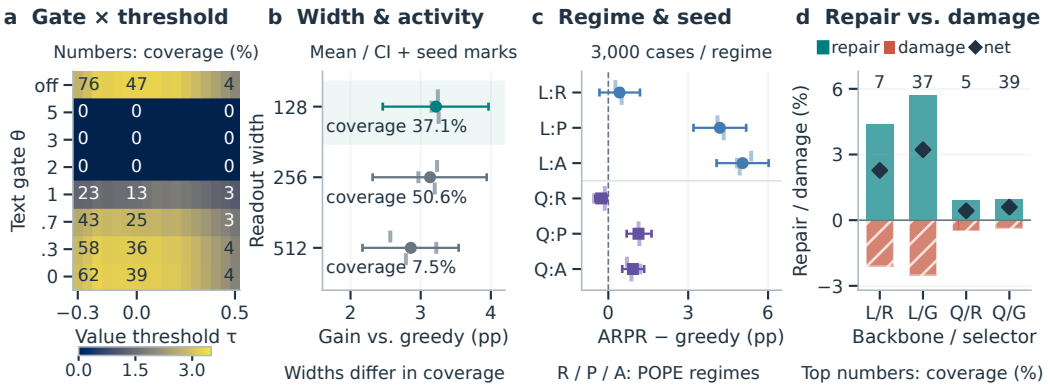


Figure 5: **Sensitivity and repair costs.** **(a)** Gate gain/coverage (off: no gate). **(b)** Width policies with paired intervals. **(c)** POPE split gains. **(d)** Repairs/damage; diamonds mark net gain. L/Q: LLaVA/Qwen; R/G: risk/ARPR.

Table 5: **Complete internal-control accounting.** POPE and AMBER accuracies are percentages. AMBER BA uses attributes+relations only; "All" and invalid-answer rate (Inv.) use all 14,216 queries. MMHal uses independently trained caption-reward policies and the masked action, not the binary transfer policy. Learned rows average three seeds; no open-response risk policy was trained. LLaVA hidden/risk widths are 256; ARPR and all Qwen binary heads use 128. These rows are component controls, not additional literature baselines.

	POPE	AMBER-D			MMHal	
Control	Mean acc.	All acc.	BA	Inv.	Score	HR
LLaVA-1.5-7B						
Original	85.23	71.64	71.77	0.08	2.021	0.667
Always act	70.86	69.06	55.10	7.59	1.927	0.667
Risk selector	87.51	72.95	72.17	0.57	—	—
Hidden value	88.49	73.90	72.43	0.84	2.038	0.660
ARPR	88.83	77.57	75.85	0.50	2.312	0.602
Qwen3.5-9B						
Original	89.89	84.46	81.38	7.27	3.406	0.573
Always act	89.51	83.60	79.19	9.65	3.417	0.583
Risk selector	90.31	84.61	81.48	7.22	—	—
Hidden value	90.50	84.44	81.20	7.50	3.438	0.569
ARPR	90.90	86.48	83.20	5.85	3.924	0.429

Table 6: Verified published context on LLaVA-1.5. These numbers are quoted from the indicated papers, not generated by our code. POPE F1 entries are reported aggregate values, without inventing unreported split-level scores. Decoding and output budgets differ from our unified setting.

Benchmark	Method	Metric 1	Metric 2	Source / setting
POPE	Greedy	82.2	–	DeCo Table 3, greedy
POPE	DoLa	83.2	–	DeCo Table 3, greedy
POPE	DeCo	86.7	–	DeCo Table 3, greedy
POPE	OPERA	85.4	–	DeCo Table 3, beam
POPE	VCD	83.1	–	DeCo Table 3, nucleus
MMHal	Greedy	1.59	0.72	CEI Table 3, 512 tokens
MMHal	VCD	1.96	0.64	CEI Table 3, 512 tokens
MMHal	Dynamic CEI	2.42	0.57	CEI Table 3, 512 tokens
MMHal	OPERA	2.41	0.57	CEI Table 3, 512 tokens
MMHal	M3ID	2.14	0.61	CEI Table 3, 512 tokens
MMHal	AvisC	2.19	0.59	CEI Table 3, 512 tokens
MMHal	CAAC	1.67	0.64	CEI Table 3, 512 tokens
MMHal	Static CEI	2.46	0.56	CEI Table 3, 512 tokens

POPE: Metric 1 is F1 (%), Metric 2 is unreported here. MMHal: Metric 1 is score, Metric 2 is hallucination rate. Sources: Wang et al. (2025) and Fazli et al. (2026). AIR's reported benchmarks include POPE, CHAIR and MM-Vet; we do not invent a published MMHal value.

In Figure 5a–b, widths 128/256 yield gains of 3.59/3.47 pp and coverages of 38.5/50.6%. These are complete development-selected policies, not a width comparison at equal coverage. Gates $\theta \geq 2$ disable every action in the displayed grid.

Additional diagnostics. Figures 8–10 report representation-specific risk and reward supervision, repair/damage curves, memory stress and seed-specific MMHal changes. Representation parameter counts (millions) are 0.05/3.43/3.49/3.49/6.87/6.77 for gap, hidden, memory-only, residual, hybrid and wide-hidden at width 256. The selected width-128 hybrid has 4.94M. Intervals condition on three fitted models.

The additional Qwen memory test scores all 5,171 distinct POPE queries with the three frozen selected checkpoints. A cyclic next-image permutation replaces the entire memory and its padding mask; zeroing preserves the original length. Intact rescoring reproduces cached decisions exactly. The decoder hidden state, logit features, action outcomes and thresholds are fixed throughout.

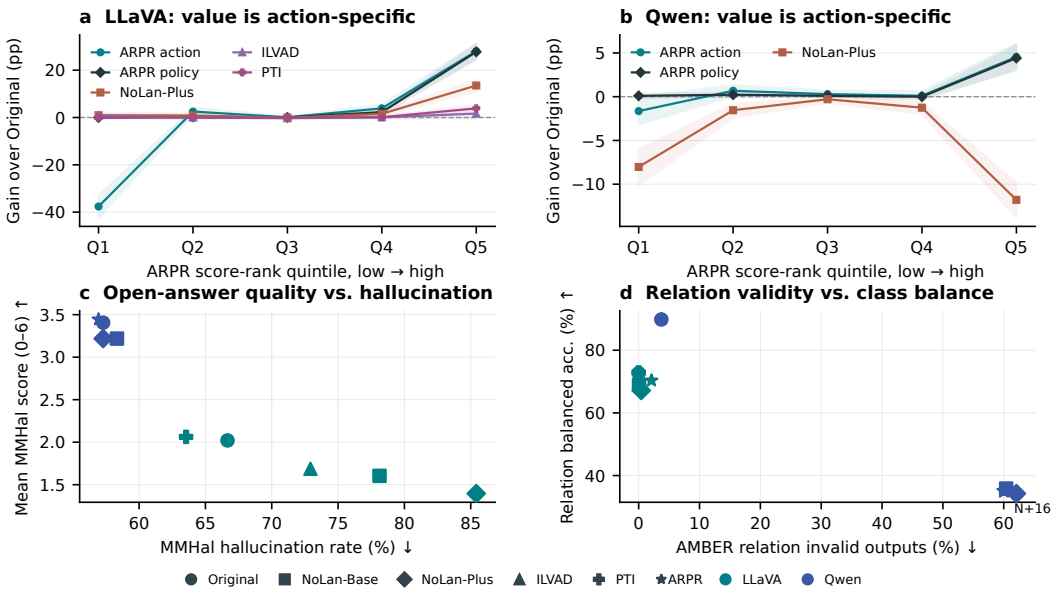


Figure 6: **Action specificity and format robustness against recent baselines. (a–b)** Effects relative to greedy by frozen ARPR score-rank quintile on 5,171 unique POPE queries, separately for LLaVA and Qwen. Bands resample images; ranks are descriptive and never set a test threshold. The ARPR fixed action, its selected policy, NoLan-Plus, ILVAD and PTI are distinct operators. **(c)** Mean MMHal score versus hallucination rate on all 96 cases. **(d)** AMBER relation balanced accuracy versus invalid-response rate under the common three-token limit; the hollow Qwen NoLan-Plus marker shows the separately labeled exploratory 16-token diagnostic. NoLan-Base is shown as an additional frozen variant.

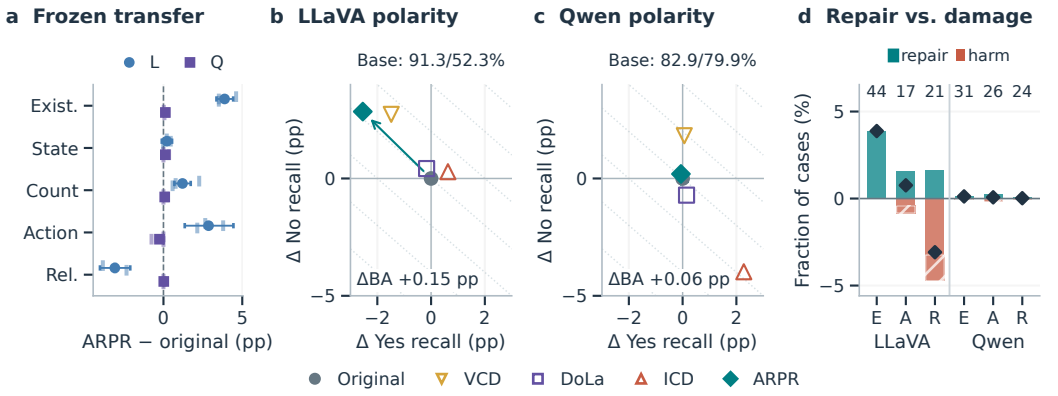


Figure 7: **Frozen AMBER transfer. (a)** ARPR gains: paired 95% image intervals and seed ticks (L/Q). **(b–c)** Recall changes over 9,292 attribute/relation cases; subtitles: Original recalls, arrows: ARPR, diagonals: 1-pp BA change. Shared baselines only. **(d)** Repairs/damage over full E/A/R subsets; diamonds: net gain, numbers: coverage, hatching: invalid outputs. Means include all three frozen policies; invalid answers remain errors.

Table 7: **Polarity and output validity for recent baselines.** POPE pools all 9,000 questions (Acc., yes/no recall in %); AMBER A/R BA is balanced accuracy on 9,292 attribute/relation cases, and R invalid is the unparsed-response rate on 1,664 relation cases. MMHal scores all 96 questions; HR is the rate with judge score below 3. These columns expose format and class trade-offs hidden by overall accuracy.

	POPE			AMBER-D		MMHal	
Method	Acc.	Yes rec.	No rec.	A/R BA	R invalid	Score	HR ↓
LLaVA-1.5-7B							
Original	85.23	90.73	79.73	71.77	0.00	2.021	66.7
NoLan-Base	86.91	89.11	84.71	72.76	0.06	1.604	78.1
NoLan-Plus	87.11	88.84	85.38	70.68	0.42	1.396	85.4
ILVAD	85.82	91.13	80.51	71.13	0.06	1.688	72.9
PTI	85.56	89.73	81.38	71.94	0.00	2.062	63.5
ARPR (ours)	89.83	87.45	92.20	75.85	1.50	2.312	60.2
Qwen3.5-9B							
Original	89.89	86.73	93.04	81.38	3.73	3.406	57.3
NoLan-Base	86.12	85.58	86.67	65.55	60.40	3.219	58.3
NoLan-Plus	86.21	85.78	86.64	65.04	62.08	3.219	57.3
ARPR (ours)	91.59	87.32	95.86	83.20	2.50	3.924	42.9

Unlike LLaVA's fixed-length mixtures, variable-length Qwen memory is tested at the intact and full-perturbation endpoints only. This is an incremental readout diagnostic, not image-corruption robustness of the backbone. Additional split and outcome-retention analyses use saved predictions and do not retrain or retune any policy.

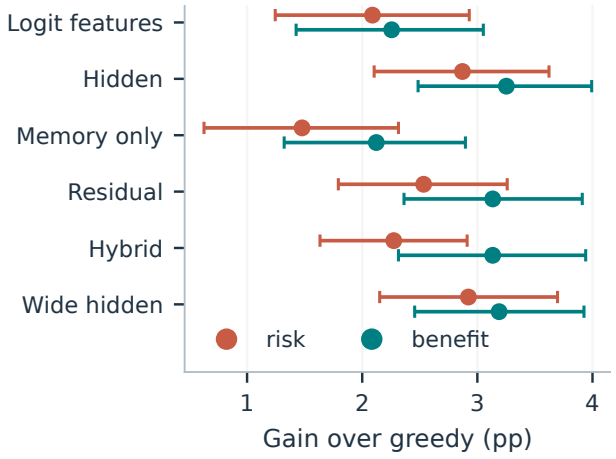


Figure 8: **Representation-specific risk and reward supervision on LLaVA.** Width-256 policies share the same correction action and training budget. Points show gain over greedy decoding; whiskers are paired image-bootstrap 95% intervals. Parameter counts are given in the text.

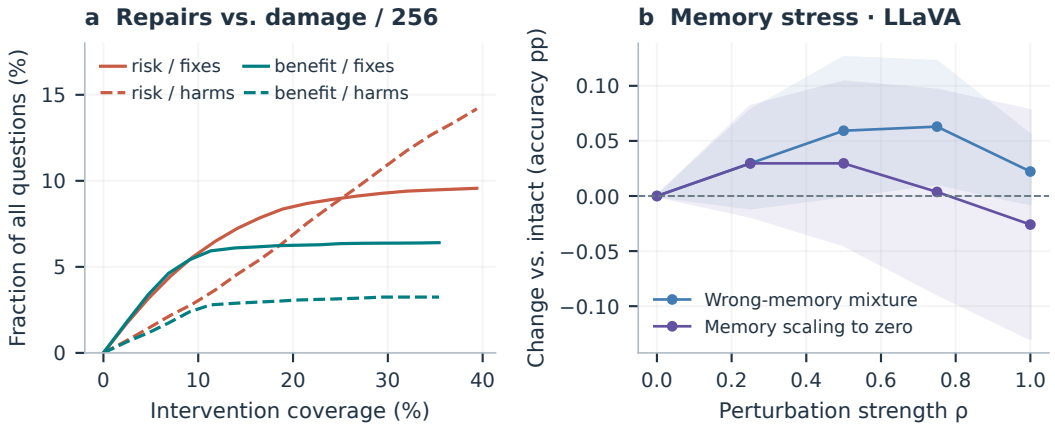


Figure 9: **Damage accounting and partial memory perturbations.** **(a)** Solid/dashed curves show repairs/damage, respectively, under matched width-256 risk/benefit ranking. **(b)** LLaVA uses $M_\rho = (1-\rho)M + \rho M_{\text{wrong}}$ or $(1-\rho)M$; bands are paired image-bootstrap intervals. Hidden states and logit features remain intact.

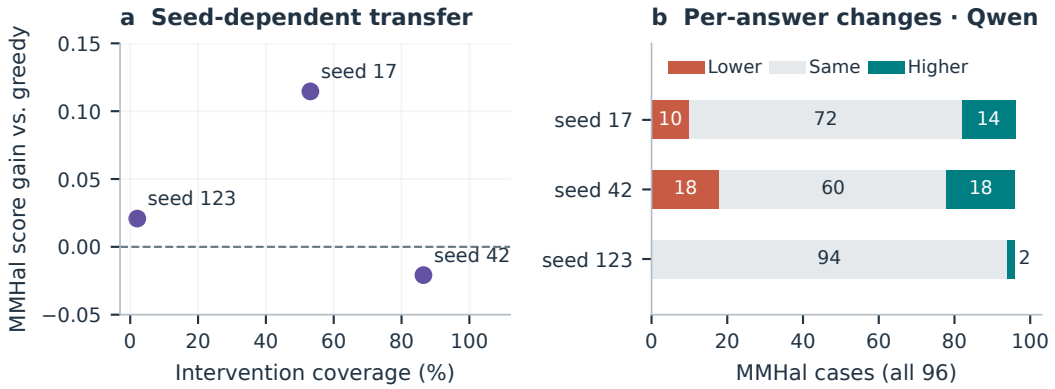


Figure 10: **Qwen MMHal transfer is seed-dependent.** **(a)** Each point is a fixed development-selected seed policy, not a sensitivity sweep. **(b)** Case counts whose cached judge score rises, falls or stays unchanged relative to greedy. All 96 cases remain in every row. No best seed is selected.

Risk-conditioned outcomes and evaluation units. Figure 11 crosses three predicted error-risk tertiles with benefit-score quartiles defined within each tertile. Both scores are averages from frozen, matched width-256 assessors; bins are set before consulting correction outcomes. The matrix separates high-risk questions for which the same fixed action is beneficial from those for which it is harmful. It is a descriptive stratification, not a policy. Budget comparisons rank 5,171 distinct POPE questions and reuse each action across its occurrences in the 9,000 official cases; equal unique-query budgets can therefore have slightly different case-level coverage. Paired intervals use 5,000 image-cluster draws. Image-level difficulty strata count Original's errors among the 18 questions for each test image; these strata do not gate any method.

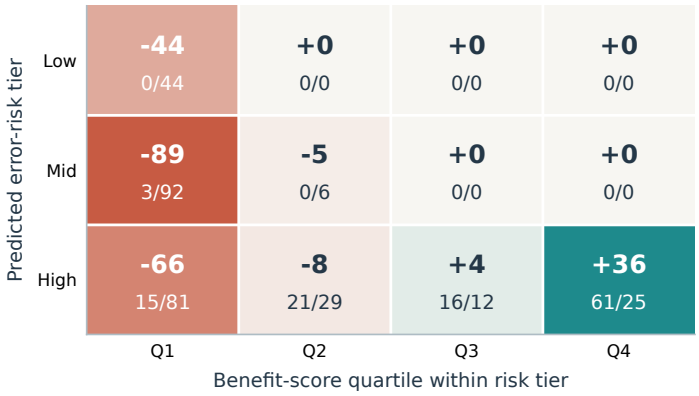


Figure 11: **Conditional fixed-action outcomes within risk tiers.** Within each predicted error-risk tier, cases are split by benefit-score quartile; each cell has about 750 official POPE cases. Large numbers show net gain in percentage points, small numbers repair/damage rates. These bins describe scores; they do not select an operating threshold.

Scope of the visualizations. These diagnostics use 5,000 image-cluster draws. Budget allocation uses frozen scores and a stable ordering of unique queries, not test correctness. Score-quantile bins are descriptive and do not select thresholds. Neither score correlation nor visual separation alone establishes a causal contribution from memory.

F Qualitative Case Studies

F.1 POPE: Repairing Binary Errors Without Forcing Correction

These outcome-selected POPE cases illustrate three repairs (A–C) and one damage-avoiding decision (D), not a representative success rate. Images, questions and labels are original; all three seeds make the shown decision. NoLan-Plus answers are replayed under the common three-token decoding setting.

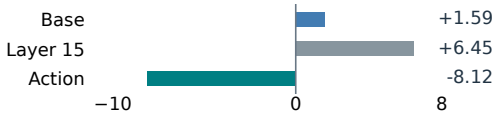
A. Repair: false dog presence
LLaVA-1.5-7B; COCO 196053



Q: Is there a dog in the image?

Reference	**No**
Greedy	`Yes, there`
VCD / DeCo	`Yes, there`
CEI / AIR	`Yes, there`
NoLan-Plus	`Yes, the`
ILVAD	`Yes, there`
ARPR (all seeds)	**`No, there`**

PRM / seed	17	42	123
$\hat{u}$	+0.364	+0.426	+0.002
τ	+0.01	+0.14	-0.06
Decision	Act	Act	Act



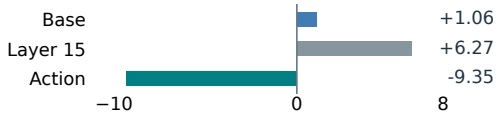
B. Repair: false horse presence
LLaVA-1.5-7B; COCO 493753



Q: Is there a horse in the image?

Reference	**No**
Greedy	`Yes, there`
VCD / DeCo	`Yes, there`
CEI / AIR	`Yes, there`
NoLan-Plus	`Yes, there`
ILVAD	`No`
ARPR (all seeds)	**`No, there`**

PRM / seed	17	42	123
$\hat{u}$	+0.238	+0.292	+0.030
τ	+0.01	+0.14	-0.06
Decision	Act	Act	Act



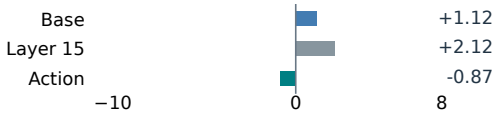
C. Repair: false chair presence
Qwen3.5-9B; COCO 482275



Q: Is there a chair in the image?

Reference	**No**
Greedy	`Yes, there`
VCD	`Yes, there`
NoLan-Plus	`Looking at the` (invalid)
Fixed action	`No, there`
ARPR (all seeds)	**`No, there`**

PRM / seed	17	42	123
$\hat{u}$	+0.093	+0.249	+0.168
τ	+0.04	-0.09	+0.01
Decision	Act	Act	Act



D. Preserve: chair is present
LLaVA-1.5-7B; COCO 12333



Q: Is there a chair in the image?

Reference	**Yes**
Greedy	`Yes, there`
Fixed action	`No, there`
NoLan-Plus	`Yes, there`
Action reward	$\Delta = -1$ (damage)
ARPR (all seeds)	**`Yes, there`**

PRM / seed	17	42	123
$\hat{u}$	-0.099	-0.118	-0.284
τ	+0.01	+0.14	-0.06
Decision	Keep	Keep	Keep

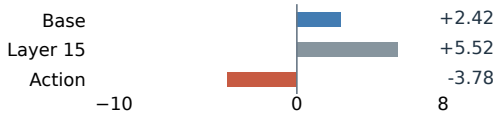


Figure 12: **Original images and replay-backed decisions.** Exact three-token repairs (A–C) and one withheld harmful action (D); ILVAD also repairs B. Mini-plots show pre/post-action margins.

F.2 MMHAL: FREE-FORM CORRECTIONS AND RATING AMBIGUITY

Unlike POPE's binary answers, MMHal asks for open-ended text. These Qwen3.5-9B cases were selected *after* inspecting the 96-case outcomes: they illustrate individual changes, not their frequency. Images and questions come from the released benchmark. Model text below is excerpted; Markdown emphasis is removed, and complete saved answers are retained in the audit. Parenthesized numbers are fixed MMHal judge ratings (0–6), not PRM scores; ratings below 3 count as hallucinated. The text changes below use Qwen. The selected hybrid PRM has development-set thresholds $(\tau_{17}, \tau_{42}, \tau_{123}) = (-0.01, -0.02, -0.01)$.

E. Correcting a utensil identity
MMHal-050; comparison



Q. Compare the utensils in the picture.
Reference. Blue-handled knife, yellow-handled spoon, red-handled fork (metal heads, plastic handles).
Greedy (1). *a spoon, a fork, and another spoon (or possibly a larger spoon)*
VCD (1). *fork, spoon, and what appears to be a spork or hybrid fork-spoon*
ARPR 17/42 (2/2). *a knife, spoon, and fork*; knife blue, spoon yellow, fork red.
Change. The left utensil is recognized as a knife, matching the visible blade and handle-color ordering. Seed 123 keeps Greedy (1).

Seed	17	42	123
$\hat{u}$	+0.001	+0.022	-0.031
Decision	Act	Act	Keep

F. Rejecting visible hands
MMHal-057; adversarial



Q. Describe the hands wearing the gloves.
Reference. Only a pair of black gloves is visible; no hands are shown.
Greedy (1). *gloves worn on hands*; *The hands are open and slightly spread apart*
VCD (0). *The palms and backs of the hands are smooth and untextured*
ARPR 17/42 (2/2). *a pair of black gloves, not hands*; *the hands themselves are not visible*.
Change. The answer rejects visible hands, yet still says the gloves are *worn on hands*; this is a partial, internally imperfect correction. Seed 123 keeps Greedy (1).

Seed	17	42	123
$\hat{u}$	+0.005	+0.022	-0.027
Decision	Act	Act	Keep

Figure 13: **MMHal text changes with visible grounding cues.** Original benchmark images and questions are shown. E corrects an object identity; F removes an unsupported visible-hands claim but leaves an inconsistency. VCD and Greedy are saved Qwen outputs under the same open answer setting. Seeds 17 and 42 act, whereas seed 123 preserves Greedy. Both selected ratings remain below the benchmark's non-hallucinated cutoff of 3. The cases do not isolate a separate causal contribution from patch memory.

G. Avoiding an unsupported weather guess
MMHal-077; environment



Q. What was the weather like when this photo was taken?
Reference. When this photo was taken, it was already night and the sky was dark.
Greedy (1). *cold and possibly snowy or overcast*
VCD (1). *the weather was cold or chilly*
ARPR 17/42 (6/6). *cannot determine the weather* from the indoor image; the answer also notes darkness outside the windows.
Change. The answer stops inferring temperature or precipitation from furnishings. Its high judge rating does not mean it exactly matches the reference's nighttime description. Seed 123 keeps Greedy (1).

Seed	17	42	123
$\hat{u}$	-0.008	+0.024	-0.040
Decision	Act	Act	Keep

H. Reference–rating ambiguity
MMHal-095; other



Q. What is shown in the mirror?
Official reference. The mirrow shows the reflection of the washing machine. (Original spelling retained.)
Greedy (2). *reflection of the stacked washer and dryer*
VCD (1). *reflection of the stacked washer and dryer*
ARPR 17/42 (1/1). *reflects the dryer*; it identifies the upper unit in the reflection.
Limit. The mirror visibly reflects the upper machine, but the reference names the washing machine. The judge's drop from 2 to 1 cannot by itself decide if visual grounding was damaged. Seed 123 keeps Greedy (2).

Seed	17	42	123
$\hat{u}$	+0.014	+0.045	-0.018
Decision	Act	Act	Keep



I. Removing invented tableware *(MMHal-049; adversarial)*
Q. How many people are eating in this kitchen?
Reference (condensed). No one is eating; the image shows a dining table, four chairs and no people.
Greedy (2). *There are no people eating*; also asserts *plates, cups, and cutlery arranged on the table*.
VCD (6). Also answers *0*, but claims *plates and cups* on the table.
ARPR, seed 42 (6). *There are no people eating*; describes the table as *covered with a tablecloth*, without claiming tableware.
What changes. Both answers give the correct count, 0. The visible table has a cloth but no discernible plates, cups or cutlery. Seed 42 acts ($\hat{u} = +0.014 > \tau_{42} = -0.02$); seeds 17/123 keep Greedy ($\hat{u} = -0.025/-0.037$, both rated 2). VCD's 6 despite invented tableware shows that the judge rating does not catch every detail.

Figure 14: **Free-form changes and evaluation limits.** G is a cautious response to a weather question whose released reference describes nighttime rather than weather. H exposes a possible conflict between the reference and the visible reflection; it is not counted as a verified repair. I keeps the correct zero-person answer while removing Greedy's invented plates, cups and cutlery; only seed 42 acts. VCD also invents tableware despite its 6 rating. Ratings use the same fixed cached judge. These outcome-selected examples do not estimate the frequency of each correction pattern across MMHal.