FROM DETECTION TO CONTROL: HALLUCINATION HEADS AS TEXT-SIDE ACTUATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models remain prone to object hallucination, generating object mentions unsupported by visual evidence. Recent work shows that intervening on particular attention heads can mitigate hallucination, but why these heads constitute effective control points remains unclear. We show that raw text-side reliance in hallucination-related heads is only weakly diagnostic at the token level, even though interventions on selected heads affect hallucinated predictions much more strongly than grounded ones. Hallucinated object predictions are more text-reliant on average, but their distributions overlap substantially with those of grounded predictions. In contrast, attenuating selected text-side pathways disproportionately reduces hallucinated-object probabilities while minimally affecting grounded predictions. We therefore reinterpret these heads as **text-side actuators**: pathways whose text-side contribution exerts selectively stronger influence on unsupported continuations. Building on this finding, we introduce **DEACT**, a training-free inference-time method with model-specific calibration that identifies actuator heads using complementary pathway-strength and hallucination-contrastive statistics and regulates their text-side contribution according to the current decoding state. Across multiple LVLM architectures, DEACT consistently reduces object hallucination while maintaining grounded-generation quality. More broadly, our results show that weak **diagnostic separability** can coexist with strong **causal controllability**, suggesting that effective hallucination mitigation need not depend on reliable online failure detection.

1 INTRODUCTION

Large vision-language models (LVLMs) have achieved remarkable progress in multimodal perception and generation (Liu et al., 2023; Lin et al., 2024; Dai et al., 2023; Li et al., 2023a), yet they remain susceptible to *object hallucination*, producing object mentions that are linguistically plausible but unsupported by visual evidence (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023; Gunjal et al., 2024). A growing body of work mitigates this failure at inference time by modifying decoding distributions, strengthening visual signals, preserving image-token evidence, or redirecting attention toward visually relevant regions (Leng et al., 2024; Huang et al., 2024; Liu et al., 2024c; Yin et al., 2025; Jiang et al., 2025; Zhao et al., 2026). More recently, attention-based intervention methods have shown that suppressing a subset of hallucination-related heads can substantially reduce unsupported generation (Yang et al., 2025). However, why these heads constitute effective intervention targets remains unclear: statistical association with hallucination does not establish whether a head provides a reliable diagnostic signal or instead forms a pathway through which language-side context influences unsupported generation.

A natural hypothesis is that hallucination-related heads act as online detectors. Yet strong textual reliance is not specific to hallucination. Prior work has linked hallucination to excessive language-side influence or insufficient visual grounding (Leng et al., 2024; Huang et al., 2024; Liu et al., 2024c; Yin et al., 2025; Jiang et al., 2025; Zhao et al., 2026), while textual context also supports syntax, coherence, referential consistency, and grounded generation. We find that raw text-side reliance in these heads is only weakly diagnostic at the token level. Hallucinated object predictions are more text-reliant on average, but their distributions overlap substantially with those of grounded predictions. In contrast, suppressing selected text-side pathways disproportionately decreases hallucinated-object probabilities while minimally affecting grounded predictions, and this asymmetry persists across

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.

matched control pools and a reproduced prior hallucination-head pool (Yang et al., 2025). We therefore interpret these heads as **text-side actuators**: pathways whose text-side contribution exerts selectively stronger influence on unsupported continuations. This observation separates **diagnostic separability** from **causal controllability**: a pathway can be difficult to use for online diagnosis while remaining useful for targeted intervention.

Motivated by this finding, we introduce **DEACT** (**D**ynamic **E**xponential **ACT**uator Suppression), a training-free inference-time method with model-specific calibration that identifies actuator heads and regulates their text-side contribution during decoding. Unlike static head suppression (Yang et al., 2025), DEACT separates *where* to intervene from *when* and *how strongly* to do so, using each actuator's bounded text-reliance state as a control variable. Across multiple LVLM architectures and evaluation settings, DEACT consistently reduces object hallucination while maintaining grounded-generation quality. On LLaVA-1.5-7B, DEACT reduces CHAIR_s from 54.6 to 29.6 and CHAIR_i from 14.8 to 7.3 while improving object-level F1.

Our contributions are summarized as follows:

- We show that raw text-side reliance provides only weak token-level discrimination between hallucinated and grounded object predictions, despite higher average reliance for hallucinated objects.
- Through controlled text-side interventions and matched head-pool comparisons, we show that selected heads exert disproportionately stronger influence on hallucinated predictions, motivating their interpretation as **text-side actuators**.
- We propose **DEACT**, a training-free inference-time method with model-specific calibration that combines actuator localization with state-dependent regulation, and demonstrate consistent hallucination reduction across LVLM architectures.

2 Related Work

2.1 Hallucination in Large Vision-Language Models

Hallucination remains a persistent failure mode in large vision-language models, particularly in open-ended generation, where linguistic plausibility can diverge from visual evidence. Early work on image captioning formalized object hallucination and introduced annotation-based evaluation protocols (Rohrbach et al., 2018), while subsequent benchmarks extended evaluation to instruction-tuned LVLMs and broader multimodal settings (Li et al., 2023b; Wang et al., 2023; Gunjal et al., 2024). Prior studies commonly associate hallucination with insufficient visual grounding and excessive language-side influence (Leng et al., 2024; Huang et al., 2024; Liu et al., 2024c; Yin et al., 2025). However, these analyses primarily characterize when hallucination occurs rather than the internal computational pathways through which unsupported generation is mediated.

2.2 Training-Free Hallucination Mitigation

A substantial line of work mitigates hallucination at inference time without updating model parameters. Visual Contrastive Decoding (Leng et al., 2024) contrasts predictions under intact and perturbed visual inputs, while OPERA (Huang et al., 2024) regulates over-trusted generation patterns. Other methods intervene more directly on multimodal information flow: PAI (Liu et al., 2024c) strengthens attention to visual tokens, ClearSight (Yin et al., 2025) enhances intermediate visual representations, TARAC (Jiang et al., 2025) preserves visual evidence across decoding, and VGA (Zhao et al., 2026) guides attention toward visually relevant regions. These approaches primarily improve grounding by strengthening, preserving, or redirecting visual evidence. In contrast, our work examines whether unsupported predictions exhibit selective dependence on particular *text-side* pathways.

2.3 Attention-Head Analysis and Causal Intervention

Recent work has examined hallucination through the internal structure of transformer attention. AD-HH (Yang et al., 2025) uses modular attribution and intervention to identify hallucination-related heads and shows that suppressing their text-side contribution can reduce object hallucination. Related methods dynamically suppress visually weak heads (Sarkar et al., 2025), analyze head-level

sensitivity to visual context (He et al., 2025), or leverage such sensitivity to construct causal interventions over decoder representations (Ji et al., 2026). More recent mechanistic studies further identify attention heads with causal roles in specific hallucination regimes and show that targeted interventions can alter hallucinated behavior (Rudman et al., 2026; Jiang et al., 2026).

Our work addresses a complementary question: *what makes such heads useful intervention targets?* Rather than interpreting head-level association or intervention sensitivity as evidence of online hallucination detection, we distinguish **diagnostic separability** from **causal controllability**. We show that raw text-side reliance is only weakly discriminative, even when the corresponding pathways provide strong selective intervention leverage. This motivates our interpretation of hallucination-related heads as **text-side actuators** and leads to DEACT, which regulates these pathways according to their current operating state rather than treating their activity as a hallucination prediction.

3 From Detection to Actuation

Before introducing DEACT, we examine what makes hallucination-related attention heads useful intervention targets. We first test whether their text-side behavior supports reliable online diagnosis, and then ask whether perturbing these pathways differentially affects hallucinated and grounded predictions. The results motivate the actuator view underlying our method.

3.1 Is Text-Side Reliance a Reliable Hallucination Signal?

We first test whether text-side reliance provides a useful online signal of hallucination. At decoding step t, we partition the available context into the system prefix $\mathcal{S}$, image tokens $\mathcal{I}$, and post-image textual context $\mathcal{T}$:

$$\mathcal{S} = [0, p_{\text{img_start}}), \qquad \mathcal{I} = [p_{\text{img_start}}, p_{\text{img_end}}], \qquad \mathcal{T} = (p_{\text{img_end}}, t). \tag{1}$$

Here, $\mathcal{I}$ represents visual evidence, whereas $\mathcal{T}$ contains the autoregressive textual context accumulated after the image tokens.

For layer l and head h, we define the bounded text-reliance score

$$r_{t,l,h}^{\text{online}} = \frac{\sum_{j\in\mathcal{T}} A_{t,l,h,j}}{\sum_{j\in\mathcal{T}} A_{t,l,h,j} + \sum_{j\in\mathcal{I}} A_{t,l,h,j} + \epsilon}, \tag{2}$$

which measures the fraction of combined text-and-image attention assigned to the post-image textual context. We aggregate this score over the selected head pool:

$$\bar{r}_t^{\text{online}} = \frac{1}{|\mathcal{H}_{\text{act}}|} \sum_{(l,h)\in\mathcal{H}_{\text{act}}} r_{t,l,h}^{\text{online}}. \tag{3}$$

Hallucinated object predictions are more text-reliant on average: $\bar{r}_t^{\text{online}} = 0.883$, compared with 0.822 for grounded objects. However, the two distributions overlap substantially, yielding an AUC of only 0.594 (Figure 1). High text reliance therefore characterizes a broader text-dominant regime rather than a hallucination-specific state.

A threshold sweep yields the same conclusion. At $\tau = 0.70$, the score recovers 96.7% of hallucinated object steps but also flags 84.5% of grounded ones. Raising the threshold to $\tau = 0.90$ reduces the grounded false-positive rate to 22.9%, but hallucination recall falls to 30.6%. Thus, no simple threshold provides a favorable token-level detection trade-off. Additional results are provided in Appendix B.1.

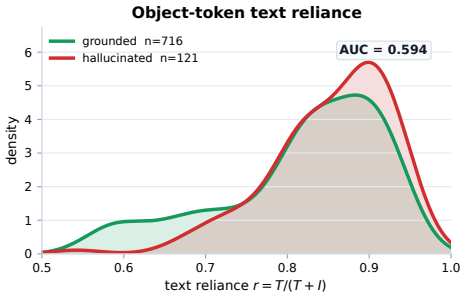


Figure 1: **Text-side reliance is weakly diagnostic of hallucination.** Grounded and hallucinated object predictions overlap substantially, yielding an AUC of 0.594.

3.2 Selected Heads Provide Selective Intervention Leverage

We next ask whether hallucinated and grounded predictions differ in their *dependence* on selected text-side pathways. We use *text-side actuator* to denote a head whose text-side suppression affects hallucinated predictions more strongly than grounded ones.

To avoid circularity, heads are selected on a 500-image calibration split, whereas the intervention probe uses a disjoint 500-image COCO split, from which we sample 200 hallucinated and 200 grounded object tokens.

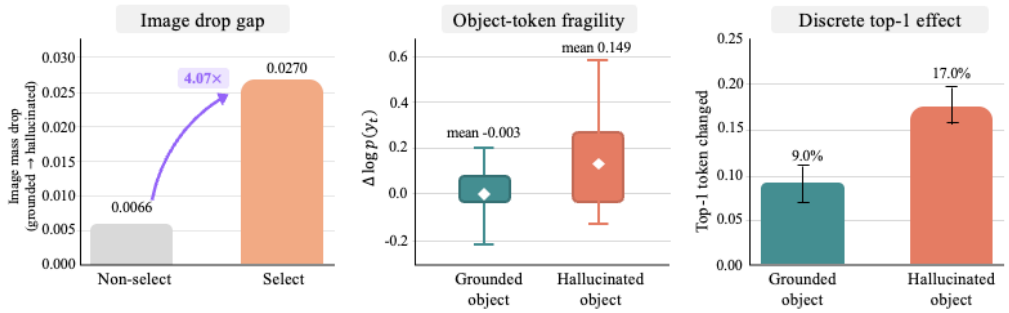


Figure 2: **Observational shift and intervention sensitivity of selected heads. Left:** Selected heads show a larger reduction in image-token attention from grounded to hallucinated object steps than non-selected heads within L9–L16 (0.027 vs. 0.007; $4.1\times$). **Center:** Under hard text-side suppression, the log-probability of hallucinated objects decreases by 0.149 on average, while grounded objects show essentially no decrease (-0.003). **Right:** The top-1 prediction changes for 17.0% of hallucinated probes and 9.0% of grounded probes.

Figure 2 shows that the selected heads exhibit both an observational shift in visual attention and a clear asymmetry under intervention. Suppressing their text-side contribution produces a larger log-probability response and more frequent top-1 changes for hallucinated than grounded object predictions. The relevant actuator property is therefore not clean separation in the pre-intervention state, but *differential sensitivity under intervention*.

3.3 Actuation Is Selective, Not Merely Suppressive

The intervention effect could arise trivially if any sufficiently text-active head were sensitive to suppression. We therefore compare the selected actuator pool against layer-matched random, text-mass-only, contrastive-only, and reproduced AD-HH heads (Yang et al., 2025) under the same hard text-side intervention.

For each pool, Δ_H and Δ_G denote the mean log-probability drops for hallucinated and grounded objects, respectively. We report

$$\mathrm{NormSel} = \frac{\Delta_H - |\Delta_G|}{\Delta_H + |\Delta_G| + \epsilon}, \qquad \mathrm{Damage}_{G/H} = \frac{|\Delta_G|}{\Delta_H + \epsilon},$$

where higher NormSel and lower $\mathrm{Damage}_{G/H}$ indicate greater hallucination-selective intervention leverage.

Table 1 shows that rank fusion achieves the highest normalized selectivity and lowest grounded-object damage. Text-mass selection perturbs both prediction types, while contrastive-only selection is less selective, indicating that pathway strength and hallucination-contrastive bias provide complementary signals. AD-HH produces a substantially larger effect on hallucinated objects, but also much greater grounded-object degradation. The rank-fused pool instead concentrates its intervention effect on hallucinated predictions while largely sparing grounded ones. Thus, neither high text-side activity nor large perturbation magnitude alone identifies useful control points; effective actuation depends on selective intervention leverage.

Table 1: **Head-pool control under identical text-side suppression.** Rank fusion provides the strongest hallucination-selective intervention profile.

Head pool	Δ_H	Δ_G	NormSel↑	Damage$_{G/H}$ ↓
Random	0.078	0.013	0.705	0.173
Text-mass	0.237	0.063	0.580	0.266
Contrastive	0.182	0.032	0.699	0.177
AD-HH	**1.374**	0.389	0.559	0.283
Rank-fused	0.149	-0.003	**0.959**	**0.021**

We additionally localize selectivity across depth. Figure 3 shows that early layers are nearly insensitive to text-side perturbation, whereas deeper layers produce larger absolute effects on both hallucinated and grounded predictions. L9–L16 yields the largest hallucinated–grounded response gap (0.019), compared with 0.001, 0.014, and 0.012 for L0–L8, L17–L24, and L25–L31, respectively, while keeping the grounded-object effect near zero. The layers most responsive to perturbation are therefore not necessarily the most selective intervention targets.

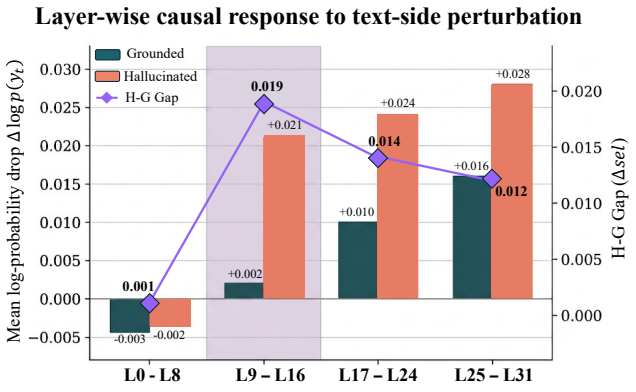


Figure 3: **Layer-wise response to text-side perturbation.** Early layers are largely insensitive, while later layers affect both prediction types. L9–L16 yields the largest response gap with near-zero grounded-object effect.

3.4 From Actuator Localization to Dynamic Regulation

These analyses localize *where* intervention is useful, but not *when* it should be applied. The same actuator may contribute to either unsupported or grounded generation depending on the decoding context. We therefore use its current bounded text-reliance state to modulate intervention strength during generation. This motivates DEACT, which combines offline actuator localization with state-dependent regulation.

4 Dynamic Exponential Actuator Suppression

Building on the preceding analysis, we introduce DEACT (**D**ynamic **E**xponential **ACT**uator Suppression), a training-free inference-time method with model-specific calibration for regulating text-side actuator pathways. DEACT separates three decisions conflated by static suppression: *where*, *when*, and *how strongly* to intervene. Offline, it identifies actuator heads using complementary measures of text-side pathway strength and hallucination-contrastive bias. Online, it compares each selected head's current text-reliance state with a position-dependent reference and maps the result to a suppression strength. Only post-image text-side attention is attenuated; model parameters and non-text attention weights remain unchanged. Figure 4 summarizes the procedure.

4.1 Actuator Pool Identification

The offline stage identifies heads that combine substantial text-side pathway strength with hallucination-contrastive bias. For layer l and head h, we first measure post-image text-side attention mass:

$$M^{\text{text}}_{t,l,h} = \sum_{j \in \mathcal{T}} A_{t,l,h,j}, \tag{4}$$

where $\mathcal{T}$ denotes the post-image textual context defined in Section 3. Averaging over calibration steps gives

$$I_{\text{text}}(l,h) = \mathbb{E}_{t \in \mathcal{D}_{\text{cal}}} \left[M^{\text{text}}_{t,l,h} \right]. \tag{5}$$

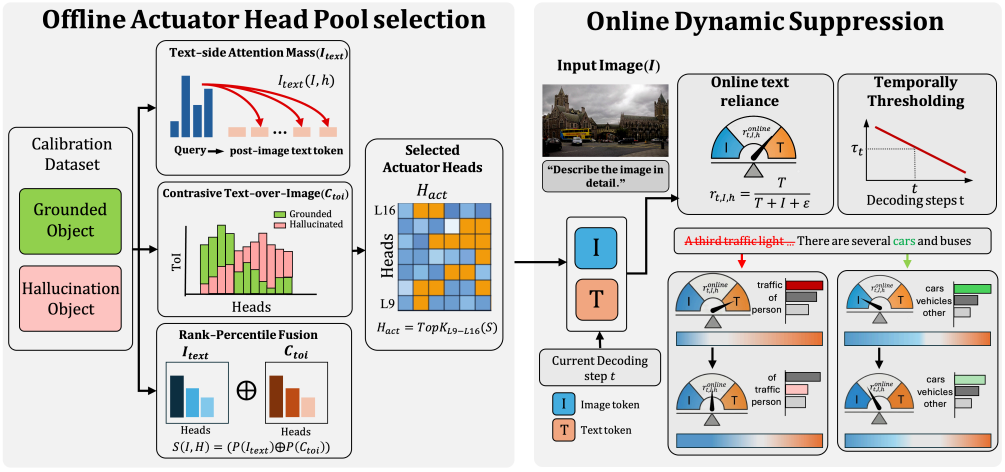


Figure 4: **Overview of DEACT.** Offline, actuator heads are identified by rank-fusing text-side pathway strength and hallucination-contrastive text-over-image bias. Online, their post-image text-side contribution is attenuated according to the current bounded text-reliance state and a position-dependent reference. Model parameters and non-text attention weights remain unchanged.

I_{text} measures the strength of the text-side pathway, but does not by itself capture whether that pathway is preferentially associated with hallucinated predictions.

We therefore define the text-over-image ratio

$$r^{\text{toi}}_{t,l,h} = \frac{\sum_{j\in\mathcal{T}} A_{t,l,h,j}}{\sum_{j\in\mathcal{I}} A_{t,l,h,j} + \epsilon}, \tag{6}$$

where $\mathcal{I}$ denotes the image-token region. Let $\mathcal{D}^H_{\text{cal}}$ and $\mathcal{D}^G_{\text{cal}}$ denote hallucinated and grounded object-token steps in the calibration set, respectively. Their difference defines the hallucination-contrastive bias:

$$C_{\text{toi}}(l,h) = \mathbb{E}_{t\in\mathcal{D}^H_{\text{cal}}}\left[r^{\text{toi}}_{t,l,h}\right] - \mathbb{E}_{t\in\mathcal{D}^G_{\text{cal}}}\left[r^{\text{toi}}_{t,l,h}\right]. \tag{7}$$

We combine pathway strength and contrastive bias through their percentile ranks:

$$S(l,h) = \frac{1}{2}\left[\mathcal{P}(I_{\text{text}}(l,h)) + \mathcal{P}(C_{\text{toi}}(l,h))\right], \tag{8}$$

where $\mathcal{P}(\cdot)$ denotes percentile rank among candidate heads. The actuator pool is

$$\mathcal{H}_{\text{act}} = \text{TopK}_{(l,h)\in\mathcal{L}_{\text{act}}} S(l,h), \tag{9}$$

where $\mathcal{L}_{\text{act}}$ is the calibrated actuator layer range. Both the layer range and selected head pool are fixed before held-out evaluation.

4.2 State-Dependent Exponential Regulation

Actuator selection determines *where* to intervene, but the appropriate strength can vary across decoding steps. We therefore reuse the bounded text-reliance score $r^{\text{online}}_{t,l,h}$ defined in Eq. 2 as the online state of each selected head. The fixed actuator score $S(l,h)$ captures head-level intervention relevance, whereas the online state modulates intervention strength during generation.

Because the post-image textual region grows throughout autoregressive decoding, the operating range of text reliance changes with generation position. We use a position-dependent reference to incorporate this shift. Let $s \in \{0,\ldots,T_{\max}-1\}$ denote the generated-token index, where $T_{\max}$ is the generation horizon. We define

$$\tau_s = \tau_{\text{hi}} + \frac{s}{T_{\max}-1}\left(\tau_{\text{lo}} - \tau_{\text{hi}}\right). \tag{10}$$

The reference therefore interpolates from τ_{hi} at the beginning of generation to τ_{lo} at later positions. Its endpoints are calibrated from the empirical distribution of the bounded text-reliance score:

$$\tau_{\text{hi}} \approx Q_{0.66}\left(r^{\text{online}}_{t,l,h}\right), \qquad \tau_{\text{lo}} \approx Q_{0.33}\left(r^{\text{online}}_{t,l,h}\right), \tag{11}$$

where Q_α denotes the α-quantile over selected heads and calibration steps. For LLaVA-1.5-7B, this yields $\tau_{\text{hi}} = 0.90$ and $\tau_{\text{lo}} = 0.80$ after rounding. The decreasing reference makes the controller progressively more permissive at later positions, where aggregate text-side reliance is empirically higher.

For the decoding step t corresponding to generation index s, we define

$$g_{t,l,h} = \exp\left(q\left(r^{\text{online}}_{t,l,h} - \tau_s\right)\right), \tag{12}$$

where $q > 0$ controls response sharpness. The suppression strength is then

$$\delta_{t,l,h} = \text{clip}\left(S(l,h)\, g_{t,l,h}, 0, 1\right). \tag{13}$$

Suppression therefore increases when a highly ranked actuator enters a more text-reliant state relative to the reference at its current generation position.

4.3 Direct Text-Side Attenuation

For each $(l,h) \in \mathcal{H}_{\text{act}}$, we attenuate only its post-image text-side attention:

$$\widetilde{A}_{t,l,h,j} = (1 - \delta_{t,l,h}) A_{t,l,h,j}, \qquad j \in \mathcal{T}, \tag{14}$$

while leaving the remaining attention weights unchanged:

$$\widetilde{A}_{t,l,h,k} = A_{t,l,h,k}, \qquad k \notin \mathcal{T}. \tag{15}$$

The modified weights are used directly for value aggregation:

$$\widetilde{\mathbf{o}}_{t,l,h} = \sum_j \widetilde{A}_{t,l,h,j} \mathbf{v}_{l,h,j}. \tag{16}$$

We do not renormalize or redistribute the removed text-side mass. Consequently,

$$\sum_j \widetilde{A}_{t,l,h,j} = 1 - \delta_{t,l,h} \sum_{j \in \mathcal{T}} A_{t,l,h,j}, \tag{17}$$

so attenuation can reduce both the targeted text-side value contribution and the overall magnitude of the head output. This update avoids imposing an additional assumption about where the removed attention mass should be reassigned. Appendix C.2 compares direct attenuation with several mass-preserving alternatives.

5 Experiments

We evaluate DEACT across multiple LVLM architectures and benchmarks, examining hallucination reduction, grounded-generation preservation, generalization, and actuator selection.

5.1 Experimental Setup

Models and datasets. We evaluate DEACT on LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). For each architecture, we use 500 MS COCO images (Lin et al., 2014) for calibration and 500 disjoint images for held-out evaluation. The calibration split determines the actuator layer range, head-level statistics, actuator pool, and controller configuration; all choices are fixed before held-out evaluation. We further evaluate on the generative split of AMBER (Wang et al., 2023) to assess generalization beyond COCO.

Table 2: **Cross-model comparison on COCO captioning.** C_s and C_i abbreviate $CHAIR_s$ and $CHAIR_i$, respectively. Lower is better for CHAIR; higher is better for F1.

	LLaVA-1.5-7B			LLaVA-1.5-13B			LLaVA-NeXT-7B			Qwen2.5-VL-7B		
Method	$C_s \downarrow$	$C_i \downarrow$	F1↑	$C_s \downarrow$	$C_i \downarrow$	F1↑	$C_s \downarrow$	$C_i \downarrow$	F1↑	$C_s \downarrow$	$C_i \downarrow$	F1↑
Greedy	54.6	14.8	76.4	52.6	14.1	76.4	30.8	8.1	72.9	32.0	8.4	74.7
PAI	35.4	9.1	75.7	40.2	10.8	76.9	28.0	8.2	70.4	–[†]	–[†]	–[†]
PAI-CD	30.2	7.8	76.5	37.0	9.9	77.3	28.4	9.5	71.8	–[†]	–[†]	–[†]
VAF	52.4	14.5	75.8	46.0	12.4	77.7	26.4	8.9	69.3	33.2	8.4	75.2
TARAC	43.4	12.1	76.6	42.6	11.6	76.9	30.8	8.9	69.9	33.0	9.3	74.6
VGA	35.0	9.5	76.6	40.2	10.9	**78.8**	26.4	8.0	69.3	33.4	8.9	**75.4**
DEACT (Ours)	**29.6**	**7.3**	**77.2**	**32.0**	**7.4**	78.2	**21.6**	**7.9**	**73.2**	**28.8**	**7.9**	74.2

[†] We could not obtain a reliable Qwen2.5-VL-7B reproduction from the publicly available implementation.

Evaluation metrics. On COCO, we report $CHAIR_s$ and $CHAIR_i$ following the standard object-hallucination protocol (Rohrbach et al., 2018), together with object-level precision, recall, and F1 to account for trivial reductions obtained by omitting objects. We additionally report BLEU-1 (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005) to assess caption quality. On AMBER, we report CHAIR, Cover, Hal, and Cog following the benchmark protocol (Wang et al., 2023).

Baselines. We compare against greedy decoding and representative training-free methods, including PAI and its contrastive-decoding variant PAI-CD (Liu et al., 2024c), VAF (Yin et al., 2025), TARAC (Jiang et al., 2025), and VGA (Zhao et al., 2026). We report results only for model–method combinations with reliable reproductions from publicly available implementations. In particular, PAI and PAI-CD are not reported on Qwen2.5-VL-7B. We separately compare against AD-HH (Yang et al., 2025), the most directly related head-level intervention method, on the LLaVA-1.5 family.

Implementation details. DEACT operates entirely at inference time without updating model parameters. The actuator layer range, head pool, and controller are calibrated once per architecture and remain fixed during held-out generation. Attention regions follow each model's native multimodal token layout. Full configurations and calibration details are provided in Appendix A.

5.2 Main Results on COCO

Table 2 shows that DEACT consistently reduces object hallucination while maintaining competitive grounded-generation quality. On LLaVA-1.5-7B, $CHAIR_s$ decreases from 54.6 to 29.6 and $CHAIR_i$ from 14.8 to 7.3, while F1 improves from 76.4 to 77.2. The same pattern extends across architectures: DEACT achieves the lowest $CHAIR_s$ and $CHAIR_i$ on all four evaluated models, while F1 remains close to or above greedy decoding. These results indicate that text-side actuator regulation transfers across architectures and provides a mechanism complementary to approaches that primarily strengthen, preserve, or redirect visual evidence.

5.3 Comparison with Static Head Suppression

Table 3 compares DEACT with AD-HH (Yang et al., 2025). Static suppression already improves substantially over greedy decoding, confirming that hallucination-related heads provide useful intervention points. DEACT further reduces both CHAIR metrics on LLaVA-1.5-7B and LLaVA-1.5-13B while improving F1. Because the two methods differ in both actuator localization and regulation strategy, this comparison does not isolate the effect of dynamic regulation alone; rather, it demonstrates the benefit of combining selective actuator identification with state-dependent suppression.

5.4 Caption Quality and Object Preservation

Table 4 examines whether hallucination reduction comes at the expense of caption quality or grounded-object generation. DEACT achieves the highest BLEU-1, METEOR, and object-level

Table 3: **Static head suppression vs. DEACT.** Lower CHAIR and higher F1 are better.

Method	LLaVA-1.5-7B			LLaVA-1.5-13B		
	$C_s \downarrow$	$C_i \downarrow$	F1↑	$C_s \downarrow$	$C_i \downarrow$	F1↑
Greedy	54.6	14.8	76.4	52.6	14.1	76.4
AD-HH	39.4	10.1	76.6	40.6	9.7	77.2
DEACT	**29.6**	**7.3**	**77.2**	**32.0**	**7.4**	**78.2**

Table 4: **Caption quality and object preservation on LLaVA-1.5-7B.** Higher is better.

Method	BLEU-1↑	METEOR↑	F1↑	Prec.↑
Greedy	17.4	17.9	76.4	73.5
PAI	17.8	18.5	75.7	82.5
PAI-CD	17.8	18.4	76.5	**85.7**
VAF	18.2	18.4	75.8	74.4
TARAC	18.4	18.4	76.6	78.7
AD-HH	17.5	18.0	76.6	80.3
DEACT	**19.1**	**19.1**	**77.2**	85.0

Table 5: **Results on the AMBER generative benchmark.** Lower is better for CHAIR, Hal, and Cog; higher is better for Cover.

Method	LLaVA-1.5-7B				LLaVA-NeXT-7B				Qwen2.5-VL-7B			
	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓
Greedy	7.8	**51.8**	35.7	4.3	9.8	**62.5**	54.9	4.1	5.0	64.7	27.4	1.9
PAI	6.2	48.5	29.9	2.6	6.5	55.1	32.8	2.1	–†	–†	–†	–†
PAI-CD	5.9	48.8	28.6	2.2	6.1	53.5	**31.6**	**1.8**	–†	–†	–†	–†
VAF	8.1	51.4	35.5	4.0	9.6	62.2	53.8	4.0	4.9	**64.8**	27.2	1.4
TARAC	5.8	49.3	28.5	2.9	6.7	60.3	37.5	3.4	5.1	64.2	27.0	1.5
VGA	**4.8**	49.6	24.5	2.2	8.8	57.1	48.3	2.6	4.6	64.6	25.4	1.5
DEACT (Ours)	**4.8**	48.4	**22.5**	**2.1**	**5.7**	56.9	39.3	2.4	**4.4**	62.5	**23.9**	**1.2**

† We could not obtain a reliable Qwen2.5-VL-7B reproduction from the publicly available implementation.

F1 among the compared methods while maintaining high precision. These results indicate that its hallucination reduction is not driven solely by indiscriminate suppression of object mentions.

5.5 Generalization to AMBER

Table 5 evaluates generalization beyond COCO. DEACT reduces CHAIR relative to greedy decoding across all evaluated architectures, from 7.8 to 4.8, 9.8 to 5.7, and 5.0 to 4.4, respectively. It also reduces both Hal and Cog on all three models, although the associated coverage trade-off varies across architectures and is most pronounced on LLaVA-NeXT-7B. These results support cross-benchmark generalization while highlighting architecture-dependent trade-offs between hallucination suppression and object coverage.

5.6 Ablation Study

We further test whether the gains can be explained by arbitrary head intervention. A layer-matched random pool yields only modest improvement, whereas the rank-fused actuator pool substantially reduces CHAIR metrics while improving F1, indicating that pathway identity is critical. Additional ablations on head selection, pool size, attention-update strategy, controller sharpness, intervention depth, and dynamics are provided in Appendix C, with qualitative examples in Appendix D.

6 Conclusion

We revisit hallucination-related attention heads in large vision-language models and find a clear asymmetry between diagnosis and intervention. Raw text-side reliance only weakly distinguishes hallucinated from grounded object predictions, whereas suppressing selected text-side pathways disproportionately affects hallucinated continuations. These findings motivate our interpretation of such heads as *text-side actuators* and lead to DEACT, a training-free inference-time method with model-specific calibration that combines actuator localization with state-dependent regulation. Across multiple LVLM architectures and benchmarks, DEACT consistently reduces object hallucination while maintaining competitive grounded-generation quality. More broadly, our results show that weak **diagnostic separability** can coexist with strong **causal controllability**, suggesting that effective hallucination mitigation need not depend on reliable online failure detection.

AI USE STATEMENT

Generative AI tools were used exclusively as writing aids during manuscript preparation, primarily for language polishing, phrasing, and improving the clarity and organization of the presentation. AI-generated content was not treated as scientific evidence and was not used to generate experimental results or replace empirical analysis. The research questions, methodological design, implementation, experimental protocol, interpretation of results, and scientific claims were determined and verified by the authors. All AI-assisted text was subsequently reviewed, revised where necessary, and approved by the authors, who take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work studies inference-time mitigation of object hallucination in large vision-language models. The experiments use publicly available pretrained models and established benchmark datasets, and do not involve human-subject experiments, collection of private user data, or processing of personally identifiable information.

The proposed method modifies internal attention computation only during inference and does not introduce additional training data or external user information. Because the objective of the work is to reduce visually unsupported generation, we do not identify application-specific ethical risks beyond the broader limitations and potential misuse associated with existing LVLMs. In particular, DEACT reduces but does not eliminate hallucination, and its outputs should therefore not be interpreted as guaranteed factual or visually grounded predictions in high-stakes settings.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce the reported results in the main paper and appendix. These include the definition of the intervention operator, actuator-head scoring procedure, model-specific actuator ranges and pool sizes, controller parameters, calibration procedure, evaluation metrics, and the settings used in all reported ablations.

For each evaluated architecture, calibration is performed on a designated 500-image MS COCO split, and all actuator-head indices and controller parameters are fixed before evaluation on a disjoint 500-image held-out split. The appendix reports the resulting model-specific configurations and further documents the hard causal probe, head-pool controls, attention-update alternatives, pool-size and controller-sharpness sensitivity, layer localization, and realized controller dynamics. These details are provided to make both the end-to-end results and the mechanistic analyses independently reconstructable.

Baseline methods were reproduced from publicly available implementations and evaluated in the same experimental environment whenever a reliable reproduction was available. Model–method combinations for which we could not obtain a reliable reproduction are explicitly marked as unavailable rather than reported using incomparable results. The source code, evaluation scripts, model-specific configurations, and implementation details necessary to reproduce the experiments will be released upon acceptance.

REFERENCES

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.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18135–18143, 2024.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

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.

Jinrui Jiang, Zhangtai Wu, Zhen Wu, and Xinyu Dai. Causal evidence for attention head imbalance in modality conflict hallucination. *arXiv preprint arXiv:2605.19250*, 2026.

Lei Jiang, Chunzhao Xie, Tongxuan Liu, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, et al. Tarac: Mitigating hallucination in lvlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*, 2025.

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.

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

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689. IEEE, 2024.

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.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

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

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, 2024c.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

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.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Carsten Eickhoff, Ritambhara Singh, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision–language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41894–41912, 2026.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12492–12511, 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.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 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.

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.

A Model-Specific Configuration

DEACT is calibrated independently for each architecture rather than transferring a single intervention configuration across models. The appropriate intervention depth, number of useful actuator heads, and distribution of online text-reliance states vary across architectures and model scales, motivating model-specific calibration.

For each model, calibration determines four components: the actuator layer range $\mathcal{L}_{\text{act}}$, actuator-pool size K, exponential controller sharpness q, and temporal reference endpoints τ_{hi} and τ_{lo}. The layer range determines *where* intervention is permitted, the actuator pool determines *which* heads are regulated, and the controller parameters govern *when* and *how strongly* their text-side contribution is attenuated.

Table 6 summarizes the resulting configurations. LLaVA-1.5-7B and LLaVA-1.5-13B use L9–L16, while the calibrated ranges shift to L11–L16 for LLaVA-NeXT-7B and L20–L22 for Qwen2.5-VL-7B. Pool size also varies substantially, from $K = 33$ for Qwen2.5-VL-7B to $K = 200$ for LLaVA-1.5-13B. The temporal reference endpoints are likewise derived from each model's empirical text-reliance distribution rather than transferred from a single reference architecture.

Calibration is performed once for each architecture on the designated calibration split. The actuator layer range, selected head indices, actuator scores, and controller parameters are then fixed for all held-out evaluations; no held-out caption information is used to modify them during evaluation. In our implementation, processing the 500-image COCO calibration split requires approximately one hour on a single NVIDIA A5000 GPU.

Table 6: **Model-specific DEACT configurations.** Each actuator range, pool, and controller schedule is calibrated independently and fixed before held-out evaluation.

Model	$\mathcal{L}_{\text{act}}$	K	q	$\tau_{\text{hi}}/\tau_{\text{lo}}$
LLaVA-1.5-7B	L9–L16	100	10	0.90 / 0.80
LLaVA-1.5-13B	L9–L16	200	10	0.90 / 0.80
LLaVA-NeXT-7B	L11–L16	100	10	0.94 / 0.84
Qwen2.5-VL-7B	L20–L22	33	8	0.96 / 0.87

B Additional Diagnostic and Intervention Analysis

We provide additional analyses supporting the diagnostic and intervention results in Section 3. We first examine whether thresholding the bounded text-reliance score can yield a useful online detector. We then specify the hard intervention probe, compare the rank-fused actuator pool with matched alternatives, and report the numerical layer-wise results used for actuator localization.

B.1 Threshold Sensitivity of Text-Side Reliance

The main analysis shows that hallucinated object predictions have higher text-side reliance on average, but overlap substantially with grounded predictions. To test whether useful discrimination nevertheless emerges at a particular operating point, we threshold the aggregated score $\bar{r}_t^{\text{online}}$ over the same object-token population.

Table 7 shows a strong precision–recall trade-off. At $\tau = 0.70$, the score recovers 96.7% of hallucinated object steps but also flags 84.5% of grounded steps. Raising the threshold reduces grounded false positives only at the cost of rapidly declining hallucination recall: recall falls to 30.6% at $\tau = 0.90$ and 1.7% at $\tau = 0.95$.

Together with the AUC of 0.594 reported in the main paper, these results show that the score captures a broad text-dominant operating regime rather than a sharply separable hallucination state. DEACT therefore uses $r_{t,l,h}^{\text{online}}$ as an online state for pathways localized offline through intervention analysis.

Table 7: **Threshold sensitivity of the aggregated text-reliance score.** No threshold provides a favorable trade-off between hallucination recall and grounded-object false positives.

τ	Prec.↑	Recall↑	F1↑	Grounded FPR↓
0.70	0.162	0.967	0.278	0.845
0.80	0.166	0.810	0.275	0.689
0.85	0.174	0.579	0.268	0.464
0.90	0.184	0.306	0.230	0.229
0.95	0.154	0.017	0.030	0.015

B.2 Intervention Probe Definition

We provide the full specification of the hard text-side intervention used in Section 3.2. This probe is intentionally simpler than the final DEACT controller: it is designed to measure sensitivity under a fixed perturbation operator rather than to optimize end-to-end generation.

For each selected head (l, h), we compute its post-image text-side attention mass

$$M^{\text{text}}_{t,l,h} = \sum_{j \in \mathcal{T}} A_{t,l,h,j}. \tag{18}$$

When

$$M^{\text{text}}_{t,l,h} \geq \tau_{\text{hard}}, \qquad \tau_{\text{hard}} = 0.4, \tag{19}$$

we remove the head's post-image text-side contribution:

$$A'_{t,l,h,j} = \begin{cases} 0, & j \in \mathcal{T}, \\ A_{t,l,h,j}, & j \notin \mathcal{T}. \end{cases} \tag{20}$$

If $M^{\text{text}}_{t,l,h} < \tau_{\text{hard}}$, the head remains unchanged. The intervention is applied after attention softmax and before value aggregation, without renormalizing the remaining attention weights.

For each originally generated target object token y_t, we measure

$$\Delta \log p(y_t) = \log p_{\text{base}}(y_t) - \log p_{\text{probe}}(y_t), \tag{21}$$

where a positive value indicates that suppression decreases the probability of the original object token.

Because this intervention reduces the attention-row sum, it can affect both the targeted text-side value contribution and the magnitude of the resulting head output. We therefore treat it as a direct text-side attenuation operator, not as a mass-preserving reallocation of attention. All head pools below are evaluated with exactly the same operator, so their relative differences reflect sensitivity under a common perturbation.

The probe is performed post hoc on held-out greedy-decoding outputs and is not used for head selection, layer localization, or controller calibration. The samples used to measure intervention responses are therefore separate from those used to construct the actuator pool.

B.3 Head-Pool Control Study

The intervention probe shows that the selected heads provide useful leverage, but does not by itself establish that rank fusion identifies unusually selective pathways. We therefore compare the actuator pool with four alternative head sets under the same hard text-side suppression operator.

The *layer-matched random* pool tests whether intervention anywhere within L9–L16 is sufficient. The *text-mass-only* pool tests whether strong post-image textual attention alone identifies useful heads. The *contrastive-only* pool isolates the hallucinated–grounded text-over-image contrast without explicitly favoring heads with substantial text-side pathway strength. Finally, the reproduced *AD-HH* pool provides a comparison with a prior hallucination-head selection strategy.

Using the notation of Section 3.3, Δ_H and Δ_G denote the mean log-probability drops for hallucinated and grounded object probes. For completeness, we report

$$\text{NormSel} = \frac{\Delta_H - |\Delta_G|}{\Delta_H + |\Delta_G| + \epsilon}, \tag{22}$$

and

$$\mathrm{Damage}_{G/H} = \frac{|\Delta_G|}{\Delta_H + \epsilon}. \tag{23}$$

Higher NormSel and lower $\mathrm{Damage}_{G/H}$ indicate more selective effects on hallucinated predictions.

Table 8: **Head-pool control study under the same hard text-side suppression probe.** The rank-fused actuator pool achieves the strongest normalized selectivity and lowest relative grounded-object damage.

Head pool	Δ_H	Δ_G	NormSel↑	$\mathrm{Damage}_{G/H}$ ↓
Layer-matched random	0.078	0.013	0.705	0.173
Text-mass only	0.237	0.063	0.580	0.266
Contrastive only	0.182	0.032	0.699	0.177
AD-HH pool	**1.374**	0.389	0.559	0.283
Rank-fused actuator pool	0.149	-0.003	**0.959**	**0.021**

The results distinguish intervention magnitude from intervention selectivity. AD-HH produces by far the largest decrease in hallucinated-object probability, but also substantially affects grounded objects. Text-mass-only selection likewise perturbs both groups, showing that strong text-side activity alone is insufficient to isolate selective control points.

Contrastive-only selection improves selectivity, but remains below the rank-fused pool. Rank fusion retains a meaningful effect on hallucinated predictions while driving the grounded-object response close to zero, yielding the highest NormSel and lowest relative damage. Its advantage is therefore not maximal perturbation magnitude, but concentration of the intervention effect on hallucinated predictions.

The two ranking components play complementary roles. Pathway strength favors heads with a substantive text-side contribution, whereas contrastive bias penalizes pathways that are similarly text-active during grounded and hallucinated generation. Their combination favors heads that are both active and differentially sensitive to intervention.

B.4 Layer-Wise Localization

We report the numerical results underlying the layer-localization analysis in the main paper. For each layer band, we apply the same text-side perturbation and measure the mean log-probability drop separately for hallucinated and grounded object probes. Using the notation above, we define

$$\Delta_{\mathrm{sel}} = \Delta_H - \Delta_G. \tag{24}$$

Table 9: **Layer-band selectivity under controlled text-side perturbation.** L9–L16 yields the largest hallucinated–grounded response gap while keeping the grounded-object effect near zero.

Layer band	Δ_H	Δ_G	Δ_{sel}
L0–L8	-0.002	-0.003	0.001
L9–L16	0.021	0.002	**0.019**
L17–L24	0.024	0.010	0.014
L25–L31	0.028	0.016	0.012

Early layers respond little to the perturbation. At greater depth, the hallucinated-object effect grows, but so does the effect on grounded predictions. L9–L16 provides the largest differential response ($\Delta_{\mathrm{sel}} = 0.019$) while keeping grounded-object change near zero. Actuator localization therefore favors differential sensitivity rather than either raw attention magnitude or maximal perturbation response.

C Additional Ablation Studies

We next examine the principal design choices underlying DEACT. Unless otherwise specified, experiments use LLaVA-1.5-7B with the configuration reported in Table 6. We study actuator-head

selection, treatment of removed text-side attention mass, actuator-pool size, controller sharpness, intervention depth, and realized controller behavior.

The reported configuration is determined by the calibration procedure and fixed before held-out evaluation. The sweeps below are post-hoc sensitivity analyses used to characterize the behavior of individual design components; they are not used to select the reported held-out configuration.

Because lower hallucination scores can also be obtained by omitting valid objects, we interpret CHAIR jointly with grounded-object preservation, including precision, recall, and F1 where applicable.

C.1 End-to-End Actuator-Head Selection

Appendix B.3 evaluates head identity under a controlled token-level intervention. We additionally test whether this difference persists during end-to-end generation with the full DEACT controller.

We compare the rank-fused actuator pool with a layer-matched random pool while holding the intervention range fixed. If the gain arose primarily from perturbing arbitrary heads within L9–L16, the random pool would be expected to produce a similar result.

Table 10: **Effect of actuator-head selection on LLaVA-1.5-7B.** The rank-fused actuator pool substantially outperforms a layer-matched random pool during end-to-end generation.

Head selection	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	F1↑
Greedy	54.6	14.8	76.4
Layer-matched random	46.8	12.0	76.9
Rank-fused actuator pool	**29.6**	**7.3**	**77.2**

The random pool yields only a modest improvement over greedy decoding, whereas the rank-fused pool sharply reduces both CHAIR metrics while also improving F1. Because the two interventions operate over the same layer range, this gap cannot be attributed to intervention depth alone. Together with the controlled probe, the result shows that head identity materially affects end-to-end hallucination mitigation.

C.2 Attention-Update Strategy

The default DEACT operator attenuates post-image text-side attention without redistributing the removed mass. The resulting post-softmax attention row is therefore not constrained to sum to one. This directly weakens the targeted text-side contribution, but can also alter the overall magnitude of the corresponding head output.

To characterize this design choice, we compare direct attenuation with four mass-preserving alternatives. *System redistribution* assigns the removed mass to the system-prefix region, *vision redistribution* reallocates it to image tokens, and *system+vision redistribution* distributes it across both non-text regions. *Full renormalization* instead rescales the remaining attention weights to restore unit mass while preserving their relative proportions.

Table 11: **Effect of the attention-update strategy.** Different redistribution rules expose a trade-off between hallucination reduction and preservation of grounded-object generation.

Update strategy	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	F1↑	Prec.↑
System redistribution	**24.0**	**7.2**	72.7	**86.3**
Vision redistribution	36.8	9.9	77.7	81.0
System+vision redistribution	36.8	9.7	**78.2**	81.1
Full renormalization	37.4	9.7	78.0	81.1
Direct attenuation	29.6	7.3	77.2	85.0

System redistribution yields the lowest raw hallucination scores, but F1 falls to 72.7. Its combination of high precision and substantially lower F1 is consistent with a more conservative object-generation regime, illustrating why CHAIR alone is insufficient to evaluate an intervention rule.

Redistributing mass to image tokens better preserves object-level accuracy, but weakens hallucination reduction. System+vision redistribution and full renormalization show a similar pattern, with F1 near 78 but substantially higher CHAIR values than direct attenuation. Direct attenuation occupies a different operating point, reaching 29.6 CHAIR_s and 7.3 CHAIR_i with F1 of 77.2 and precision of 85.0.

These controls also bound the interpretation of the default operator. Because direct attenuation reduces the post-softmax row sum, its effect may combine removal of the targeted text-side contribution with a reduction in head-output magnitude. None of the tested mass-preserving alternatives reproduces the same operating point, but these comparisons do not fully disentangle those two effects. We therefore treat direct attenuation as the intervention operator evaluated by DEACT, without attributing its entire effect exclusively to text-side mass removal.

C.3 Actuator-Pool Size

The actuator-pool size K controls how broadly intervention is applied within the calibrated layer range. Small K restricts regulation to the highest-ranked heads, whereas larger values progressively include lower-ranked pathways.

Table 12: **Effect of actuator-pool size K on LLaVA-1.5-7B.** Overly broad intervention reduces hallucination scores at the cost of substantial recall degradation.

Config.	$\mathrm{C}_s \downarrow$	$\mathrm{C}_i \downarrow$	Prec.↑	Rec.↑	F1↑
Greedy	54.6	14.8	73.5	**79.5**	76.4
$K = 20$	47.2	13.2	76.6	78.7	77.6
$K = 50$	39.6	10.8	80.2	75.3	**77.7**
$K = 100$	29.6	7.3	85.0	70.7	77.2
$K = 150$	15.8	**4.2**	**91.9**	59.3	72.0
$K = 200$	**14.8**	5.3	91.8	54.9	68.7

Increasing K from 20 to 100 steadily lowers both CHAIR metrics while precision rises from 76.6 to 85.0. Recall decreases as the intervention broadens. Beyond $K = 100$, this trade-off becomes pronounced: recall falls to 59.3 at $K = 150$ and 54.9 at $K = 200$, with corresponding F1 scores of 72.0 and 68.7.

The calibrated LLaVA-1.5-7B configuration uses $K = 100$, before the sharp recall degradation observed for $K \geq 150$. The post-hoc sweep also shows that $K = 50$ yields slightly higher F1, whereas $K = 100$ provides substantially stronger hallucination reduction. Pool size therefore controls the trade-off between suppression strength and grounded-object recall rather than producing uniform gains as more heads are included.

C.4 Controller Sharpness

The parameter q controls how rapidly the exponential gate responds to the difference between the current text-reliance state and its position-dependent reference. Smaller values yield smoother changes in suppression strength, whereas larger values approach the clipping boundary more rapidly.

Table 13: **Sensitivity to exponential sharpness q on LLaVA-1.5-7B.** Performance remains stable across a broad range of controller sharpness values.

q	$\mathrm{CHAIR}_s \downarrow$	$\mathrm{CHAIR}_i \downarrow$	F1↑	Prec.↑
4	31.2	7.9	77.3	84.5
6	29.8	7.7	77.3	84.9
8	**29.6**	7.4	**77.4**	**85.0**
10	**29.6**	**7.3**	77.2	**85.0**
12	29.8	7.4	**77.4**	84.9

Performance varies little over $q \in \{4, 6, 8, 10, 12\}$. Moving from $q = 4$ to $q = 8$ reduces CHAIR_s from 31.2 to 29.6 and CHAIR_i from 7.9 to 7.4 while leaving F1 nearly unchanged. Differences beyond $q = 8$ are small across all reported metrics.

The calibrated configuration uses $q = 10$. The post-hoc sensitivity sweep shows that nearby values produce similar operating points: $q = 8$ and $q = 12$ yield slightly higher F1, while $q = 10$ gives the lowest CHAIR_i and matches the best CHAIR_s and precision. The end-task behavior is therefore not highly sensitive to precise tuning of q.

C.5 Intervention Depth

The intervention range specifies which layers are eligible for actuator regulation. For LLaVA-1.5-7B, its calibrated range follows the localization analysis reported in Table 9: L9–L16 maximizes the hallucinated–grounded response gap while keeping the grounded-object response close to zero.

The same localization procedure is applied independently to each architecture. As shown in Table 6, the selected range shifts to L11–L16 for LLaVA-NeXT-7B and L20–L22 for Qwen2.5-VL-7B. These differences support architecture-specific localization rather than transferring a fixed absolute depth range across models.

C.6 Calibration-Time Temporal Shift

The position-dependent reference is motivated by a systematic shift in the bounded text-reliance score over generation. On the calibration split, the median score across selected actuator heads increases from 0.748 in the first quarter of generation to 0.929 in the final quarter. The fraction of selected head–step pairs with $r^{\mathrm{online}}_{t,l,h} > 0.90$ likewise rises from 23.0% to 64.1%.

As examined in Appendix C.8, this shift is closely associated with growth of the post-image textual region. Aggregate text-side attention increases over generation even though average attention per visible text token decreases.

In the calibrated controller, $\tau_{\mathrm{hi}} > \tau_{\mathrm{lo}}$, so the position-dependent reference decreases with generated-token index. Intervention consequently becomes more permissive later in generation, alongside the empirically increasing region-level text reliance. The schedule is therefore a deliberate position-dependent component of the controller rather than a stationary threshold applied throughout decoding.

C.7 Realized Controller Dynamics

The exponential gate and clipping operation specify the controller mathematically, but do not reveal its realized operating regime. We therefore instrument the LLaVA-1.5-7B evaluation runs and collect statistics over all selected head–step pairs.

Table 14: **Realized controller statistics on LLaVA-1.5-7B.** Statistics are computed over all selected head–step pairs in the instrumented generation runs.

Metric	Value
Selected head–step count	1,094,600
Scaled rate	96.6%
Exact saturation, $\delta = 1$	56.0%
Near-zero retained scale, $1 - \delta \leq 0.05$	57.3%
Mean δ	0.702
Median δ	1.000
Mean post-update row sum	0.602

The realized controller is substantially sharper than a uniformly mild reweighting mechanism. Of the 1,094,600 selected head–step pairs, 96.6% undergo nontrivial scaling, 56.0% reach exact saturation ($\delta = 1$), and 57.3% retain at most 5% of their original post-image text-side attention. The median suppression strength is 1.0 and the mean is 0.702.

The mean post-update attention-row sum is 0.602. Because removed text-side mass is not redistributed, this statistic reflects both the targeted attenuation and the associated reduction in the magnitude of the modified head output. Empirically, DEACT therefore operates as a controller with a broad continuous transition region but frequent near-complete suppression.

Intervention strength also changes markedly with generation position.

Table 15: **Controller behavior across generated-token index bins.** Suppression becomes progressively stronger at later generation positions.

Token index	Saturation	Mean δ	Row sum
0–15	14.0%	0.297	0.864
16–31	33.5%	0.535	0.714
32–63	55.5%	0.728	0.585
64–95	74.6%	0.857	0.501
96+	87.4%	0.936	0.450

Saturation rises from 14.0% over generated-token indices 0–15 to 33.5% over 16–31 and 55.5% over 32–63. It reaches 74.6% over 64–95 and 87.4% after index 96. Over the same bins, mean suppression strength increases from 0.297 to 0.936, while the mean post-update row sum decreases from 0.864 to 0.450. These measurements confirm that the realized controller becomes substantially more aggressive at later positions.

C.8 Temporal Decomposition of Text Reliance

To identify what drives the temporal shift above, we decompose selected-head text reliance as generation proceeds.

Table 16: **Temporal decomposition of selected-head text reliance.** Aggregate post-image text mass increases with decoding progress, whereas average mass per visible text token decreases.

Token index	Text mass	Per-token mass	Text reliance $\bar{r}^{\text{online}}$	Saturation
0–15	0.278	0.01318	0.620	14.0%
16–31	0.415	0.01117	0.759	33.5%
32–63	0.500	0.00829	0.844	55.5%
64–95	0.544	0.00590	0.887	74.6%
96+	0.567	0.00466	0.918	87.4%

Aggregate post-image text mass increases from 0.278 over indices 0–15 to 0.567 after index 96. In contrast, average attention mass per visible text token decreases from 0.01318 to 0.00466. The increase is therefore primarily a region-level effect: as the generated prefix expands, the text region contains more positions and accumulates greater total attention even though the average contribution of each individual text token decreases.

The bounded text-reliance score also rises from 0.620 to 0.918 across the same bins. Selected heads thus allocate an increasing fraction of their combined text-and-image attention to the post-image text region as a whole while distributing that mass across an expanding textual context.

This decomposition explains why generation position is included in the online controller. The operating range of the bounded text-reliance score shifts as the available textual context grows, so a single stationary reference would not represent the same operating regime throughout generation.

D Qualitative Results

Figure 5 presents representative COCO examples comparing greedy decoding, static AD-HH suppression, and DEACT with LLaVA-1.5-7B. The examples complement aggregate metrics by showing how the methods differ at the level of individual generations.

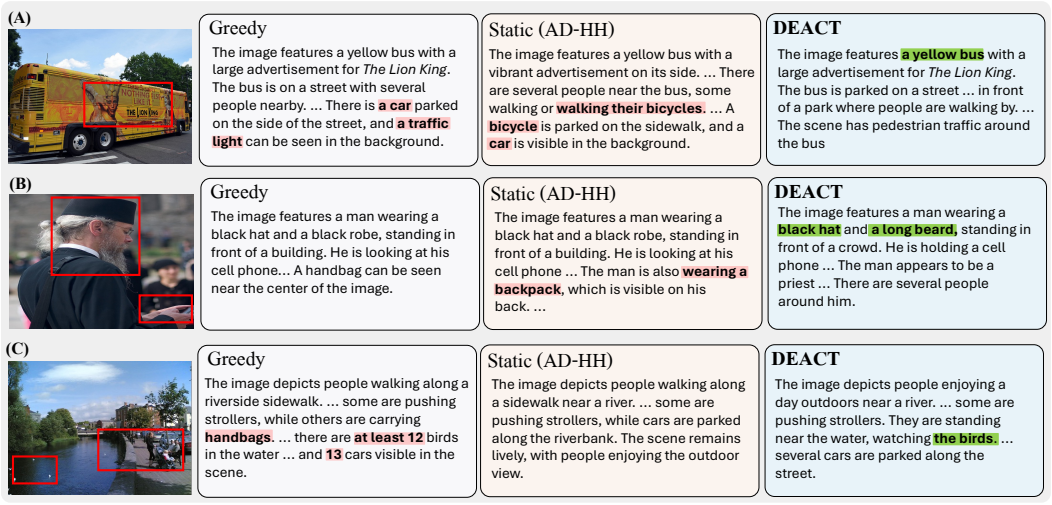


Figure 5: **Representative comparisons among greedy decoding, static AD-HH suppression, and DEACT on COCO captioning with LLaVA-1.5-7B.** In the bus scene, DEACT removes CHAIR-counted hallucinations retained or introduced by greedy decoding and AD-HH while preserving annotated objects. In the second scene, AD-HH introduces a hallucinated backpack absent under greedy decoding and DEACT. In the riverside scene, DEACT recovers the annotated object *bird*, which static suppression omits, without adding another CHAIR-counted hallucination.

In the first example, greedy decoding and static AD-HH retain or introduce object mentions counted as hallucinations under the COCO CHAIR protocol. DEACT removes these mentions while retaining the major annotated entities, illustrating that hallucination reduction need not come from globally preventing object generation.

The second example illustrates a different behavior of static suppression. AD-HH introduces a hallucinated *backpack* that appears in neither the greedy nor the DEACT output. Suppressing a fixed head set therefore does not guarantee monotonic improvement for every generation, consistent with the motivation for conditioning intervention strength on the current decoding state.

The third example highlights grounded-content preservation. Static suppression omits the annotated object *bird*, whereas DEACT retains it without introducing another CHAIR-counted hallucination. This is consistent with the object-level F1 analysis: useful mitigation should reduce unsupported object mentions without obtaining gains solely through omission.

These examples follow the same annotation-based protocol as the quantitative evaluation. CHAIR determines object hallucination using COCO object annotations and synonym matching (Rohrbach et al., 2018); the examples should therefore be understood as representative outcomes under that protocol rather than independent human judgments of visual plausibility.

E LIMITATIONS

Our study focuses on object hallucination in open-ended image captioning. Although end-to-end behavior is evaluated across several LVLM architectures and on AMBER, the most detailed intervention analyses are conducted on LLaVA-1.5-7B. Whether the same head-level organization and response patterns extend to substantially different multimodal architectures, tasks, or forms of hallucination remains open.

DEACT also requires model-specific calibration. The effective actuator range, useful pool size, and distribution of online text-reliance states vary across architectures, so a single intervention configuration is not assumed to transfer universally. The method is training-free in the sense that it requires no parameter updates, but it is not calibration-free and requires a one-time architecture-specific calibration stage before held-out inference.

Finally, the default attenuation operator is non-renormalized. It can therefore affect both the targeted post-image text-side contribution and the magnitude of the corresponding head output. The mass-preserving controls in Appendix C.2 show that several natural redistribution strategies do not reproduce the same operating point, but they do not fully separate pathway-specific attenuation from generic head-output magnitude effects. A finer causal decomposition of how these local interventions propagate through downstream residual computation remains future work.