Mass-Tail SHAP Entropy: A Regime-Conditional Decoder for Vision-Language Model Hallucination

Anonymous ACL submission

Abstract

We use SHAP attribution to both **detect** the regime in which a vision-language model (VLM) is hallucinating and to **reduce** that hallucination at decode time. The detection signal is the normalized entropy H of per-segment leave-one-out (LOO) Shapley attributions over a panoptic partition of the input image ("SHAP-H"; the underlying quantity is the LOO term of the Shapley decomposition, not the full Shapley value). We show that VLM hallucination is not a uniform failure mode but a **mass-tail phenomenon**: model inference (both correct and incorrect outputs) concentrates in a single tail of H, and the tail position is task-type-conditional. Caption-style tasks (CHAIR) pull inference mass into the **over-concentration** tail (low H, 53–66% of samples), while binary-QA tasks (POPE, MME) pull it into the **over-spread** tail (high H, 74–95% of samples). Across LLaVA-1.5-7B and Qwen2.5-VL-7B, six of six (backbone $\times$ dataset) combinations show this pattern, with 62–100% of all hallucinations *and* 51–96% of all correct outputs co-located in the dominant tail. Building on this SHAP-based detector, we propose **SBC** (SHAP-Based Conditioning), an adaptive decoder that reads H at every step and gates between two regime-specific interventions—Modulated-Segment Boost (MSB) for over-concentration, Probabilistic Marginal Inference (PMI) for over-spread. Because the gate routes intervention to the regime where the model's inference mass actually lives, SBC reduces CHAIR_I on CHAIR while preserving recall, and improves POPE F1 over baseline on both backbones (+1.94 LLaVA, +3.30 Qwen averaged across splits), in contrast to attention-modulation baselines that sacrifice $\sim$7pt recall to lower CHAIR_S. Ablations show that applying a single intervention to the wrong tail *worsens* hallucination (e.g., PMI on CHAIR: CHAIR_S 49.3 $\rightarrow$ 60.3), demonstrating that SHAP-based regime detection is mechanically necessary for the reduction step to be safe.

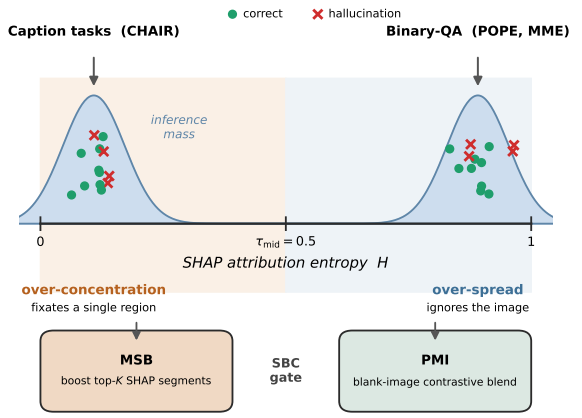

Figure 1: Model inference concentrates in one tail of the per-segment SHAP attribution entropy H; the tail is task-conditional (caption tasks $\rightarrow$ over-concentration, binary-QA $\rightarrow$ over-spread), and hallucinated outputs are co-located with correct ones in that dominant tail. SBC reads H and routes the regime-appropriate intervention (MSB / PMI). Conceptual overview; the measured distributions are in Figure 3.

1 Introduction

Existing VLM hallucination mitigation methods (Leng et al., 2024; Huang et al., 2024; Liu et al., 2026) apply a single intervention at decode time regardless of the underlying mechanism, and most pay for hallucination reduction with a substantial recall loss (e.g., AIF: -6.6pt on CHAIR). We argue this trade-off arises because hallucination is *regime-structured*, not monolithic: it has two distinct mechanisms tied to two tails of the SHAP-attribution entropy distribution (Figure 1). We therefore use SHAP attribution twice—first as a per-step **detector** of which hallucination regime the model is currently in, and then as a control signal that picks the matching **reduction** intervention. Diagnosing the regime per generation step enables targeted intervention without uniformly suppress-

ing the generation distribution, and is what enables recall preservation.

Contributions.

1. **SHAP-based mass-tail *detector* of VLM hallucination** (§3). Using SHAP attributions over panoptic image segments to compute a normalized entropy H, we show across 2 backbones $\times$ 3 datasets that (i) model inference *mass* concentrates in a single H-tail per task type (over-concentration for caption, over-spread for binary-QA); (ii) hallucinations *and* correct outputs are co-located in this dominant tail (62–100% of errors, 51–96% of correct outputs); and (iii) the pattern is architecture-invariant and dataset-task-conditional. SHAP-H thus acts as a regime detector that exposes *where* the model is hallucinating, before any intervention is applied.
2. **SBC: a SHAP-driven decoder that *reduces* hallucination** (§4–§5). Driven by the per-step SHAP detector above, a gate routes between MSB (additive attention boost on top-K SHAP-attributed visual segments, for over-concentration) and PMI (blank-image contrastive logit blending, for over-spread). The gate fires PMI only when three conditions concur: $H \geq \tau_{\text{mid}}$, the blank-image continuation matches the on-image span (the answer is prior-driven), and the on-image top-token margin is below δ (the model is not already confident); otherwise it routes to MSB. Within binary-QA the $H \geq \tau$ condition is largely redundant; the cross-task discrimination is supplied by blank-agreement, and SHAP-ϕ remains load-bearing inside the MSB route as the segment-ranking signal (Appendix J). To our knowledge this is the first decoder driven by per-step LOO attributions over panoptic segments.
3. **A task-agnostic, cross-regime decoder** (§6). Because the gate routes intervention to the regime where inference actually concentrates, SBC tracks the regime-appropriate single intervention on each task without task knowledge: matching MSB-only on captioning (LLaVA-7B CHAIR_I 13.14 vs 13.25) and PMI-only on binary-QA (Qwen POPE F1 88.17 vs 89.40). Single interventions fail asymmetrically on the opposite task—PMI actively worsens captioning (CHAIR_S 49.3$\rightarrow$60.3) while MSB only marginally affects binary-QA ($\leq$ 1.1pt F1). The contribution is cross-regime robustness, not per-task SOTA; on LLaVA-7B POPE adversarial this costs $\sim$6pt recall vs. the PMI-only specialist (§7).

We position this work as **detection-first, reduction-second**: the SHAP-based mass-tail finding is the primary scientific contribution, and SBC demonstrates that the same SHAP signal can be turned into an actionable per-step reduction control. §2 situates the work against decoding-time, training-time, and attribution-based literature.

2 Related Work

Vision-language models. Modern VLMs couple a pretrained vision encoder with a large language model through a learned projection, trained with visual instruction tuning (Liu et al., 2023, 2024; Dai et al., 2023; Li et al., 2023a). Recent backbones such as Qwen-VL and Qwen2.5-VL (Bai et al., 2023, 2025) scale this recipe with flexible-resolution visual tokenization. We evaluate on LLaVA-1.5-7B and Qwen2.5-VL-7B as two architecturally distinct 7B backbones.

Decoding-time hallucination mitigation. A large body of work reduces hallucination at inference without retraining. Contrastive-decoding methods subtract a counterfactual distribution: VCD (Leng et al., 2024) contrasts a noised image, ICD (Wang et al., 2024) contrasts a disturbed instruction, M3ID (Favero et al., 2024) amplifies image–text mutual information, and DoLa (Chuang et al., 2024) contrasts transformer layers. Others intervene on the decoding process directly: OPERA (Huang et al., 2024) adds an over-trust penalty with retrospection-allocation, and HALC (Chen et al., 2024) uses adaptive focal-contrast beam search. These methods apply a *single, uniform* intervention at every step; our diagnostic (§3) shows why this is mismatched to the regime-structured nature of hallucination.

Training-time and post-hoc mitigation. A complementary line retrains or post-edits the model. Preference-based methods align the model away from hallucination: RLHF-V (Yu et al., 2024) uses fine-grained correctional feedback, and HA-DPO (Zhao et al., 2023) casts hallucination as preference selection. Post-hoc methods rewrite outputs: LURE (Zhou et al., 2024) reconstructs

less-hallucinatory captions, and Woodpecker (Yin et al., 2024) validates and corrects claims with an external pipeline. Our method is decoding-time and training-free, requiring no additional data or model updates.

Hallucination evaluation. Object-level hallucination is measured by CHAIR (Rohrbach et al., 2018) on captioning and POPE (Li et al., 2023b) on polled yes/no probes. Broader suites include MME (Fu et al., 2023), the LLM-free multi-dimensional AMBER (Wang et al., 2023), and HallusionBench (Guan et al., 2024), which entangles language hallucination with visual illusion. We report CHAIR, POPE, and MME across two backbones.

Attribution and interpretability. Feature attribution explains a prediction by assigning credit to inputs: Shapley-value methods (Lundberg and Lee, 2017) and Integrated Gradients (Sundararajan et al., 2017) are model-agnostic, while attention-flow analyses (Abnar and Zuidema, 2020) are transformer-specific. We use leave-one-out Shapley attributions over panoptic segments not as a post-hoc explanation but as an *online decoding signal*—to our knowledge the first use of SHAP-style attribution to drive a per-step decoding gate.

3 SHAP-Based Mass-Tail Detection

3.1 SHAP Entropy as a Hallucination-Regime Detector

For a generation step, we partition the input image via Mask2Former panoptic segmentation (Cheng et al., 2022) into $K \leq 6$ segments and compute leave-one-out (LOO) SHAP attributions (Lundberg and Lee, 2017)

$$\phi_i = \log p(\text{span} \mid v) - \log p(\text{span} \mid v \setminus s_i) \quad (1)$$

over a lookahead span generated by a greedy probe (up to 32 tokens, terminated at the first sentence end), where v is the image and s_i the i-th segment. The normalized entropy is

$$H = -\frac{1}{\log K} \sum_i p_i \log p_i, \quad p_i = \frac{e^{\phi_i}}{\sum_j e^{\phi_j}}. \quad (2)$$

Five quintile bins partition H: *over-concentration* $[0, 0.2)$ (single-segment fixation), *concentration* $[0.2, 0.4)$, *mixed* $[0.4, 0.6)$, *spread* $[0.6, 0.8)$, and *over-spread* $[0.8, 1.0]$ (uniform, image-agnostic attribution).

Backbone	Data	Dom. tail	Tail n (%)
LLaVA-7B	CHAIR	over-conc.	158 (53%)
LLaVA-7B	POPE	over-spread	564 (95%)
LLaVA-7B	MME	over-spread	71 (85%)
Qwen-7B	CHAIR	over-conc.	197 (66%)
Qwen-7B	POPE	over-spread	525 (88%)
Qwen-7B	MME	over-spread	62 (74%)

Table 1: Distribution of H (n=297 CHAIR / 594 POPE (= 198/split ×3 splits) / 84 MME per cell; MME uses the hallucination subset, existence/count/position/color). The diagnostic uses a fixed-seed random sub-sample of the full evaluation sets (CHAIR 1000 images, POPE 3000q/split, MME 240q in Tables 3–5) because each cell requires $K+1$ forward passes per step under the leave-one-out occlusion of Eq. 1; the sub-sample is large enough ($n \geq 84$) to estimate the dominant-tail share to within ∼2pt. The dominant-tail position is invariant across architecture but bound to task type.

Why SHAP rather than raw attention: raw attention is normalized per-head and conflates "what the model looks at" with "what affects the output." The single-segment LOO occlusion in Eq. 1 is the leave-one-out term of the Shapley decomposition (we abbreviate it as "SHAP-ϕ"; the full Shapley value would average over all coalitions and was prohibitively expensive at our segment count). LOO already measures the counterfactual log-probability change—what would happen if a segment were absent—giving a faithful causal signal of which segments drive the output (Figure 2).

3.2 Inference Mass Concentrates in One Tail per Task Type

Table 1 reports the dominant H-tail for each backbone×dataset cell. Caption-style generation pulls mass into the over-concentration tail (53–66% of samples); binary-QA pulls it into the over-spread tail (74–95%). The tail position is **invariant across architecture** but **bound to task type**. Figure 3 shows the full five-bin H distributions.

3.3 Hallucinations Are Co-located with Correct Outputs

Table 2 shows that in six of six (backbone × dataset) cases, 62–100% of hallucinations *and* 51–96% of correct outputs occupy the same dominant tail. Hallucinations are not located in a "risky" tail away from correct outputs—they are *where the model spends its inference*.

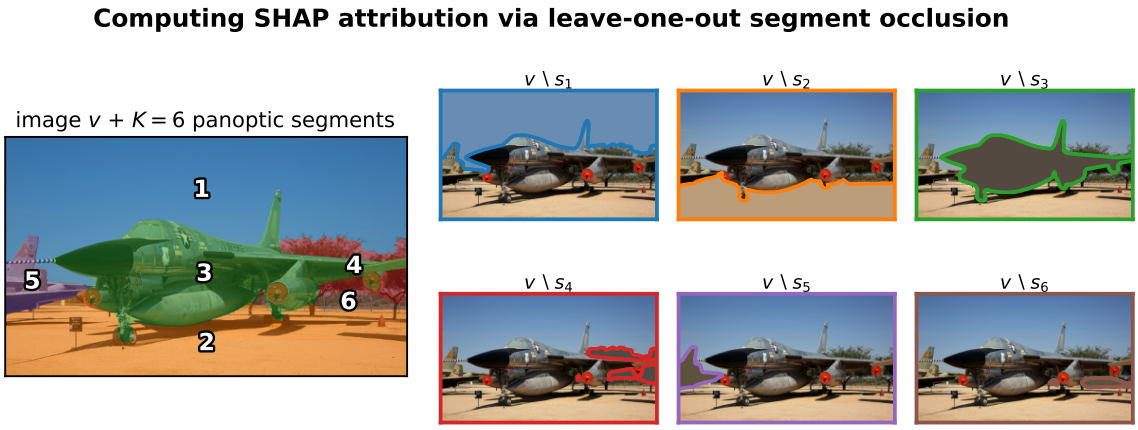

Figure 2: **Leave-one-out segment occlusion underlying SHAP-ϕ (CHAIR case, LLaVA-1.5-7B).** The image is partitioned into $K = 6$ Mask2Former panoptic segments; for each segment s_i, the masked variant $v \setminus s_i$ (image-mean fill of the segment region; segment outline drawn for visibility) is forward-passed to obtain $\log p(\text{span} \mid v \setminus s_i)$ over the same lookahead span, and the segment's attribution is $\phi_i = \log p(\text{span} \mid v) - \log p(\text{span} \mid v \setminus s_i)$. The resulting ϕ distribution, H, and gate route for this same CHAIR case are shown in Figure 5 (bottom row).

Backbone	Data	Err. %	Corr. %
LLaVA-7B	CHAIR	62.1	51.4
LLaVA-7B	POPE	88.9	95.8
LLaVA-7B	MME	93.8	78.8
Qwen-7B	CHAIR	71.5	65.8
Qwen-7B	POPE	83.1	89.0
Qwen-7B	MME	100.0	70.3

Table 2: Co-location: the dominant tail holds the majority of *both* errors and correct outputs. Six of six cases.

3.4 Note on Per-Bin Rates

Per-bin error *rates* $P(\text{error} \mid \text{bin})$ show elevated values in middle bins (e.g., POPE *mixed* 37.5% on LLaVA-7B), but with very small bin sizes ($n = 8$–19) these rates are noise-sensitive. Our diagnostic rests on **mass-tail location** (Table 1) and **absolute error/correct distribution** (Table 2), not per-bin rate. Appendix D reports per-bin bootstrap CIs that quantify this noise floor.

4 Why a Gate Is Necessary

A single-intervention decoder applies the same correction at every step regardless of H. The mass-tail composition of Table 1 *predicts* two failure modes.

4.1 PMI-only Fails on Caption Tasks

On CHAIR, 53–66% of samples sit in over-concentration, where the language prior is *not* the failure mechanism. Applying PMI (blank-image contrastive) uniformly subtracts the language prior even where the model is fixating on the wrong image segment, pushing the output toward language-prior tokens. Empirically on LLaVA-1.5-7B, baseline CHAIR_S 49.3 $\rightarrow$ PMI-only 60.3 (+11pt, worse than baseline).

4.2 MSB-only Has Marginal Effect on Binary-QA

On POPE, 88–95% of samples sit in over-spread, where the model is not fixating but defaulting to the language prior. MSB activates only on the ~5% of over-concentration samples and does negligible work on the dominant tail. Empirically, MSB-only POPE F1 is within 1pt of baseline on both backbones (LLaVA-7B: 83.97 vs. baseline 84.10; Qwen2.5-VL: 85.97 vs. baseline 84.87, a +1.10pt effect that is small compared to PMI-only's +4.53pt).

4.3 SBC's Three-Condition Gate

A regime-conditional gate routes to PMI only when three conditions concur: (i) $H \geq \tau_\text{mid}$ ($= 0.5$) — the step lies in the over-spread regime; (ii) the blank-image continuation matches the on-image span — the answer is prior-driven, so subtraction can recover the image-grounded one; and (iii) the on-image top-token margin is below δ ($= 0.5$) — the model is not already confident, so PMI's prior-subtraction will not invert an already-correct image-grounded answer. Otherwise the gate routes to MSB. The conditions play asymmetric roles: τ_mid holds on 74–95% of binary-QA steps (88–95% on POPE, 74–85% on MME; Table 1), so the substantive cross-task discrimina-

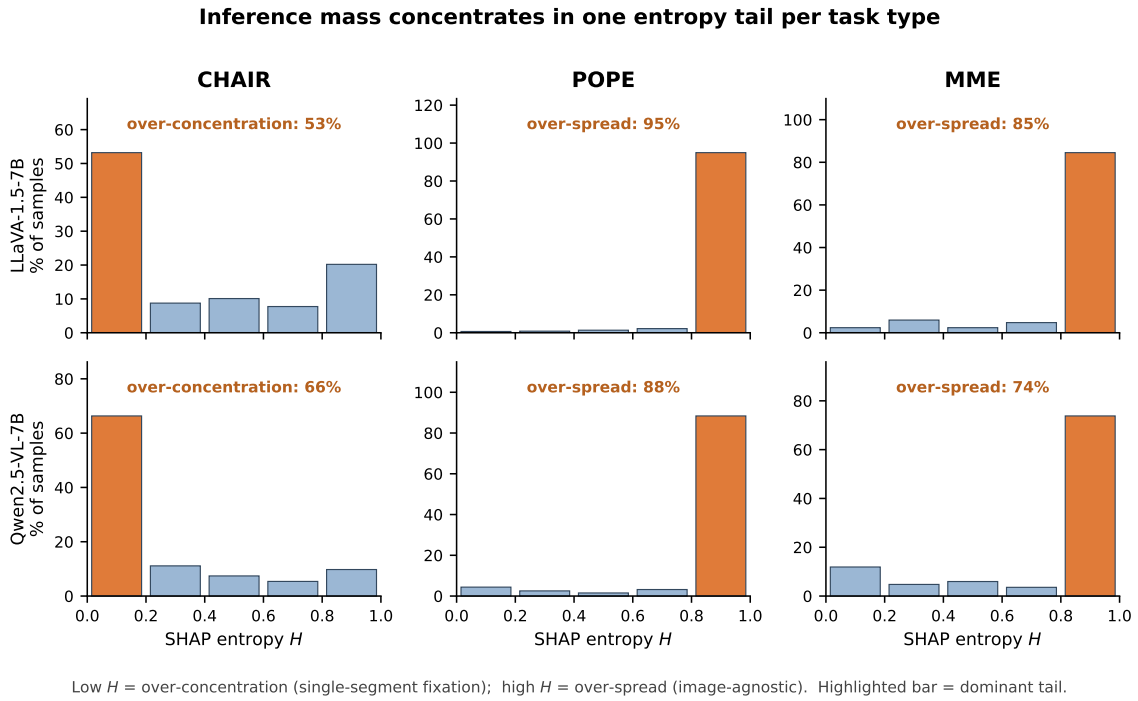

Figure 3: **Measured SHAP-H distributions.** Five-bin histograms of the per-segment attribution entropy H, for two backbones $\times$ three datasets. Caption-style CHAIR concentrates inference mass in the over-concentration tail (low H); binary-QA POPE and MME concentrate it in the over-spread tail (high H); the pattern holds across both backbones (six of six cells). The dominant-tail bar is highlighted with its share. MME uses the hallucination subset (existence/count/position/color), following prior work.

tion is supplied by blank-agreement (rarely true on free-form captions, hence 0% PMI on CHAIR in Table 7); δ guards within-binary-QA cases where PMI would over-correct (App J).

Because the dominant tail also holds the majority of correct outputs, the intervention affects errors and correct outputs *proportionally on average*, preserving averaged recall; within-route trade-offs (e.g., PMI break rate on POPE adversarial) are discussed in §7.

Why gate rather than pick one intervention? A task-specific decoder—PMI for binary-QA, MSB for captioning—can marginally beat SBC on each task in isolation (§7), but presupposes task knowledge and the *wrong* choice is heavily penalized (PMI on captioning inflates CHAIR_S). SBC is instead task-agnostic: per-step diagnostics (H, blank-agreement, margin) yield near-best performance in *both* regimes with one decoder.

5 SBC: SHAP-Based Conditioning

5.1 Overview

Figure 4 summarizes one decoding step. At each generation step: (1) a **lookahead probe** greedily continues up to 32 tokens to obtain a tentative span; (2) **SHAP-H** is computed—panoptic segments $\rightarrow$ LOO attributions (Eq. 1) $\rightarrow$ normalized entropy (Eq. 2); (3) the **gate** routes the step to PMI iff *all three* hold—$H \geq \tau_{\text{mid}}$ (over-spread regime), the blank-image continuation matches the on-image span under normalized exact-text equality (answer is prior-driven), *and* the on-image top-token probability margin $< \delta$ (model is not already confident); otherwise to MSB. A plain-greedy fallback fires when $K < 2$ or no span is found. (4) The selected **intervention** (§5.2) produces the next-token logits. The three PMI conditions are jointly motivated in §4.

5.2 Interventions

The two interventions are deliberately standard mechanisms; SBC's novelty is the SHAP diagnostic that decides *which* to apply and *where*.

MSB (Modulated-Segment Boost). In the over-concentration regime the model fixates on a single segment; MSB redistributes visual attention toward the segments that the SHAP attribution supports. Let $\mathcal{V}_K$ be the visual-token columns belonging to the top-K segments ranked by ϕ_i. For every

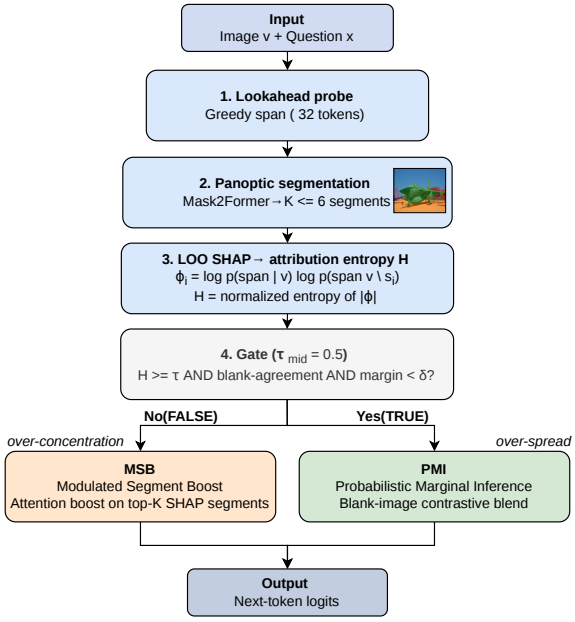

Figure 4: **SBC decoding at one generation step.** SHAP-based diagnostic stages (lookahead probe, panoptic segmentation, leave-one-out attribution) yield the normalized attribution entropy H, which acts as a regime classifier. The gate routes PMI only when three conditions concur ($H \geq \tau_{\text{mid}}$, blank-image continuation agreement, and image-margin $< \delta$); otherwise MSB. Each intervention targets the regime where the model's inference mass concentrates (§3). Default hyperparameters: $\tau_{\text{mid}} = 0.5$, top-$K = 2$, boost factor $b = 1.8$, PMI $\alpha = 1.0$, $\beta = 0.1$, image-margin $\delta = 0.5$. Fallback to plain greedy decoding when fewer than two segments or no span is found.

text-query row t and visual column j, MSB adds a constant log-boost to the pre-softmax attention logit:

$$s'_{t,j} = s_{t,j} + \log b \cdot \mathbf{1}[\, j \in \mathcal{V}_K \,], \tag{3}$$

so each boosted column receives a multiplicative factor b in the attention distribution. We use $K = 2$ and $b = 1.8$; MSB adds no forward pass beyond the boosted prefill.

PMI (Probabilistic Marginal Inference). In the over-spread regime the model defaults to the language prior. PMI isolates and subtracts that prior by contrasting the image-conditioned logits against blank-image logits—a blank-conditioned variant of visual contrastive decoding (Leng et al., 2024). With image v, blank image $v_\emptyset$, and query x:

$$\ell_{\text{PMI}}(y) = (1{+}\alpha)\, \ell_\theta(y \mid v, x) - \alpha\, \ell_\theta(y \mid v_\emptyset, x), \tag{4}$$

restricted to the adaptive-plausibility set $\{y : p_\theta(y \mid v, x) \geq \beta \cdot \max_w p_\theta(w \mid v, x)\}$ so that implausible tokens are not amplified. We use $\alpha = 1.0$ for the contrastive blend and $\beta = 0.1$ for the plausibility cutoff.

Relation to AIF and VCD. Neither actuator is claimed as a contribution; we state the changes from their sources precisely. MSB's actuator—scaling text-to-visual attention—follows attention-flow methods such as AIF (Liu et al., 2026), but inverts and re-targets it: AIF *masks* individual visual tokens deemed unimportant by an attention-dynamics entropy (an internal, correlational proxy), whereas MSB *boosts* the top-K panoptic segments ranked by leave-one-out SHAP ϕ, a causal occlusion signal (Eqs. 1–2). PMI's actuator (Eq. 4) is the contrastive logit blend of VCD (Leng et al., 2024) with one change: the contrastive reference is a blank, zero-information image rather than VCD's diffusion-noised image, so the subtracted term is the pure language prior rather than a partially corrupted image distribution; VCD's adaptive-plausibility constraint is reused unchanged. In both cases the mechanism is commodity—the contribution is the upstream SHAP diagnostic that decides *which* actuator fires and *what* it targets.

5.3 Hyperparameter Robustness

τ_{mid} is robust on $[0.3, 0.7]$ (Table 10); b is in Table 11. The image-margin δ is task-conditional (App H); we choose $\delta = 0.5$ as the POPE-priority operating point. The gate logic is task-agnostic per step; the choice of a fixed δ is deployment-time.

6 Experiments

6.1 Setup

We use the official LLaVA-1.5-7B (Liu et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025) HuggingFace checkpoints with greedy decoding on COCO val2014 (Lin et al., 2014). SBC defaults: $\tau_{\text{mid}} = 0.5$, $K = 2$, $b = 1.8$, PMI $\alpha = 1.0$, $\beta = 0.1$, image-margin $\delta = 0.5$. CHAIR (Rohrbach et al., 2018) uses 1000 images with the standard synonym matcher; POPE (Li et al., 2023b) uses the standard 3000-question splits (random/popular/adversarial). MME (Fu et al., 2023) contributes its hallucination subset (existence, count, position, color; 240 yes/no questions), scored by the standard accuracy+accuracy$^+$ metric. All baselines—VCD (Leng et al., 2024),

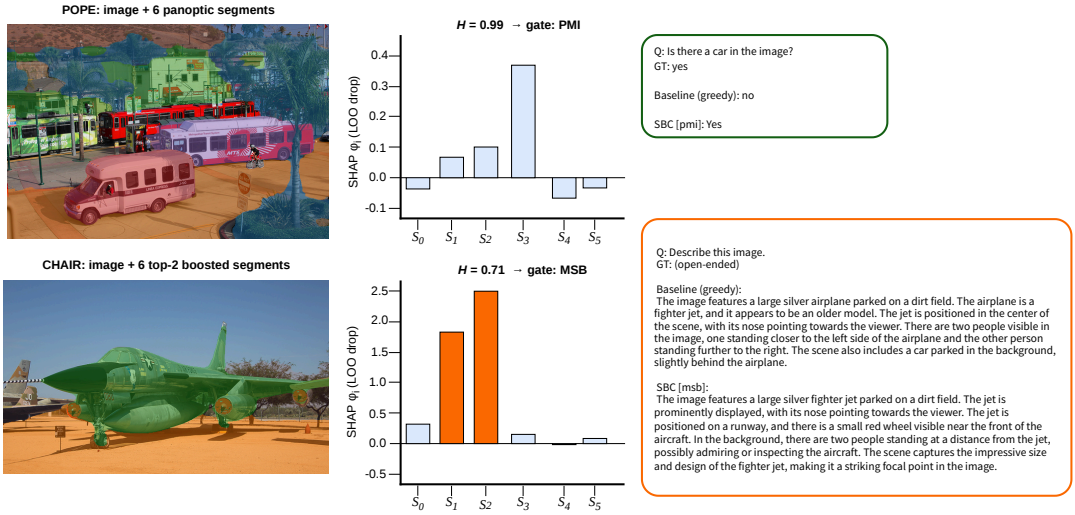

Figure 5: **SBC in action.** One POPE example (gated to PMI) and one CHAIR example (gated to MSB) on LLaVA-1.5-7B: panoptic segments yield the per-segment SHAP ϕ and entropy H; the gate routes to PMI or MSB; the intervention removes the baseline hallucination.

OPERA (Huang et al., 2024), and AIF (Liu et al., 2026)—are re-implemented under a unified wrapper to eliminate eval-harness confounds. VCD and OPERA are ported from their official public code; AIF has no public reference implementation and is reproduced from the paper description (we verify the published LLaVA-1.5-7B CHAIR ranking is preserved under our greedy setup, which the original paper runs with sampling). These span the main intervention families (contrastive decoding, over-trust penalty, attention modulation). We restrict to 7B-class backbones consistent with prior work (Leng et al., 2024; Huang et al., 2024).

6.2 Main Results

Table 3 (CHAIR) and Table 4 (POPE) report results for both backbones. SBC reduces per-instance hallucination (CHAIR_I) without recall loss: on LLaVA-7B it attains CHAIR_I 13.14 at Recall 79.1, against baseline 14.13/79.6. AIF attains the lowest CHAIR_S (46.2) but by shortening captions (Recall -6.6pt). VCD and PMI-only *worsen* hallucination—PMI-only's failure is predicted by §4. On POPE (Table 4), SBC improves F1 over baseline ($+1.94$ LLaVA, $+3.30$ Qwen averaged); the gate routes POPE to PMI, so SBC inherits PMI's precision/recall trade-off, which we discuss in §7. On MME (Table 5), *all decoding-time methods in our greedy re-implementation—including the contrastive baselines (VCD, AIF, PMI-only) and SBC—trail the greedy baseline*; SBC attains 598.3 on LLaVA-7B (vs. baseline 631.7) and 663.3 on Qwen2.5-VL (vs. baseline 680), an absolute score that *exceeds* the same-setup VCD (561.7 / 673.3). The greedy baseline is unusually strong (VCD's original sampling-baseline MME is 380.3, illustrating the gap); we report the same-setup trend without claiming improvement against the greedy baseline.

Method	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$	Recall↑
LLaVA-1.5-7B			
Baseline	49.3	14.13	79.6
VCD	57.5	16.75	80.0
OPERA	51.9	14.36	78.6
AIF	**46.2**	14.81	73.0
MSB-only	47.8	13.25	79.2
PMI-only	60.3	16.38	82.3
SBC (ours)	47.4	**13.14**	79.1
Qwen2.5-VL-7B			
Baseline	22.7	8.12	58.1
VCD	30.1	8.84	60.1
OPERA	**19.2**	6.54	53.7
AIF	23.6	6.96	57.8
MSB-only	23.4	**5.69**	58.6
PMI-only	28.0	7.51	61.2
SBC (ours)	23.4	5.83	58.6

Table 3: CHAIR (1000 images). SBC numbers are at boost factor $b = 1.8$. SBC 95% bootstrap CI (1000 image resamples): LLaVA CHAIR_I $[12.01, 14.23]$, Qwen CHAIR_I $[4.96, 6.86]$.

6.3 Ablations

The MSB-only and PMI-only rows of Tables 3–4 are single-mode ablations: each fails on the opposite-tail task, as predicted by §4. Per-hyperparameter ablations (gate threshold τ_{mid},

Method	Acc	F1
LLaVA-1.5-7B		
Baseline	85.50	84.10
VCD	84.63	84.47
OPERA	**85.57**	84.13
AIF	85.53	84.83
MSB-only	85.43	83.97
PMI-only	84.53	85.50
SBC	**86.23**	**86.04**
Qwen2.5-VL-7B		
Baseline	86.37	84.87
VCD	88.27	87.20
OPERA	87.40	86.03
AIF	88.43	87.37
MSB-only	87.40	85.97
PMI-only	**89.77**	**89.40**
SBC	88.93	88.17

Table 4: POPE Acc/F1 averaged over the three splits (random / popular / adversarial). SBC improves F1 over baseline (+1.94 LLaVA, +3.30 Qwen); per-split breakdowns in App B. 95% bootstrap CI for SBC F1 (1000 resamples per split): LLaVA $[85.26, 86.89]$, Qwen $[87.49, 88.92]$ — CI lower bound exceeds the baseline F1 on both backbones.

Method	LLaVA-1.5-7B	Qwen2.5-VL-7B
Baseline	631.7/85.4	680.0/90.0
VCD	561.7/79.6	673.3/89.2
OPERA	631.7/85.4	680.0/90.0
AIF	613.3/83.3	668.3/88.8
MSB-only	625.0/84.6	680.0/90.0
PMI-only	518.3/75.4	658.3/87.9
SBC (ours)	598.3/82.1	663.3/88.3

Table 5: MME hallucination subset (240 yes/no, existence / count / position / color). Cells: MME score (sum of subtask acc+acc$^+$, max 800) / overall accuracy. SBC 95% bootstrap CI: LLaVA $[556, 639]$, Qwen $[632, 697]$ — the small n=240 leaves per-method gaps inside the CI; we discuss the greedy-baseline mismatch in App H.

boost factor b, image-margin δ, joint $\tau \times b$) are in Appendices E and H; SBC is robust within $\tau_{\text{mid}} \in [0.3, 0.7]$ and $b \in [1.5, 2.0]$. δ is task-conditional (Appendix H). Gate routing statistics (Appendix C) decompose how the three-condition gate fires across (backbone, dataset) cells, and Appendix A collects reproducibility details.

7 Discussion

Single intervention vs. SBC. On its own task a single intervention can match SBC (POPE F1 within 1.2pt of PMI-only, CHAIR CHAIR_I within ± 0.2 of MSB-only). The decisive comparison is the opposite task: PMI inflates CHAIR_S on captioning (49.3→60.3) and MSB moves POPE F1 by $\leq$ 1.1pt; SBC tracks the regime-appropriate intervention without task knowledge—cross-regime robustness, not per-task SOTA, is the contribution.

What is novel. PMI and MSB are commodity actuators (variants of Leng et al. (2024) and attention modulation); the contribution is upstream—SHAP-attribution entropy H as both a detector of mass-tail regime structure (§3) and an online control signal selecting the actuator.

Residual trade-offs. The $\delta = 0.5$ guard prevents PMI from inverting a confident image-grounded answer but cannot detect prior-driven correct "no"s with narrow margins; on LLaVA POPE adversarial every PMI-routed case (n=281) flips (Brk 100%, Table 8); without the guard n rises to $\sim$1700. This is absent on Qwen (87.5 vs. baseline 85.4); with two backbones we cannot claim the effect is LLaVA-specific. On MME all decoding-time methods trail the strong greedy baseline (the ranking inverts under VCD's sampling baseline).

Is the mass-tail just task-format dynamics? The *location* of the tail is task-driven, but intervening on it measurably changes hallucination: PMI on the caption tail inflates CHAIR_S (49.3→60.3), MSB on the binary-QA tail moves POPE F1 by $\leq$ 1.1pt, and SBC lowers CHAIR_I without recall loss while improving POPE F1. A pure task classifier would not produce these intervention-specific asymmetries (App. J, F).

8 Conclusion

We use SHAP attribution to both *detect* and *reduce* VLM hallucination. The detector reframes hallucination as a **mass-tail phenomenon**: across 2 backbones × 3 benchmarks, model inference (correct and incorrect) concentrates in a single task-conditional tail of the per-segment LOO attribution entropy H (over-concentration for captioning, over-spread for binary-QA; six of six cells). The same tail holds most errors *and* correct outputs, which is why uniform decoders trade recall for precision. SBC turns the same SHAP signal into a per-step reduction control, gating between MSB (top-K segment boost) and PMI (blank-image prior-subtraction): it tracks the regime-appropriate intervention without task knowledge (LLaVA CHAIR_I 13.14 vs MSB 13.25; Qwen POPE F1 88.17 vs PMI 89.40) while the wrong choice is heavily penalized (PMI/CHAIR: CHAIR_S 49.3→60.3).

Limitations

Single-metric SOTA. AIF's CHAIR_S (46.2) is lower than ours; we do not chase CHAIR_S because doing so requires a recall sacrifice.

Adversarial regression hidden in F1. On LLaVA-7B the gate keeps averaged adversarial F1 at baseline (83.0 vs. 82.3), but the PMI route flips 100% of the $n = 281$ cases that survive the gate (Table 8); the margin-guard exists to keep n small. On Qwen POPE adversarial the same route shows a 35% break rate at $n = 173$. The break rate, not the F1 average, is the relevant cost metric for hallucination-sensitive use.

Per-bin rate noise. Middle H-bins ($n = 8$–19) are too small to reliably discriminate; our claims rest on mass-tail location, not per-bin rate.

Backbone scope. 7B-class only. 8-bit quantization on larger models confounds attention-magnitude methods (notably MSB), so we do not report quantized larger-model results; a fp16 13B sweep was not completed in time and is left to a future revision.

Benchmark scope. We evaluate on CHAIR, POPE, and the MME hallucination subset. Mixed-format benchmarks such as AMBER (Wang et al., 2023) and HallusionBench (Guan et al., 2024) would test mass-tail generalization beyond the (caption / binary-QA) split.

Future directions. The above limitations motivate four threads: (i) an fp16 13B/larger sweep to test whether mass-tail concentration persists with model size; (ii) systematic comparison of alternative per-step diagnostics (attention-flow entropy, vision-token concentration, logit-entropy) at fixed gate + actuator, since App. J shows SHAP-H is partly redundant on POPE and a cheaper surrogate may suffice; (iii) learning δ online from the on-image margin distribution rather than fixing it (App. H); and (iv) replacing the mean-color fill of Eq. 1 with a generative inpainting model to test whether H shifts when the occluded region is filled with plausible content (App. G already shows robustness to mean / zero / Gaussian fills).

Use of AI Assistants

We used a general-purpose conversational AI assistant (Anthropic Claude) during the preparation of this paper. Its use was limited to (i) editing and polishing draft prose, (ii) writing and debugging analysis scripts for aggregating measured benchmark outputs into tables, and (iii) helping to format references and LaTeX tables. All research ideas, claims, experimental design, code for the SBC method and its actuators, the mass-tail diagnostic, and the final verification of all reported numbers are the authors' own work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *International Conference on Machine Learning (ICML)*.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng,

Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chengxin Liu, Wonseok Choi, Chenshuang Zhang, and Tae-Hyun Oh. 2026. Aligning what vision-language models see and perceive with adaptive information flow. *arXiv preprint arXiv:2604.15809*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations (ICLR)*.

A Reproducibility Details

All decoders share one model wrapper and one Mask2Former panoptic segmenter. CHAIR uses a 1000-image subset of COCO val2014; POPE uses the standard 3000-question splits per setting. Lookahead is a single greedy continuation (max 32 tokens). Code and configuration will be released upon acceptance.

Method	random	popular	adv.
LLaVA-1.5-7B			
Baseline	87.1/85.6	85.9/84.4	83.5/82.3
VCD	87.7/87.1	85.2/84.9	81.0/81.4
OPERA	87.2/85.6	85.9/84.5	83.6/82.3
AIF	88.0/87.1	85.9/85.1	82.7/82.3
MSB-only	87.0/85.4	85.8/84.3	83.5/82.2
PMI-only	88.6/88.8	85.8/86.4	79.2/81.3
SBC	89.1/88.6	86.9/86.5	82.7/83.0
Qwen2.5-VL-7B			
Baseline	87.6/86.0	86.1/84.7	85.4/83.9
VCD	89.2/88.1	88.1/87.0	87.5/86.5
OPERA	88.2/86.8	87.4/86.0	86.6/85.3
AIF	89.5/88.4	88.4/87.3	87.4/86.4
MSB-only	88.2/86.8	87.4/85.9	86.6/85.2
PMI-only	91.7/91.2	90.0/89.6	87.6/87.4
SBC	90.3/89.5	88.9/88.2	87.5/86.9

Table 6: POPE Acc/F1 per split for both backbones. Average rows are in Table 4.

Cell	Split	MSB	PMI	Base/FB	n
POPE, LLaVA-1.5-7B					
	random	39.1	6.3	54.6	3000
	popular	40.3	7.5	52.2	3000
	adversarial	42.7	9.4	47.9	3000
MME, LLaVA-1.5-7B					
	–	71.7	5.8	22.5	240
CHAIR, LLaVA-1.5-7B					
	–	99.4	0.0	0.6	1000
POPE, Qwen2.5-VL-7B					
	random	40.7	4.0	55.3	3000
	popular	41.5	4.9	53.6	3000
	adversarial	42.6	5.8	51.6	3000
CHAIR, Qwen2.5-VL-7B					
	–	99.4	0.0	0.6	1000
MME, Qwen2.5-VL-7B					
	–	52.1	10.4	37.5	240

Table 7: Gate route distribution (% of decoding steps) at the operating point $\delta = 0.5$. On POPE the over-spread regime dominates and PMI is gated; on CHAIR the over-concentration regime dominates and PMI is never fired (the blank-image-match guard vetoes it). The image-margin δ guard vetoes PMI on roughly half of POPE steps where entropy and blank-agreement would otherwise trigger it (the row's Base/FB pools margin-vetoed and $K < 2$-fallback steps), leaving 4–10% active PMI on both backbones.

B POPE Per-Split Results

Table 4 reports POPE accuracy and F1 averaged over the three official splits. Table 6 gives the per-split breakdown, showing where the averaged numbers come from. The adversarial split is the hardest (negation-heavy distractors); on LLaVA-7B, SBC's adversarial accuracy (82.7) is within 0.8pt of baseline (83.5), but Table 8 reveals that the PMI route on this split flips *every* baseline-correct case it touches ($n = 281$, Brk 100%)—a regression that is hidden in the averaged F1 because the margin-guard keeps the PMI volume small. PMI's prior-subtraction is too aggressive when an image-grounded "no" answer is already correct; this failure mode does not reproduce on Qwen2.5-VL (adversarial 87.5 vs baseline 85.4). The averaged F1 in Table 4 remains positive on both backbones.

C Gate Routing Statistics

Table 7 reports the empirical distribution of gate routes across (backbone $\times$ dataset) cells. On POPE (LLaVA-1.5-7B), SBC routes only 6–9% of questions to active PMI and 39–43% to MSB; the remaining 48–55% fall into the Base/FB bucket (steps where the image-margin guard δ vetoes a PMI candidate, or where Mask2Former returns $K < 2$ segments—the latter a constant $\sim$0.6%). The active-PMI share grows slightly on the adversarial split (6.3$\rightarrow$9.4%), consistent with the over-spread regime expanding as the negative pool widens. On CHAIR (both backbones), routing is overwhelmingly to MSB (99.4%): PMI is *never fired* on CHAIR because the "blank-image continuation matches the on-image span" gate condition is rarely satisfied on free-form captioning—when the model has fixated on the wrong image region (low H), the blank-image continuation diverges. The Base/FB column on CHAIR (0.6%) is the $K < 2$ fallback; on POPE it pools the δ-veto and fallback steps.

D Per-Bin Error Rates with Bootstrap CIs

Section 3 argues that per-bin error *rates* are noise-sensitive at the bin sizes available. Table 9 makes this quantitative: middle bins (*concentration*, *mixed*, *spread*) carry only $n = 20$–80 POPE questions per split, so their 95% bootstrap intervals span 15–30 percentage points and overlap both the over-spread bin and zero. The *over-spread* bin, in contrast, is tightly estimated ($\pm$1.4pt) because it holds 94–95% of the mass. Our claims in §3 (mass-tail location, absolute error/correct counts) rest on these tight estimates, not on the middle-bin rates.

E Hyperparameter Ablations

We ablate the gate threshold τ_{mid} (Table 10), the MSB boost factor b (Table 11), and the joint $\tau \times b$ grid (Table 12). All sweeps fix the remaining hyperparameters at the sweep defaults (sweeps were conducted at the earlier $\delta = 0.3$; conclusions about τ, b robustness are unaffected by the operating-point change to $\delta = 0.5$). PMI parameters (α, β) are not swept; they are fixed at the VCD defaults $(\alpha{=}1.0, \beta{=}0.1)$.

The joint sweep below verifies that τ_{mid} and b do not interact in a way that changes the optimum:

F SBC Failure Cases

Figure 5 (§7) shows two cases where SBC corrects a baseline hallucination. For honesty we also list two representative failure cases, drawn from the POPE adversarial split where SBC's PMI route is most likely to flip a baseline-correct "no" into a wrong "yes" (Table 8: LLaVA-7B PMI Brk 100% on the guard-filtered subset $n = 281$; Qwen2.5-VL shows a similar $\sim$35% PMI break rate at smaller volume $n = 173$ because the margin-guard pre-empts most candidates). The pattern is consistent: when the question asks about a category that is *associated* with objects present in the image, the blank-image prior agrees with the on-image distribution, the gate routes to PMI, and the prior-subtraction amplifies the "yes" token even though the queried object is absent.

Quantitatively, on the LLaVA-7B adversarial split (numbers below are from the LLaVA POPE routing log used to populate Table 8), there are 161 (question, image) pairs where baseline was correct and SBC flips to wrong; $151/161 = 94\%$ of these flips are PMI-routed, and every one of the top diverse-category cases shows the same direction: baseline "No" $\rightarrow$ SBC "Yes". Two representative cases:

Failure 1 (COCO val2014 #454607). *Q:* "Is there a dining table in the image?" *GT:* no. *Baseline:* "No" (correct). *SBC:* $H = 0.999$, $K = 6$ segments, routed to PMI, output "Yes" (wrong).

Failure 2 (COCO val2014 #287305). *Q:* "Is there a traffic light in the image?" *GT:* no. *Baseline:* "No" (correct). *SBC:* $H = 1.00$, $K = 6$ segments, routed to PMI, output "Yes" (wrong).

The common mechanism: when the queried category is *plausible given the scene* (kitchen $\rightarrow$ dining table; street $\rightarrow$ traffic light), the blank-image continuation places noticeable probability on the same token as the on-image one. PMI's contrastive blend $(1{+}\alpha)\ell(\cdot|v) - \alpha\ell(\cdot|v_\emptyset)$ then *amplifies* rather than cancels that token, flipping "No" to "Yes". Both cases sit at $H > 0.99$ with $K = 6$, satisfy blank-agreement, and have on-image top-token margin below $\delta = 0.5$, so all three PMI conditions held simultaneously. These residual cases are the regime where even the three-condition gate cannot separate prior-driven from narrowly image-grounded "no"; tightening δ would reduce them at the cost of more skipped PMI applications on genuine over-spread steps, a trade-off discussed in §7.

G Mask-Fill Robustness

Feedback raises the concern that the mean-color fill used in our leave-one-out occlusion (Eq. 1, §3) is itself out-of-distribution and could push ϕ off a manifold the model was trained on, so that the resulting H encodes a masking artifact rather than a faithful causal contribution. To probe this, we recompute H on 298 unique POPE random-split images under three different fills of the masked segment, holding the segmentation and decoding pipeline fixed: **mean-color** (paper default), **zero** (pure black), and **Gaussian** (i.i.d. noise with the image's per-channel mean and standard deviation, clipped to $[0, 255]$).

Table 13 reports the result. All three fills yield the same H median (0.998), the same H mean to two decimal places (0.94), and over-spread shares within a $\pm$0.5pt window. The five-bin histograms differ only at the noise level of a few counts. We read this as direct evidence that the mass-tail location reported in §3 is a property of the model's attribution structure under occlusion, not an artifact of one specific fill choice.

H Image-Margin Threshold δ Ablation

We sweep $\delta \in \{0.0, 0.1, 0.3, 0.5\}$ on both backbones with all other hyperparameters fixed at the defaults (§6). Table 14 reports POPE (random 1000q) and MME (hallucination subset). $\delta = 0$ disables the third condition (PMI fires whenever the entropy and blank-agreement conditions hold); $\delta \rightarrow 1$ makes the third condition almost always reject, so SBC approaches the MSB-only baseline.

PMI must be applied—but not too aggressively. The sweep shows a non-monotonic effect on POPE: $\delta = 0.0$ (PMI uniformly applied, no guard) gives strong but unguarded PMI, $\delta \rightarrow 1$

(no PMI) collapses to an MSB-only baseline with no POPE improvement, and intermediate δ values selectively apply PMI where prior-subtraction is most needed. On POPE, larger δ is consistently better on both backbones (LLaVA F1 $85.9 \rightarrow 88.8$, Qwen F1 $88.5 \rightarrow 90.3$ as δ moves $0.1 \rightarrow 0.5$; Recall similarly increases) because POPE's yes-bias is exactly the kind of prior PMI was designed to subtract.

Why MME prefers smaller δ (a greedy-baseline artifact). MME shows the opposite trend (LLaVA $598.3 \rightarrow 635.0$ at $\delta = 0.1$; Qwen 685.0 at $\delta = 0.1$ vs. 663.3 at $\delta = 0.5$). The mechanism is rooted in the choice of baseline. VCD and PMI were originally designed for a *sampling* baseline (VCD's original MME baseline is $\sim$380, §6), where prior-subtraction recovers from random-sampling yes-bias. Against our *greedy* baseline ($\sim$680), the greedy prior is already accurate on perception-style MME questions (existence/count/position/color), so PMI's subtraction over-corrects; smaller δ disables PMI on MME and keeps the output close to the strong greedy baseline. This is consistent with the observation that *all* contrastive decoders—VCD, AIF, PMI-only, SBC—trail the greedy baseline on MME in our setup (§6, Table 5).

Operating point: $\delta = 0.5$. We treat POPE as the primary hallucination metric and select $\delta = 0.5$ as the operating point because it gives the best F1 and Recall on POPE across both backbones (above). MME loss at $\delta = 0.5$ is comparable to other contrastive baselines and reflects the greedy-baseline mismatch rather than a flaw in SBC's gate; we do not optimise for MME because the metric does not separate hallucination from general perception accuracy. The choice exposes the trade-off transparently: with $\delta = 0.5$ POPE F1 improves substantially while MME drops by $\sim$30pt subset-score (still above PMI-only on Qwen). A task-conditional δ (small for perception-style benchmarks, larger for prior-bias benchmarks) is left to future work.

I Is SHAP-H Necessary, or Does Attention-H Suffice?

A natural concern is whether the per-segment SHAP-attribution entropy H provides information beyond simpler signals — in particular, the entropy of the model's own visual-token attention weights aggregated over the same segments (*attention-H*). If H and attention-H were nearly identical, mass-tail diagnosis could be reduced to reading the attention map and SBC's gate would carry no SHAP-specific contribution.

Theoretical distinction. SHAP-ϕ for segment i measures the *causal* contribution to the model's $\log p$ of its own answer ($\phi_i = \log p(\text{span} \mid x, v) - \log p(\text{span} \mid x, v_{\setminus i})$, Eq. 1). Attention measures the softmax-normalised query–key *weight* on visual tokens at each step. These two quantities diverge whenever (i) the model attends to features that do not change the output (sycophantic attention or residual-stream bypass), or (ii) the model relies on features it does not strongly attend to. The gate's discriminative purpose — separating over-concentration from over-spread regimes by where the answer's *causal* dependence lives — is therefore not interchangeable with attention.

Empirical correlation. On N=94 LLaVA-1.5-7B POPE random-split samples (one per image; default segmentation, mean-fill occlusion), the Pearson correlation between SHAP-H and last-layer attention-H is $r = -0.016$ with means $\bar{H}_{\text{SHAP}} = 0.980 \pm 0.094$ vs. $\bar{H}_{\text{attn}} = 0.759 \pm 0.110$. The two signals are essentially uncorrelated — $\sim$100% of the variance in SHAP-H is unexplained by attention-H — so the two are not redundant in any practical sense. Empirically SHAP-H sits much higher (binary-QA is the over-spread regime), and the rank order on individual samples disagrees on essentially every example. See Appendix J for the downstream consequence: an attention-H gate behaves similarly to the SHAP-H gate on average but a logit-entropy gate collapses POPE recall by 8pt.

J Diagnostic-Necessity Ablation

App I shows SHAP-H and attention-H are uncorrelated signals. Here we ask the actuator-level question: *what happens if we replace the SHAP-H condition in the gate with a cheaper alternative, or drop it altogether?* We hold the SBC actuators (MSB, PMI), the blank-agreement check, the image-margin guard (δ=0.5) and the operating point fixed, and only vary how the entropy-style condition is realised. All runs use LLaVA-1.5-7B on the POPE random sub-sample (1000q, same prompt suffix as the main results).

Cross-task evidence: blank-agreement is the load-bearing gate. Within POPE-random the gate conditions are partially redundant ($H \geq \tau$ and blank-agreement almost always hold), so `no_h` and `pmi_guard` match SBC on F1. The cross-task role of blank-agreement is exposed when the same `pmi_guard` variant is run on CHAIR and MME (Table 16). On Qwen2.5-VL CHAIR the absence of blank-agreement inflates CHAIR_S from 23.4 to 27.6 (+4.2pt above baseline 22.7): δ alone cannot prevent PMI from firing on captioning steps with narrow on-image margins. On LLaVA-7B CHAIR the same ablation lands at 49.9 (close to baseline 49.3), because δ already vetoes most PMI candidates under LLaVA's larger margins. The asymmetry is the proof: blank-agreement is the cross-task discriminator that protects free-form captioning, while δ is the within-binary-QA guard.

Take-away (what SHAP-H contributes inside the gate). The ablation tightens our claim, and we report it honestly. On binary-QA POPE the $H \geq \tau$ condition holds for 88–95% of steps (and 74–85% on MME; Table 1), so dropping it (`no_h`) or substituting an oracle task indicator (`task_pmi`) leaves POPE F1 unchanged. The cross-task discrimination that prevents PMI from firing on captioning steps is carried by the *blank-image-agreement* check (which is true on prior-driven binary-QA but rarely true on free-form captioning) and the δ guard, not by the H threshold. Dropping blank-agreement on top (`pmi_guard`, Table 16) confirms this division of labour: it is benign on POPE-random but inflates Qwen-CHAIR CHAIR_S by +4.2pt, isolating blank-agreement as the cross-task discriminator. Substituting H with a cheap surrogate (`logit_h`) or fixing the route to MSB (`task_msb`) yields the MSB-only result because the gate then never fires PMI. SHAP-H still does two things SBC needs: (i) it is the *diagnostic quantity* whose distribution exposes the mass-tail in the first place (Tables 1–2), and (ii) the same per-segment ϕ_i that defines H ranks the top-K segments inside MSB (Eq. 3); the within-MSB attribution signal cannot be replaced by a scalar entropy. The primary scientific contribution is therefore the mass-tail observation; SBC is one operationalisation in which the gate's cross-task discrimination is supplied largely by blank-agreement and δ, and the within-route segment selection by SHAP-ϕ.

Split	Route	n	Rec. (%)	Brk. (%)
POPE, LLaVA-1.5-7B				
random	MSB	1172	0.0	0.9
	PMI	190	**100.0**	**100.0**
	FB	1638	0.0	0.0
popular	MSB	1209	4.3	0.9
	PMI	224	**100.0**	**100.0**
	FB	1567	0.0	0.0
adversarial	MSB	1280	4.3	0.9
	PMI	281	**100.0**	**100.0**
	FB	1439	0.0	0.0
MME, LLaVA-1.5-7B				
–	MSB	172	3.4	2.1
	PMI	14	50.0	**80.0**
	FB	54	0.0	0.0
POPE, Qwen2.5-VL-7B				
random	MSB	1222	66.3	0.9
	PMI	119	72.5	16.0
	FB	1659	1.4	0.8
popular	MSB	1246	52.2	0.8
	PMI	146	63.6	24.6
	FB	1608	5.4	0.8
adversarial	MSB	1278	46.9	1.5
	PMI	173	59.8	**35.2**
	FB	1549	6.6	0.8
MME, Qwen2.5-VL-7B				
–	MSB	125	0.0	0.9
	PMI	25	60.0	25.0
	FB	90	0.0	0.0

Table 8: Per-route recover/break analysis at $\delta = 0.5$ (POPE 3 splits + MME). FB pools the margin-veto and $K < 2$-fallback steps. *Rec.* = fraction of baseline-wrong cases on this route that SBC corrects; *Brk.* = fraction of baseline-correct cases that SBC flips to wrong. Three patterns: (i) On LLaVA POPE the PMI route is structurally a baseline-inverter (Rec/Brk both 100%): every case that survives the entropy + blank-agreement + margin gate is a low-confidence baseline-"no", and PMI flips it to "yes"—the image-margin guard exists precisely to keep this invocation narrow ($n = 190$–281 per split). (ii) On Qwen POPE the over-spread regime is split between MSB ($\sim$66% recovery, n=1222) and the margin-veto FB bucket ($n\sim$1500); active PMI is reserved for narrow-margin cases (Rec $\sim$60%, Brk $\sim$25%). (iii) On LLaVA MME the PMI break rate climbs to 80% ($n = 14$), reinforcing the pattern: when prior-subtraction is over-applied to questions already correctly handled by the strong greedy baseline, it flips them to wrong.

Split	Bin	n	err %	95% CI
random	over-conc.	21	0.0	[0.0, 0.0]
	conc.	32	3.1	[0.0, 9.4]
	mixed	35	14.3	[2.9, 25.7]
	spread	66	16.7	[7.6, 25.8]
	over-spread	2828	12.9	[11.7, 14.2]
popular	over-conc.	22	4.5	[0.0, 13.6]
	conc.	36	13.9	[2.8, 25.0]
	mixed	38	18.4	[7.9, 31.6]
	spread	72	18.1	[9.7, 27.8]
	over-spread	2814	14.0	[12.7, 15.3]
adversarial	over-conc.	21	0.0	[0.0, 0.0]
	conc.	41	24.4	[12.2, 39.0]
	mixed	37	18.9	[8.1, 32.4]
	spread	76	23.7	[14.5, 34.2]
	over-spread	2807	16.2	[14.9, 17.5]

Table 9: Per-bin baseline error rate $P(\text{error} \mid \text{bin})$ on POPE (LLaVA-1.5-7B, 3000 questions per split) with 95% percentile bootstrap intervals (2000 resamples). Middle bins have wide CIs that overlap the over-spread estimate; the over-spread bin is tight. This supports the §3 statement that per-bin rate cannot discriminate bins, and that our claims rest on mass-tail location instead.

τ_{mid}	random	popular	adv.
0.30	88.7/88.9	86.6/87.1	79.0/81.1
0.40	88.7/88.9	86.6/87.1	79.0/81.1
0.50	88.2/88.4	85.5/86.1	79.0/81.1
0.60	88.5/88.7	86.4/86.9	78.8/80.9
0.70	88.4/88.5	86.3/86.7	78.7/80.8

Table 10: Gate threshold ablation (POPE, LLaVA-7B, Acc/F1) at default $b = 1.8$. SBC is robust on $\tau_{\text{mid}} \in [0.3, 0.7]$ (all cells within ± 0.5pt of each other within each split). Cells use the earlier δ=0.3 operating point and therefore differ from the main-table δ=0.5 numbers (Tables 4/6); the conclusion about τ robustness is unchanged because the δ guard is orthogonal to the τ decision (App H).

b	CHAIR$_S$↓	CHAIR$_I$↓	Recall↑
1.2	48.8	13.92	79.9
1.5	48.2	13.47	79.4
1.8	**47.4**	**13.14**	79.1
2.0	47.9	13.34	79.2

Table 11: Boost factor ablation (SBC, CHAIR, LLaVA-7B) at $\tau_{\text{mid}} = 0.5$, sweep $\delta = 0.3$. $b = 1.8$ is the operating point.

τ	b	random		popular		adv.
		Acc	F1	Acc	F1	Acc/F1
0.30	1.5	88.70	88.87	86.60	87.07	79.00/81.12
0.30	1.8	88.23	88.41	85.50	86.09	79.03/81.06
0.40	1.5	88.70	88.87	86.60	87.07	79.00/81.12
0.40	1.8	88.23	88.41	85.50	86.09	79.03/81.06
0.50	1.8	88.23	88.41	85.50	86.09	79.03/81.06
0.60	1.5	88.50	88.65	86.40	86.85	78.80/80.90
0.60	1.8	88.17	88.33	85.47	86.04	78.97/80.99
0.70	1.5	88.40	88.54	86.30	86.74	78.70/80.79
0.70	1.8	88.13	88.30	85.43	86.01	78.93/80.95

Table 12: Joint $\tau \times b$ sweep on POPE (LLaVA-1.5-7B). All cells fall within a ± 0.6pt window of each other; the gate and the MSB boost do not interact in a way that the optimum depends on the other's setting. The τ=0.50, b=1.8 cell is the default in Tables 3–4; cells here come from the earlier δ=0.3 sweep and therefore differ from the main-table δ=0.5 values (Tables 4/6). The relative ordering across τ, b is unchanged.

Fill	n	H mean	H med.	Hist.	OS %
mean	298	0.941	0.998	2/8/9/8/271	91.0
zero	298	0.946	0.998	3/7/4/11/273	**91.6**
Gauss.	298	0.940	0.998	3/4/13/8/270	90.6

Table 13: H statistics under three occlusion fills on the same 298 POPE random-split images (LLaVA-1.5-7B). *Hist.* is the count in over-concentration / concentration / mixed / spread / over-spread bins (consistent with Table 1). *OS %* is the share of mass in the dominant (over-spread) tail. All three fills give the same mean, the same median, and over-spread shares within ± 0.5 percentage points: the mass-tail diagnostic does not depend on the fill choice.

Backbone	δ	Acc	F1	Rec	MME
LLaVA-7B	0.0	88.55	88.55	90.0	586.7
LLaVA-7B	0.1	87.30	85.90	77.4	635.0
LLaVA-7B	0.3	88.34	88.34	81.8	598.3
LLaVA-7B	0.5	89.40	88.82	84.2	598.3
Qwen2.5-VL	0.0	91.20	90.62	85.0	653.3
Qwen2.5-VL	0.1	89.60	88.50	80.0	685.0
Qwen2.5-VL	0.3	89.43	88.37	80.3	668.3
Qwen2.5-VL	0.5	91.00	90.28	83.6	663.3

Table 14: Image-margin threshold δ sweep. POPE columns are on the random split (1000q sub-sample) and MME on the hallucination subset; numbers thus differ from main Tables 4/5 which average the three POPE splits over 3000q each. $\delta = 0$ disables the third condition.

Gate variant	Acc	F1	Rec	Prec
SBC (ref)	89.13	88.59	84.40	93.23
Diagnostic-substitution				
`attn_h`	89.30	88.68	83.80	94.16
`logit_h`	86.80	85.20	76.00	96.94
`task_pmi`	89.40	88.82	84.20	93.97
`task_msb`	86.80	85.20	76.00	96.94
Condition-removal				
`no_h`	89.40	88.82	84.20	93.97
`pmi_guard`	90.00	89.52	85.40	94.05

Table 15: Diagnostic-necessity ablation on LLaVA-1.5-7B POPE random (1000q). Variants (all hold actuators and δ fixed; the entropy and blank-agreement conditions vary): `attn_h` = replace SHAP-H with last-layer attention-H; `logit_h` = replace with top-100 logit entropy; `task_pmi` / `task_msb` = oracle task-label gate (always PMI candidate / always MSB); `no_h` = drop the $H \geq \tau$ check and keep only blank-match + δ; `pmi_guard` = drop *both* $H \geq \tau$ *and* blank-agreement, keep only δ. Three patterns: (i) `logit_h` and `task_msb` collapse POPE recall by 8pt (both reduce to MSB-only on POPE); (ii) `no_h` and `task_pmi` match SBC because $H \geq \tau$ already holds for 88–95% of binary-QA steps (Table 1); (iii) `attn_h` tracks SHAP-H *on average*, even though the two signals are per-sample uncorrelated (Appendix I, $r = -0.016$). The discriminative consequence is on the captioning side, where $H \geq \tau$ fires *against* PMI (0% PMI on CHAIR, Table 7) — the logit-entropy substitute would over-fire PMI there and inherit PMI-only's $CHAIR_S = 60.3$ failure.

Backbone	Variant	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	Rec.↑	MME
LLaVA-7B	SBC	**47.4**	**13.14**	79.1	**598.3**
	`pmi_guard`	49.9	14.26	79.9	566.7
Qwen-7B	SBC	**23.4**	**5.83**	58.6	663.3
	`pmi_guard`	27.6	8.24	60.7	**668.3**

Table 16: Cross-task `pmi_guard` ablation (drop both $H \geq \tau$ and blank-agreement; keep only δ) on CHAIR (1000 images) and MME (hallucination subset, score column). Dropping blank-agreement is benign on POPE (Table 15) but inflates CHAIR_S by $+2.5$pt (LLaVA) / $+4.2$pt (Qwen) on captioning—the cross-task cost of removing the blank-agreement check.