WHEN VISUAL TRUTH LOSES CONTROL: DIAGNOSING AND PRESERVING GROUNDED PREDICTIONS IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable multimodal generation requires visual evidence to remain influential throughout decoding, yet existing studies mainly attribute hallucination to insufficient visual perception or excessive language priors, leaving its internal evolution underexplored. We uncover a fundamental failure mode, termed *Grounded Control Reversal* (GCR), where a visually grounded preference emerges in intermediate representations but is subsequently weakened or overturned before the final answer is produced. Through layer-wise preference tracing and same-image internal interventions on Qwen2.5-VL-7B across HallusionBench and POPE-Adversarial, we show that reversals arise from the combined effects of attenuated visual pathways, competing textual pathways, and feed-forward transformations. Based on this insight, we introduce ReCAP, a training-free decoding framework that identifies a persistent visually grounded intermediate state and restores its lost predictive influence through a minimum-norm projection in hidden space. The intervention selectively recovers the grounded preference while preserving unrelated residual components. Our findings provide a causal perspective on multimodal hallucination, showing that failures can arise not from the absence of visual understanding, but from the subsequent loss of control over already acquired visual evidence.

1 INTRODUCTION

Multimodal large language models (MLLMs) combine visual representations with powerful language decoders, enabling image description, visual question answering, and multimodal reasoning (Liu et al., 2024a; Bai et al., 2025). Despite these capabilities, their outputs can still contradict visible content, including by describing objects that are absent from the image (Rohrbach et al., 2018; Li et al., 2023). Such failures are commonly associated with insufficient visual grounding or the growing influence of language priors during decoding. Yet the final answer reveals only the endpoint of this process: a model may have already formed a correct, visually supported prediction before ultimately producing the wrong answer. This raises a different question: *what happens to a grounded prediction after it has already emerged inside the decoder?*

Recent training-free methods improve grounding by perturbing visual inputs (Leng et al., 2024), refining visual evidence (An et al., 2025; Park et al., 2025a), strengthening visual attention (Yin et al., 2025; Ji et al., 2026), or reinjecting visual information into intermediate computation (Zou et al., 2024). Related studies further show that intermediate representations can contain useful visual information. In particular, DeCo dynamically selects preceding-layer predictions and integrates them into the final distribution for decoding correction (Wang et al., 2025), while SAVER and VGA exploit intermediate visual signals to guide generation (Li et al., 2026; Zhao et al., 2026). These findings suggest that hallucination can arise not only from failing to acquire visual evidence, but also from failing to preserve its influence.

We study this failure through the layerwise competition between the correct answer and the model's eventual incorrect answer. On HallusionBench and POPE-Adversarial, we find a recurring trajectory in which the correct candidate remains preferred across consecutive intermediate layers, receives measurable support from visual computation, and is then overtaken before the final prediction. We term this phenomenon *Grounded Control Reversal* (GCR). Rather than treating an incorrect output as

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.

a static decoding error, GCR exposes a temporal transition in which a grounded prediction first gains and then loses control of the final decision.

To understand where this control is lost, we trace computation after the grounded intermediate state along three residual-update routes: visual attention, textual attention, and the feed-forward network (FFN). Intervention-based attribution shows that reversals can arise through different combinations of these routes, even though their effects ultimately converge on the same final residual state. This observation turns hallucination mitigation into a more focused problem: instead of separately correcting every possible takeover mechanism, can we preserve the grounded candidate at the shared decision point where these effects are finally expressed?

Motivated by this view, we propose ReCAP, *Reversal-aware Control Anchor Projection*, a training-free decoding method for preserving grounded intermediate preferences. ReCAP first identifies a persistent intermediate candidate whose preference is aligned with visual support. When that preference is reversed by the final layer, the method restores an anchor-derived pairwise advantage through a closed-form minimum-norm projection of the final residual state. In this way, the correction targets the lost decision direction directly while keeping the intervention as small as possible.

We evaluate ReCAP on Qwen2.5-VL-7B and LLaVA-1.5-7B across binary visual judgments, open-ended captioning, and multimodal benchmarks. The experiments examine hallucination mitigation and the connection between the diagnosed reversal mechanism and the proposed recovery strategy.

Our contributions are threefold:

- We identify and operationalize *Grounded Control Reversal*, where a visually supported intermediate prediction is subsequently overturned during later decoding.
- We provide an intervention-based account of how visual attention, textual attention, and FFN updates contribute to the loss of grounded prediction advantage.
- We introduce ReCAP, a training-free minimum-norm projection that restores a selected grounded preference at the final decision state.

2 Related Work

Mechanisms of Hallucination in MLLM Decoding. Recent studies increasingly examine hallucination through the internal computation of multimodal models. iTaD links hallucinated segments to reduced attention to image tokens and exploits inter-layer attention differences for decoding (Xu et al., 2025). ClearSight identifies middle-layer visual–text fusion as an effective intervention point (Yin et al., 2025). CausalLens performs sensitivity-guided head interventions to strengthen visually reliable internal pathways (Ji et al., 2026). PAS instead analyzes dependence on generated preliminary text and uses it as a signal for hallucination detection (Hoang-Xuan et al., 2025). Collectively, these studies show that hallucination is closely associated with how visual and textual information are redistributed during decoding. They further suggest that relevant failure signals can emerge across layers, attention pathways, and previously generated context, rather than being confined to the final output distribution.

Training-Free Hallucination Mitigation. A broad line of work mitigates hallucination at inference time without additional model training. VCD contrasts output distributions from original and distorted visual inputs (Leng et al., 2024); AGLA combines global and local visual evidence (An et al., 2025); SECOND selectively incorporates multiscale patches to strengthen object-relevant perception (Park et al., 2025a); and ConVis constructs contrastive signals from semantically reconstructed images (Park et al., 2025b). Other methods intervene directly on internal computation: ASCD steers attention through contrastive branches (Wang et al., 2026), MemVR reinjects visual memory through FFNs (Zou et al., 2024), and VGA uses visual-token semantics to guide attention toward relevant regions (Zhao et al., 2026). These methods broadly follow two complementary directions: modifying the evidence presented to the model or reshaping how the decoder uses that evidence internally. Together, they demonstrate that stronger visual grounding can be achieved through input contrast, attention modulation, memory reinjection, and feature-level guidance during inference.

Intermediate Representations for Decoding and Steering. Another line of work exploits information encoded in intermediate representations. DeCo dynamically selects preceding-layer predictions to correct the final output (Wang et al., 2025), while SAVER uses style-aware visual activations from earlier layers to guide output revision (Li et al., 2026). However, existing approaches primarily use intermediate information as a correction signal, without explicitly characterizing how a grounded prediction evolves after it emerges. A key question therefore remains: can a visually supported candidate acquire stable predictive control in intermediate layers and subsequently lose that control before the final readout? Answering this question requires tracing not only where useful intermediate signals appear, but also how their predictive influence changes across subsequent computation.

3 Grounded Control Reversal in MLLMs

We diagnose Grounded Control Reversal (GCR) in three steps: we first test whether the correct answer gains a persistent intermediate-layer advantage before being overturned, then verify whether this advantage is visually supported, and finally attribute its post-anchor loss to visual attention, textual attention, and feed-forward computation.

3.1 Identifying Layerwise Control Reversal

We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and collect 300 incorrect predictions from POPE-Adversarial (Li et al., 2023) and 300 from HallusionBench (Guan et al., 2024). These subsets are used only for mechanism analysis; full-set performance is reported in Section 5.

For each example, let $h_l \in \mathbb{R}^d$ be the residual state after block l at the prediction position. We read out the correct answer c and the model's final incorrect competitor b using

$$z_l = W_U \,\mathrm{Norm}_f(h_l), \qquad m_l = z_l(c) - z_l(b), \tag{1}$$

where $m_l > 0$ indicates that the correct candidate controls the pairwise prediction.

To avoid unstable single-layer readouts, we require the advantage to persist across two consecutive layers, defining $\mathcal{E} = \{l : 2 \le l < L,\ m_{l-1} > 0,\ m_l > 0\}$. A candidate reversal occurs when $\mathcal{E} \neq \varnothing$ but $m_L < 0$. We select the strongest stable point as $A = \arg\max_{l \in \mathcal{E}} m_l$ and define the subsequent control loss as $\mathrm{CL} = m_A - m_L$.

Table 1: Layerwise and visually grounded reversals among 300 diagnostic errors per benchmark. G_{vis} and CL are averaged over grounded cases.

Dataset	GCR	G-GCR	G_{vis}	CL
POPE-Adversarial	184 (61.3%)	144 (48.0%)	1.62	3.50
HallusionBench	164 (54.7%)	124 (41.3%)	1.78	4.10

As shown in Table 1, 61.3% of POPE errors and 54.7% of HallusionBench errors exhibit this trajectory. Thus, in a substantial fraction of failures, the correct answer is already preferred at intermediate layers and is overturned only later. A stricter margin-based analysis is provided in Appendix B.

3.2 Testing Visual Support of the Intermediate Preference

A prediction reversal alone does not establish that the intermediate correct preference is grounded in the image. We therefore perform a same-image internal intervention that suppresses visual-value transmission before the anchor.

For head r in block l, let $\alpha_{l,r,p,i}$ be the attention weight from prediction position p to key i, and $v_{l,r,i}$ its value vector. Splitting source positions into visual tokens $\mathcal{V}$ and remaining tokens $\mathcal{T}$, their projected residual contributions are $a_l^X = \sum_r W_{l,r}^O \sum_{i \in X} \alpha_{l,r,p,i} v_{l,r,i}$ for $X \in \{\mathcal{V}, \mathcal{T}\}$, with $a_l = a_l^V + a_l^T$.

For each candidate reversal, we rerun the same image and prompt to the fixed anchor A while removing visual contributions to nonvisual query positions throughout blocks $1{:}A$. If the resulting

anchor margin is m_A^{cut}, we define the visual-dependence effect as $G_{\text{vis}} = m_A - m_A^{\text{cut}}$. A reversal with $G_{\text{vis}} > 0$ is considered visually grounded, since suppressing the visual route weakens the correct intermediate preference.

This identifies 144 grounded reversals on POPE-Adversarial and 124 on HallusionBench, corresponding to 48.0% and 41.3% of all diagnostic errors. Their mean G_{vis} values are 1.62 and 1.78 logit units, while the subsequent control losses reach 3.50 and 4.10. These results show that visual computation can establish the correct intermediate prediction even though later computation ultimately overturns it. Additional controlled interventions are reported in Appendix B.

3.3 Attributing the Post-Anchor Reversal

We next trace how the grounded advantage is lost between the anchor and the final layer. For the pre-normalized decoder, the prediction-position update is $r_l = h_{l-1} + a_l^V + a_l^T$ and $h_l = r_l + f_l$, where $f_l = F_l(\text{Norm}_{2,l}(r_l))$. The post-anchor computation is therefore partitioned into three residual-update routes: visual attention V, textual attention T, and FFN F. These routes exhaust the residual additions at the analyzed position without assuming that each route carries a purely visual, linguistic, or parametric semantic source.

Starting from the factual anchor state h_A, we replay blocks $A+1{:}L$ under all $2^3 = 8$ combinations of these routes. For $S \subseteq N = \{V, T, F\}$, let $v(S)$ be the final correct–incorrect margin when only routes in S are enabled. Because their effects interact nonlinearly, we use Shapley attribution:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(2-|S|)!}{3!} \left[v(S \cup \{j\}) - v(S) \right]. \tag{2}$$

Here, $\phi_j < 0$ means that route j contributes to reducing the correct candidate's advantage. Since $v(\varnothing) = m_A$ and $v(N) = m_L$, Shapley efficiency gives $\phi_V + \phi_T + \phi_F = m_L - m_A < 0$, allocating the entire observed control loss among the three routes.

Table 2: Mean interventional Shapley contributions over grounded reversals. Negative values reduce the correct–incorrect margin.

Dataset	ϕ_V	ϕ_T	ϕ_F	$m_L - m_A$
POPE-Adversarial	-0.54	-1.02	-1.94	-3.50
HallusionBench	-0.69	-1.61	-1.80	-4.10

Table 2 shows that all three routes contribute negatively on average, with textual-attention and FFN updates accounting for larger portions of the loss in these cohorts. The dominant route nevertheless varies across examples. For later mechanism-specific evaluation, we define the normalized negative contribution as $R_j = [-\phi_j]_+ / \sum_k [-\phi_k]_+$ and assign a mixed label when $\max_j R_j < 0.5$.

Together, these experiments establish the complete GCR chain: a correct prediction gains a persistent intermediate advantage, this advantage is supported by visual computation, and subsequent residual updates can reverse it before the final readout. Because all three pathways ultimately act on the same final residual state, their shared endpoint motivates the unified restoration strategy in Section 4.

4 ReCAP: Reversal-Aware Control Anchor Projection

The GCR diagnosis reveals a specific failure process: a visually supported prediction can first gain control at an intermediate layer and then lose that control as later visual-attention, textual-attention, and FFN updates accumulate toward the final state. This observation suggests a different mitigation strategy from continually strengthening visual evidence throughout decoding. If the model has already formed a grounded preference, the key problem is to identify that preference without gold supervision, determine whether it has truly been reversed, and restore only the lost decision component.

Motivated by this view, we propose **ReCAP**, a training-free decoding method that operationalizes the reverse process of GCR. At each decoding step, ReCAP proceeds in three conceptual stages. It first searches the model's own intermediate trajectory for a *grounded control anchor*: a candidate

that is persistent across layers and locally supported by visual computation. It then checks whether this preference has actually been overturned by the final layer, so that already-consistent predictions are left untouched. Finally, when a reversal is detected, ReCAP restores an anchor-derived amount of pairwise control through the minimum residual displacement required to cross the corresponding decision boundary. Figure 1 summarizes this process.

4.1 Recovering a Grounded Control Anchor

The first challenge is to recover, *without knowing the correct answer*, the type of intermediate state identified by the diagnosis. An arbitrary high-scoring intermediate token is insufficient: intermediate readouts may fluctuate across layers, and a high score need not be supported by the image. We therefore require an anchor to satisfy three properties suggested directly by the GCR analysis: it must remain plausible at the final decision, persist across neighboring layers, and receive candidate-specific visual support.

Plausible alternatives. At generation step t, all quantities below refer to the current prediction position, and we omit t for clarity. Let $h_l \in \mathbb{R}^D$ denote the residual state after layer l, with $z_l = W_U \,\mathrm{Norm}_f(h_l)$. Let

$$b = \arg\max_w z_L(w)$$

be the vanilla final prediction. Because ReCAP aims to recover a preference that was lost rather than introduce an unrelated token, we restrict the search to alternatives that remain plausible under the final model distribution:

$$\mathcal{C} = \mathrm{TopK}(z_L, K).$$

For each candidate layer $l \in \mathcal{L}_c$, we read out only this shortlist and define

$$a_l = \arg\max_{w \in \mathcal{C}} z_l(w), \qquad M_l = z_l(a_l) - z_l(b). \tag{3}$$

Here a_l is the provisional anchor candidate and M_l measures how strongly the intermediate state prefers it over the final competitor. Restricting the search to $\mathcal{C}$ both keeps the candidate tied to the final decoding context and avoids full-vocabulary readouts at every intermediate layer.

Visual support rather than confidence alone. A positive M_l indicates an intermediate preference, but GCR specifically concerns a *visually grounded* preference. We therefore ask whether the visual residual update at layer l pushes the state toward a_l relative to b.

For the RMS-normalized, bias-free readout used here (Zhang & Sennrich, 2019), write

$$s(h) = \sqrt{D^{-1}\|h\|_2^2 + \epsilon_{\mathrm{rms}}}, \qquad \mathrm{Norm}_f(h) = \frac{\gamma \odot h}{s(h)}, \qquad d_{ab} = \gamma \odot (u_a - u_b), \tag{4}$$

where γ is the final normalization gain and u_w is the unembedding vector associated with token w. The vector d_{ab} is the normal of the pairwise a-versus-b decision boundary in pre-normalization residual space. During anchor search, we set $a = a_l$ and recompute this direction for each provisional pair.

Let a_l^V denote the visual-attention residual contribution defined in Equation (A61). We measure whether this update agrees with the current candidate preference by

$$G_l = \frac{d_{a_l b}^\top a_l^V}{\|d_{a_l b}\|_2 \, \|a_l^V\|_2 + \epsilon_g}. \tag{5}$$

Thus, $G_l > 0$ means that the visual update moves the residual state toward a_l relative to b. Importantly, G_l is tied to the same candidate competition that will later be restored.

Persistence and reversal. The diagnosis further shows that meaningful grounded control is a trajectory property rather than a single-layer spike. ReCAP therefore considers layer l eligible only when

$$a_l = a_{l-1} \neq b, \qquad M_l > 0, \qquad M_{l-1} > 0, \qquad G_l > 0.$$

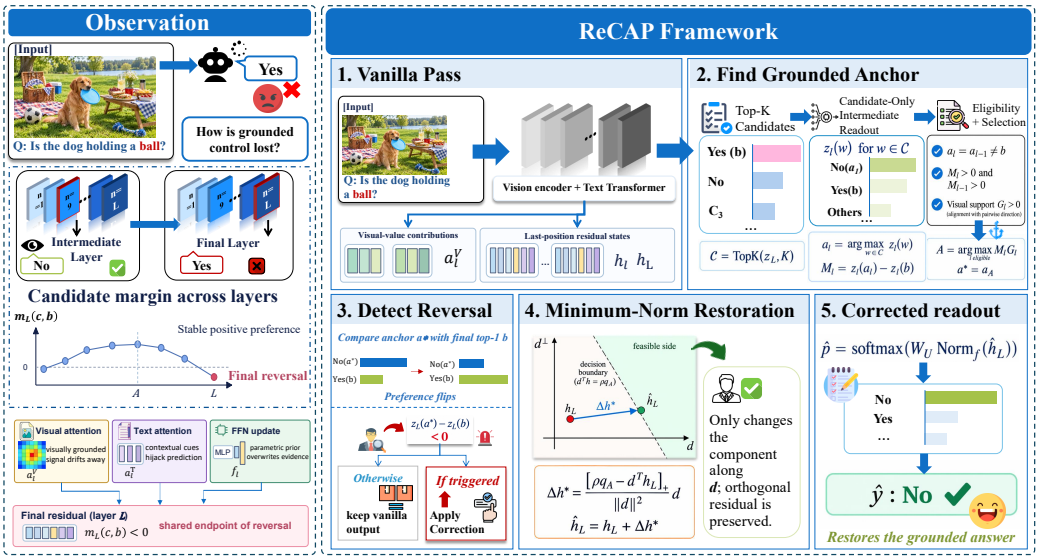


Figure 1: Overview of the GCR diagnosis and ReCAP decoding framework.

Among eligible layers, we select

$$A = \arg\max_{l \text{ eligible}} M_l G_l, \qquad a^\star = a_A. \tag{6}$$

This choice favors a candidate that is simultaneously strong, persistent, and visually aligned.

The anchor alone does not justify an intervention. The defining event in GCR is that a grounded preference is subsequently *reversed*. ReCAP therefore acts only when

$$z_L(a^\star) < z_L(b).$$

If no eligible anchor exists, or if its preference survives to the final layer, the vanilla output is retained. This reversal-aware trigger is what turns the intermediate state from a generic early-layer signal into evidence of an actual loss of grounded control.

4.2 Restoring Only the Lost Control

Once a reversal is detected, the remaining question is how much of the intermediate state should be restored. Copying or interpolating the entire anchor state would also reintroduce components unrelated to the selected decision. Instead, ReCAP uses the pairwise geometry to recover only the control that was lost.

Fix $a = a^\star$, b as the vanilla competitor, and $d = d_{ab}$. For any residual state h,

$$z_h(a) - z_h(b) = \frac{d^\top h}{s(h)}. \tag{7}$$

Because $s(h) > 0$, the sign of the pairwise preference depends only on $d^\top h$. The anchor state therefore supplies a natural reference for how much control the grounded candidate previously possessed:

$$q_A = d^\top h_A > 0.$$

Rather than forcing the final state back to h_A, we ask for the smallest change that restores a fraction $\rho \in (0, 1]$ of this anchor-derived control:

$$\min_{\Delta h} \frac{1}{2}\|\Delta h\|_2^2 \quad \text{subject to} \quad d^\top(h_L + \Delta h) \geq \rho q_A. \tag{8}$$

The constraint specifies *what* preference must be recovered, while the objective limits *how much* of the final representation is altered.

Proposition 1 (Minimum-norm pairwise restoration)**.** *For $d \neq 0$, Equation* (8) *has the unique solution*

$$\Delta h^\star = \frac{[\rho q_A - d^\top h_L]_+}{\|d\|_2^2}\, d, \qquad \hat{h}_L = h_L + \Delta h^\star. \tag{9}$$

For a triggered reversal and $\rho > 0$, the corrected readout $\hat{z} = W_U \operatorname{Norm}_f(\hat{h}_L)$ satisfies $\hat{z}(a) > \hat{z}(b)$.

Here $[x]_+ = \max(x, 0)$. The update is parallel to d, the normal of the pairwise decision boundary, and therefore gives the shortest displacement that satisfies the restoration constraint. The KKT proof is provided in Appendix A; numerically degenerate directions are skipped.

This geometry also makes explicit what ReCAP leaves unchanged. Let

$$P_\perp = I - \frac{dd^\top}{\|d\|_2^2}.$$

Since $\Delta h^\star \parallel d$,

$$P_\perp \hat{h}_L = P_\perp h_L. \tag{10}$$

Thus, ReCAP modifies only the residual component associated with the selected a-versus-b competition while preserving every component orthogonal to that direction. Generation then continues from the corrected full-vocabulary distribution

$$\hat{p} = \operatorname{softmax}\Big(W_U \operatorname{Norm}_f(\hat{h}_L)\Big).$$

The guarantee is deliberately pairwise: other vocabulary logits may still change through normalization and unembedding. We therefore evaluate full-vocabulary answer recovery and distributional distortion separately in Section 5.

4.3 One Shared Endpoint for Different Takeover Routes

Section 3 shows that grounded control can be lost through different mixtures of visual-attention, textual-attention, and FFN updates. Rather than classifying these routes online, ReCAP targets their shared endpoint in the final residual state. Starting from the selected anchor,

$$d^\top h_L = d^\top h_A + \sum_{l=A+1}^{L} d^\top \left(a_l^V + a_l^T + f_l\right), \tag{11}$$

where $f_l = F_l(\operatorname{Norm}_{2,l}(r_l))$. Thus, although the route contributions vary across examples, their combined effect is summarized by the same scalar endpoint $d^\top h_L$. The diagnostic attribution explains *how* control is lost, whereas ReCAP directly restores the resulting endpoint deficit without route-specific decoding or online causal attribution.

Computationally, ReCAP reuses a single standard transformer trajectory, caches candidate-layer states and visual contributions, and applies the closed-form $O(D)$ projection only when a reversal is triggered. The full procedure is given in Algorithm A1, with implementation details in Appendix C.

5 Experiments

Our evaluation connects aggregate hallucination performance with the mechanism that motivates ReCAP. We additionally test anchor reliability, the geometry of the correction, preservation of useful output content, and total inference cost.

5.1 Experimental Setup

We evaluate Qwen2.5-VL-7B-Instruct and LLaVA-1.5-7B (Bai et al., 2025; Liu et al., 2024a) on POPE, HallusionBench, CHAIR, VLMBias, PhD, MMStar, and MMBench (Li et al., 2023; Guan et al., 2024; Rohrbach et al., 2018; Vo et al., 2025; Liu et al., 2025; Chen et al., 2024; Liu et al., 2024b). Baselines include VCD, DeCo, SAVER, MemVR, CausalLens, and VGA. All methods share backbone-specific preprocessing, prompts, token limits, and greedy decoding; ReCAP uses $K = 10$, $\rho = 0.1$, two-layer persistence, and candidate blocks $\lceil 0.25L \rceil : \lfloor 0.875L \rfloor$. Diagnostic cohorts are used only for analysis, with no diagnostic gold labels used by ReCAP or for hyperparameter selection. Detailed benchmark protocols and configuration are provided in Appendix C.2.

5.2 Hallucination Mitigation Across Tasks

Table 3: Comparison of hallucination mitigation methods on two backbones. **Bold** and underlined values mark the best and second-best results within each backbone, respectively. Vanilla denotes the unmodified backbone; Bias denotes the bias rate. Arrows indicate the direction of better performance.

Method	CHAIR		POPE		Hall.	VLMBias		PhD	
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Acc.↑	F1↑	qAcc↑	Acc.↑	Bias↓	Acc.↑	Index↑
Backbone: **Qwen2.5-VL-7B**									
Vanilla	26.53	9.32	84.82	84.02	68.12	22.67	60.27	54.12	48.22
VCD	24.50	8.92	85.72	85.22	68.53	23.31	**57.76**	57.90	52.99
DeCo	25.83	7.96	84.97	85.03	67.24	22.67	60.49	54.98	48.98
SAVER	26.27	12.43	83.62	85.43	69.30	21.84	61.82	54.71	48.84
MemVR	23.87	7.10	86.49	84.59	69.81	22.70	60.17	53.32	47.66
CausalLens	24.13	7.43	87.40	87.70	70.15	23.42	59.34	53.65	46.34
VGA	27.40	8.75	86.52	86.41	70.06	23.10	58.45	52.79	47.27
ReCAP (Ours)	**23.17**	**6.71**	**89.38**	**88.93**	**71.47**	**24.32**	59.12	**59.03**	**54.55**
Backbone: **LLaVA-1.5-7B**									
Vanilla	51.60	14.20	86.31	84.53	49.16	16.20	43.28	62.38	53.17
VCD	54.40	15.83	86.73	86.60	48.21	15.31	45.46	57.32	52.34
DeCo	39.00	10.65	87.01	85.78	51.06	16.87	**42.83**	63.23	53.46
SAVER	43.67	11.67	85.42	85.32	52.31	15.63	46.45	58.90	52.80
MemVR	50.83	14.81	84.39	82.11	47.12	**18.42**	43.26	60.46	51.29
CausalLens	47.80	12.37	85.67	86.46	53.08	16.39	44.90	59.51	54.07
VGA	39.37	**10.43**	84.38	82.37	47.72	17.21	45.83	61.27	50.62
ReCAP (Ours)	**38.27**	11.82	**88.20**	**86.94**	**53.65**	17.46	44.58	**64.76**	**54.53**

Table 3 shows that ReCAP improves most hallucination metrics relative to vanilla decoding on both backbones. On Qwen2.5-VL-7B, it raises POPE F1 from 84.02 to 88.93 and HallusionBench qAcc from 68.12 to 71.47 while reducing $CHAIR_i$ from 9.32 to 6.71. The gains also transfer to LLaVA-1.5-7B, where ReCAP reduces $CHAIR_s$ from 51.60 to 38.27 and improves PhD accuracy from 62.38 to 64.76; the VLMBias bias rate is the main exception.

5.3 Recovery Across the Three Takeover Routes

To more comprehensively analyze the recoverability of grounded control reversal across different hallucination settings and takeover routes, we construct an expanded evaluation set from POPE-Adversarial, HallusionBench, PhD-ccs, and VLMBias, containing 574 grounded reversals in total. Across these cases, the final Top-K shortlist contains the gold answer in 516 cases (89.9%), and ReCAP triggers an intervention in 463 cases (80.7%). Among the triggered cases, the selected anchor matches the gold answer in 381 cases (82.3%), of which 334 are successfully recovered as the final prediction (87.7%). This yields an overall end-to-end recovery rate of 58.19% ($334/574$), as reported in Table 4. ReCAP improves recovery across all takeover-route groups, although mixed-route cases remain more challenging. Importantly, a correct anchor does not necessarily guarantee full-vocabulary recovery, since the projection restores a pairwise preference while other vocabulary candidates may still dominate the corrected distribution.

Table 4: End-to-end recovery (%) on 574 grounded reversals grouped by the dominant vanilla takeover route.

Method	V (190)	T (126)	F (216)	Mixed (42)	All (574)
DeCo	45.26	33.33	36.11	28.57	37.98
MemVR	38.95	46.03	44.44	42.85	42.86
VGA	33.68	49.21	41.66	33.33	40.07
ReCAP	**58.95**	**61.90**	**56.48**	**52.38**	**58.19**

5.4 Component Analysis

Table 5: Ablations on Qwen2.5-VL-7B; correction variants use matched anchors and triggers.

Variant	POPE ↑	Hall. ↑	C_s ↓	C_i ↓
w/o visual support	85.51	64.38	27.60	10.24
Fixed anchor layer	85.47	67.50	28.97	9.79
Always-on anchor revision	73.42	57.36	32.57	13.57
Hidden-state interpolation	86.54	69.73	24.87	7.59
Matched-anchor logit addition	87.82	70.13	25.13	7.37
Matched-anchor pairwise logit bias	86.46	70.18	25.07	8.32
Full ReCAP	**89.38**	**71.47**	**23.17**	**6.71**

Table 5 isolates the main design choices. Removing visual support increases interventions and reduces accuracy, while a fixed anchor loses the benefit of instance-dependent depth. For correction geometry, hidden-state interpolation, whole-distribution logit addition, and a minimum-norm *pairwise logit bias* reuse the same anchors and triggers. This matched comparison tests the contribution of residual-space geometry independently of anchor selection. The full method consistently achieves the strongest joint factuality in this comparison. The always-on variant additionally acts when a supported preference has weakened without a sign reversal; its greater intervention rate further illustrates the value of the selective trigger.

5.5 Content Retention, General Capability, and Cost

Table 6: Content retention, general capability, and decoding cost. Obj. R is caption object recall (%); Len. is mean output length; TPOT is measured in ms/token and memory in GiB.

Method	Obj. R ↑	Len.	TPOT ↓	GiB ↓	MMStar ↑	MMBench ↑
Backbone: **Qwen2.5-VL-7B**						
Vanilla	63.74	219.73	24.18	16.27	64.35	82.64
DeCo	58.99	223.61	26.13	16.23	62.51	80.61
VGA	62.11	207.04	26.86	16.26	62.36	81.77
ReCAP (Ours)	63.32	230.69	25.37	16.23	64.93	82.05
Backbone: **LLaVA-1.5-7B**						
Vanilla	77.83	117.66	30.35	14.63	34.37	65.43
DeCo	76.68	115.07	35.10	14.42	33.16	63.70
VGA	76.01	111.74	32.95	14.41	35.23	64.51
ReCAP (Ours)	78.20	113.73	32.04	14.01	34.80	64.98

Together with the CHAIR reductions in Table 3, Table 6 shows that ReCAP improves factuality while largely preserving content coverage and general capability. Object recall, output length, MMStar, and MMBench remain close to vanilla on both backbones. The additional decoding cost is modest: TPOT changes from 24.18 to 25.37 ms/token on Qwen and from 30.35 to 32.04 ms/token on LLaVA, while peak memory remains comparable or slightly lower.

6 Conclusion

We study grounded control reversal: an image-supported intermediate answer preference can be overturned by later decoder computation. A three-route intervention framework accounts for the change through visual attention, textual attention, and FFN updates. Their convergence on the final residual state motivates ReCAP, which identifies a supported candidate and restores its pairwise control with a minimum-norm projection. The evaluation connects hallucination mitigation to anchor reliability, mechanism-specific recovery, content retention, and inference cost. Grounded control retention offers a focused principle for making visual evidence more decisive at the point of generation.

AI USE STATEMENT

Generative AI tools were used to assist with software-code editing and debugging, algebraic checking of mathematical derivations, and grammar and phrasing in manuscript drafts. All AI-assisted code and mathematical content were independently reviewed and verified by the authors. The authors take responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work concerns the reliability of image-conditioned generation using public benchmarks and pretrained models. Hallucination mitigation can improve factual alignment, while residual errors remain possible and especially consequential in high-stakes applications. Benchmark improvements should be interpreted within the evaluated tasks and should be accompanied by appropriate human oversight in deployment. Dataset licenses, image privacy constraints, and model-use terms apply to any redistribution.

REPRODUCIBILITY STATEMENT

The main text defines the diagnosis, anchor selection, trigger, and projection. Appendix A provides proofs and the exact readout assumptions. Appendices B, C, and D specify the diagnostic protocol, route interventions, implementation details, and recovery analysis. The appendix also reports the default hyperparameters, benchmark protocols, and computational costs needed to reproduce the evaluations. All comparisons use deterministic tie-breaking, fixed diagnostic candidate pairs, and shared route assignments to ensure consistent evaluation conditions.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 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.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Nhat Hoang-Xuan, Minh Vu, My T Thai, and Manish Bhattarai. Pas: Prelim attention score for detecting object hallucinations in large vision–language models. *arXiv preprint arXiv:2511.11502*, 2025.

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.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

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.

Zhaoxu Li, Chenqi Kong, Yi Yu, Qiangqiang Wu, Xinghao Jiang, Ngai-Man Cheung, Bihan Wen, Alex Kot, and Xudong Jiang. Saver: mitigating hallucinations in large vision-language models via style-aware visual early revision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35617–35625, 2026.

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 (CVPR)*, pp. 26296–26306, June 2024a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19857–19866. IEEE, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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*, 2025a.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6434–6442, 2025b.

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.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

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*, volume 2025, pp. 13712–13736, 2025.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1571–1590, 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 IEEE/CVF conference on computer vision and pattern recognition*, pp. 14625–14634, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

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.

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 Mathematical Details

This section provides the mathematical details underlying the residual-space correction used by ReCAP. We first make explicit the pairwise geometry induced by the RMS-normalized output head, then prove the minimum-norm restoration stated in Proposition 1, and finally characterize precisely which components are preserved by the correction and which quantities may still change.

Throughout this section, all quantities refer to one decoding step at the current prediction position. We omit the step index for clarity. Let $h_l \in \mathbb{R}^D$ denote the residual state after decoder block l, and let L denote the final decoder block.

A.1 Exact Pairwise Geometry under RMSNorm

We begin by making explicit the readout geometry used by ReCAP. For the RMS-normalized, bias-free output head considered in the main text, define

$$s(h) = \sqrt{\frac{\|h\|_2^2}{D} + \epsilon_{\mathrm{rms}}}, \qquad \mathrm{Norm}_f(h) = \frac{\gamma \odot h}{s(h)}, \tag{A1}$$

where $\gamma \in \mathbb{R}^D$ is the learned gain of the final RMSNorm and $\epsilon_{\mathrm{rms}} > 0$ is its numerical constant. Let $u_w \in \mathbb{R}^D$ denote the unembedding vector associated with vocabulary token w. The logit of token w at residual state h is therefore

$$z_h(w) = u_w^\top \mathrm{Norm}_f(h) = \frac{u_w^\top(\gamma \odot h)}{s(h)}. \tag{A2}$$

For notational convenience, define the gain-adjusted unembedding vector

$$\bar{u}_w = \gamma \odot u_w. \tag{A3}$$

Then Equation (A2) becomes

$$z_h(w) = \frac{\bar{u}_w^\top h}{s(h)}. \tag{A4}$$

Consider an anchor candidate a and the vanilla final competitor b. Define their pairwise readout direction as

$$d_{ab} = \bar{u}_a - \bar{u}_b = \gamma \odot (u_a - u_b). \tag{A5}$$

Subtracting their logits gives

$$\begin{aligned} z_h(a) - z_h(b) &= \frac{\bar{u}_a^\top h}{s(h)} - \frac{\bar{u}_b^\top h}{s(h)} \\ &= \frac{(\bar{u}_a - \bar{u}_b)^\top h}{s(h)} \\ &= \frac{d_{ab}^\top h}{s(h)}. \end{aligned} \tag{A6}$$

Since $\epsilon_{\mathrm{rms}} > 0$, we have $s(h) > 0$ for every $h \in \mathbb{R}^D$. Hence

$$\mathrm{sign}\big(z_h(a) - z_h(b)\big) = \mathrm{sign}\big(d_{ab}^\top h\big). \tag{A7}$$

Equation (A7) is the key geometric identity used by ReCAP. In particular, the pairwise zero-margin boundary between a and b is

$$\mathcal{H}_{ab} = \left\{ h \in \mathbb{R}^D : d_{ab}^\top h = 0 \right\}, \tag{A8}$$

which is a hyperplane in the pre-normalization residual space with normal vector d_{ab}.

The two corresponding open half-spaces have a direct decoding interpretation:

$$d_{ab}^\top h > 0 \iff z_h(a) > z_h(b), \qquad d_{ab}^\top h < 0 \iff z_h(a) < z_h(b). \tag{A9}$$

Thus, restoring the pairwise preference does not require directly optimizing the normalized logits. It is sufficient to move the pre-normalization residual state across the corresponding linear decision boundary.

Anchor-derived control strength. Let h_A denote the selected anchor state and let

$$M_A = z_A(a) - z_A(b) > 0 \tag{A10}$$

be its normalized pairwise logit margin. From Equation (A6),

$$M_A = \frac{d_{ab}^\top h_A}{s(h_A)}. \tag{A11}$$

Therefore,

$$q_A := d_{ab}^\top h_A = s(h_A) M_A > 0. \tag{A12}$$

The quantity q_A is the pairwise control strength of the anchor in the *pre-normalization residual space*. Consequently, the ReCAP target ρq_A is not identical to requiring a final logit margin of ρM_A. Because the RMSNorm denominator depends on the corrected state, the normalized margin after intervention is also affected by the residual norm.

This distinction is important for the optimization below: ReCAP restores a fraction of the anchor's residual-space control rather than copying its complete normalized output distribution.

A.2 Proof of the Minimum-Norm Restoration

We now prove Proposition 1. Fix the selected anchor a, the vanilla competitor b, and write $d = d_{ab}$ for brevity. Let

$$\tau = \rho q_A, \qquad \rho \in (0, 1]. \tag{A13}$$

ReCAP seeks the smallest Euclidean perturbation to the final residual state h_L that restores the anchor-derived pairwise control:

$$\begin{aligned} \min_{\Delta h \in \mathbb{R}^D} \quad & \frac{1}{2}\|\Delta h\|_2^2 \\ \text{s.t.} \quad & d^\top (h_L + \Delta h) \geq \tau. \end{aligned} \tag{A14}$$

The feasible region is a closed half-space. Because $d \neq 0$, it is nonempty: for sufficiently large $c > 0$, $h_L + cd$ is strictly feasible. The objective is strictly convex and the constraint is affine, so the optimization problem has a unique global minimizer and the KKT conditions are both necessary and sufficient.

Lagrangian. Rewrite the constraint as

$$\tau - d^\top h_L - d^\top \Delta h \leq 0. \tag{A15}$$

Introducing the dual variable $\lambda \geq 0$, the Lagrangian is

$$\mathcal{L}(\Delta h, \lambda) = \frac{1}{2}\|\Delta h\|_2^2 + \lambda \left(\tau - d^\top h_L - d^\top \Delta h\right). \tag{A16}$$

The KKT conditions are

$$\begin{aligned} &\text{stationarity:} && \nabla_{\Delta h} \mathcal{L} = \Delta h - \lambda d = 0, && \text{(A17)} \\ &\text{primal feasibility:} && d^\top (h_L + \Delta h) \geq \tau, && \text{(A18)} \\ &\text{dual feasibility:} && \lambda \geq 0, && \text{(A19)} \\ &\text{complementary slackness:} && \lambda \left(\tau - d^\top h_L - d^\top \Delta h\right) = 0. && \text{(A20)} \end{aligned}$$

From stationarity,

$$\Delta h = \lambda d. \tag{A21}$$

Hence every optimal correction must lie along the normal direction d of the pairwise decision boundary.

There are two possible cases.

Case 1: the original state is already feasible. If

$$d^\top h_L \geq \tau, \tag{A22}$$

then $\Delta h = 0$ satisfies the constraint. Since the objective is nonnegative for every Δh, this achieves its global minimum value of zero. Thus

$$\Delta h^\star = 0. \tag{A23}$$

Case 2: the constraint is violated. Suppose

$$d^\top h_L < \tau. \tag{A24}$$

Then $\Delta h = 0$ is infeasible, so the optimum must satisfy $\lambda > 0$. By complementary slackness, the constraint is active:

$$d^\top (h_L + \Delta h^\star) = \tau. \tag{A25}$$

Using Equation (A21),

$$d^\top h_L + \lambda \|d\|_2^2 = \tau, \tag{A26}$$

and therefore

$$\lambda = \frac{\tau - d^\top h_L}{\|d\|_2^2} > 0. \tag{A27}$$

Substituting this into Equation (A21) gives

$$\Delta h^\star = \frac{\tau - d^\top h_L}{\|d\|_2^2} d. \tag{A28}$$

Combining the active and inactive cases yields

$$\boxed{\Delta h^\star = \frac{[\tau - d^\top h_L]_+}{\|d\|_2^2} d}, \qquad [x]_+ = \max(x, 0). \tag{A29}$$

Using $\tau = \rho q_A$,

$$\boxed{\Delta h^\star = \frac{[\rho q_A - d^\top h_L]_+}{\|d\|_2^2} d}, \qquad \hat{h}_L = h_L + \Delta h^\star, \tag{A30}$$

which is exactly the update stated in Proposition 1.

Because the objective of Equation (A14) is strictly convex, this solution is unique.

Minimum displacement. From Equation (A29),

$$\begin{aligned} \|\Delta h^\star\|_2 &= \frac{[\tau - d^\top h_L]_+}{\|d\|_2^2} \|d\|_2 \\ &= \frac{[\tau - d^\top h_L]_+}{\|d\|_2}. \end{aligned} \tag{A31}$$

This is precisely the Euclidean distance from h_L to the feasible half-space

$$\mathcal{F} = \{h : d^\top h \geq \tau\}. \tag{A32}$$

Thus ReCAP is not merely a perturbation parallel to the preference direction; it is the orthogonal projection of the final residual state onto the nearest point satisfying the required pairwise control.

Pairwise restoration after a triggered reversal. A ReCAP intervention is triggered only when the selected anchor candidate has been overtaken by the vanilla final prediction:

$$z_L(a) - z_L(b) < 0. \tag{A33}$$

By Equation (A7),

$$d^\top h_L < 0. \tag{A34}$$

At the anchor, $M_A > 0$, so Equation (A12) gives $q_A > 0$. Since $\rho > 0$,

$$\tau = \rho q_A > 0. \tag{A35}$$

Therefore,

$$d^\top h_L < 0 < \tau, \tag{A36}$$

which implies that the constraint is necessarily active for every triggered reversal. Hence

$$d^\top \hat{h}_L = \rho q_A. \tag{A37}$$

Applying Equation (A6) at the corrected state gives

$$\begin{aligned} \hat{z}(a) - \hat{z}(b) &= \frac{d^\top \hat{h}_L}{s(\hat{h}_L)} \\ &= \frac{\rho q_A}{s(\hat{h}_L)} \\ &= \rho \frac{s(h_A)}{s(\hat{h}_L)} M_A. \end{aligned} \tag{A38}$$

Every factor on the final line is positive: $\rho > 0$, $s(h_A) > 0$, $s(\hat{h}_L) > 0$, and $M_A > 0$. Therefore

$$\boxed{\hat{z}(a) > \hat{z}(b)}. \tag{A39}$$

This proves the pairwise restoration guarantee in Proposition 1.

Numerically degenerate directions. The mathematical solution assumes $d \neq 0$. The denominator in Equation (A30) is therefore exactly $\|d\|_2^2$; the optimization problem itself does not contain an additive stabilizer. In implementation, a numerically degenerate direction is skipped before applying the projection rather than changing the optimization objective by replacing the denominator with $\|d\|_2^2 + \epsilon$.

A.3 Scope of the Preservation Guarantee

The minimum-norm solution also gives a simple geometric preservation property. Let

$$P_\perp = I - \frac{dd^\top}{\|d\|_2^2} \tag{A40}$$

denote the orthogonal projector onto the subspace perpendicular to d.

Since

$$\Delta h^\star = \alpha d \tag{A41}$$

for some $\alpha \geq 0$, we have

$$P_\perp \Delta h^\star = \alpha P_\perp d = 0. \tag{A42}$$

Therefore,

$$\begin{aligned} P_\perp \hat{h}_L &= P_\perp (h_L + \Delta h^\star) \\ &= P_\perp h_L. \end{aligned} \tag{A43}$$

Equivalently, for every vector $v \in \mathbb{R}^D$ satisfying $v^\top d = 0$,

$$v^\top \hat{h}_L = v^\top h_L. \tag{A44}$$

Thus the intervention modifies only the one-dimensional residual component parallel to the selected pairwise direction d. All residual coordinates in the orthogonal complement of d are unchanged.

Importantly, this is a statement about the *pre-normalization residual representation*. It does not imply that every unrelated vocabulary logit or probability remains unchanged.

To see this explicitly, let

$$\hat{h}_L = h_L + \alpha d. \tag{A45}$$

For an arbitrary vocabulary token w,

$$\begin{aligned}\hat{z}(w) &= \frac{\bar{u}_w^\top (h_L + \alpha d)}{s(\hat{h}_L)} \\ &= \frac{\bar{u}_w^\top h_L}{s(\hat{h}_L)} + \frac{\alpha \bar{u}_w^\top d}{s(\hat{h}_L)}.\end{aligned} \tag{A46}$$

Since

$$z_L(w) = \frac{\bar{u}_w^\top h_L}{s(h_L)}, \tag{A47}$$

we obtain

$$\boxed{\hat{z}(w) = \frac{s(h_L)}{s(\hat{h}_L)} z_L(w) + \frac{\alpha \bar{u}_w^\top d}{s(\hat{h}_L)}}. \tag{A48}$$

Equation (A48) reveals two ways in which a nonselected vocabulary token can change.

First, all logits are affected by the common RMSNorm rescaling factor $s(h_L)/s(\hat{h}_L)$. Second, a token whose gain-adjusted unembedding vector is not orthogonal to d receives an additional token-specific change proportional to $\bar{u}_w^\top d$.

Therefore,

$$P_\perp \hat{h}_L = P_\perp h_L \tag{A49}$$

should not be interpreted as

$$\hat{z}(w) = z_L(w) \tag{A50}$$

for tokens $w \notin \{a, b\}$, nor does it imply preservation of the entire softmax distribution.

Pairwise restoration versus full-vocabulary recovery. For the corrected anchor a to become the full-vocabulary top-1 token, we require

$$\hat{z}(a) \geq \hat{z}(w) \qquad \forall w \in \mathcal{W}, \tag{A51}$$

where $\mathcal{W}$ denotes the full vocabulary. By the same RMSNorm identity,

$$\hat{z}(a) - \hat{z}(w) = \frac{(\bar{u}_a - \bar{u}_w)^\top \hat{h}_L}{s(\hat{h}_L)}. \tag{A52}$$

Hence Equation (A51) is equivalent to

$$(\bar{u}_a - \bar{u}_w)^\top \hat{h}_L \geq 0 \qquad \forall w \in \mathcal{W}. \tag{A53}$$

ReCAP explicitly enforces only the constraint associated with $w = b$:

$$(\bar{u}_a - \bar{u}_b)^\top \hat{h}_L = d^\top \hat{h}_L > 0. \tag{A54}$$

It does not impose Equation (A53) simultaneously for every other vocabulary item. Consequently, another token $w \neq b$ may still remain above, or overtake, the selected anchor after the correction.

This distinction motivates reporting three different empirical quantities:

1. whether the selected a-versus-b pair is restored;
2. whether the selected anchor becomes the full-vocabulary top-1 token;
3. whether the resulting decoded answer is correct under the benchmark evaluator.

These quantities should not be treated as interchangeable.

Minimum residual displacement versus distributional change. Finally, the optimality established in Equation (A14) is specific to the Euclidean geometry of the residual state:

$$\Delta h^\star = \arg\min_{\Delta h} \|\Delta h\|_2 \tag{A55}$$

among residual perturbations satisfying the specified pairwise constraint. It does not imply

$$\Delta h^\star = \arg\min_{\Delta h} D_{\mathrm{KL}}\big(p_L \,\|\, p_{\Delta h}\big), \tag{A56}$$

nor does it imply minimum total variation or minimum change to every individual vocabulary probability.

Accordingly, the "minimum-norm" property of ReCAP refers precisely to the magnitude of the residual-space intervention under the chosen pairwise restoration constraint. Full-vocabulary recovery and distributional distortion are empirical properties and are evaluated separately.

B Diagnostic Protocol and Robustness

This section specifies the sampling, intervention, and replay procedures used for the Grounded Control Reversal (GCR) diagnosis. The goal is to separate four questions that are conceptually distinct: whether a correct intermediate preference persists, whether that preference depends on visual computation, whether the observed effect is specific to the visual route rather than a generic perturbation effect, and how the subsequent loss of control is allocated across the three residual-update routes.

B.1 Diagnostic Cohorts and Fixed Candidate Pairs

The mechanism analysis conditions on the 300 incorrect predictions per benchmark specified in the main text. For the default setting, 184 POPE-Adversarial examples and 164 HallusionBench examples satisfy the layerwise reversal criterion. After the visual-dependence test, 144 and 124 examples, respectively, remain as grounded reversals. These 268 examples define the grounded diagnostic cohort used by the supplementary mechanism-specific analyses in this appendix.

For binary diagnostics, we use the model's native chat template with the fixed instruction *"Answer the question with exactly one word: Yes or No."* The tokenization of the canonical answer forms is verified at the actual assistant generation boundary and their token IDs are recorded.

For each diagnostic example, let c denote the gold answer token and let b denote the model's factual incorrect final answer token. Both tokens are fixed throughout the layerwise analysis. In particular, the competitor is not reselected independently at each layer. Thus,

$$m_l = z_l(c) - z_l(b) \tag{A57}$$

always measures the evolution of the same pairwise competition.

Gold labels are used only to construct this offline diagnostic. Neither c nor any ground-truth route label is available to the online ReCAP decoder.

B.2 Persistence and Positive-Margin Robustness

The default GCR definition requires the correct candidate to be preferred over the final incorrect competitor for two consecutive layers. To test whether the phenomenon depends on isolated or near-zero intermediate readouts, we generalize the criterion using a persistence length r and positive-margin threshold $\delta \geq 0$.

Define

$$\mathcal{E}_{r,\delta} = \{l : r \leq l < L,\ m_{l-j} > \delta \text{ for all } j = 0, \ldots, r-1\}\,. \tag{A58}$$

A candidate reversal is identified when

$$\mathcal{E}_{r,\delta} \neq \varnothing \qquad \text{and} \qquad m_L < 0. \tag{A59}$$

The corresponding diagnostic anchor is

$$A = \arg\max_{l \in \mathcal{E}_{r,\delta}} m_l. \tag{A60}$$

Ties are resolved toward the later layer, following the deterministic rule used in the diagnostic implementation.

The default setting is $r = 2$ and $\delta = 0$. Increasing r requires the correct preference to survive for a longer contiguous span, whereas increasing δ excludes weak positive margins close to the pairwise decision boundary. Table A1 reports both the number of candidate reversals and the subset that additionally passes the visual-dependence test.

Table A1: Robustness of the GCR diagnosis under stronger persistence and positive-margin requirements. GCR denotes layerwise candidate reversals and G-GCR denotes the subset passing the visual-dependence test. Counts are measured among the 300 diagnostic errors from each benchmark rather than over the full benchmark.

		POPE-Adversarial		HallusionBench	
r	δ	GCR	G-GCR	GCR	G-GCR
1	0.00	217	163	198	145
2	0.00	184	144	164	124
3	0.00	168	133	149	115
2	0.10	174	136	155	118
2	0.25	156	122	142	109
3	0.25	143	114	131	101

As expected, stronger persistence and margin requirements reduce the number of detected cases. The purpose of this analysis is not to replace the default operational definition, but to test whether the reported phenomenon disappears once weak or short-lived intermediate preferences are excluded.

B.3 Exact Residual-Route Decomposition

We next specify the residual decomposition used by both the visual-dependence intervention and the post-anchor route replay.

At decoder block l and prediction position p, let $\mathcal{V}$ denote the positions corresponding to image features and let $\mathcal{T}$ contain all remaining unmasked source positions, including system tokens, textual instructions, previously generated tokens, and non-image special tokens.

For query head r, let $\alpha_{l,r,p,i}$ be the native attention weight from prediction position p to source position i, and let $v_{l,r,i}$ denote the corresponding value vector. The projected contribution from source group $\mathcal{X} \in \{\mathcal{V}, \mathcal{T}\}$ is

$$a_l^X = \sum_r W_{l,r}^O \left(\sum_{i \in \mathcal{X}} \alpha_{l,r,p,i} v_{l,r,i} \right), \qquad X \in \{V, T\}. \tag{A61}$$

The complete attention update satisfies

$$a_l = a_l^V + a_l^T. \tag{A62}$$

The partition is applied after computing attention with the model's original softmax denominator over all unmasked source positions. Visual and nonvisual contributions are therefore not independently renormalized. For grouped-query attention, query heads use the backbone's native mapping to key/value heads.

For a pre-normalized sequential attention–FFN block, the prediction-position update is

$$r_l = h_{l-1} + a_l^V + a_l^T, \qquad f_l = F_l(\mathrm{Norm}_{2,l}(r_l)), \qquad h_l = r_l + f_l. \tag{A63}$$

Consequently, after a factual anchor A,

$$h_L - h_A = \sum_{l=A+1}^{L} \left(a_l^V + a_l^T + f_l \right). \tag{A64}$$

Thus visual attention, nonvisual attention, and the FFN exhaust the additive residual updates at the analyzed prediction position under the stated architecture. This is an architectural decomposition of residual updates, not a claim that the three routes correspond to semantically pure sources. For example, values at visual-token positions may already contain textual context, while nonvisual positions may contain visual information integrated at earlier layers.

B.4 Visual-Value Intervention and Matched Controls

A layerwise reversal alone does not establish that the intermediate correct preference depends on visual computation. We therefore perform a same-image internal intervention on every candidate reversal while holding its candidate pair and diagnostic anchor fixed.

For a candidate reversal with factual anchor A, we replay the same image and prompt through blocks $1{:}A$. At each affected block, the projected visual contribution to nonvisual query positions is attenuated according to

$$a_l^V \longleftarrow (1-\eta)a_l^V, \qquad \eta \in [0,1]. \tag{A65}$$

The full visual-value cut corresponds to $\eta = 1$.

All downstream states affected by the intervention are recomputed. Neither the candidate pair (c, b) nor the factual anchor A is reselected after intervention. If $m_A^{(\eta)}$ denotes the intervened anchor margin, the default visual-dependence effect is defined at $\eta = 1$ as

$$G^{\text{vis}} = m_A - m_A^{(1)}. \tag{A66}$$

A candidate reversal is classified as visually grounded when $G^{\text{vis}} > 0$.

To distinguish visual-route dependence from generic sensitivity to hidden-state perturbation, we evaluate three matched controls at the same layers and query positions:

1. **Sham replay:** recompute the trajectory without modifying any residual contribution;
2. **Norm-matched random perturbation:** remove a perturbation with the same norm as the corresponding visual contribution but along a seeded random direction;
3. **Norm-matched nonvisual perturbation:** apply the same perturbation magnitude along the local nonvisual-attention direction.

Zero-norm control directions are skipped and recorded. Importantly, all control effects are computed on the complete candidate-reversal population *before* conditioning on $G^{\text{vis}} > 0$. This prevents the grounding criterion itself from inducing an artificially positive average intervention effect.

Table A2: Anchor-margin reduction under visual-value suppression and matched controls, evaluated before the $G^{\text{vis}} > 0$ grounding gate. The cohorts contain 184 POPE-Adversarial and 164 HallusionBench candidate reversals. Brackets denote image-cluster bootstrap 95% confidence intervals. Effects are reported in logit units.

Intervention	POPE-Adversarial	HallusionBench
Sham replay	$0.00\ [0.00, 0.00]$	$0.00\ [0.00, 0.00]$
Norm-matched random direction	$0.12\ [-0.12, 0.26]$	$0.16\ [-0.08, 0.32]$
Norm-matched nonvisual direction	$0.58\ [0.20, 0.83]$	$0.54\ [0.24, 0.77]$
Visual-value cut	$1.18\ [0.91, 1.39]$	$1.25\ [0.89, 1.46]$

A positive G^{vis} establishes that the fixed intermediate preference is sensitive to the specified visual-value route. The matched controls test whether this effect is stronger than the effect of perturbations with comparable magnitude. This intervention does not, by itself, localize the supporting image region or establish that every component transmitted through visual-token positions is semantically visual.

B.5 Post-Anchor Coalition Replay and Route Assignment

After establishing a visually grounded reversal, we attribute the subsequent control loss using interventions applied only after the factual anchor. Let

$$N = \{V, T, F\} \tag{A67}$$

denote the visual-attention, nonvisual-attention, and FFN routes. We evaluate all eight coalitions $S \subseteq N$.

The replay begins from the factual anchor state h_A at the current prediction position. Earlier prefix positions retain their factual states and cached keys and values. At each block $l > A$, the current prediction-position query and its current-position key/value are recomputed from the intervened trajectory.

For coalition S, define

$$r_l^S = h_{l-1}^S + \mathbf{1}[V \in S] a_l^{V,S} + \mathbf{1}[T \in S] a_l^{T,S}, \tag{A68}$$

$$h_l^S = r_l^S + \mathbf{1}[F \in S] F_l\big(\mathrm{Norm}_{2,l}(r_l^S)\big)\,. \tag{A69}$$

Enabled branches are recomputed from the current intervened state. Route gates are applied to the projected branch outputs, and attention over the remaining source positions is not renormalized after a route is disabled.

Two endpoint identities serve as implementation checks. With all post-anchor routes disabled,

$$h_L^{\varnothing} = h_A, \tag{A70}$$

whereas enabling all three routes must reproduce the factual endpoint up to numerical tolerance:

$$h_L^N \approx h_L. \tag{A71}$$

Let

$$v(S) = z_L^S(c) - z_L^S(b) \tag{A72}$$

denote the final fixed-pair margin under coalition S. The Shapley contribution of route $j \in N$ is

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{j\}) - v(S)\right]. \tag{A73}$$

Because

$$v(\varnothing) = m_A, \qquad v(N) = m_L, \tag{A74}$$

Shapley efficiency gives

$$\sum_{j \in \{V,T,F\}} \phi_j = m_L - m_A. \tag{A75}$$

For mechanism-specific evaluation, define the normalized negative contribution of route j as

$$R_j = \frac{[-\phi_j]_+}{\sum_{k \in \{V,T,F\}} [-\phi_k]_+}. \tag{A76}$$

A case is assigned to route

$$j^\star = \arg\max_j R_j \tag{A77}$$

when

$$R_{j^\star} \geq 0.5. \tag{A78}$$

Otherwise it is labeled *Mixed*. Exact ties are resolved using a fixed deterministic rule recorded in the implementation.

Route labels are computed once from the vanilla factual trajectory and are then frozen. All recovery methods are evaluated using the same route assignment; route labels are never recomputed after a method changes the generated state. Consequently, route-specific recovery measures how different decoders behave on the same pre-existing failure groups rather than allowing each method to redefine the groups on which it is evaluated.

C Algorithm and Reproducible Implementation

This section gives the complete online decoding procedure and the implementation details required to reproduce ReCAP. All operations are performed at the current generation position using states from the model's standard forward trajectory.

C.1 One-Step ReCAP Decoding

Algorithm A1 summarizes one generation step. The final-model shortlist is reused for all intermediate candidate readouts. Let $W_U[C]$ denote the rows of the output head corresponding to candidate set C.

Algorithm A1 ReCAP at one decoding step

Require: Image-conditioned prefix, frozen model, candidate budget K, restoration strength ρ, searched layers $\mathcal{L}_c$
Ensure: Next-token distribution

Run the standard forward trajectory and obtain h_L, z_L, and $p_L = \mathrm{softmax}(z_L)$
Cache h_l and visual residual contributions a_l^V for $l \in \mathcal{L}_c$, together with the predecessor of the first searched layer
$b \leftarrow \arg\max_w z_L(w)$
$C \leftarrow \mathrm{TopK}(z_L, K)$
for each cached layer l **do**
    $z_l(C) \leftarrow W_U[C]\,\mathrm{Norm}_f(h_l)$
    $a_l \leftarrow \arg\max_{w \in C} z_l(w)$
    $M_l \leftarrow z_l(a_l) - z_l(b)$
end for
$\mathcal{E} \leftarrow \varnothing$
for $l \in \mathcal{L}_c$ **do**
    if $a_l = a_{l-1} \neq b$ **and** $M_l > 0$ **and** $M_{l-1} > 0$ **then**
        $d_l \leftarrow \gamma \odot (u_{a_l} - u_b)$
        $G_l \leftarrow \dfrac{d_l^\top a_l^V}{\|d_l\|_2 \|a_l^V\|_2 + \epsilon_g}$
        if $G_l > 0$ **and** $\|d_l\|_2^2 > \epsilon_d$ **then**
            $\mathcal{E} \leftarrow \mathcal{E} \cup \{l\}$
        end if
    end if
end for
if $\mathcal{E} = \varnothing$ **then**
    return p_L
end if
$A \leftarrow \arg\max_{l \in \mathcal{E}} M_l G_l$ ▷ ties resolved toward later depth
$a \leftarrow a_A, \quad d \leftarrow d_A$
if $z_L(a) \geq z_L(b)$ **then**
    return p_L
end if
$q_A \leftarrow d^\top h_A$
$\alpha \leftarrow \dfrac{[\rho q_A - d^\top h_L]_+}{\|d\|_2^2}$
$\hat{h}_L \leftarrow h_L + \alpha d$
$\hat{z} \leftarrow W_U\,\mathrm{Norm}_f(\hat{h}_L)$
if $\hat{h}_L$ or $\hat{z}$ contains a non-finite value **then**
    return p_L
end if
return $\mathrm{softmax}(\hat{z})$

The vanilla readout z_L is first used to construct the shortlist, identify the final competitor, and test whether an eligible intermediate preference has been reversed. When the trigger is active, ReCAP modifies the final residual state h_L and recomputes the full-vocabulary output distribution from $\hat{h}_L$.

The correction does not modify previously cached keys or values. After the corrected next token is selected, autoregressive generation continues normally from the resulting prefix.

C.2 Evaluation Protocol and Default Configuration

We evaluate Qwen2.5-VL-7B-Instruct and LLaVA-1.5-7B with their native image processors and shared prompts, token limits, and greedy decoding. POPE uses the MSCOCO Random, Popular, and Adversarial splits with both answer polarities and 3,000 questions per split. HallusionBench uses image-bearing questions and native grouped metrics. CHAIR is computed on a fixed 500-image COCO subset with sentence- and instance-level hallucination, object recall, and output length; all reported CHAIR values are averaged over three experimental runs. MMBench uses its multiple-choice protocol and CircularEval, while MMStar uses vision-indispensable, low-leakage samples. VLMBias retains native accuracy and bias metrics, and PhD retains native accuracy and index definitions. Official baseline defaults are used when available; transferred layer ranges are specified by normalized depth.

Table A3 lists the parameters that directly determine ReCAP decoding. The same K, ρ, persistence rule, and normalized search interval are used across the primary backbones.

Table A3: Default ReCAP decoding configuration.

Setting	Specification
Candidate budget	$K = 10$
Restoration strength	$\rho = 0.1$
Persistence	Two adjacent layers with the same candidate
Searched layers	$\lceil 0.25L \rceil : \lfloor 0.875L \rfloor$
Additional cached layer	Predecessor of the first searched layer
Visual-support criterion	$G_l > 0$
Support constant	$\epsilon_g = 10^{-8}$
Degenerate-direction threshold	$\|d_l\|_2^2 \le \epsilon_d$, $\epsilon_d = 10^{-12}$
Projection arithmetic	FP32
Decoding	Greedy decoding
Candidate-score tie	Lower token ID
Anchor-score tie	Later layer

The model's native final RMSNorm gain γ is used in the pairwise direction $d_l = \gamma \odot (u_{a_l} - u_b)$. The threshold ϵ_d is used only to identify numerically degenerate directions; the projection denominator remains $\|d_l\|_2^2$.

C.3 Visual-Contribution Extraction

ReCAP uses the projected visual residual contribution a_l^V when computing the support score G_l. For prediction position p, this contribution is

$$a_l^V = W_l^O \,\mathrm{Concat}_r \left(\sum_{i \in \mathcal{V}} \alpha_{l,r,p,i} v_{l,r,i} \right), \tag{A79}$$

where $\mathcal{V}$ indexes visual-token positions, $\alpha_{l,r,p,i}$ is the native attention weight, and $v_{l,r,i}$ is the corresponding value vector.

For attention backends that expose the required attention row, Equation (A79) is evaluated directly from the current forward pass. For fused implementations such as FlashAttention, we reconstruct only the attention row associated with the current prediction query. The reconstruction uses the model's native queries, cached keys and values, positional transformation, attention mask, and grouped-query head mapping.

The softmax is evaluated over all valid source positions. The visual-token entries are then used to form the visual value sum, followed by the model's native output projection W_l^O. A small audit batch is evaluated with an explicit-attention backend to verify the reconstructed a_l^V numerically.

C.4 Online Computation and Memory

Let $m = |\mathcal{L}_c|$, and let N denote the current prefix length, D the residual dimension, H_q the number of query heads, d_h the head dimension, and $|\mathcal{W}|$ the vocabulary size.

The candidate-only intermediate readouts require $O(mKD)$ additional computation. When the visual contribution is reconstructed from a fused attention backend, the current-query attention rows require approximately $O(mH_qNd_h)$ operations, together with the corresponding output projections. The residual correction itself requires $O(D)$, and a triggered correction recomputes one full-vocabulary output head at cost $O(|\mathcal{W}|D)$.

Caching the searched residual states and their visual contributions requires $O(mD)$ additional vector storage beyond the model's native KV cache.

Thus each decoding step uses one transformer trajectory, with candidate readouts, visual-support extraction, and the optional final-state correction computed from that trajectory. End-to-end latency measurements include all of these online operations.

D Anchor Reliability and Recovery Analysis

This section examines the two empirical conditions that connect the GCR diagnosis to successful ReCAP decoding. We first test whether the online visual-support score predicts the intervention-based visual dependence used in the diagnosis. We then trace recovery through candidate coverage, triggering, anchor selection, and full-vocabulary decoding, followed by a paired analysis over the complete POPE evaluation set.

D.1 Validating the Online Visual-Support Proxy

The diagnostic experiment measures visual dependence through the same-image intervention

$$G^{\text{vis}} = m_A - m_A^{\text{cut}}, \tag{A80}$$

whereas online ReCAP uses the candidate-specific alignment

$$G_A = \frac{d^\top a_A^V}{\|d\|_2 \|a_A^V\|_2 + \epsilon_g}. \tag{A81}$$

We evaluate their relationship on a separate 400-example audit set, balanced across POPE-Adversarial and HallusionBench and containing both correct and incorrect vanilla predictions. For each eligible example, we select one persistent provisional candidate using the final shortlist and intermediate margin before applying the visual-support gate. The candidate pair and anchor are then fixed for all scores and interventions.

For each example, the binary reference label is

$$y_{\text{vis}} = \mathbf{1}[G^{\text{vis}} > 0]. \tag{A82}$$

We compare three online scores: total visual attention mass, the raw candidate-direction contribution $d^\top a_A^V$, and the normalized alignment G_A. Table A4 reports their ability to identify intervention-supported cases.

Table A4: Prediction of intervention-based visual dependence on the 400-example pre-gate audit set. AUROC and AUPRC use $\mathbf{1}[G^{\text{vis}} > 0]$ as the reference label.

Score	AUROC	AUPRC
Visual attention mass	0.62	0.71
Raw directional contribution $d^\top a_A^V$	0.73	0.81
Normalized alignment G_A	**0.80**	**0.86**

The candidate-specific scores provide stronger correspondence with the intervention-based reference than aggregate visual attention. The normalized alignment G_A gives the strongest ranking performance and is therefore used as the online visual-support signal in ReCAP.

D.2 Full-Set Recovery and Correct-Answer Retention

We finally evaluate paired outcome transitions over all 9,000 Qwen POPE questions. Each example is classified according to its vanilla and ReCAP correctness, producing the transition matrix in Table A5.

Table A5: Paired correctness transitions on the complete 9,000-question Qwen POPE evaluation set.

Vanilla outcome	ReCAP correct	ReCAP incorrect	Total
Correct	7,482	152	7,634
Incorrect	562	804	1,366
Total	8,044	956	9,000

The vanilla model is correct on $7634/9000 = 84.82\%$ of the evaluation set, while ReCAP is correct on $8044/9000 = 89.38\%$.

Among initially correct predictions, ReCAP retains

$$\frac{7482}{7634} = 98.01\% \tag{A83}$$

and changes $152/7634 = 1.99\%$ to incorrect outcomes. Among the 1,366 initial errors, it recovers

$$\frac{562}{1366} = 41.14\%. \tag{A84}$$

The paired transition balance is therefore

$$100 \times \frac{562 - 152}{9000} = 4.56 \tag{A85}$$

percentage points, matching the aggregate POPE accuracy increase from 84.82% to 89.38%.

E Matched Correction Geometry

The component analysis in the main text compares several correction rules using the same selected anchors and trigger decisions. This section examines these corrections at a fixed decoding step, focusing on the geometry of the intervention and the resulting change in the output distribution.

For each audited triggered step, we freeze the image, generated prefix, anchor layer A, selected candidate a, vanilla competitor b, and trigger decision. Let

$$m_L = z_L(a) - z_L(b), \qquad M_A = z_A(a) - z_A(b) > 0, \tag{A86}$$

and let

$$\mu_R = \hat{z}_R(a) - \hat{z}_R(b) > 0 \tag{A87}$$

denote the pairwise margin produced by ReCAP.

E.1 Matched Correction Rules

Residual interpolation. We interpolate between the final state and the selected anchor state:

$$h_I = (1 - \lambda)h_L + \lambda h_A. \tag{A88}$$

The coefficient

$$\lambda = \frac{\rho q_A - d^\top h_L}{q_A - d^\top h_L} \tag{A89}$$

gives

$$d^\top h_I = \rho q_A. \tag{A90}$$

Residual interpolation and ReCAP therefore satisfy the same anchor-derived constraint in residual space.

Anchor-logit addition. We add the complete anchor-layer logit vector to the final logits:

$$z_D = z_L + \beta z_A, \tag{A91}$$

where

$$\beta = \frac{\mu_R - m_L}{M_A}. \tag{A92}$$

The resulting pairwise margin is

$$z_D(a) - z_D(b) = \mu_R. \tag{A93}$$

Pairwise logit bias. A direct pairwise correction modifies only the selected anchor and competitor logits:

$$z_B = z_L + \frac{\mu_R - m_L}{2}(e_a - e_b), \tag{A94}$$

where e_w is the vocabulary basis vector associated with token w. This correction also attains

$$z_B(a) - z_B(b) = \mu_R. \tag{A95}$$

The first comparison matches ReCAP in residual-space control, while the two logit-space corrections match its achieved normalized pairwise margin.

E.2 Residual Displacement and Distributional Change

For residual-space interventions, we measure the relative state change

$$r_h = \frac{\|h' - h_L\|_2}{\|h_L\|_2}. \tag{A96}$$

For all correction rules, distributional change is measured from the same image and generated prefix using

$$D_{\mathrm{KL}}(p_L \| p') = \sum_{w \in \mathcal{W}} p_L(w) \log \frac{p_L(w)}{p'(w)}, \tag{A97}$$

where $p_L = \mathrm{softmax}(z_L)$ and p' is the corrected full-vocabulary distribution.

Table A6: Matched correction audit at fixed prefixes, anchors, and triggers. Residual interpolation and ReCAP share the target $d^\top h' = \rho q_A$; the logit-space controls match ReCAP's achieved pairwise margin μ_R. Relative norm is reported for residual-space interventions.

Correction	Matched quantity	Rel. norm	$D_{\mathrm{KL}}(p_L \| p')$
Residual interpolation	$d^\top h' = \rho q_A$	0.071	0.028
Anchor-logit addition	$\mu = \mu_R$	–	0.032
Pairwise logit bias	$\mu = \mu_R$	–	0.005
ReCAP	$d^\top h' = \rho q_A$	**0.014**	0.009

Under the shared residual-space target, ReCAP requires a smaller relative hidden-state displacement than interpolation, consistent with the minimum-norm solution in Appendix A. The logit-space controls provide a complementary view of output-space change: pairwise logit bias produces the smallest KL divergence, while anchor-logit addition changes the full anchor distribution.

All distributional quantities are computed at identical prefixes. End-to-end task performance for these matched correction rules is reported in the main component analysis.

F Hyperparameter Sensitivity Analysis

F.1 Sensitivity to Candidate Budget K

The candidate budget K determines the size of the final-distribution shortlist $\mathcal{C} = \mathrm{TopK}(z_L, K)$ from which intermediate anchor candidates are recovered. A very small K may exclude a grounded

candidate that has lost substantial probability by the final layer, whereas an unnecessarily large K admits increasingly weak alternatives and increases the cost of candidate-only intermediate readouts. We vary

$$K \in \{3, 5, 10, 20, 50\},$$

while fixing $\rho = 0.1$ and using two-layer persistence.

Table A7: Sensitivity to candidate budget K on Qwen2.5-VL-7B-Instruct. All other ReCAP hyperparameters are fixed to their default values. ↓ indicates lower is better and ↑ indicates higher is better.

K	POPE F1 ↑	Hall. qAcc ↑	CHAIR$_i$ ↓
3	87.84	70.41	7.43
5	**89.12**	70.26	7.05
10	88.93	**71.47**	**6.71**
20	86.78	69.31	7.83
50	86.37	68.88	8.10

Performance is strongest at small-to-moderate candidate budgets. $K = 5$ gives the highest POPE F1, whereas $K = 10$ gives the best HallusionBench qAcc and CHAIR$_i$. The smaller $K = 3$ may exclude grounded alternatives that have lost substantial final-layer probability, while $K = 20$ and $K = 50$ degrade all three metrics. Larger candidate sets also introduce lower-probability alternatives into anchor search and increase the $O(|\mathcal{L}_c|KD)$ candidate-readout cost. We therefore use $K = 10$, which provides the best joint performance across the three benchmarks.

F.2 Sensitivity to Restoration Strength ρ

The restoration coefficient ρ controls how much of the anchor-derived pairwise control is recovered after a reversal is detected. Recall that ReCAP solves

$$\min_{\Delta h} \frac{1}{2}\|\Delta h\|_2^2 \qquad \text{s.t.} \qquad d^\top(h_L + \Delta h) \geq \rho q_A,$$

where $q_A = d^\top h_A > 0$ is the pairwise control strength at the selected anchor. Thus, ρ does not determine whether an intervention is triggered; it determines the target margin imposed after a reversal has already been identified. We evaluate

$$\rho \in \{0.02, 0.05, 0.10, 0.20, 0.40, 0.80\},$$

with $K = 10$ and two-layer persistence.

Table A8: Sensitivity to restoration strength ρ. The intervention trigger and selected anchor are determined independently of ρ; only the amount of restored pairwise control changes. Bold metric values mark the best result in each column, while bold ρ denotes the default setting.

ρ	POPE F1 ↑	Hall. qAcc ↑	CHAIR$_i$ ↓
0.02	86.39	69.63	8.35
0.05	88.54	**72.37**	6.94
0.10	**88.93**	71.47	6.71
0.20	87.85	71.39	**6.62**
0.40	85.42	68.92	9.08
0.80	84.56	67.84	10.61

The results favor moderate restoration strengths but show metric-specific optima. $\rho = 0.05$ achieves the highest HallusionBench qAcc, $\rho = 0.10$ gives the highest POPE F1, and $\rho = 0.20$ gives the lowest CHAIR$_i$. Very small restoration at $\rho = 0.02$ is insufficient across all three metrics, whereas aggressive restoration at $\rho = 0.4$ and 0.8 causes consistent degradation because the stronger constraint requires a larger residual displacement and can induce greater changes to non-target logits. We use $\rho = 0.10$ as the balanced default because it is optimal on POPE and remains competitive on HallusionBench and CHAIR.

F.3 Sensitivity to Layer Persistence

The default ReCAP anchor criterion requires the same provisional candidate to maintain a positive advantage over the final competitor across two adjacent layers. We generalize this condition to n-layer persistence. For an endpoint layer l, a candidate is persistence-eligible when

$$a_l = a_{l-1} = \cdots = a_{l-n+1} \neq b, \qquad M_{l-j} > 0 \quad \forall j \in \{0, \ldots, n-1\},$$

together with the original visual-support requirement $G_l > 0$. Increasing n therefore makes anchor selection more conservative: transient intermediate preferences are progressively excluded, but genuinely grounded candidates that persist for only a limited portion of the trajectory may also be rejected.

We test

$$n \in \{1, 2, 3, 4, 5\},$$

while retaining $K = 10$ and $\rho = 0.1$.

Table A9: Sensitivity to the number of consecutive layers required for candidate persistence. $n = 1$ removes the cross-layer persistence requirement, whereas larger values impose increasingly strict trajectory consistency.

Persistence n	POPE F1 ↑	Hall. qAcc ↑	CHAIR$_i$ ↓
1	87.46	69.82	7.58
2	**88.93**	**71.47**	**6.71**
3	88.61	71.22	6.89
4	87.13	69.73	8.16
5	87.69	70.21	7.89

Removing persistence ($n = 1$) degrades all three metrics relative to the two-layer default, consistent with the possibility that a single-layer preference reflects a transient readout fluctuation. Three-layer persistence remains competitive, indicating that many useful anchors extend beyond the minimum interval. However, stricter requirements at $n = 4$ and $n = 5$ reduce anchor coverage and lower performance, with $n = 4$ producing the weakest overall results. Two-layer persistence therefore provides the best balance between rejecting isolated fluctuations and retaining visually supported reversals.

F.4 Overall Robustness

Across the three sensitivity studies, the strongest operating region is concentrated around $K = 5$–10, $\rho = 0.05$–0.20, and two to three persistence layers. The degradation patterns correspond to the intended roles of the parameters: K trades candidate coverage against shortlist selectivity, ρ trades recovery strength against intervention magnitude, and persistence trades anchor reliability against anchor availability. We therefore retain the common setting

$$K = 10, \qquad \rho = 0.1, \qquad n = 2$$

for all main experiments rather than selecting benchmark-specific values.