BEYOND AVERAGE GAINS: REPLAY-BASED POST-INTERVENTION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free mitigation methods can reduce multimodal hallucinations on average, but fixed, always-on use still leaves a practical problem: an intervention can improve aggregate performance while turning some originally correct outputs into wrong ones. We study this as post-intervention risk control, where the goal is to identify harmful method-induced changes after generation and apply the smallest possible correction. We find that standard generation-time signals, including confidence, margin, entropy, selective prediction, and self-consistency, provide little information about whether an intervention was warranted. In contrast, replaying the committed intervention output through the unmodified frozen backbone exposes a substantially stronger signal for separating harmful from helpful changes. Based on this observation, we propose RAPIC, a lightweight replay-based control layer that selectively retains or reverts intervention outputs. In our primary LLaVA-1.5 + VGA setting, RAPIC improves accuracy by +1.21, +1.61, and +0.67 points on MSCOCO, AOKVQA, and GQA, respectively, with paired image-clustered bootstrap confidence intervals excluding zero. Repeated calibration shows that changed-answer labeling is more efficient than uniform sampling.

1 INTRODUCTION

Multimodal large language models (MLLMs) still produce outputs that do not match the image in both generative and discriminative settings (Rohrbach et al., 2018; Li et al., 2023; Liu et al., 2024a). This has led to many training-free mitigation methods that modify inference without retraining, including attention guidance (Liu et al., 2024d; Zhao et al., 2026), contrastive decoding (Leng et al., 2024), and related decoding-time control methods (Huang et al., 2024). These approaches are attractive because they can improve hallucination-related performance while leaving the underlying model unchanged.

But average benchmark gains do not mean that a fixed mitigation method is reliable on every input. A method can improve aggregate performance and still turn some baseline answers that were originally correct into wrong ones. In answer-level settings, this risk is concentrated in cases where the intervention changes the baseline prediction: some changes correct baseline errors, while others corrupt answers that were already correct. Average accuracy reports only the balance between these two effects and can therefore hide a non-trivial tail of harmful replacements.

We study this gap as *post-intervention risk control*: after a fixed mitigation method has acted, can we identify which of its output changes should be trusted? A natural approach is to use uncertainty signals available during generation. However, we find that confidence, margin, entropy, selective prediction, and sampling-based self-consistency provide little useful separation between harmful and helpful intervention-induced changes. These signals describe how confidently the intervened model produces an answer, but not whether the intervention itself was warranted. In contrast, replaying the committed intervention output through the original frozen backbone exposes substantially stronger structure: harmful changes tend to occupy a distinct high-risk region in replay-consistency space. This suggests that useful information about intervention risk can be more readily exposed once the intervention has committed to an output.

Based on this observation, we introduce **RAPIC** (**R**isk-**A**ware **P**ost-**I**ntervention **C**ontrol), a lightweight control layer over fixed hallucination mitigation methods. RAPIC focuses on method-induced changed outputs, replays the committed intervention answer through the same frozen back-

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.

bone, and uses replay-consistency signals to decide whether to retain the intervention output or fall back to the baseline answer. The routing policy is calibrated on a small labeled discovery set and then frozen; no labels are used at application time. Importantly, our experiments indicate that the main benefit comes from the replay information itself rather than from a particular fusion rule or learned router.

Our strongest evidence is in discrete-answer multimodal settings, where changed outputs and conservative fallback are both well defined. Across POPE benchmarks, mitigation methods, and VLM backbones, RAPIC reduces intervention-induced harm while preserving useful corrections in its primary settings. We further analyze statistical uncertainty, calibration sensitivity, stronger routing baselines, and cases where replay separation is weak. We therefore view RAPIC not as a universal wrapper that improves every intervention, but as a concrete instance of a broader principle: method-induced output changes should themselves be treated as risk-bearing decisions that can be audited after generation.

Our contributions are summarized as follows:

- We formulate *post-intervention risk control* as a distinct reliability problem, showing that aggregate mitigation gains can conceal harmful method-induced output changes.
- We show that conventional generation-time uncertainty provides limited control over these changes, whereas replaying the committed output through the frozen backbone exposes substantially stronger harmful-versus-helpful separation.
- We introduce RAPIC, a lightweight replay-based controller, and characterize when it succeeds or abstains across datasets, backbones, and mitigation methods, together with statistical and calibration analyses.

2 Related Work

Multimodal Hallucination and Evaluation. Multimodal hallucination has been studied through object-level captioning metrics such as CHAIR (Rohrbach et al., 2018), binary object-existence evaluation with POPE (Li et al., 2023), and broader benchmarks including MMHal-Bench (Sun et al., 2024), AMBER (Wang et al., 2023), and THRONE (Kaul et al., 2024). Later work further considers open-set evaluation and broader hallucination taxonomies (Matsuda et al., 2024; Tu et al., 2025; Liu et al., 2025; 2024a). While these benchmarks primarily summarize dataset-level behavior, we study the residual sample-level risk after a mitigation method has already acted.

Inference-Time Hallucination Mitigation. Hallucination can be reduced through additional training (Liu et al., 2023; Sun et al., 2024; Yu et al., 2025; Sarkar et al., 2024; Xiao et al., 2025; Chen et al., 2025; Yoon et al., 2025; Yang et al., 2025) or by modifying inference without parameter updates. Training-free methods include decoding control (Huang et al., 2024; Leng et al., 2024), attention-guided intervention (Liu et al., 2024d; Zhao et al., 2026; An et al., 2025), and related signal-amplification or routing methods (Yin et al., 2025; Wang et al., 2025; Che et al., 2025; Park et al., 2025; Zhu et al., 2025; Zou et al., 2024; Dong et al., 2025). RAPIC assumes such an intervention is fixed and instead controls the output changes it induces.

Post-Hoc Correction and Selective Risk Control. Post-hoc methods verify or modify an already generated response, including Woodpecker (Yin et al., 2024) and related intervention or hallucination detection approaches (Peng et al., 2025; Prabhakaran et al., 2025; Zhou et al., 2023; Zhao et al., 2024). Selective prediction instead studies abstention or deferral under uncertainty (Geifman & El-Yaniv, 2017; 2019; Xin et al., 2021; Yoshikawa & Okazaki, 2023; Srinivasan et al., 2024; Kogilathota et al., 2026). RAPIC differs from existing methods by restricting control to intervention-induced changes and using same-backbone replay to choose between retaining the intervention output and reverting to the baseline.

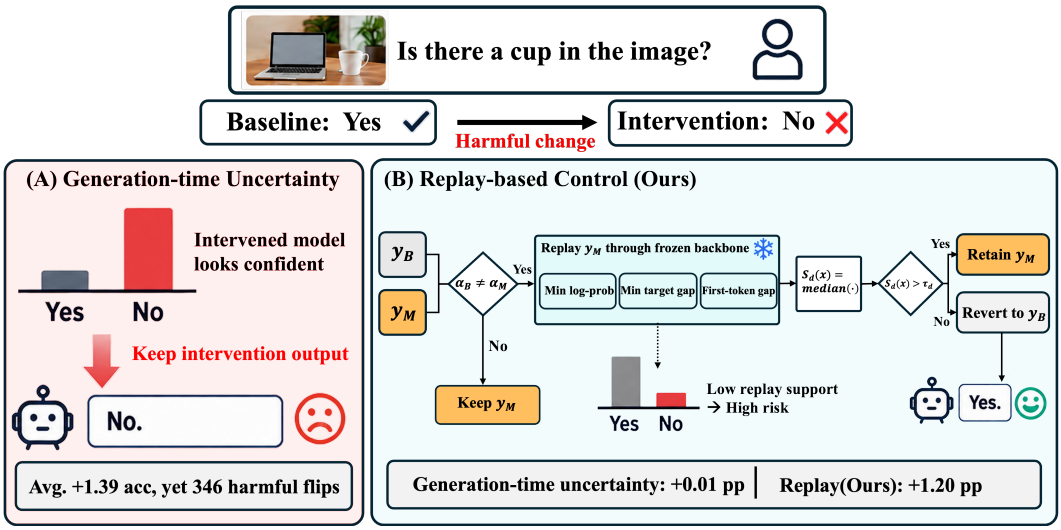


Figure 1: **Overview of RAPIC. (a)** A fixed mitigation method can introduce both helpful and harmful output changes, so aggregate gains may still conceal sample-level regressions. **(b)** RAPIC operates only on intervention-induced changes, replays the committed intervention output through the frozen backbone, and uses replay-consistency features to decide whether to retain the intervention output or revert to the baseline output. The aggregate statistic in (a) is computed on the full MSCOCO set, whereas the routing comparison in (b) is reported on the matched held-out subset.

3 Proposed Method

We study post-intervention control over a fixed hallucination mitigation method. Given a baseline output and an intervention-modified output, our goal is to act only on intervention-induced changes and to apply the smallest available correction when risk is high. Figure 1 shows the overall pipeline.

3.1 Post-Intervention Risk Formulation

Training-free mitigation methods are typically applied as fixed inference-time rules, but their sample-level risk is concentrated in the cases where the intervention changes the baseline decision. Let $x = (I, q)$ denote an image–question pair, and let $y_B(x)$ and $y_M(x)$ denote the outputs of the baseline model and a fixed mitigation method, respectively. After deterministic answer parsing, we write the corresponding predictions as $a_B(x)$ and $a_M(x)$. We define a *route candidate* as any example for which

$$a_B(x) \neq a_M(x). \tag{1}$$

If the two predictions agree, the intervention has not changed the answer and no routing decision is required.

On a labeled discovery set, a changed answer is *helpful* when the baseline is wrong and the intervention is correct, and *harmful* when the baseline is correct and the intervention is wrong. For a fallback set S of changed candidates, let $H(S)$ and $G(S)$ denote the numbers of harmful and helpful changes selected for fallback. Replacing the intervention output with the baseline output on S changes benchmark accuracy by

$$\Delta_{\text{fb}}(S) = \frac{H(S) - G(S)}{N}, \tag{2}$$

where N is the number of evaluation examples. Thus, selective fallback is useful only when harmful changes are recovered faster than helpful corrections are discarded.

At application time, these labels are unavailable. The goal of RAPIC is therefore to estimate which changed outputs are risky using only information available from the frozen model, the multimodal input, and the committed intervention output. Rather than predicting correctness for every sample,

RAPIC addresses the narrower decision induced by the intervention itself: whether a changed output should be retained or conservatively reverted to the baseline.

3.2 Replay-Consistency Routing

For each route candidate, RAPIC re-evaluates the committed intervention answer under the original frozen backbone. Let

$$y_M = (w_1, \ldots, w_T) \tag{3}$$

denote the intervention output, and let $\mathcal{P} \subseteq \{1, \ldots, T\}$ denote its content-token positions. Under teacher forcing, the backbone assigns each target token w_t a conditional probability

$$p_B(w_t \mid I, q, w_{<t}), \tag{4}$$

together with a target logit $z_t(w_t)$ and logits for alternative tokens. RAPIC extracts three replay-consistency signals:

$$\begin{aligned} f_{\mathrm{LP}} &= \min_{t \in \mathcal{P}} \log p_B(w_t \mid I, q, w_{<t}), \\ f_{\mathrm{gap}} &= \min_{t \in \mathcal{P}} \left[z_t(w_t) - \max_{v \neq w_t} z_t(v) \right], \\ f_{\mathrm{first}} &= z_1(w_1) - \max_{v \neq w_1} z_1(v). \end{aligned} \tag{5}$$

The first signal measures the weakest token-level support for the committed answer, while the latter two measure its margin over the strongest competing token. Low support or a small target gap indicates that the intervention output is less consistent with the original model when replayed against the same multimodal input.

We calibrate the two transition directions, $a_B = \text{yes} \to a_M = \text{no}$ and $a_B = \text{no} \to a_M = \text{yes}$, separately because their harmful–helpful composition and replay-score distributions can differ substantially. For each direction d, the three replay signals are oriented so that larger values indicate greater harmful-change risk and standardized using discovery-set statistics. RAPIC then forms a scalar score

$$s_d(x) = \mathrm{median}\left(u_{1,d}(x), u_{2,d}(x), u_{3,d}(x)\right), \tag{6}$$

where $u_{k,d}$ denotes the oriented and standardized replay feature.

Given a frozen threshold τ_d for transition direction d, the final routing rule is

$$y_R(x) = \begin{cases} y_B(x), & a_B(x) \neq a_M(x) \ \wedge \ s_d(x) \geq \tau_d, \\ y_M(x), & \text{otherwise.} \end{cases} \tag{7}$$

If no fallback threshold has positive utility on the discovery set for a direction, RAPIC selects NOOP and retains the intervention output for that direction. Thus, replay determines the ordering of risky changed outputs, while discovery calibration determines whether and where that ordering should trigger conservative fallback.

3.3 Discovery Calibration and Frozen Application

RAPIC uses a small labeled discovery set only to calibrate the routing policy. In practice, we first run the baseline and mitigation method without labels, collect examples whose parsed answers differ, and annotate only a subset of these changed candidates. This focuses supervision on the samples for which a routing decision is actually required.

For each transition direction d, the replay features are oriented and standardized using the corresponding discovery candidates. We then sweep thresholds over the discovery score distribution. Let

$$S_d(\tau) = \{x : s_d(x) \geq \tau\} \tag{8}$$

denote the candidates selected for fallback at threshold τ. Using the fallback utility defined in Sec. 3.1, RAPIC selects

$$\tau_d^* = \arg\max_{\tau \in \mathcal{T}_d} \Delta_{\mathrm{fb}}(S_d(\tau)), \tag{9}$$

where candidate thresholds must select at least five discovery examples ($|S_d(\tau)| \geq 5$). The candidate policy set also includes NOOP, which retains all intervention outputs. If no valid threshold improves over NOOP on the discovery set, RAPIC abstains from routing in that direction.

After calibration, the feature orientations, normalization statistics, direction-specific thresholds, and any NOOP decisions are frozen. Held-out routing then uses only the image–question pair, the baseline and intervention outputs, and same-backbone replay logits; no evaluation labels are used. Thus, RAPIC requires labeled supervision during deployment calibration, but its per-example control decisions are label-free once the policy has been fixed.

3.4 Object-Level Residual Repair for Captioning

Whole-output fallback is often too coarse for captioning, where a generated sentence may contain a small number of unsupported object mentions while the remaining content is still useful. We therefore use the same post-intervention control principle as a narrower object-level repair. Given a mitigated caption $c_M(x)$, RAPIC extracts object mentions from a benchmark-aligned vocabulary and probes each object o with the same frozen model. Using the next-token logits z_{yes} and z_{no}, we define its visual support as

$$p(o \mid x) = \frac{\exp(z_{\text{yes}})}{\exp(z_{\text{yes}}) + \exp(z_{\text{no}})}. \tag{10}$$

If the least-supported mention falls below a validation-calibrated threshold, RAPIC locally suppresses that object during re-decoding while otherwise preserving the mitigation output. This component is intended as a bounded residual-repair mechanism for unsupported object mentions rather than a general caption-correction system.

4 Experiments

4.1 Experimental Setup

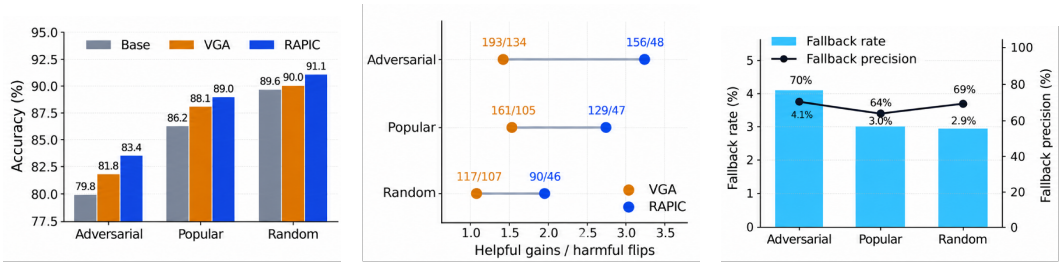
Benchmarks and models. We evaluate RAPIC primarily on POPE-style object-existence VQA (Li et al., 2023), using POPE-MSCOCO (Lin et al., 2014), POPE-AOKVQA (Schwenk et al., 2022), and POPE-GQA (Hudson & Manning, 2019). Each benchmark contains 9,000 examples divided into adversarial, popular, and random subsets. We evaluate three VLM backbones: LLaVA-1.5-7B (Liu et al., 2024b), LLaVA-NeXT Llama3-8B (Liu et al., 2024c), and Qwen2.5-VL-7B (Bai et al., 2025). VGA (Zhao et al., 2026) is our primary mitigation method, while PAI-attn (Liu et al., 2024d) and VAF (Yin et al., 2025) are used to evaluate whether the same post-intervention control procedure remains useful across different inference-time interventions.

Calibration and evaluation. For each backbone–method pair, RAPIC is calibrated on labeled changed-answer candidates from a discovery split and then frozen before held-out evaluation. Unless otherwise noted, the POPE controller is calibrated on MSCOCO discovery data and applied unchanged to the held-out benchmarks. Evaluation labels are never used for held-out routing. We report accuracy as the primary task metric together with route-level statistics that decompose intervention-induced changes into harmful and helpful cases. Full F1 results and implementation details are provided in the appendix.

Statistical evaluation. For the primary LLaVA-1.5 + VGA setting, we report paired image-clustered bootstrap confidence intervals with 10,000 replicates for the accuracy difference between RAPIC and the always-on intervention, resampling images together with all associated questions. We additionally evaluate calibration stability over repeated discovery splits and compare RAPIC with routing baselines under matched candidate sets, fallback actions, and supervision.

4.2 Main Results on POPE

We first examine the primary LLaVA-1.5 + VGA setting, where the effect of post-intervention control can be measured directly at the changed-answer level. Figure 2(a) shows that RAPIC improves accuracy over the always-on intervention across all three POPE-MSCOCO splits. Figure 2(b) shows a corresponding increase in the gain-to-harm ratio, indicating that useful corrections increasingly



(a) Split-wise accuracy (b) Gain-to-harm ratio (c) Fallback precision and rate

Figure 2: **Post-intervention control on POPE-MSCOCO with LLaVA-1.5 + VGA.** RAPIC improves split-wise accuracy and the gain-to-harm ratio while maintaining selective fallback with relatively high precision.

Table 1: Primary LLaVA-1.5 + VGA results on POPE. Δ is RAPIC$-$VGA accuracy with image-clustered 95% CIs; Changed/Sel. H/G denote all/reverted harmful/helpful changes.

Dataset	Base	VGA	+RAPIC	Δ Acc. [95% CI]	+RAPIC F1	Changed H/G	Sel. H/G/Net
MSCOCO	85.22	86.61	87.82	$+1.21$ $[0.63, 1.80]$	87.81	346/471	205/96/$+109$
AOKVQA	78.98	81.04	82.66	$+1.61$ $[1.09, 2.17]$	84.58	319/505	200/55/$+145$
GQA	76.61	80.08	80.74	$+0.67$ $[0.21, 1.13]$	83.10	257/569	126/66/$+60$

outweigh harmful intervention-induced changes. Fallback remains selective (Figure 2(c)): only 2.9–4.1% of all examples are reverted, with fallback precision between 64% and 70%.

Table 1 reports the corresponding full-benchmark results with statistical uncertainty. RAPIC improves VGA by +1.21, +1.61, and +0.67 percentage points on MSCOCO, AOKVQA, and GQA, respectively, with paired image-clustered bootstrap 95% confidence intervals excluding zero in all three cases. The route-level decomposition further shows that these gains relate to positive net recovery of intervention-induced errors, supporting the same behavioral pattern observed in Figure 2.

4.3 Across Backbones and Mitigation Methods

We next test whether post-intervention control is specific to a single backbone or mitigation method. Table 2 reports all evaluated backbone–method–dataset combinations. RAPIC should not be interpreted as a wrapper that necessarily improves every intervention. Instead, the results reveal three recurring behaviors. When replay-consistency separates harmful from helpful changes, RAPIC performs selective correction. When discovery calibration finds no useful fallback region, the policy remains NOOP. When an intervention substantially degrades the baseline, the same mechanism can approach broad rollback.

The strongest and most consistent results occur for LLaVA-1.5, where all VGA, PAI-attn, and VAF settings improve after control. LLaVA-NeXT also benefits from VGA, while PAI-attn resolves to NOOP and VAF enters a rollback-dominant regime that recovers much of the intervention damage. Qwen2.5-VL exhibits weaker separation: RAPIC improves several settings but produces small negative changes in a few cases. These failures are small in magnitude but important, suggesting that replay-consistency is most useful when the underlying harmful and helpful populations are sufficiently separable.

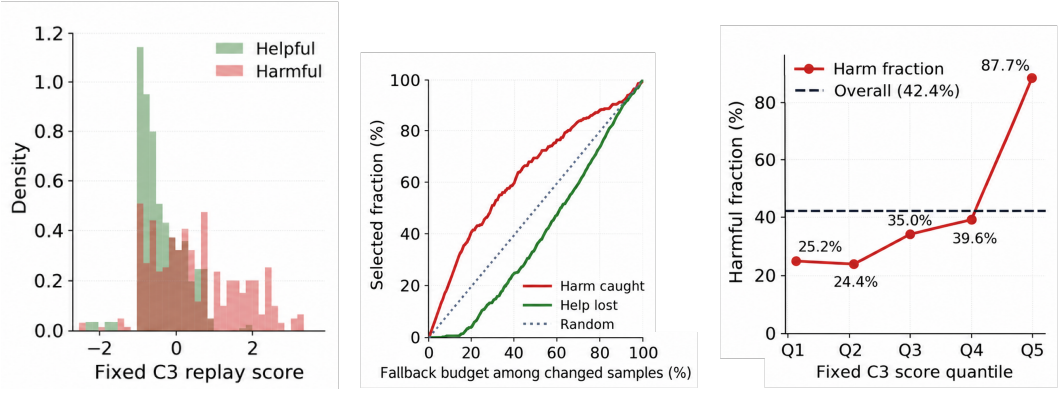
4.4 Why Replay Rather Than Generic Uncertainty?

The gains above depend on whether harmful intervention-induced changes can be distinguished from helpful corrections after generation. Figure 3(a) shows a distributional shift between helpful and harmful changes in replay-consistency space. Figure 3(b) shows that ranking changed candidates by replay score catches harmful changes faster than random selection while sacrificing helpful changes

Table 2: Post-intervention control across POPE settings. Harm is the harmful-change rate; Δ_R is RAPIC−intervention accuracy, with positive values in bold.

Method	Backbone	Data	Changed	Int. H/G/Gain	Harm (%)		Fb.%	Fb.Prec.	Δ_R
					Before	After			
VGA	LLaVA-1.5	MCO	817	346/471/+125	3.84	1.57	3.34	68.1	+1.21
		AOK	824	319/505/+186	3.54	1.32	2.83	78.4	+1.61
		GQA	826	257/569/+312	2.86	1.46	2.13	65.6	+0.67
	LLaVA-NeXT	MCO	330	142/188/+46	1.58	1.41	0.20	83.3	+0.13
		AOK	393	164/229/+65	1.82	1.32	0.56	90.0	+0.44
		GQA	492	178/314/+136	1.98	1.31	0.76	88.2	+0.58
	Qwen2.5-VL	MCO	135	107/28/−79	1.19	0.07	1.40	80.2	+0.84
		AOK	191	101/90/−11	1.12	0.14	1.76	55.7	+0.20
		GQA	125	57/68/+11	0.63	0.11	1.06	49.5	−0.01
PAI-attn	LLaVA-1.5	MCO	749	430/319/−111	4.78	2.04	4.07	67.2	+1.40
		AOK	616	395/221/−174	4.39	1.78	3.14	83.0	+2.08
		GQA	560	350/210/−140	3.89	0.93	3.46	85.5	+2.46
	LLaVA-NeXT	MCO	101	41/60/+19	0.46	0.46†	0.00	–	+0.00
		AOK	125	44/81/+37	0.49	0.49†	0.00	–	+0.00
		GQA	105	31/74/+43	0.34	0.34†	0.00	–	+0.00
VAF	LLaVA-1.5	MCO	874	381/493/+112	4.23	2.91	2.22	59.5	+0.42
		AOK	925	357/568/+211	3.97	2.90	1.50	71.1	+0.63
		GQA	947	295/652/+357	3.28	2.43	1.29	65.5	+0.40
	LLaVA-NeXT	MCO	403	244/159/−85	2.71	0.00	4.44	61.0	+0.98
		AOK	350	296/54/−242	3.29	0.02	3.87	84.5	+2.67
		GQA	387	330/57/−273	3.67	0.04	4.26	85.1	+2.99
	Qwen2.5-VL	MCO	224	47/177/+130	0.52	0.46	0.07	100.0	+0.07
		AOK	284	86/198/+112	0.96	0.92	0.09	37.5	−0.02
		GQA	345	79/266/+187	0.88	0.84	0.13	25.0	−0.07

MCO = MSCOCO, AOK = AOKVQA. Fb.% is measured over all 9,000 examples. Fb.Prec. is undefined for NOOP. † Discovery calibration selects NOOP, leaving the intervention output unchanged. PAI-attn on Qwen2.5-VL is omitted because the corresponding official configuration is not publicly available.



(a) Helpful vs. harmful (b) Fallback budget curve (c) Harm enrichment by score bin

Figure 3: **Replay-consistency diagnosis of changed candidates for LLaVA-1.5 + VGA.** Harmful intervention-induced changes concentrate toward the high-risk replay region, enabling replay-based ranking to identify harmful changes ahead of random fallback.

more slowly. The enrichment is strongest in the highest-risk region: the harmful fraction reaches 87.7% in the top score quintile (Figure 3(c)), compared with 42.4% overall.

Table 3: Routing-signal and replay-fusion ablations. Left: matched LLaVA-1.5 + VGA MSCOCO subset ($N = 8{,}996$); right: unweighted mean over 24 backbone–method–dataset POPE settings.

Routing signal		**Replay fusion**	
Router	Δ vs. VGA (pp)	Fusion	Mean gain (pp)
Best decode-time scalar	$+0.011$	Fixed median	$+0.8116$
Selective prediction	-0.011	Logistic regression	$+0.8014$
Self-consistency ($K = 5$)	0.000	MLP	$+0.8111$
Joint uncertainty logistic	$+0.111$	XGBoost	$+0.7630$
Replay C3 (median)	$+1.201$		

Table 4: Sensitivity to the labeled discovery budget over 20 repeated splits.

Labeled budget	25	50	100	250	500
Raw examples: full-NOOP	100%	90%	85%	20%	0%
Changed candidates: full-NOOP	–	0%	0%	–	–

A natural question is whether this structure is specific to replay or can be recovered from conventional uncertainty signals. Table 3 compares routing strategies under the same changed-candidate set, fallback action, and discovery supervision. Individual generation-time baselines remain near zero. Even a supervised logistic router combining the available generation-time uncertainty features yields only $+0.11$ points, compared with $+1.20$ points for replay-based routing on the same held-out subset. The difference is also not explained by a more expressive fusion rule. Combining the same replay features with logistic regression, a small MLP, or shallow XGBoost yields mean gains close to the fixed median across the evaluated POPE settings. Together, Figure 3 and Table 3 suggest that much of the improvement stems from information exposed by post-intervention replay rather than generic generation-time uncertainty or routing-model capacity.

4.5 Calibration Robustness and Label Budget

Because RAPIC uses labeled discovery calibration, we explicitly measure how much supervision is required and how sensitive the resulting policy is to the discovery sample. Uniformly labeling raw benchmark examples is inefficient: only about 9% of examples become route candidates in the primary setting, so small random discovery sets often contain too few informative changes.

At 25, 50, and 100 uniformly sampled labels (Table 4), a full-NOOP controller is selected in 100%, 90%, and 85% of splits, respectively. This behavior largely disappears once supervision is targeted to changed candidates. Labeling 50 changed candidates yields no full-NOOP split across 20 repetitions and an average gain of $+0.88$ points, with a range of $+0.31$ to $+1.57$. With 100 changed-candidate labels, both routing directions become active. These results do not remove the need for task-specific calibration, but suggest that the relevant supervision budget is better measured in intervention-induced decisions than in uniformly sampled raw examples.

4.6 Beyond Binary Object-Existence VQA

We further evaluate 4-way multiple-choice prediction on MMStar (Chen et al., 2024) using 500 discovery and 1,000 held-out examples (Table 5). PAI changes too few answers for reliable routing, while SID (Huo et al., 2025) and VGA yield nearly balanced harmful and helpful fallback counts. VAF produces the clearest asymmetric recovery: RAPIC reverts 55 harmful versus 37 helpful changes, recovering $+1.8$ points. Its confidence interval still includes zero, so we interpret this result as a boundary-condition analysis rather than a broad performance claim.

4.7 Object-Level Captioning Control

Finally, we evaluate the bounded generative variant on COCO-CHAIR (Rohrbach et al., 2018; Lin et al., 2014). Rather than replacing the whole caption, RAPIC suppresses the least-supported object

Table 5: RAPIC on MMStar 4-way multiple-choice evaluation. Interv. damage is the change caused by the mitigation method relative to its baseline. Net is selected harmful minus selected helpful changes. PAI uses a different runner and its baseline is not directly comparable to the others.

Method	Interv. damage	Changed	Sel. H/G	Net	Δ_R [95% CI]
PAI	-0.2	15	–	$+1$	$+0.1\,[-0.4, +0.6]$
SID	-0.6	137	41/38	$+3$	$+0.3\,[-1.4, +2.1]$
VGA	-0.7	131	30/31	-1	-0.1
VAF	-2.8	277	55/37	$+18$	$+1.8\,[-0.1, +3.7]$

Table 6: Object-level residual suppression on COCO-CHAIR with LLaVA-1.5 + VGA. K is the maximum number of least-supported object mentions suppressed during re-decoding.

Setting	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall↑	Precision↑	F1↑
VGA	32.20	8.90	72.81	82.15	77.20
+ RAPIC ($K = 1$)	30.20	8.04	72.62	83.57	77.71
+ RAPIC ($K = 2$)	28.80	7.92	72.04	84.47	77.76
+ RAPIC ($K = 3$)	**28.60**	**7.63**	72.37	84.43	**77.94**

mentions while preserving the remaining generation. Table 6 reports the primary LLaVA-1.5 + VGA setting. Suppressing one unsupported mention reduces CHAIR_s from 32.20 to 30.20 and CHAIR_i from 8.90 to 8.04 with little recall change. Increasing the suppression depth to $K = 3$ further reduces them to 28.60 and 7.63, respectively, while recall remains within one point of the intervention baseline. Since the operating points are not strictly monotonic and broader cross-method and cross-backbone results are more mixed (Appendix D.2), K should be viewed as a precision–recall tradeoff rather than a universally ordered setting.

This experiment addresses only unsupported object mentions within a benchmark-aligned vocabulary. RAPIC does not insert missing objects or perform unrestricted caption rewriting, and we therefore treat this result as a bounded generative extension rather than evidence for a general post-generation correction system. We additionally compare against the released Woodpecker post-hoc correction pipeline (Yin et al., 2024) on MMHal-Bench; the results and sample-level analysis are provided in Appendix D.4.

5 Conclusion

We study multimodal hallucination mitigation from the perspective of post-intervention risk: an always-on intervention can improve average performance while still introducing harmful output changes. We find that conventional generation-time uncertainty offers little signal about these changes, whereas replaying the committed intervention output through the frozen backbone exposes a much stronger risk signal. Based on this observation, RAPIC selectively retains or reverts intervention-induced changes using replay-consistency signals and a discovery-calibrated frozen policy. Across POPE datasets, mitigation methods, and VLM backbones, RAPIC reduces intervention-induced harm in its strongest settings, while repeated calibration under limited labeled supervision and broader evaluations clarify when selective control is useful and when the controller should abstain. Targeting discovery labels to changed candidates also makes calibration substantially more label-efficient than uniform sampling. Our evidence is strongest in discrete-answer settings and still requires task-specific labeled calibration, but within this scope it suggests that inference-time mitigation should be judged not only by average gains, but also by whether its changes remain identifiable and controllable at the sample level.

Limitation. While RAPIC is most effective on discrete-answer multimodal tasks and currently relies on task-specific labeled calibration, extending it to broader output spaces and developing replay signals that transfer across tasks without target-task labels are promising directions for future work. Likewise, its generative extension focuses on bounded repair of unsupported object mentions, and the benefit of replay consistency varies with the backbone and intervention, which motivates viewing RAPIC as a calibrated control layer rather than a universal correction mechanism.

AI USE STATEMENT

Generative AI tools were used solely as writing aids during manuscript preparation, including drafting and revising portions of the manuscript, language polishing, phrasing, and improving the clarity and organization of the presentation. AI-generated content was not used as scientific evidence, nor was it used to produce experimental results or substitute for empirical analysis. The research questions, methodological design, implementation, experimental protocol, interpretation of results, and scientific claims were developed and verified by the authors. All AI-assisted text was subsequently reviewed and revised as needed by the authors, who take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work studies reliability in multimodal hallucination mitigation using existing models and evaluation benchmarks, and does not involve human subjects or the collection of personal data. RAPIC is intended as a research method for analyzing and controlling intervention-induced errors, not as a guarantee of factual correctness or safety in high-stakes applications. As with other hallucination-mitigation methods, residual model errors may remain after applying the proposed controller.

REPRODUCIBILITY STATEMENT

The main paper specifies the post-intervention routing formulation, replay-consistency features, and discovery calibration procedure used by RAPIC. The appendix provides additional implementation details, evaluation protocols, calibration settings, routing-baseline configurations, and complete results omitted from the main text for space. We also report repeated calibration experiments to quantify sensitivity to the discovery split and paired-bootstrap confidence intervals to quantify statistical uncertainty. All reported results are produced using fixed evaluation protocols described in the paper and appendix.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29915–29926, 2025.

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

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21635–21644, 2025.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2534–2544, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

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*, pp. 13418–13427, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 24272–24295, 2025.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27228–27238, 2024.

Sai Akhil Kogilathota, Sripadha Vallabha EG, 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.

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*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

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

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

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19857–19866, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024d.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3570–3586, 2024.

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.

Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 635–646, 2025.

Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. Vade: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14949–14965, 2025.

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.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*, 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Yahan Tu, Rui Hu, and Jitao Sang. Ode: Open-set evaluation of hallucinations in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19836–19845, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25543–25551, 2025.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, 2021.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14635–14645, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Dokyoon Yoon, Youngsook Song, and Woomyoung Park. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4200–4208, 2025.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2017–2028, 2023.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19985–19995, 2025.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32582–32591, 2026.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1615–1624. IEEE, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

A Additional Method and Experimental Details

A.1 Shared Calibration Protocol

RAPIC follows a calibrate-then-freeze protocol. For discriminative VQA, the replay-consistency feature set is fixed before evaluation, while a labeled discovery split is used only to determine feature orientations, normalization statistics, direction-specific fallback thresholds, and possible NOOP decisions. For generative captioning, validation selects the object-support threshold and the parameters used for local residual suppression.

At held-out application time, RAPIC uses only the multimodal input, the baseline and intervention outputs, and same-backbone replay or support-probe logits. Evaluation annotations are used only for reporting metrics and are never used for per-example routing or repair decisions. Thus, RAPIC requires labeled supervision for task-specific calibration, but its frozen application-time decisions do not require evaluation labels.

A.2 Discriminative Controller Details

Changed-answer candidates. For an image–question pair $x_i = (I_i, q_i)$, let y_i^B and y_i^M denote the outputs of the baseline model and a fixed mitigation method, respectively. A deterministic parser maps them to task answers a_i^B and a_i^M. RAPIC operates only on

$$\mathcal{C} = \{x_i : a_i^B \neq a_i^M\}. \tag{11}$$

The changed candidates are divided into the two transition directions $Y \to N$ and $N \to Y$, since their harmful–helpful composition and replay-score distributions can differ substantially.

On labeled discovery candidates, a change is *harmful* when the baseline answer is correct and the intervention answer is wrong, and *helpful* when the baseline is wrong and the intervention is correct. For a fallback set $S \subseteq \mathcal{C}$, let $H(S)$ and $G(S)$ denote the numbers of harmful and helpful changes selected for fallback. Replacing the intervention output with the baseline on S changes benchmark accuracy by

$$\Delta_{\mathrm{fb}}(S) = \frac{H(S) - G(S)}{N}, \tag{12}$$

where N is the number of evaluation examples.

Replay-consistency features. For an intervention output $y_M = (w_1, \ldots, w_T)$ and its content-token positions $\mathcal{P}$, we replay the committed sequence under teacher forcing through the original frozen backbone. RAPIC uses the same three fixed features in all discriminative experiments:

$$f_{\mathrm{LP}} = \min_{t \in \mathcal{P}} \log p_B(w_t \mid I, q, w_{<t}), \tag{13}$$

$$f_{\mathrm{gap}} = \min_{t \in \mathcal{P}} \left[z_t(w_t) - \max_{v \neq w_t} z_t(v) \right], \tag{14}$$

and

$$f_{\mathrm{first}} = z_1(w_1) - \max_{v \neq w_1} z_1(v). \tag{15}$$

For each transition direction d, each feature is oriented so that larger values indicate greater harmful-change risk and standardized using discovery statistics. The fixed RAPIC score is

$$s_d(x) = \mathrm{median}\left(u_{1,d}(x), u_{2,d}(x), u_{3,d}(x)\right). \tag{16}$$

Threshold selection and NOOP. For each direction d, thresholds are swept over the discovery score distribution. Let

$$S_d(\tau) = \{x : s_d(x) \geq \tau\}. \tag{17}$$

The frozen threshold is selected to maximize discovery fallback utility over thresholds that select at least five discovery candidates within the corresponding transition direction:

$$\tau_d^* = \arg \max_{\substack{\tau \in \mathcal{T}_d \\ |S_d(\tau)| \geq 5}} \Delta_{\mathrm{fb}}(S_d(\tau)). \tag{18}$$

The candidate policy set also contains NOOP, which retains every intervention output. If no valid threshold improves over NOOP, RAPIC abstains from routing in that direction.

After calibration, feature orientations, normalization statistics, τ_d^*, and any NOOP decisions are frozen. The held-out output is

$$y_R(x) = \begin{cases} y_B(x), & a_B(x) \neq a_M(x) \ \wedge \ s_d(x) \geq \tau_d^*, \\ y_M(x), & \text{otherwise.} \end{cases} \tag{19}$$

A.3 Object-Level Captioning Controller

For captioning, whole-output fallback can remove correct content together with a local hallucinated mention. RAPIC therefore uses an object-level residual repair. Given a mitigated caption $c_M(x)$, we extract visually verifiable object mentions from a benchmark-aligned vocabulary. For each candidate object o, the same frozen backbone is queried with a binary support probe. If z_{yes} and z_{no} are the corresponding next-token logits, we define

$$p(o \mid x) = \frac{\exp(z_{\text{yes}})}{\exp(z_{\text{yes}}) + \exp(z_{\text{no}})}. \tag{20}$$

RAPIC ranks candidate mentions by this support score and locally suppresses the least-supported mentions whose support falls below a validation-calibrated threshold. If $\mathcal{T}(o)$ denotes the token set associated with object o and $b < 0$ the suppression bias, the repaired output is produced by re-decoding under the same intervention while applying the negative bias only to $\mathcal{T}(o)$. The procedure therefore targets unsupported object mentions rather than replacing the full caption.

B Additional POPE Results

B.1 Full Accuracy and F1 Results

Table 7 reports the complete per-dataset accuracy and F1 results corresponding to the aggregate and route-level results in the main paper. PAI-attn is reported only for backbones for which the corresponding official intervention configuration is publicly available.

B.2 Fallback Decomposition

Table 8 reports the exact number of RAPIC-selected fallbacks and their harmful/helpful decomposition. Positive $H - G$ corresponds directly to recovered benchmark accuracy.

B.3 Direction-Specific Calibration

Table 9 shows the primary VGA / LLaVA-1.5 controller separately for the two answer-transition directions. The score distributions and useful fallback operating points differ strongly between $Y \to N$ and $N \to Y$, motivating direction-specific calibration.

Table 10 lists the corresponding frozen direction-specific thresholds for all evaluated backbone–method pairs.

C Statistical Robustness and Routing Ablations

C.1 Paired Bootstrap Confidence Intervals

For the primary LLaVA-1.5 + VGA setting, we quantify uncertainty in the held-out accuracy improvement using 10,000 paired bootstrap replicates. We report both example-wise and image-clustered resampling. In the example-wise analysis, evaluation examples are sampled with replacement, with the same sampled indices applied jointly to VGA and RAPIC outputs. For image-clustered resampling, images are sampled with replacement and all questions associated with each sampled image are included together, preserving within-image dependence.

Table 7: Full POPE accuracy and F1 results. Base denotes the vanilla backbone, Meth. the always-on intervention, and +R RAPIC. Δ is the gain of RAPIC over the corresponding intervention.

			Accuracy ↑				F1 ↑			
Method	Backbone	Data	Base	Meth.	+R	Δ	Base	Meth.	+R	Δ
VGA	LLaVA-1.5	MCO	85.22	86.61	87.82	+1.21	85.99	86.59	87.81	+1.22
		AOK	78.98	81.04	82.66	+1.61	82.29	83.01	84.58	+1.57
		GQA	76.61	80.08	80.74	+0.67	80.66	82.48	83.10	+0.62
	LLaVA-NeXT	MCO	89.17	89.68	89.81	+0.13	89.26	89.41	89.55	+0.14
		AOK	85.23	85.96	86.40	+0.44	86.84	87.07	87.45	+0.38
		GQA	83.19	84.70	85.28	+0.58	85.15	85.96	86.45	+0.49
	Qwen2.5-VL	MCO	83.77	82.89	83.73	+0.84	80.88	79.63	80.85	+1.22
		AOK	85.00	84.88	85.08	+0.20	83.96	83.75	84.09	+0.34
		GQA	84.96	85.08	85.07	-0.01	84.24	84.33	84.40	+0.07
PAI-attn	LLaVA-1.5	MCO	85.22	83.99	85.39	+1.40	85.99	84.85	85.62	+0.77
		AOK	78.98	77.04	79.12	+2.08	82.29	80.59	81.87	+1.27
		GQA	76.61	75.06	77.51	+2.46	80.66	79.50	80.98	+1.48
	LLaVA-NeXT	MCO	89.17	89.38	89.38	+0.00	89.26	89.36	89.36	+0.00
		AOK	85.23	85.64	85.64	+0.00	86.84	87.05	87.05	+0.00
		GQA	83.19	83.67	83.67	+0.00	85.15	85.44	85.44	+0.00
VAF	LLaVA-1.5	MCO	85.22	86.47	86.89	+0.42	85.99	86.47	86.70	+0.23
		AOK	78.98	81.32	81.96	+0.63	82.29	83.15	83.62	+0.47
		GQA	76.61	80.58	80.98	+0.40	80.66	82.81	83.02	+0.21
	LLaVA-NeXT	MCO	89.17	88.22	89.20	+0.98	89.26	88.82	89.30	+0.48
		AOK	85.23	82.54	85.21	+2.67	86.84	84.96	86.82	+1.86
		GQA	83.19	80.16	83.14	+2.99	85.15	83.11	85.12	+2.00
	Qwen2.5-VL	MCO	83.77	85.21	85.28	+0.07	80.88	82.96	83.05	+0.09
		AOK	85.00	86.24	86.22	-0.02	83.96	85.68	85.67	-0.01
		GQA	84.96	87.03	86.97	-0.07	84.24	86.67	86.62	-0.05

MCO = MSCOCO, AOK = AOKVQA.

Table 8: Detailed fallback decomposition on POPE. Sel. H/G/Net reports selected harmful changes, selected helpful changes, and their difference $H - G$.

Method	Backbone	Data	Base	Meth.	+R	Fallback	Sel. H/G/Net
VGA	LLaVA-1.5	MCO	85.22	86.61	87.82	301	205/96/+109
		AOK	78.98	81.04	82.66	255	200/55/+145
		GQA	76.61	80.08	80.74	192	126/66/+60
	LLaVA-NeXT	MCO	89.17	89.68	89.81	18	15/3/+12
		AOK	85.23	85.96	86.40	50	45/5/+40
		GQA	83.19	84.70	85.28	68	60/8/+52
	Qwen2.5-VL	MCO	83.77	82.89	83.73	126	101/25/+76
		AOK	85.00	84.88	85.08	158	88/70/+18
		GQA	84.96	85.08	85.07	95	47/48/−1
PAI-attn	LLaVA-1.5	MCO	85.22	83.99	85.39	366	246/120/+126
		AOK	78.98	77.04	79.12	283	235/48/+187
		GQA	76.61	75.06	77.51	311	266/45/+221
	LLaVA-NeXT	MCO	89.17	89.38	89.38	0	–/–/0
		AOK	85.23	85.64	85.64	0	–/–/0
		GQA	83.19	83.67	83.67	0	–/–/0
VAF	LLaVA-1.5	MCO	85.22	86.47	86.89	200	119/81/+38
		AOK	78.98	81.32	81.96	135	96/39/+57
		GQA	76.61	80.58	80.98	116	76/40/+36
	LLaVA-NeXT	MCO	89.17	88.22	89.20	400	244/156/+88
		AOK	85.23	82.54	85.21	348	294/54/+240
		GQA	83.19	80.16	83.14	383	326/57/+269
	Qwen2.5-VL	MCO	83.77	85.21	85.28	6	6/0/+6
		AOK	85.00	86.24	86.22	8	3/5/−2
		GQA	84.96	87.03	86.97	12	3/9/−6

Table 11 reports the resulting 95% confidence intervals. Image-clustered resampling produces wider intervals, as expected from treating correlated questions from the same image jointly, but all three

Table 9: Direction-specific calibration and held-out routing for VGA / LLaVA-1.5 across POPE datasets. Hrec and Grec denote harmful and helpful fallback recall, respectively.

Data	Dir.	τ	Apply H/G/Net	Fallback	Sel. H/G/Net	Hrec	Grec
MCO	Y→N	0.557	272/393/−121	154	133/21/+112	48.90	5.34
MCO	N→Y	-2.580	74/78/−4	147	72/75/−3	97.30	96.15
AOK	Y→N	0.557	264/469/−205	169	147/22/+125	55.68	4.69
AOK	N→Y	-2.580	55/36/+19	86	53/33/+20	96.36	91.67
GQA	Y→N	0.557	198/539/−341	105	69/36/+33	34.85	6.68
GQA	N→Y	-2.580	59/30/+29	87	57/30/+27	96.61	100.00

Table 10: Frozen direction-specific thresholds for the evaluated backbone–method pairs. NOOP indicates that discovery calibration found no net-positive fallback policy for that direction.

Method	Backbone	$\tau_{Y\to N}$	$\tau_{N\to Y}$	Active directions
VGA	LLaVA-1.5	0.557	-2.580	both
VGA	LLaVA-NeXT	1.673	-2.072	both
VGA	Qwen2.5-VL	-0.895	-0.159	both
PAI-attn	LLaVA-1.5	NOOP	-2.117	N→Y only
PAI-attn	LLaVA-NeXT	NOOP	NOOP	none
VAF	LLaVA-1.5	2.128	-1.103	both
VAF	LLaVA-NeXT	NOOP	-4.192	N→Y only
VAF	Qwen2.5-VL	1.422	NOOP	Y→N only

intervals remain strictly above zero. The point estimates are unchanged, indicating that the primary RAPIC gains are not an artifact of example-level resampling assumptions.

C.2 Repeated Calibration and Label Efficiency

The main controller is calibrated on changed-answer candidates rather than uniformly sampled benchmark examples. We quantify why this distinction matters using 20 repeated discovery splits.

In the primary VGA / LLaVA-1.5 setting, approximately 9% of raw POPE examples become changed-answer candidates. Consequently, uniformly sampled small discovery sets frequently contain too few informative intervention decisions. With raw-example budgets of 25, 50, 100, 250, and 500 labels, the full-NOOP rates across 20 repeated splits are 100%, 90%, 85%, 20%, and 0%, respectively.

Targeting supervision directly to changed candidates is substantially more sample-efficient. With 50 labeled changed candidates, none of the 20 repeated splits produces a full-NOOP controller; the mean held-out gain is $+0.88$ points, with gains ranging from $+0.31$ to $+1.57$. With 100 labeled changed candidates, both transition directions become active.

These results do not eliminate task-specific labeled calibration. Rather, they show that the relevant labeling unit is the intervention-induced decision rather than an arbitrary raw benchmark example.

C.3 Matched Routing Baselines

We compare replay-based routing against conventional uncertainty signals using the same candidate set, fallback action, and discovery supervision. To ensure answer parity across all methods, the comparison uses 8,996 held-out examples, of which 815 are changed-answer candidates. The decode-time scalar is selected on discovery data from log-probability, margin, and entropy. Self-consistency uses $K = 5$ sampled generations and vote agreement as the routing signal. We additionally train a logistic router over the available generation-time uncertainty features using the same discovery supervision. Even this supervised combination yields only $+0.111$ points, whereas replay-based routing yields $+1.201$ points (Table 3).

Table 11: Paired-bootstrap uncertainty for RAPIC relative to VGA in the primary LLaVA-1.5 + VGA setting. Image-clustered resampling samples images with replacement and includes all associated questions jointly.

Dataset	Δ Acc. (pp)	Example-wise 95% CI	Image-clustered 95% CI
MSCOCO	+1.21	[0.83, 1.59]	[0.63, 1.80]
AOKVQA	+1.61	[1.27, 1.97]	[1.09, 2.17]
GQA	+0.67	[0.37, 0.97]	[0.21, 1.13]

Table 12: Direct comparison of individual replay-based selectors for VGA / LLaVA-1.5 across POPE datasets. NOOP retains the intervention output, All reverts all changed answers, and Fixed C3 denotes the proposed median replay score. All entries are accuracy (%).

Dataset	NOOP	All	Random	LP-min	Target-gap	First-gap	Entropy	Fixed C3
MSCOCO	86.61	85.22	86.10	**87.88**	87.82	87.62	87.00	87.82
AOKVQA	81.04	78.98	80.40	82.63	**82.66**	82.32	81.73	**82.66**
GQA	80.08	76.61	79.27	80.71	80.74	**80.88**	80.58	80.74

C.4 Replay Feature and Fusion Ablations

We first test each replay signal directly as a scalar selector. Table 12 shows that replay-based features consistently outperform random fallback and full rollback, but no individual signal is uniformly strongest across datasets.

We additionally replace the fixed median aggregation with learned fusion rules over exactly the same three replay features. As summarized in Table 3, increasing router capacity does not materially improve the aggregate gain. This supports the interpretation that the useful information comes mainly from post-intervention replay rather than from the expressive capacity of the fusion model.

C.5 Controlled Image-Flip Analysis

To test whether replay responds to visual evidence rather than merely to the surface form of the committed answer, we conduct a controlled image-flip analysis. For each changed candidate, we hold the question and committed intervention answer fixed while altering the image-side target presence. Across 815 changed candidates and three repeated constructions, this produces 4,890 paired evaluations.

As shown in Table 13, the dominant $Y \to N$ branch, which contains 664 of the 815 changed candidates, exhibits strong alignment between replay risk and image-side target presence. The $N \to Y$ branch behaves differently and even reverses under the controlled flip.

D Additional Evaluations and Boundary Analyses

D.1 MMStar Multiple-Choice Evaluation

For MMStar, we evaluate 4-way multiple-choice prediction with 500 discovery examples and 1,000 disjoint held-out examples. RAPIC controls only cases where the intervention changes the parsed multiple-choice answer. Unlike binary POPE, a changed candidate can also be neutral when both the baseline and intervention predictions are incorrect but differ. Such candidates remain in the direction-specific calibration pool, including score statistics and threshold sweeps, but contribute zero to $H(S) - G(S)$.

The held-out results are reported in Table 5 of the main paper. The experiment is best interpreted as a boundary analysis rather than as a uniform generalization result. PAI changes too few answers for reliable routing; SID and VGA have nearly balanced harmful and helpful selected changes; and VAF provides the clearest asymmetric recovery, although its confidence interval still includes zero.

Table 13: Direction-specific replay behavior under the controlled image-flip analysis. AUROC is reported separately for the discovery data and the controlled image-flip construction.

Direction	Changed candidates	Discovery AUROC	Controlled-flip AUROC
Y→N	664	0.842	0.9998
N→Y	151	0.52–0.58	0.0026

Table 14: Cross-method and cross-backbone COCO-CHAIR results for the base interventions and their RAPIC-controlled variants. CHAIR_s and CHAIR_i measure hallucination frequency and are lower-is-better; Recall, Precision, and F1 are higher-is-better.

Backbone	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall↑	Precision↑	F1↑
LLaVA-1.5	Vanilla	56.60	15.30	80.38	73.09	76.57
	VGA	32.20	8.90	72.81	82.15	77.20
	PAI	48.00	14.08	79.17	74.96	77.01
	VAF	51.80	15.49	78.24	72.88	75.47
	RAPIC + VGA	30.20	8.04	72.62	83.57	77.71
	RAPIC + PAI	32.00	8.89	73.55	83.10	78.03
	RAPIC + VAF	30.80	8.73	72.99	82.99	77.67
Qwen2.5-VL	Vanilla	27.20	7.66	60.57	84.85	70.68
	VGA	28.20	8.93	61.68	84.65	71.36
	PAI	28.00	8.62	62.36	85.15	71.99
	VAF	27.40	7.56	62.36	84.79	71.87
	RAPIC + VGA	26.60	8.41	62.24	85.19	71.93
	RAPIC + PAI	27.80	8.67	62.48	84.32	71.78
	RAPIC + VAF	26.80	8.41	62.42	84.38	71.76

Cross-task calibration transfer. We additionally test whether a routing policy calibrated in one task can be transferred without target-task discovery labels. A POPE-calibrated controller applied directly to MMStar selects no candidates and therefore changes accuracy by 0.0 points. Calibrating on MMBench-Dev and applying the policy to MMStar produces a $+0.6$ point change with a 95% confidence interval of $[-0.7, +1.9]$, but reverts all 186 changed candidates. We therefore interpret this behavior as broad rollback rather than evidence for calibration-free cross-task transfer.

D.2 Additional COCO-CHAIR Results

The main paper reports the suppression-depth ablation for the primary LLaVA-1.5 + VGA setting. Table 14 provides the broader cross-method and cross-backbone comparison from the original evaluation.

On LLaVA-1.5, residual suppression reduces CHAIR_s relative to each evaluated intervention; recall is largely preserved for VGA but decreases for PAI and VAF. The Qwen2.5-VL results are more mixed and generally smaller in magnitude, reinforcing that the captioning controller should be interpreted as a bounded residual repair rather than a uniformly improving wrapper.

Increasing the suppression depth strengthens hallucination suppression in the primary LLaVA-1.5 + VGA setting, as reported in Table 6 of the main paper. However, the mechanism remains one-sided: it removes unsupported mentions but cannot recover objects omitted from the original caption.

D.3 Captioning Threshold Sensitivity

Table 15 examines sensitivity to the object-support threshold τ_p for VGA on COCO-CHAIR. Performance varies across operating points, with no single threshold uniformly dominating all metrics. This supports treating τ_p as a validation-calibrated operating point rather than a universal fixed threshold.

D.4 Comparison with Post-Hoc Correction

To examine whether the aggregate-versus-sample-level gap also appears in a substantially different correction paradigm, we evaluate the released Woodpecker verification-and-regeneration pipeline on MMHal-Bench with LLaVA-1.5. Table 16 reports both aggregate and sample-level outcomes.

Table 15: Sensitivity of the object-support threshold τ_p for VGA on COCO-CHAIR with LLaVA-1.5 and Qwen2.5-VL. Cs and Ci denote CHAIR_s and CHAIR_i, and R, P, and F1 denote recall, precision, and F1.

	LLaVA-1.5					Qwen2.5-VL				
τ_p	Cs	Ci	R	P	F1	Cs	Ci	R	P	F1
0.20	31.60	8.74	72.74	82.19	77.18	27.40	8.95	60.82	84.90	70.87
0.30	31.40	8.56	72.81	82.55	77.37	27.00	8.94	61.19	85.12	71.20
0.40	32.20	8.90	72.81	82.15	77.20	26.60	8.50	60.94	85.29	71.09
0.50	30.80	8.37	72.68	82.99	77.50	26.80	8.36	61.80	85.18	71.63
0.60	30.80	8.28	72.68	83.23	77.60	26.60	8.41	62.24	85.19	71.93
0.70	30.40	8.26	72.50	83.31	77.53	26.80	8.49	62.73	84.94	72.16
0.80	30.20	8.04	72.62	83.57	77.71	28.20	9.06	62.86	83.91	71.87

Table 16: Woodpecker post-hoc correction on MMHal-Bench with LLaVA-1.5. Improve / Degrade / Same counts sample-level score changes relative to the vanilla model.

Setting	Mean score↑	Hallucination↓	Improve / Degrade / Same
Vanilla	3.052	53.13%	–
+ Woodpecker	3.063	51.04%	35 / 33 / 28

Although Woodpecker slightly improves both aggregate metrics, its sample-level effects are nearly balanced: 35 of 96 samples improve, 33 degrade, and 28 remain unchanged. Thus, aggregate improvement can coexist with substantial sample-level regressions even for an external verification-and-regeneration pipeline.

Woodpecker and RAPIC should not be interpreted as directly interchangeable systems. Woodpecker verifies concepts using external components and regenerates the response, whereas RAPIC controls changes introduced by a fixed intervention using same-backbone replay.

We additionally evaluated Woodpecker on COCO-CHAIR. Relative to the vanilla LLaVA-1.5 caption, Woodpecker changes CHAIR_s from 56.60 to 53.80 and CHAIR_i from 15.30 to 14.74, while recall decreases from 80.41 to 77.63. The regenerated captions are also approximately 72% longer. These results reinforce that post-hoc regeneration and RAPIC represent different correction regimes and should not be read as a direct head-to-head performance comparison.

E Efficiency and Compute

E.1 Replay Frequency and Latency

Replay is required only when the mitigation method changes the parsed answer. Table 17 reports a deployment analysis on 1,000 samples for VGA / LLaVA-1.5.

The baseline-first implementation generates the baseline answer unconditionally but invokes teacher-forced replay only for changed-answer candidates. Because 93.70% of samples do not require replay in this analysis, average latency falls from 2.444 to 1.836 seconds per sample relative to an always-replay implementation. This analysis uses a separate 1,000-sample subset, so its change and fallback rates differ from the main evaluation.

E.2 Compute Environment

All experiments are inference-time evaluations; RAPIC does not fine-tune the VLM backbones. Experiments were run on an internal GPU cluster using NVIDIA RTX A6000 48GB GPUs, with one GPU used for each evaluation run. The software environment uses Python 3.10, PyTorch, Transformers, and the official or publicly available implementations of the evaluated models and mitigation methods when applicable.

Table 17: Deployment latency for VGA / LLaVA-1.5 over 1,000 samples. Baseline-first RAPIC computes replay only for changed-answer candidates, whereas Always-replay replays every sample.

Quantity	Value
Samples evaluated	1,000
Answer changed	6.30%
Replay computed	6.30%
Route to baseline	0.10%
Method-only latency	0.955 s/sample
Baseline-first RAPIC	1.836 s/sample
Always-replay	2.444 s/sample
Speedup vs. always-replay	1.331×
Latency saved	24.86%

A full POPE evaluation contains 9,000 examples per dataset and is executed separately for MSCOCO, AOKVQA, and GQA. Across the project, the dominant compute cost is repeated VLM inference rather than parameter optimization.

F Limitations and Scope

The strongest empirical evidence for RAPIC remains in discrete-answer multimodal tasks, where method-induced changes and conservative fallback admit a clear answer-level interpretation. The MMStar results show that the same formulation can be applied to multi-way prediction, but they do not establish uniform gains outside binary VQA. In particular, routing is useful only when the underlying intervention produces a sufficiently informative and asymmetric population of harmful and helpful changes.

RAPIC also requires task-specific labeled discovery calibration. Application after calibration is label-free on individual evaluation examples, but the current evidence does not support calibration-free transfer across tasks. Indeed, direct POPE-to-MMStar transfer selects no candidates, while MMBench-to-MMStar calibration tends toward broad rollback. Developing replay signals that transfer across tasks without target-task labels remains an open problem.

The generative extension is narrower still. The current controller operates on benchmark-aligned object mentions and performs local suppression of unsupported objects. It cannot insert missing objects, recover omitted information, or perform unrestricted grounded rewriting. The captioning experiments should therefore be interpreted as evidence for bounded residual repair rather than a general post-generation correction framework.

Finally, replay consistency is not a universal detector of intervention harm. Its usefulness depends on the backbone, mitigation method, and transition direction. Some settings produce strong selective-correction regimes, others resolve to NOOP, and severely degraded interventions can induce rollback-dominant policies. These behaviors motivate treating RAPIC as a calibrated control layer rather than as a guarantee that every intervention can be improved.