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

Jaegwang Shin, Suan Lee*
School of Computer Science, Semyung University, Jecheon, Republic of Korea
worhkdwkd3@semyung.ac.kr, suanlee@semyung.ac.kr

Abstract

Decoding-time mitigation of vision-language model (VLM) hallucination applies one fixed intervention and typically trades recall for lower hallucination. We trace this trade-off to treating hallucination as a single failure mode. Measuring the normalised entropy H of leave-one-out attributions over panoptic image segments, we find that a VLM's inference mass concentrates in one task-conditional tail of H: captioning in the **over-concentration** tail (low H, 53–66%), binary question-answering in the **over-spread** tail (high H, 74–95%). The pattern holds in 6/6 (backbone $\times$ benchmark) cells on LLaVA-1.5-7B and Qwen2.5-VL-7B, survives exact Shapley recomputation, reproduces within AMBER, and is shared by three further backbones. Because errors and correct outputs share the tail—H is a regime signal, not a hallucination detector—an intervention matched to the wrong tail actively hurts: blank-image contrastive decoding raises captioning CHAIR_S from 49.3 to 60.3. The regime can be read without attribution at inference, from blank-image agreement and on-image token margin; SBC (Segment-Based Conditioning) uses them to route between segment boosting and probabilistic marginalisation. With one shared configuration, SBC lowers CHAIR_I while preserving recall and raises POPE F1 on both backbones ($+1.95$ / $+3.31$), at 1.7–1.9$\times$ greedy cost on captioning and 4.9$\times$ on single-token QA.

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 CHAIR recall). We argue that this trade-off is not a property of decoding-time intervention as such, but a symptom of treating hallucination as one uniform failure mode: the failure has two distinct regimes, a decoder must know which one it is in, and the regime can be read without any attribution at inference time.

To test that, we measure *where a VLM's inference mass actually lives*. Partitioning an image into panoptic segments and occluding one at a time gives a per-segment leave-one-out (LOO) attribution ϕ_i, whose normalised entropy H says whether the answer depends on a single segment (low H, *over-concentration*) or on none in particular (high H, *over-spread*). Across two backbones and three benchmarks we find a sharp, reproducible structure (§3): inference mass—correct and hallucinated outputs alike—sits in *one* tail of H, and which tail is fixed by task type, not architecture. Captioning is an over-concentration regime, binary QA an over-spread one, and errors occupy no separate "risky" region: they are simply where the model spends its inference.

The consequence is measurable. An intervention designed for one regime is not merely useless on the other but harmful: blank-image contrastive decoding (PMI), the right tool in the over-spread regime, raises CHAIR_S from 49.3 to 60.3 on captioning (§4). A decoder therefore has to know which regime it is in—and that, not the choice of actuator, is where the leverage is.

Reading the regime needs no attribution at inference. Two cheap signals suffice: whether the blank-image continuation agrees with the on-image one, and the on-image token margin. Attribution re-enters only inside one route, where ϕ ranks the segments that the boost actuator (MSB) targets. The resulting decoder, SBC (Segment-Based Conditioning), improves CHAIR_I with recall preserved and POPE F1 on both backbones under one shared configuration (§6), and its extra work fires *once per response* rather than per token,

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3466–3490
November 6-10, 2026 ©2026 Association for Computational Linguistics

so the cost amortises on long-form generation.

Contributions.

1. **A regime structure of VLM hallucination.** Using the entropy H of leave-one-out attributions over panoptic segments, we show that inference mass concentrates in one task-conditional tail of H (over-concentration for captioning, over-spread for binary QA), and that hallucinated and correct outputs are co-located in that tail. The pattern holds in 6/6 (backbone $\times$ benchmark) cells, is invariant to the attribution estimator (LOO vs. exact Shapley), to occlusion fill, and to segment count, and reproduces within a mixed-format benchmark (AMBER) and across three further backbones, including a third family (§3).
2. **Mismatched intervention is harmful, not merely ineffective.** Regime-specific interventions fail asymmetrically on the opposite regime (PMI on captioning: CHAIR_S $49.3 \rightarrow 60.3$; MSB on binary QA: $\leq$ 1.1pt). This reframes the recall loss of uniform decoding-time methods as a regime mismatch (§4).
3. **Regime-matched decoding with an attribution-free router.** We show that the regime can be read at inference from two cheap signals (blank-image agreement and on-image margin), and that the H threshold is dispensable: removing it reproduces the LLaVA results and the CHAIR results on both backbones, and is neutral-to-positive on Qwen POPE ($+0.08$–0.20 F1) (§4.2). The resulting decoder, SBC, improves CHAIR_I with recall preserved and POPE F1 on both backbones under one shared configuration, at 1.7–1.9$\times$ greedy cost on long-form captioning and $4.9\times$ on single-token QA—$3.5\times$ with attention ranking, against $1.9\times$ for VCD (§6).

The empirical structure is the primary claim; SBC demonstrates that it is actionable without H on the inference path.

2 Related Work

Vision-language models and decoding-time mitigation. Modern VLMs couple a pretrained vision encoder with an LLM through a learned connector (Li et al., 2023a), often trained with visual instruction tuning (Liu et al., 2023, 2024; Dai et al., 2023); Qwen-VL and Qwen2.5-VL (Bai et al., 2023, 2025) scale this recipe, the latter with native dynamic resolution. A large body of work then reduces hallucination at inference without retraining. Contrastive-decoding methods subtract a counterfactual distribution—VCD (Leng et al., 2024) a noised image, ICD (Wang et al., 2024b) a disturbed instruction, M3ID (Favero et al., 2024) image–text mutual information, DoLa (Chuang et al., 2024) transformer layers—while OPERA (Huang et al., 2024) adds an over-trust penalty and HALC (Chen et al., 2024) an adaptive focal-contrast beam search. All apply one *uniform* intervention, whatever the failure regime. Our diagnostic (§3) says why that is the wrong shape for the problem, and §4 shows such an intervention is actively harmful on the regime it was not designed for. Training-time and post-hoc alternatives exist (Yu et al., 2024; Zhao et al., 2023; Zhou et al., 2024; Yin et al., 2024); ours is decoding-time and training-free. We evaluate with CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023b), MME (Fu et al., 2025) and AMBER (Wang et al., 2023), the last because it holds a generative and a discriminative format over the *same* images—the direct test of task-conditionality (§3.5).

Attribution and interpretability. Feature attribution explains a prediction by assigning credit to inputs: Shapley-value methods such as SHAP (Lundberg and Lee, 2017) and gradient-based Integrated Gradients (Sundararajan et al., 2017) are architecture-agnostic, while attention-flow analyses (Abnar and Zuidema, 2020) are transformer-specific. These are post-hoc, offline explanations of individual predictions. We use them in two roles that we keep separate throughout: as an *analysis* instrument, the LOO attribution entropy H is what exposes the regime structure of §3, computed offline and not needed at inference; as an *actuator input*, the per-segment attribution ϕ ranks the segments MSB boosts, and only there. To our knowledge, the mass-tail structure itself—inference mass and errors co-located in a single task-conditional attribution-entropy tail—is a new empirical characterisation of VLM hallucination.

3 A Regime Structure of VLM Hallucination

3.1 Attribution Entropy over Panoptic Segments

We partition the input image via Mask2Former panoptic segmentation (Cheng et al., 2022) into

$K \leq 6$ segments and compute leave-one-out (LOO) attributions

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

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

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

Throughout, H **denotes the normalised entropy of the leave-one-out segment attributions of Eq. 1**—the leave-one-out term of the Shapley decomposition, not the full Shapley value; §3.5 shows that recomputing ϕ with exact Shapley values (all 2^K coalitions) yields the same diagnostic. One value of H is computed per response, over the lookahead span, not per generation step.

Five equal-width 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).

Why occlusion rather than raw attention: raw attention is normalised per-head and conflates "what the model looks at" with "what affects the output." The single-segment occlusion in Eq. 1 measures the counterfactual log-probability change—what would happen if a segment were absent—giving a counterfactual occlusion signal of which segments drive the output (Figure 3, App. A). The two quantities are nearly uncorrelated on POPE, where H is near its ceiling: the correlation between H and last-layer attention entropy is $r = -0.016$ (App. P).

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 **consistent across all evaluated backbones** but **bound to task type**. Figure 1 shows the full five-bin H distributions.

3.3 Hallucinations Are Co-located with Correct Outputs

The last two columns of Table 1 show that in 6/6 cells, 62–100% of hallucinations *and* 51–96% of correct outputs occupy the same dominant tail. Hallucinations do not sit in a "risky" tail away from correct outputs—they are *where the model spends its inference*.

Backbone	Data	Dom. tail	n (%)	Err	Corr
LLaVA-7B	CHAIR	over-conc.	158 (53)	62.1	51.4
LLaVA-7B	POPE	over-spread	564 (95)	88.9	95.8
LLaVA-7B	MME	over-spread	71 (85)	93.8	78.8
Qwen-7B	CHAIR	over-conc.	197 (66)	71.5	65.8
Qwen-7B	POPE	over-spread	525 (88)	83.1	89.0
Qwen-7B	MME	over-spread	62 (74)	100.0	70.3

Table 1: Distribution of H (n=297 CHAIR / 594 POPE / 84 MME per cell; MME uses the hallucination subset). The diagnostic runs on a fixed-seed sub-sample of the full evaluation sets because each cell costs K+1 forward passes per response, giving the dominant-tail share to within ∼2pt on POPE and ∼8pt on MME (95% binomial). *Err*/*Corr*: percentages of errors (CHAIR: of summed CHAIR_I) and of correct outputs ($1-\text{CHAIR}_I$) in that tail: in 6/6 cells it holds the majority of both.

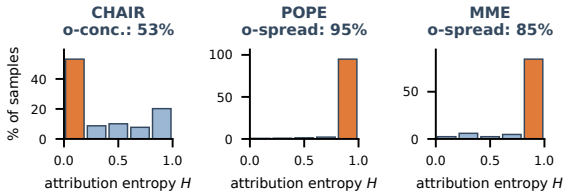


Figure 1: **Measured H distributions (LLaVA-1.5-7B).** Five-bin histograms of the LOO attribution entropy H; the highlighted bar is the dominant tail. Qwen2.5-VL-7B is the same picture (Figure 4, App. H).

3.4 H Is a Regime Diagnostic, Not a Hallucination Detector

Table 1 is easy to misread as a negative result. H does *not* separate hallucinated from correct outputs and we claim no such thing: within the dominant tail the two sit together, and a threshold on H would flag correct outputs at nearly the rate it flags errors. That co-location *is* the finding. H answers a different question—which failure *mechanism* is available to the model right now—and it is that regime label, not a per-sample score, that decides which intervention is safe (§4). Per-bin error *rates* are correspondingly uninformative at these bin sizes—middle bins hold $n = 5$–19 questions on the 594-question diagnostic sub-sample and $n = 35$–38 (mixed bin) on the full splits, with rates swinging from 37.5% to 14.3–18.9% between the two (App. E)—so our claims rest on

mass-tail location and the absolute error/correct distribution, not on per-bin rate.

3.5 Robustness of the Diagnostic

The structure survives all checks below; only a segmenter swap moves it (§6.5). *Estimator.* ϕ_i is the leave-one-out term of the Shapley decomposition, not the full Shapley value; since $K \leq 6$ the exact value is computable, and recomputing H with it on all six (backbone $\times$ dataset) cells (fixed-seed subsamples, n=94–240/cell) leaves the dominant tail unchanged in 6/6 and preserves co-location. What it does change is the per-segment *ranking*, which the MSB actuator consumes and the diagnostic does not, and the actuator tolerates perturbation of that size (App. F; §6.5). *Task format vs. benchmark.* Table 1 confounds the two, since CHAIR is a captioning corpus and POPE/MME are QA corpora. AMBER (Wang et al., 2023) separates them by posing a generative and a discriminative task over the *same* images, and the split reproduces inside that one benchmark—generative $\rightarrow$ over-concentration (51.5% LLaVA / 66.6% Qwen), discriminative $\rightarrow$ over-spread (98.3% / 81.2%), tracking the CHAIR (53/66%) and POPE (95/88%) shares (Table 16, App. G). With the images held fixed, the regime is a property of what the model is asked to do. *Backbone.* A second scale and a second generation agree: Qwen2-VL-7B (Wang et al., 2024a) at 54.5% / 86.1% and LLaVA-1.5-13B (8-bit, diagnostic only) at 50.8% / 94.6% (Table 18). A third family, InternVL3-8B (Zhu et al., 2025) (InternViT encoder, Qwen2.5 LM), gives 64.6% / 93.6%. The tail location is identical in all five evaluated backbones; only the share moves, and it stays within $\sim$3pt of the band of Table 1.

4 Regime-Matched Intervention

The structure of §3 makes a testable prediction about mitigation: a single-intervention decoder, which applies the same correction to every output, must fail on the regime its intervention was not designed for. We verify both halves of the prediction, and then build the routing rule that follows from it.

4.1 Mismatched Intervention Fails Asymmetrically

An intervention aimed at the wrong regime can be worse than wasted. PMI subtracts the language prior, which is the right move in the over-spread regime but the wrong one on captioning, where the model is fixating rather than defaulting: on LLaVA-1.5-7B it drives CHAIR_S from the baseline 49.3 to 60.3, 11pt *worse* than doing nothing. MSB boosts attention on the top-ranked segments, which does almost nothing in the over-spread regime where binary QA lives: mean POPE F1 moves by $\leq$ 1.1pt on either backbone (Qwen +1.1pt, against PMI-only's +4.6pt). The failures are asymmetric—mismatching PMI is expensive, mismatching MSB is merely inert—so a router has one job above all others: keep PMI off the over-concentration regime.

4.2 A Three-Way, Attribution-Free Router

The router reads two cheap signals—whether the blank-image continuation matches the on-image span under normalised exact-text equality (*blank-agreement*), and the on-image top-token probability margin—and has three outcomes:

1. blank-agreement *and* margin $< \delta$ $(= 0.5) \rightarrow$ **PMI**: the answer is prior-driven and the model is not confident, so subtracting the prior can recover the image-grounded answer;
2. blank-agreement *and* margin $\geq \delta \rightarrow$ **plain greedy**: the model agrees with its own prior *and* is confident, so nothing is applied—PMI here would only invert an answer that is already image-grounded;
3. no blank-agreement $\rightarrow$ **MSB**: removing the image changes the continuation, so the output is image-dependent and the over-concentration actuator applies.

A plain-greedy fallback also fires when Mask2Former returns $K < 2$ segments or no lookahead span is found. The middle outcome is not a technicality: on POPE-random it is the largest bucket (Table 10: MSB 39.1%, PMI 6.3%, greedy/fallback 54.6%), and it keeps PMI off confident answers, protecting precision at some recall cost vs. unguarded PMI (App. M).

The two signals divide labour asymmetrically. Blank-agreement is the *cross-task* discriminator: frequently true on prior-driven binary QA, never true on free-form captioning, and therefore what keeps PMI off the captioning regime (0% PMI on CHAIR, Table 10). The margin does no cross-task work; it separates outcomes (i) and (ii) within binary QA. The split also fixes the cost profile: under the pipeline of §5 neither the greedy nor the PMI route touches segmentation or occlusion, so only the MSB share pays the K occlusion passes.

Why $H \geq \tau$ is not among the conditions. Our original design added a third condition, $H \geq \tau_{\text{mid}}$ ($= 0.5$); it turns out to do almost no work. It already holds on 88–95% of POPE responses and 74–85% of MME responses (Table 1), so it is nearly always true where it is evaluated. Rerunning the full splits without it makes the size of the effect precise (Table 2), and the answer differs by cell. On captioning the condition is inert on both backbones: all 1000 CHAIR captions are bit-identical, because blank-agreement never holds there, so the route is MSB whatever H says. On LLaVA POPE it is very nearly inert: the route changes for 12 of 3000 responses per split, 11 of them $K<2$ fallbacks that are merely relabelled and decode identically, leaving exactly one differing prediction worth $\leq$0.04 Acc/F1. On Qwen POPE it is *not* inert: blank-agreement holds while $H < \tau$ on 1.8–2.0% of responses, which the three-condition router sent to MSB; the two-signal router sends them to plain greedy or PMI, changing 5–9 predictions per split and *raising* F1 by 0.08–0.20 (averaged 88.17 $\rightarrow$ 88.30). MME is unaffected on either backbone. Across all ten (backbone $\times$ benchmark $\times$ split) cells, then, the condition either does nothing or, where it acts, is mildly harmful. Replacing the entropy test with logit entropy, or fixing the route to MSB, instead collapses POPE recall by 8pt, because the router then never selects PMI; attention-H matches H on average (App. Q, P).

Blank-agreement is the load-bearing signal. Within POPE-random the conditions are partially redundant, so `pmi_guard`—which drops blank-agreement too and keeps only δ—even edges out SBC. Its cost appears elsewhere: on Qwen2.5-VL CHAIR, removing blank-agreement inflates CHAIR_S from 23.4 to 27.6, +4.2pt above SBC and 4.9pt above the untouched baseline 22.7 (Table 30, App. Q). The asymmetry is consistent with blank-agreement being the cross-task discriminator that protects free-form captioning while δ guards within binary QA; because H is low on captioning, `pmi_guard` removes two conditions at once, and the isolating comparison—`no_h` on CHAIR—is reported in Table 2. On both backbones that comparison leaves captioning bit-identical (LLaVA 47.4/13.14/79.1): blank-agreement alone keeps PMI off captioning.

The router is not an oracle task label. Because the regimes align with task format, it is fair to ask whether the router is a dataset detector in disguise. Four things say otherwise. (a) Blank-agreement is a *response-level* signal: it fires on what the model is about to say under two prefills of the same input, not on which corpus the input came from. (b) The AMBER route distributions (Table 3) rule the oracle out directly: the two discriminative sub-tasks are *identically formatted* yes/no probes drawn from the same image pool, and yet they route completely differently—MSB 43.9% vs. 7.5%, greedy 34.4% vs. 79.3%—which no function of the task format can produce. The same table explains the existence-split regression of §6.3 mechanically: on that all-negative set 79.3% of responses are routed to plain greedy and only 10.5% to PMI, so the accuracy that is lost comes from a small, identifiable slice rather than from the decoder as a whole. (c) A task-free rule is not trivially safe: keeping only δ (`pmi_guard`) inflates Qwen CHAIR CHAIR_S 23.4 $\rightarrow$ 27.6, so blank-agreement does work a margin threshold alone cannot. (d) On a format-homogeneous QA split an oracle task label (`task_pmi`) reduces to `no_h` by construction (App. Q); router and task label can only be told apart on mixed or free-form input, which is where Table 2 and AMBER report it. A pure task classifier would in any case predict none of the intervention-specific asymmetries of §4.

BB	Cell	3-cond.	no_h
LLaVA-7B	CHAIR C_S/C_I	47.40/13.14	47.40/13.14
	CHAIR Recall	79.11	79.11
	POPE random	89.13/88.59	89.17/88.63
	POPE popular	86.87/86.53	86.90/86.57
	POPE adversarial	82.70/82.99	82.73/83.03
	MME (score/acc)	598.3/82.08	598.3/82.08
Qwen-7B	CHAIR C_S/C_I	23.40/5.83	23.40/5.83
	CHAIR Recall	58.61	58.61
	POPE random	90.33/89.50	90.50/89.70
	POPE popular	88.93/88.16	89.00/88.26
	POPE adversarial	87.53/86.86	87.57/86.94
	POPE Recall	82.40	82.73
	MME (score/acc)	663.3/88.33	663.3/88.33

Table 2: Removing the $H \geq \tau$ condition, all else fixed, on the full evaluation splits (CHAIR 1000 images, POPE 3000q/split); POPE cells are Acc/F1. Every CHAIR caption is bit-identical on both backbones. On LLaVA POPE one prediction per split changes; on Qwen POPE the condition does act, on $\sim$2% of responses, and removing it is slightly *beneficial* (+0.08 to +0.20 F1). MME is unaffected on both backbones: the subset score and accuracy are unchanged and no answer changes, though on Qwen 8 of 240 routes do.

AMBER track	n	MSB	PMI	greedy	FB
generative	1004	97.4	0.0	0.0	2.6
disc. attribute	7628	43.9	20.0	34.4	1.7
disc. existence	4924	7.5	10.5	79.3	2.6

Table 3: Route distribution (% of responses) on AMBER, LLaVA-1.5-7B, under the shared configuration. *greedy* is the margin-guard exit and *FB* the $K<2$ fallback. The two discriminative sub-tasks share a format and an image pool yet route very differently, which a task-format oracle cannot reproduce.

Why route 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 penalised (PMI on captioning inflates CHAIR_S). SBC is instead task-agnostic: the router's two signals (blank-agreement, margin) are read from the model itself and yield near-best performance in *both* regimes with one decoder and one configuration.

5 SBC: Segment-Based Conditioning

5.1 Overview

Figure 2 (App. A) summarises one response. Once per generated response:

1. a **lookahead probe** greedily continues up to 32 tokens, to the first sentence end or EOS, to obtain a tentative span;
2. a **blank-image prefill** on the same prompt gives the two routing signals—blank-agreement and the on-image top-token margin—and the three-way route is decided (§4.2): PMI, plain greedy, or MSB. The blank-image logits are reused by PMI when that route is taken, so this prefill is not an extra pass;
3. **on the MSB route only**, the image is panoptically segmented and K batched occlusion passes give ϕ (Eq. 1), which selects the top-k segments MSB boosts.

The selected intervention (§5.2) then produces the next-token logits for the response.

Two consequences of this ordering are worth stating plainly. First, **the diagnostic H is not on the inference path**: it is the analysis instrument of §3, computed offline on sub-sampled evaluation sets, and the final router never reads it. Second, attribution leaves the inference path on two of the three routes—segmentation and occlusion run only after the router has chosen MSB—so the K occlusion passes are paid only on the MSB share of responses (on POPE, roughly 40%; App. D), cutting POPE cost from $7.3\times$ to $4.9\times$ greedy (§5.3). The main-table results of §6 were produced by the earlier configuration that also evaluated $H \geq \tau_{\text{mid}}$ before routing, which is why the eager SBC row of Table 4 pays for segmentation and occlusion on every response; dropping the condition changes almost nothing on any of the ten cells (§4.2, Table 2).

5.2 Interventions

The two interventions are deliberately standard mechanisms; the contribution is the regime structure 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 occlusion attribution supports. Let $\mathcal{V}_k$ be the visual-token columns belonging to the top-k segments ranked by ϕ_i. For every 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. MSB's actuator follows attention-flow methods such as AIF (Liu

	CHAIR		POPE	
Method	s/img	×	s/q	×
Baseline (greedy)	3.01	1.0	0.23	1.0
AIF	2.97	1.0	0.42	1.8
VCD	6.15	2.0	0.44	1.9
OPERA	16.83	5.6	0.47	2.0
MSB-only	4.90	1.6	—	—
PMI-only	6.57	2.2	—	—
SBC (eager)	5.05	1.7	1.65	7.3
SBC (lazy)	5.59	1.9	1.10	4.9
attn-ranked	4.83	1.6	0.79	3.5

Table 4: Wall-clock cost under one latency harness (LLaVA-1.5-7B, fp16, RTX 3090); CHAIR s/image at max 512 new tokens, POPE s/question. Baseline and published-method rows are from an earlier session; the three SBC rows are one same-session set, so their ratios are directly comparable. The eager POPE figure was 1.59 s / 6.9× in that session and 1.65 s / 7.3× here (the GPU was shared; SBC ratios use a same-session ≈0.226 s baseline), which is why we quote the three together rather than mixing sessions. POPE ratios other than SBC's are derived from the s/q values against the 0.23 s baseline; corroborating run-harness numbers are in App. J. Deferring segmentation saves nothing on CHAIR, where ∼99% of responses take the MSB route (§4.2), so the eager row is the relevant captioning cost.

et al., 2026) but inverts and re-targets it: AIF *masks* 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 ϕ, a counterfactual occlusion signal. PMI is VCD's contrastive blend (Leng et al., 2024) with one change: the reference is a blank, zero-information image rather than a diffusion-noised one, so the subtracted term approximates the language prior; VCD's adaptive-plausibility constraint is reused unchanged. The mechanisms are commodity—the contribution is the regime structure that decides which fires and what it targets.

5.3 Inference Cost

SBC's extra work fires once per *response* rather than per token, so it amortises with output length: 1.7× greedy on long-form captioning against 7.3× on single-token QA under the eager three-condition configuration, a cell measured at 6.9× in an earlier session on an unshared GPU. Deferring segmentation and occlusion to the MSB route cuts that worst case to 4.9× and ranking segments by attention cuts it to 3.5×—reductions of 33% and 52% measured in the same session, against 1.9× for VCD and 1.8× for AIF on the same workload (Table 4).

Attention-ranked variant. MSB consumes only a *ranking* of segments, not the attribution values, and the end metrics are insensitive to the ranking perturbation of the segmenter swap (§6.5). This suggests removing the K occlusion passes from the inference path altogether: rank segments by attention mass—last-layer, head-averaged attention from the final prompt token to each segment's visual tokens, from one extra attention-output prefill—and leave routing unchanged. On LLaVA CHAIR the variant reaches CHAIR_S 48.6 / CHAIR_I 13.58 / recall 79.56, against ϕ-ranked 47.4/13.14/79.11 and baseline 49.3/14.13/79.6: it recovers 56% of the CHAIR_I reduction with recall preserved, at 4.83 vs. 5.59 s/image (lazy). On binary QA the substitution is close to free on both backbones: F1 88.53/86.41/82.97 on LLaVA POPE across random / popular / adversarial against 88.6/86.5/83.0 (within 0.15pt), and 89.42/87.95/86.60 on Qwen POPE against the three-condition 89.50/88.16/86.86 (within 0.3pt, recall 82.27 vs. 82.40), at 0.79 s/question against 1.10 (lazy) and 1.65 (eager) on LLaVA and a provisional run-harness 1.04 vs. 1.53 (lazy) on Qwen. On Qwen CHAIR, unlike LLaVA CHAIR, the substitution is free: CHAIR_S 21.5 / CHAIR_I 5.91 / recall 58.93 against ϕ-ranked 23.4/5.83/58.61 and baseline 22.7/8.12/58.1—CHAIR_I within 0.08, recovering 97% of the reduction with a *lower* CHAIR_S and slightly higher recall, at 7513 s against 8720 s. Together these localise what the attribution buys. It is not needed on binary QA at all, because MSB is nearly inert there (§4) and the ranking it consumes barely matters; on captioning its value is *backbone-dependent*, load-bearing on LLaVA (+0.44 CHAIR_I when replaced) and redundant on Qwen. On MME the substitution is within ±5 subset-score points and changes sign by backbone (596.7 vs. 598.3 on LLaVA, 668.3 vs. 663.3 on Qwen), which at n=240 is inside the bootstrap interval either way. This is a claim about the actuator, not the router: the attention-ranked variant is a safe default for short answers and a backbone-dependent trade on long-form generation, so a deployment should measure it on its own backbone rather than assume either outcome. Neither lever touches the diagnostic of §3, which is offline and is not paid for at inference at all.

5.4 Hyperparameter Robustness

The image-margin δ is task-conditional and we take $\delta = 0.5$ as the POPE-priority operating point; b and the original τ_{mid} are swept in App. L, and δ in App. M. That τ_{mid} is flat on $[0.3, 0.7]$ is a further reading of why the condition can be dropped: over its useful range F1 moves $\approx$0.1pt at b=1.8.

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 on COCO val2014 (Lin et al., 2014) with greedy decoding (VCD samples on CHAIR/POPE; App. A). SBC defaults: top-k=2, $b = 1.8$, PMI $\alpha = 1.0$, $\beta = 0.1$, image-margin $\delta = 0.5$; one shared configuration for both backbones, with no per-backbone re-tuning. 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., 2025) contributes its hallucination subset (existence, count, position, color; 240 yes/no questions), scored by the standard accuracy+accuracy$^+$ metric. AMBER (Wang et al., 2023) contributes its generative and discriminative tracks (§6.3). All baselines—VCD (Leng et al., 2024), OPERA (Huang et al., 2024) and AIF (Liu et al., 2026)—are re-implemented under a unified wrapper to eliminate eval-harness confounds and span the main intervention families; provenance details are in App. A. As in prior work, all methods are compared on both 7B backbones; InternVL3-8B is compared against greedy only (App. H.1).

6.2 Main Results

Table 5 reports CHAIR and POPE for both backbones. SBC reduces per-instance hallucination (CHAIR_I) without recall loss: on LLaVA-1.5-7B 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, SBC improves F1 on every split and on the average ($+1.95$ LLaVA, $+3.31$ Qwen; Table 6, App. B); the router sends most of POPE to PMI or plain greedy, so SBC inherits PMI's precision/recall trade-off (Limitations; App. M).

	CHAIR			POPE F1		
Method	$C_S\downarrow$	$C_I\downarrow$	Rec↑	rand	pop	adv
LLaVA-1.5-7B						
Baseline	49.3	14.13	79.6	85.6	84.4	82.3
VCD	57.5	16.75	80.0	87.1	84.9	81.4
OPERA	51.9	14.36	78.6	85.6	84.5	82.3
AIF	**46.2**	14.81	73.0	87.1	85.1	82.3
MSB-only	47.8	13.25	79.2	85.4	84.3	82.2
PMI-only	60.3	16.38	82.3	**88.8**	86.4	81.3
SBC	47.4	**13.14**	79.1	88.6	**86.5**	**83.0**
Qwen2.5-VL-7B						
Baseline	22.7	8.12	58.1	86.0	84.6	83.9
VCD	30.1	8.84	60.1	88.1	87.0	86.5
OPERA	**19.2**	6.54	53.7	86.8	86.0	85.3
AIF	23.6	6.96	57.8	88.4	87.3	86.4
MSB-only	23.4	**5.69**	58.6	86.8	85.9	85.2
PMI-only	28.0	7.51	61.2	**91.2**	**89.6**	**87.4**
SBC	23.4	5.83	58.6	89.5	88.2	86.9

Table 5: Main results: CHAIR (1000 images) and POPE F1 per split (3000q each), one shared configuration for both backbones, b=1.8; SBC rows use the three-condition router, which the two-signal router matches on CHAIR and slightly exceeds on POPE (Table 2). POPE F1 is shown to 1 dp; averaged 2-dp Acc/F1 (SBC $+1.95$ LLaVA, $+3.31$ Qwen) and bootstrap CIs are in App. B; MME in App. C.

On MME we report a result that does not favour us: no greedy-decoded method at its default setting beats the greedy baseline. SBC attains 598.3 on LLaVA and 663.3 on Qwen against greedy 631.7/680.0 and same-harness VCD 561.7/673.3; the baseline is unusually strong (VCD's paper reports a 565.3 sampling baseline for LLaVA-1.5), and App. M traces the shared gap to that sampling-vs-greedy mismatch: like VCD, PMI over-corrects a strong greedy baseline, and within SBC most of the loss sits in the PMI-routed questions (on LLaVA, 14 questions with Brk 80%; Table 11). Against the sampling baseline re-measured under VCD's protocol, SBC is $+38.3$ (LLaVA) and $+28.3$ (Qwen), level with VCD-sampling on LLaVA (598.3 vs. 596.7) and behind it on Qwen (663.3 vs. 685.0) (App. C).

6.3 AMBER Decoding Results

Under the unchanged shared configuration the router also transfers to AMBER: on the generative track SBC matches baseline (13.2 vs. 13.0 LLaVA; 7.1 vs. 7.1 Qwen), and on the balanced attribute subset (7,628q) it improves F1 on both backbones (79.5 vs. 78.2; 83.7 vs. 82.2), the POPE result carrying over to a different QA corpus. On the existence split accuracy drops (LLaVA 92.3 $\rightarrow$

82.5; Qwen only $95.4 \rightarrow 93.5$): that split is all-negative—all $4{,}924$ gold answers are "no"—and $597/600$ wrong flips (both backbones pooled) are PMI-routed, which is exactly the distribution on which pushing an answer toward "yes" can only hurt. We report it as a characterised failure of the PMI actuator on all-negative probe sets rather than as noise (Table 17, App. G, O).

6.4 Generation Quality

On the full 1000-caption CHAIR outputs, length, distinct-1/2, 4-gram repetition and sentence-level CLIPScore stay within noise of baseline on LLaVA and are flat-to-better on Qwen, while the recall-sacrificing baselines shorten captions (LLaVA AIF 86.8 vs. 90.4 words; Qwen OPERA 113.8 vs. 159.4, -29%): lower CHAIR_I is not bought with fluency, detail, or diversity (App. K).

6.5 Ablations

The MSB-only and PMI-only rows of Tables 5–6 are single-mode ablations: each fails on the opposite-tail task, as predicted by §4. The router-condition ablations are in §4.2 (App. Q); dropping $H \geq \tau$ leaves CHAIR bit-identical on both backbones, moves one LLaVA POPE prediction per split, and raises Qwen POPE F1 by 0.08–0.20 (Table 2), leaving MME unchanged on both backbones. Per-hyperparameter ablations (router threshold τ_{mid}, boost factor b, image-margin δ, joint $\tau \times b$) are in Appendices L and M; SBC is robust within $\tau_{\text{mid}} \in [0.3, 0.7]$ and $b \in [1.5, 2.0]$, and δ is task-conditional. Per-cell routing statistics are in App. D, reproducibility details in App. A.

Segmentation robustness. The partition is a hyperparameter of the whole pipeline, and we vary it three ways (App. N, I). Occlusion fill (mean/zero/Gaussian) leaves the H distribution unchanged within ± 0.5pt. Segment count $K \in \{4, 6, 8\}$ leaves POPE-random F1 flat ($88.70/88.82/88.82$) and the H tails unchanged, as H is normalised by $\log K$. Swapping Mask2Former for class-agnostic SAM (Kirillov et al., 2023) leaves end metrics at noise level (CHAIR_I $13.69 \rightarrow 13.35$) and the router safe (PMI fires $0/300$ on CHAIR, since blank-agreement does not depend on the partition), but it *moves* CHAIR's dominant tail to over-spread (54.0%) while POPE's is unchanged (98.5%): the over-concentration tail is defined relative to an object-aligned partition, and SAM fragments objects into parts. We record this partition-dependence as a limitation.

7 Discussion

Single intervention vs. SBC, and what is novel. On its own task a single intervention can match SBC (POPE F1 within 1.3pt of PMI-only, CHAIR CHAIR_I within ± 0.2 of MSB-only); the decisive comparison is the opposite task, where PMI inflates CHAIR_S ($49.3 \rightarrow 60.3$) and MSB moves POPE F1 by ≤ 1.1pt. Cross-regime robustness, not per-task SOTA, is the claim. Both actuators and both routing signals are commodity; what is new is the empirical structure that makes routing the right move at all—inference mass, correct and hallucinated alike, in one task-conditional tail of H, and mismatch harmful for PMI, inert for MSB. H is the instrument that exposed it; the decoder follows.

8 Conclusion

Measuring where a VLM's inference mass lives reframes hallucination as regime-structured rather than uniform: across two backbones $\times$ three benchmarks it concentrates—correct and hallucinated outputs alike—in one task-conditional tail of the LOO attribution entropy H, and the structure survives exact-Shapley recomputation, reproduces within AMBER, and holds on three further backbones. Because the same tail holds most errors *and* most correct outputs, a mismatched intervention is harmful rather than inert (PMI on captioning: CHAIR_S $49.3 \rightarrow 60.3$). Reading the regime needs no attribution at inference—blank-image agreement and the on-image margin recover the regime—and SBC uses them to track the regime-appropriate intervention without task knowledge, at a cost that amortises over the response.

Limitations

Partition-dependence of the over-concentration tail. The tail positions of §3 are defined with respect to a partition of the image, and one of the two is partition-dependent. Under class-agnostic SAM instead of Mask2Former, POPE's over-spread tail is unchanged (98.5%) but CHAIR's dominant tail moves to over-spread (54.0%), because SAM fragments objects into parts and a caption that depends on one object then depends on several segments (§6.5, App. I). The router and end metrics are unaffected (PMI fires $0/300$ on CHAIR; CHAIR_I

$13.69 \rightarrow 13.35$), but the statement "captioning is an over-concentration regime" should be read as "under an object-aligned partition".

Generality. The diagnostic is measured on five backbones spanning three families (LLaVA, Qwen, InternVL) and two scales and holds in all of them, across three benchmarks plus AMBER's two formats over shared images, under two attribution estimators, three occlusion fills and three segment counts; the decoder is evaluated under one shared configuration with no per-backbone tuning, against all baselines on the two 7B backbones and against greedy only on InternVL3-8B (App. H.1); the two further backbones are diagnostic-only. Beyond these three families, further designs (e.g., non-ViT encoders or cross-attention fusion) remain untested.

Hyperparameter selection. b was chosen on LLaVA CHAIR (Table 25), τ_{mid} on LLaVA POPE (Table 24) and δ on POPE-random for both 7B backbones (App. M), i.e., on evaluation data with no held-out split. InternVL3-8B and AMBER inherit all three, and Qwen2.5-VL inherits b and τ_{mid}, without re-tuning.

The margin guard bounds PMI but does not remove it. The δ guard stops PMI inverting a confident image-grounded answer but cannot detect prior-driven correct "no"s with narrow margins: on LLaVA POPE adversarial every PMI-routed answer (n=281) flips (Table 11), and without the guard n rises to $\sim$1700. The mode is much weaker on Qwen (PMI Brk 35%; adversarial Acc 87.5 vs. baseline 85.4). On MME no greedy-decoded method at its default setting beats the greedy baseline (VCD with sampling does on Qwen, 685.0 vs. 680.0).

PMI breaks on all-negative probe sets. PMI subtracts the blank-image prior; on the blank-agreeing responses it is applied to, that prior already favours "no", so PMI shifts mass toward "yes". On probe sets where the gold answer is always "no" this can only hurt. AMBER's existence split is exactly such a set (4,924/4,924 "no") and SBC loses 9.8pt accuracy there on LLaVA (§6.3), every flip to a wrong answer PMI-routed (504/504); on LLaVA POPE adversarial the PMI route flips 100% of the n=281 cases that survive the guard (Table 11). The margin guard keeps the volume small but does not remove the mode. The break rate, not the F1 average, is the relevant cost metric for hallucination-sensitive use.

Language and image-domain scope. Every experiment is English, and every image is from COCO val2014 or AMBER (MS-COCO 2014 test and Unsplash photographs). We do not know whether the tail positions hold for other languages or for image domains (documents, charts, medical imagery) whose panoptic structure differs from natural photographs.

Backbone scope. The diagnostic covers three families and two scales; the decoder is compared with all baselines on the two 7B backbones only, InternVL3-8B is decoded against greedy only in a single run (App. H.1), and Qwen2-VL-7B and LLaVA-1.5-13B are diagnostic-only. Quantising larger models to 8 bits confounds attention-magnitude methods (notably MSB), which is why LLaVA-1.5-13B appears as diagnostic-only, and an fp16 13B decoding sweep was not completed.

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.

Lookahead-span dependence. ϕ is measured on the model's own first-sentence lookahead span ($\leq$32 tokens), so it attributes a tentative continuation rather than the final response; we did not ablate the span length.

Per-bin rate noise. Middle H-bins (n = 5–19) are too small to reliably discriminate; our claims rest on mass-tail location, not per-bin rate.

Cost. Single-token QA remains the worst case: $7.3\times$ greedy for the eager three-condition configuration, $4.9\times$ once segmentation is deferred to the MSB route and $3.5\times$ with attention ranking (Table 4), against $1.9\times$ for VCD. The attention-ranked variant removes the occlusion passes altogether at 0.79 s/question with no measurable POPE loss, but its cost on long-form generation is backbone-dependent—$+0.44$ CHAIR_I on LLaVA CHAIR against no penalty on Qwen (5.91 vs. 5.83)—so it cannot be recommended for captioning without a per-backbone check.

Future directions. The above motivate five threads: (i) an fp16 13B/larger decoding sweep to test whether the decoder's gains persist with model size; (ii) a partition-invariant formulation of the tail definition, since §6.5 shows the over-concentration tail is tied to object-aligned segmentation; (iii) learning δ online from the on-image margin distribution rather than fixing it (App. M); (iv) replacing the mean-colour fill of Eq. 1 with a generative inpainting model to test whether H shifts when the occluded region is filled with plau-

sible content (App. N already shows robustness to mean / zero / Gaussian fills); and (v) amortised attribution (e.g., FastSHAP-style estimators; Jethani et al., 2022) to make the ϕ-ranked MSB route cheaper on single-token QA.

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, (iii) helping to format references and LaTeX tables, and (iv) implementing and packaging code under the authors' direction (the public release and the InternVL3 model wrapper, App. H.1); all such code and every reported number were checked by the authors. Research ideas, claims and experimental design are the authors' own.

Acknowledgments

This work was supported by the Ministry of Science and ICT and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25498341).

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, et al. 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. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

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)*.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. 2022. FastSHAP: Real-time Shapley value estimation. In *International Conference on Learning Representations (ICLR)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. 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*, 67(12):220105.

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)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Reproducibility Details

Baseline provenance. VCD is ported from its official public code (α=1, β=0.1, 999 noise steps) and samples directly from the contrastive distribution on CHAIR and POPE (its paper samples on POPE and reports no CHAIR) (POPE: mean of 5 runs on LLaVA, one run on Qwen); on MME it decodes greedily in Table 8 and samples in Table 9. OPERA is a greedy adaptation of the official code: only the over-trust penalty is kept (beam retrospection-rollback is omitted), each of N_{can}=5 candidates is re-scored after one forward step, and the penalty is compressed as $\log(1+\omega)$ for the over-trust score ω (σ=50, α_{OPERA}=1, window 8). AIF had no public implementation when we ran it (official code appeared in June 2026) and is reproduced from the paper description, with the selected mask ratio capped at 0.5 (the raw selection criterion saturates at 0.9 on LLaVA). The AIF paper reports no CHAIR results; our AIF CHAIR numbers come from this re-implementation. All other decoders, including every SBC variant, are greedy. The three baselines

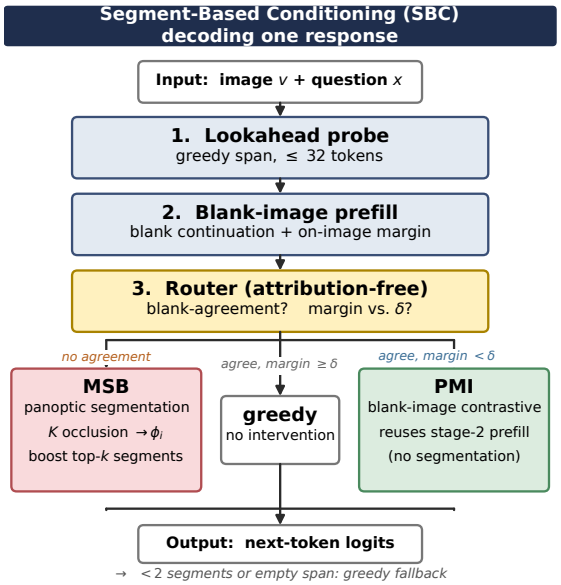


Figure 2: **SBC decoding for one response.** A lookahead probe ($\leq$ 32 tokens, to the first sentence end or EOS) yields a tentative span; a blank-image prefill supplies the two routing signals. The router has three exits: no blank-agreement $\rightarrow$ MSB (the over-concentration actuator), agreement with margin $\geq \delta \rightarrow$ plain greedy, agreement with margin $< \delta \rightarrow$ PMI (the over-spread actuator). Panoptic segmentation and the K occlusion passes that yield ϕ run on the MSB route only. Routing fires *once per response*, not per generation step. Defaults: top-$k = 2$ (of $K \leq 6$ segments), $b = 1.8$, PMI $\alpha = 1.0$, $\beta = 0.1$, $\delta = 0.5$.

span contrastive decoding, an over-trust penalty and attention modulation respectively.

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, terminated at the first sentence end), computed once per response; a sentence end stops it only after at least 4 tokens, but EOS stops it at once, so single-token answers (binary QA) give a one-token span. The blank image is the all-zero tensor in normalised pixel space (the normalisation mean colour). The margin is the on-image first-token softmax probability of the top-1 minus the top-2 token; the router decodes greedily when it exceeds δ. Segments come from the Mask2Former Swin-L COCO-panoptic checkpoint (`facebook/mask2former-swin-large-coco-panoptic`): the $\leq$6 largest segments covering $\geq$1% of the image. MSB is applied as an additive attention mask shared by all layers and heads; a visual token belongs to a segment when $\geq$30% of its patch is covered. Generation is capped at 512 new tokens on CHAIR and 64 on POPE/MME. LLaVA runs in fp16; Qwen and InternVL run in bf16. InternVL3-8B sees one 448×448 view (tiling disabled), i.e. a 16×16 grid of visual tokens, in both the diagnostic and decoding. Code, configurations, the analysis scripts (exact Shapley, quality proxies, latency) and the raw per-sample model outputs behind the tables (except the development-run CHAIR_I cells of Table 21) are available at `https://github.com/JaeGwang98/mass-tail-entropy`. Relative to the reviewed version we corrected transcription errors in Tables 24 (rows now all at b=1.8), 27 (all cells on the first 1000 random-split questions), the Qwen MME rows of Tables 10–11 (now the δ=0.5 run), and recomputed POPE averages from unrounded per-split values (+1.94/ + 3.30 $\rightarrow$ +1.95/ + 3.31); no conclusion changes.

B Per-Split Results and Bootstrap CIs

Bootstrap confidence intervals. SBC 95% percentile bootstrap CIs (seed 0; `scripts/chair_pope_bootstrap_ci.py`): CHAIR CHAIR_I (image-level bootstrap, 1000 image resamples) LLaVA $[12.04, 14.22]$, Qwen $[5.00, 6.81]$; POPE F1 (question-level bootstrap within each split, F1 averaged over splits, 1000 resamples per split) LLaVA $[85.26, 86.82]$, Qwen $[87.46, 88.94]$, both lower bounds above the corresponding baseline F1; MME (image-level bootstrap within category, 1000 resamples, seed 0, 240q) LLaVA $[552, 647]$, Qwen $[623, 705]$. Table 6 reports POPE accuracy and F1 averaged over the three official splits. Table 7 gives the per-split breakdown, showing where the averaged numbers come from. The adversarial split is the hardest (co-occurrence-based negatives); on LLaVA-1.5-7B, SBC's adversarial accuracy (82.7) is within 0.8pt of baseline (83.5), but Table 11 reveals that the PMI route on this split flips *every* baseline-correct case it touches (Brk 100%; the route holds n=281 cases)—a regression that is hidden in the averaged F1 because the margin guard keeps the PMI volume small. PMI's blank-image contrast overturns narrow-margin "no" answers even when they are already correct; this failure mode is much weaker on Qwen2.5-VL (Brk 35%; adversarial 87.5 vs. baseline 85.4). The averaged F1 in Table 6 remains positive on

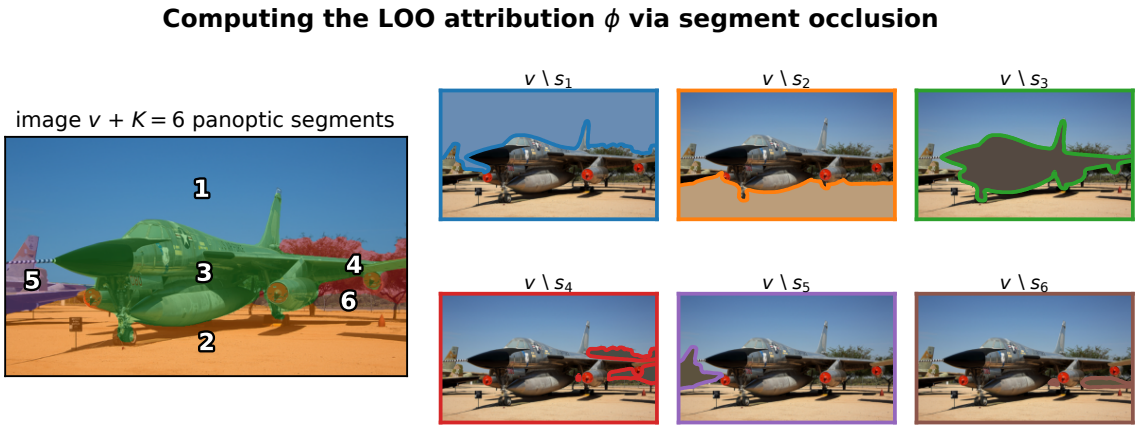


Figure 3: **Leave-one-out segment occlusion underlying ϕ (CHAIR case, LLaVA-1.5-7B).** The image is partitioned into $K = 6$ Mask2Former panoptic segments; for each segment s_i, the masked variant $v \setminus s_i$ (image-mean fill of the segment region; segment outline drawn for visibility) is forward-passed 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 route for this same CHAIR case are shown in Figure 5 (bottom row).

Method	Acc	F1
LLaVA-1.5-7B		
Baseline	85.52	84.09
VCD	84.61	84.48
OPERA	85.56	84.13
AIF	85.52	84.83
MSB-only	85.40	83.94
PMI-only	84.52	85.51
SBC	**86.23**	**86.04**
Qwen2.5-VL-7B		
Baseline	86.36	84.86
VCD	88.26	87.18
OPERA	87.43	86.04
AIF	88.40	87.34
MSB-only	87.39	85.97
PMI-only	**89.77**	**89.44**
SBC	88.93	88.17

Table 6: POPE Acc/F1 averaged over the three splits (random / popular / adversarial). SBC improves F1 over baseline (+1.95 LLaVA, +3.31 Qwen); per-split breakdowns in App. B. The 95% bootstrap CI for SBC F1 (question-level bootstrap within each split, F1 averaged over splits, 1000 resamples per split, seed 0; `scripts/chair_pope_bootstrap_ci.py`): LLaVA $[85.26, 86.82]$, Qwen $[87.46, 88.94]$—CI lower bound exceeds the baseline F1 on both backbones.

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	86.1/84.6	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 7: POPE Acc/F1 per split for both backbones. Average rows are in Table 6.

both backbones.

C MME Results

Tables 8 and 9 report MME under the greedy and sampling protocols.

D Routing Statistics

Table 10 reports the empirical distribution of routes across (backbone × dataset) cells. Counts are over *responses*: the router fires once per generated response, so the n column is the number of questions (POPE, MME) or images (CHAIR). 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 (responses 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→9.4%), consistent with

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 8: 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 (image-level bootstrap within category, 1000 resamples, seed 0; `scripts/mme_bootstrap_ci.py`): LLaVA $[552, 647]$, Qwen $[623, 705]$—the small n=240 leaves per-method gaps inside the CI (baseline 95% CI, same procedure: LLaVA $[590, 673]$, Qwen $[640, 720]$); we discuss the greedy-baseline mismatch in App. M.

Setting	LLaVA-7B	Qwen-7B
Greedy baseline	631.7	680.0
Sampling baseline	560.0	635.0
VCD (sampling)	596.7	**685.0**
SBC (greedy)	598.3	663.3

Table 9: MME under the two baseline protocols (hallucination subset, subset score). SBC is unchanged—it decodes greedily under both protocols—so the comparison is entirely about what it is measured against. Under VCD's own sampling protocol SBC clears the baseline by 38.3 (LLaVA) and 28.3 (Qwen) and matches VCD-sampling on LLaVA, while on Qwen VCD-sampling (685.0) exceeds both SBC and the greedy baseline (680.0). No greedy-decoded method at its default setting beats the greedy baseline on either backbone (Table 8).

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" condition is never satisfied on CHAIR captioning in our runs—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 responses. The MSB share is also what determines how often the deferred segmentation and occlusion passes of §5 are paid: ∼40% on POPE, ∼99% on CHAIR.

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	48.0	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	17.9	30.0	240

Table 10: Route distribution (% of responses) at the operating point $\delta = 0.5$. On POPE the over-spread regime dominates and PMI is available; on CHAIR the over-concentration regime dominates and PMI is never fired (the blank-image-match condition vetoes it). The image-margin δ guard vetoes PMI on roughly half of all POPE responses, most of the blank-agreeing ones (the Base/FB column pools margin-vetoed and $K < 2$-fallback responses), leaving 4–9% active PMI on POPE for both backbones and 6–18% on MME.

E Per-Bin Error Rates with Bootstrap CIs

§3.4 argues that per-bin error *rates* are noise-sensitive at the bin sizes available. Table 12 makes this quantitative: middle bins (*concentration*, *mixed*, *spread*) carry only $n = 32$–76 POPE questions per split, so their 95% bootstrap intervals span 9–27pt, and most overlap the over-spread estimate. The *over-spread* bin, in contrast, is tightly estimated ($\pm$1.3pt) because it holds ≈94% 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.

F Exact-Shapley Details

§3.5 summarises the estimator-robustness check; this appendix gives the per-cell detail. For each cell we recompute, on the same fixed-seed subsample, the exact Shapley value of every segment,

$$\phi_i^{\text{exact}} = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(K-|S|-1)!}{K!} \times \big[f(S \cup \{i\}) - f(S)\big], \tag{5}$$

with $f(S) = \log p(\text{span} \mid v_S)$ the log-probability of the same lookahead span under the image with only the segments in S retained; all 2^K coalitions are evaluated ($K \leq 6$, so at most 64 forward passes per sample). H is then recomputed from ϕ^{exact} by Eq. 2. The LOO estimator of Eq. 1 is the $S = N \setminus \{i\}$ term of this sum.

Per-cell rank correlation between the two ϕ vectors is Spearman 0.44–0.72 with top-1 disagreement 21–34%, and the efficiency residual is $\sim 10^{-14}$. Tables 14–15 report the full histograms and the divergence statistics. Two further observations. First, co-location survives the estimator swap: on Qwen CHAIR the share of errors in the dominant tail is 71.4% (LOO) vs. 78.6% (exact), and no cell reverses. Second, the LOO estimator does make a specific, identifiable error—it calls a segment "irrelevant" ($|\phi_i| < 0.05$) when the exact value shows a real coalition-level contribution (total exact mass $M = \sum_i \phi_i^{\text{exact}} > 0.5$)—on 0–10 samples per cell, which mildly inflates the overspread share on the binary-QA cells. This is visible as the small downward correction in the exact column of Table 13 (92.8 → 91.8, 84.6 → 80.8, 89.4 → 84.0, 75.4 → 70.0) and does not change any dominant tail.

G AMBER Details

AMBER (Wang et al., 2023) is an LLM-free hallucination benchmark with a generative track (free-form captioning; we report a response-level hallucination rate, the share of captions mentioning at least one of AMBER's per-image annotated hallucination-target (absent) objects, CHAIR_S-style, computed with our scorer rather than the official AMBER CHAIR) and a discriminative track (yes/no probes), both built over the same image pool. This makes it the direct test of whether the tail location follows the task format rather than the corpus (§3.5). The diagnostic uses $n \approx 300$ ($K \geq 2$ samples) per cell, matching the sub-sampling protocol of Table 1; the decoding runs use the full existence and attribute splits (AMBER's relation split is not evaluated).

The discriminative track separates into subtasks with very different label distributions, which matters for interpreting Table 17:

- **Existence** (4,924 questions): every gold answer is "no". A yes-leaning prior is penalised at every question and a no-leaning prior is rewarded at every question, so any method that shifts probability toward "yes"—which is what PMI does when it subtracts a blank-image prior that already favours "no"—loses accuracy monotonically. SBC loses 9.8pt on LLaVA (92.3 → 82.5) and 1.9pt on Qwen (95.4 → 93.5); the flips to a wrong answer are PMI-routed (LLaVA 504/504, Qwen 93/96). We report this split separately rather than pooling it into a discriminative average, because pooling would hide a systematic failure inside an average.
- **Attribute** (7,628 questions): balanced 50/50 yes/no. Here the same configuration improves F1 on both backbones (78.2 → 79.5 LLaVA; 82.2 → 83.7 Qwen), consistent with the POPE result.

The contrast between the two sub-tasks is itself informative: it locates the PMI failure precisely in the label distribution of the probe set, not in the image content or in the router.

H Further-Backbone Diagnostics

Table 18 collects the dominant tail and its share across the five backbones on which the diagnostic was run, spanning three families and two scales (§3.5). InternVL3-8B combines an InternViT-300M vision encoder (24 layers, 448×448 input) with a Qwen2.5-7B language model, so it is a third vision-encoder and model family but shares its LM family with Qwen2.5-VL. It is run with image tiling disabled, so that the original, occluded and blank images all map to one 448×448 view (as LLaVA-1.5 sees one fixed-size view); the protocol is otherwise identical to the Qwen2-VL-7B row (n=297 CHAIR / 598 POPE, same seed; the Table 1 rows use a separate 198q/split sub-sample, n=594).

H.1 Decoding on InternVL3-8B

Table 19 compares SBC with greedy decoding on InternVL3-8B under the configuration of Table 5, unchanged, and on the same CHAIR, POPE and MME subsets. On CHAIR, SBC lowers CHAIR_I from 7.58 to 7.13 and CHAIR_S from 27.8 to 26.6 with recall unchanged (70.1); 99.4% of captions are routed to MSB, as on the two 7B backbones (App. D). On POPE the average F1 changes by +0.16 (against +1.95/+3.31 on LLaVA/Qwen): F1 rises on the random and popular splits and falls on adversarial (−0.4), where the yes-ratio rises from 52.7% to 55.8%; PMI handles only 3.3–6.6% of POPE questions per split. On MME,

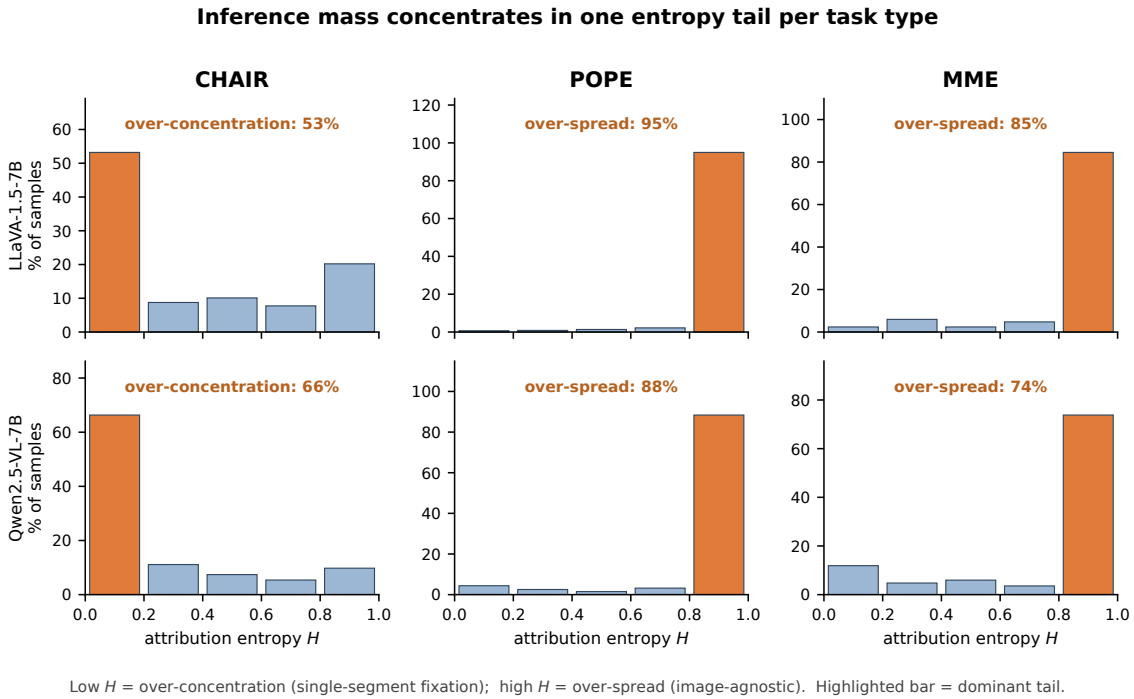


Figure 4: **Measured H distributions, both backbones.** Five-bin histograms of the LOO attribution entropy H, for two backbones $\times$ three datasets. Caption-style CHAIR concentrates inference mass in the over-concentration tail (low H); binary-QA POPE and MME concentrate it in the over-spread tail (high H); the pattern holds across both backbones (6/6 cells). The dominant-tail bar is highlighted with its share. MME uses the hallucination subset (existence/count/position/color), following prior work.

SBC scores below the greedy baseline (683.3 vs. 693.3), as on both 7B backbones (App. C). The directions thus match LLaVA and Qwen, but in this single run all three differences are within bootstrap noise: the paired 95% CIs of SBC$-$greedy include 0 (CHAIR_I -0.45 $[-1.03, +0.11]$; POPE avg F1 $+0.16$ $[-0.18, +0.51]$; MME -10.0 $[-38.3, +16.7]$; B=1000, seed 0). We therefore read this as consistent with, not as independent evidence for, the 7B results; VCD, OPERA and AIF were not run on this backbone.

I Segmentation Robustness

Tables 20–21 give the numbers summarised in §6.5. Together with the occlusion-fill ablation (App. N) they cover the three degrees of freedom in the partition: how the image is cut (K), what cuts it (segmenter), and what replaces a removed piece (fill). Two of the three leave the diagnostic untouched; the segmenter does not, and that is the honest boundary of the finding.

J Cost-Measurement Protocol

All timings in Table 4 are wall-clock, measured on one machine (LLaVA-1.5-7B, fp16, single RTX 3090) under the same evaluation wrapper used for the accuracy tables, so segmentation, tokenisation, and I/O are counted identically for every method. CHAIR timings are seconds per image at max 512 new tokens; POPE timings are seconds per question with single-token answers. Model load is excluded, and on CHAIR one warm-up image is run before timing; the Mask2Former forward pass is included in all "ours" rows.

Per response, SBC adds: one greedy lookahead prefill ($\leq$ 32 tokens, whose log-probs supply $\log p(\text{span} \mid v)$, the unmasked base term of Eq. 1); one blank-image prefill, reused as PMI's contrastive term when that route fires; and K occlusion passes (batched into one forward call on LLaVA, sequential on Qwen) on the MSB route. MSB itself adds no forward pass beyond the boosted prefill. FLOPs scale with these counts. Since all three fire once per response, the per-token amortised overhead falls as $1/|\text{output}|$, which is the entire explanation for the gap between the CHAIR column ($1.7\times$) and the POPE column ($7.3\times$, $+1.42$ s/question) of Table 4. Both figures are measured under the eager three-condition configuration, in which segmentation and occlusion

run on every response.

Deferring them to the MSB route does not change what the decoder emits: on a 40-question POPE-random sample the eager and deferred *two-signal* routers (both without $H \geq \tau_{\text{mid}}$) produce identical outputs on $40/40$ questions ($0/1000$ differing predictions on the first 1000 random-split questions) while the deferred one runs $1.45\times$ faster ($1.54 \rightarrow 1.07$ s/question on the run harness); the separate effect of dropping $H \geq \tau_{\text{mid}}$ is in Table 2. The provisional full-split figures are in Table 22; the harness-matched measurement is in Table 4.

K Generation-Quality Proxies

Table 23 answers the question of whether the CHAIR_I reduction is paid for in generation quality. Taken together with the recall column of Table 5 (SBC 79.1 vs. baseline 79.6 on LLaVA; 58.6 vs. 58.1 on Qwen), the proxies cover fluency (repetition), detail (length), diversity (distinct-n), and semantic consistency (CLIPScore). None of them moves outside noise on LLaVA, and all of them are flat or better on Qwen.

L Hyperparameter Ablations

We ablate the router threshold τ_{mid} of the original three-condition design (Table 24), the MSB boost factor b (Table 25), and the joint $\tau \times b$ grid (Table 26). All sweeps fix the remaining hyperparameters at the sweep defaults (the τ and joint $\tau \times b$ POPE sweeps were run with an earlier router that had no margin guard; conclusions about τ, b robustness are unaffected by adding the guard at $\delta = 0.5$). PMI parameters (α, β) are not swept; they are fixed at the VCD defaults (α=1.0, β=0.1). The τ sweep is retained because its flatness is part of the evidence that the H condition is dispensable (§4.2): the condition moves POPE F1 by about 0.1pt at b=1.8 (0.33pt at b=1.5, 1000q) across $\tau_{\text{mid}} \in [0.3, 0.7]$.

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

M 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 27 reports POPE (random 1000q) and MME (hallucination subset). The value $\delta = 0$ is special: it disables the margin guard (PMI fires whenever blank-agreement holds); a small $\delta > 0$ makes the guard almost always reject, so SBC approaches the MSB-only baseline (δ=0.1: LLaVA Rec 77.4).

PMI must be applied—but not too aggressively. The sweep shows a non-monotonic effect on POPE: $\delta = 0.0$ applies PMI unguarded, a small $\delta > 0$ (almost no PMI) collapses toward an MSB-only baseline with little POPE gain on LLaVA (F1 85.9), and larger δ (0.3–0.5) selectively applies PMI where it is most needed. On POPE, δ=0.5 is the best guarded setting 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 on POPE-random the blank-agreeing, narrow-margin "no"s that PMI overturns are mostly prior-driven misses.

Why MME prefers smaller δ (a greedy-baseline artefact). 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 LLaVA-1.5 MME baseline is $\sim$565, §6), where contrasting against a corrupted input recovers visual evidence lost to sampling noise. Against our *greedy* baseline (631.7 LLaVA / 680.0 Qwen), greedy decoding 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 at or slightly above the greedy baseline (635.0/685.0 vs. 631.7/680.0 at δ=0.1). At the shared operating point, no greedy-decoded method beats the greedy baseline on MME in our setup (VCD, AIF, PMI-only and SBC trail it; §6, Table 8); the re-measurement under VCD's original sampling protocol is in Table 9.

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 among the guarded settings ($\delta > 0$) on both backbones (above). MME loss at $\delta = 0.5$ is comparable to other contrastive baselines and reflects the greedy-baseline mismatch—PMI over-correcting a strong greedy baseline—which within SBC is concentrated in the PMI route (Table 11). We do not optimise for MME: its hallucination subset is perception-heavy and our greedy baseline is already strong.

The choice exposes the trade-off transparently: with $\delta = 0.5$ POPE F1 improves substantially while MME drops by 17–33 subset-score points vs. greedy (still above PMI-only on both backbones). A task-conditional δ (small for perception-style benchmarks, larger for prior-bias benchmarks) is left to future work.

N Mask-Fill Robustness

A natural concern is that the mean-colour 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 artefact rather than a counterfactual occlusion signal. 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-colour** (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 28 reports the result. All three fills yield the same H median (0.998), the same H mean to within 0.01 (0.940–0.946), and over-spread shares within a ± 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 artefact of one specific fill choice.

O SBC Failure Cases

Figure 5 shows two cases where the routed actuator corrects a baseline error; its routes come from the instrumented router described in the caption. Under the reported router the CHAIR correction holds, but on the POPE case the margin guard vetoes PMI and the output remains the greedy baseline's incorrect "no". 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 11: LLaVA-1.5-7B PMI Brk 100% on the guard-filtered subset $n = 281$; Qwen2.5-VL shows a lower $\sim$35% PMI break rate at smaller volume $n = 173$ because the margin-guard preempts 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 answer, the router selects PMI, and the contrast amplifies the image's small lift of "yes" over the blank-image prior even though the queried object is absent.

Quantitatively, on the LLaVA-1.5-7B adversarial split (numbers below are from the LLaVA POPE routing log used to populate Table 11), 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 all 151 go in the same direction: baseline "No" $\rightarrow$ SBC "Yes". Two representative cases:

Failure 1 (COCO val2014 #442809). *Q:* "Is there a dining table in the image?" *GT:* no. *Baseline:* "No" (correct). *SBC:* $H = 0.998$, $K = 6$ segments, routed to PMI, output "Yes" (wrong).

Failure 2 (COCO val2014 #478736). *Q:* "Is there a fire hydrant in the image?" *GT:* no. *Baseline:* "No" (correct). *SBC:* $H = 0.995$, $K = 5$ segments, routed to PMI, output "Yes" (wrong).

The common mechanism: when the queried category is *plausible given the scene* (a cake $\rightarrow$ dining table; traffic lights and a stop sign $\rightarrow$ fire hydrant), the image raises the probability of "Yes" relative to the blank-image prior even though the on-image argmax is still "No". PMI's contrastive blend $(1+\alpha)\ell(\cdot|v) - \alpha\ell(\cdot|v_\emptyset)$ then *amplifies* that difference, flipping "No" to "Yes". Both cases satisfy blank-agreement and have on-image top-token margin below $\delta = 0.5$, so both routing conditions held (and both sit at $H > 0.99$, so the original third condition also held). These residual cases are the regime where even the guarded router 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 responses, a trade-off discussed in App. M. AMBER's all-negative existence split (§6.3) is this failure mode measured on a probe set built entirely out of such cases.

P Is Attention Entropy a Substitute for H?

A natural concern is whether the LOO 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*).

Theoretical distinction. ϕ_i measures the *counterfactual occlusion signal*: the change in the

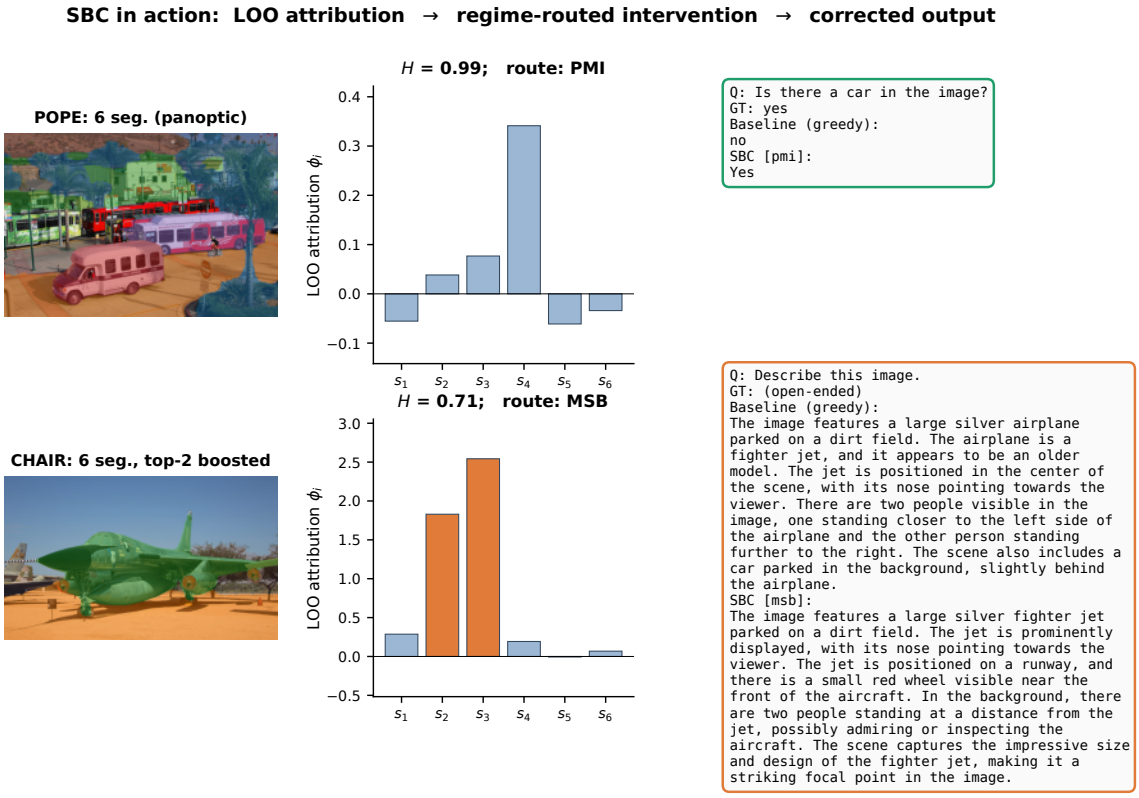


Figure 5: **SBC in action.** One POPE example (routed to PMI) and one CHAIR example (routed to MSB) on LLaVA-1.5-7B: panoptic segments yield the per-segment LOO attribution ϕ and its entropy H; the router selects PMI or MSB; PMI recovers the car the baseline missed (POPE), and MSB removes the baseline's hallucinated car (CHAIR; other errors persist). H and ϕ are shown for analysis; the routes here come from an instrumented router that checks $H \geq \tau_{\text{mid}}$ and blank-image agreement without the margin guard (reported router: §4.2); under the reported router the margin guard vetoes PMI on the POPE example, which stays "no" (App. O).

model's $\log p$ of its own answer when segment i is removed ($\phi_i = \log p(\text{span} \mid v) - \log p(\text{span} \mid v \setminus s_i)$, Eq. 1). Attention measures the softmax-normalised query–key *weight* on visual tokens. 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.

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 H and last-layer attention-H is $r = -0.016$ with means $\bar{H} = 0.980 \pm 0.094$ vs. $\bar{H}_{\text{attn}} = 0.759 \pm 0.110$. The two signals are nearly uncorrelated on this sample ($r^2 < 0.001$; Spearman $\rho = -0.11$, $p = 0.28$), although H is near its ceiling here (84% of samples have $H \geq 0.99$), which restricts its range. H also sits much higher ($H > H_{\text{attn}}$ on 97% of samples; binary QA is the over-spread regime).

Conclusion: substitutable in the router at this operating point. Two things are measured here, and we claim only those. First, Table 29 shows that an attention-H router behaves like the H router *on average* on POPE (F1 88.68 vs. 88.59 for the 3000q SBC reference; the H-free no_h gives 88.82 on the same 1000q sub-sample), so as a routing input the two are interchangeable at this operating point—unsurprising once one notes that the router uses only a threshold crossing, and on binary-QA both signals sit above their thresholds almost always. Second, the two signals are per-sample uncorrelated ($r = -0.016$) and sit at different levels (a mean offset of 0.22), so they are not one quantity measured two ways. We claim nothing beyond this: we have not recomputed the five-bin distributions of §3 under attention-H, and therefore make no claim about whether the regime structure would be visible in them. What the two measurements do support is the design of §5, in which H is an offline analysis instrument and the router reads neither it nor a substitute for it.

Q Router-Condition Substitution Ablations

The full-cell removal of the $H \geq \tau$ condition is in §4.2 (Table 2). This appendix holds the two ablations that vary *what* the condition is rather than whether it is present: Table 29 substitutes the entropy test on POPE-random, and Table 30 removes blank-agreement as well and measures the cost on CHAIR and MME.

Split	Route	n	Rec. (%)	Brk. (%)
POPE, LLaVA-1.5-7B				
random	MSB	1172	0.0	0.9
	PMI	190	**100.0**	**100.0**
	Base/FB	1638	0.0	0.0
popular	MSB	1209	4.3	0.9
	PMI	224	**100.0**	**100.0**
	Base/FB	1567	0.0	0.0
adversarial	MSB	1280	4.3	0.9
	PMI	281	**100.0**	**100.0**
	Base/FB	1439	0.0	0.0
MME, LLaVA-1.5-7B				
–	MSB	172	3.4	2.1
	PMI	14	50.0	**80.0**
	Base/FB	54	0.0	0.0
POPE, Qwen2.5-VL-7B				
random	MSB	1222	66.3	0.9
	PMI	119	72.5	16.0
	Base/FB	1659	1.4	0.8
popular	MSB	1246	52.2	0.8
	PMI	146	63.6	24.6
	Base/FB	1608	5.4	0.8
adversarial	MSB	1278	46.9	1.5
	PMI	173	59.8	**35.2**
	Base/FB	1549	6.6	0.8
MME, Qwen2.5-VL-7B				
–	MSB	125	0.0	0.9
	PMI	43	37.5	17.1
	Base/FB	72	0.0	0.0

Table 11: Per-route recover/break analysis at $\delta = 0.5$ (POPE 3 splits + MME). Base/FB (as in Table 10) pools the margin-veto and $K < 2$-fallback responses. *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 blank-agreement and the margin guard 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$66% recovery, n=1222) and the margin-veto Base/FB bucket ($n\sim1600$); active PMI is reserved for narrow-margin cases (Rec $\sim65\%$, Brk $\sim25\%$). (iii) On LLaVA MME the PMI break rate climbs to 80% ($n = 14$), reinforcing the pattern: when the PMI contrast 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 12: Per-bin baseline error rate $P(\text{error} \mid \text{bin})$ on POPE (LLaVA-1.5-7B, the 2982 of 3000 questions per split with $K \geq 2$) 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 §3.4 statement that per-bin rate cannot discriminate bins, and that our claims rest on mass-tail location instead.

Backbone	Data	Dom. tail	LOO	Exact
LLaVA-7B	CHAIR	over-conc.	50.3	52.8
LLaVA-7B	POPE-adv	over-spread	92.8	91.8
LLaVA-7B	MME	over-spread	84.6	80.8
Qwen-7B	CHAIR	over-conc.	67.7	68.7
Qwen-7B	POPE-adv	over-spread	89.4	84.0
Qwen-7B	MME	over-spread	75.4	70.0

Table 13: Dominant-tail share (% of samples) under the LOO estimator and under exact Shapley values, on fixed-seed sub-samples (n=199/194/240/99/94/240 in row order; POPE is the adversarial split; all sub-samples differ from those of Table 1, hence the offsets). The dominant tail is unchanged in 6/6 cells.

Cell	Est.	5-bin histogram	Dom. %
LLaVA CHAIR (n=199)	LOO	100/19/18/15/47	50.3
	Exact	105/15/30/17/32	52.8
LLaVA POPE-adv (n=194)	LOO	4/1/6/3/180	92.8
	Exact	4/0/4/8/178	91.8
LLaVA MME (n=240)	LOO	5/12/8/12/203	84.6
	Exact	5/14/6/21/194	80.8
Qwen CHAIR (n=99)	LOO	67/9/7/5/11	67.7
	Exact	68/10/13/7/1	68.7
Qwen POPE-adv (n=94)	LOO	1/1/3/5/84	89.4
	Exact	0/3/2/10/79	84.0
Qwen MME (n=240)	LOO	24/10/10/15/181	75.4
	Exact	20/13/15/24/168	70.0

Table 14: Five-bin H histograms (over-conc./conc./mixed/spread/ over-spread) under the LOO estimator and under exact Shapley values. The dominant bin is unchanged in every cell; the mass that moves goes to the middle bins, never across the distribution.

Cell	Spearman	Top-1 dis.	Resid.
LLaVA CHAIR	0.715	40/189 (21%)	1.4e-14
LLaVA POPE-adv	0.520	52/186 (28%)	8.9e-16
LLaVA MME	0.521	65/234 (28%)	2.2e-15
Qwen CHAIR	0.673	27/97 (28%)	2.8e-14
Qwen POPE-adv	0.538	32/94 (34%)	1.8e-15
Qwen MME	0.441	79/234 (34%)	3.6e-15

Table 15: Estimator divergence on samples with $K \geq 3$. *Spearman*: mean per-sample rank correlation between the LOO and exact ϕ vectors. *Top-1 dis.*: samples on which the highest-attribution segment—the one MSB would boost first—differs. *Resid.*: maximum efficiency residual $|f(N) - f(\emptyset) - \sum_j \phi_j^{\text{exact}}|$, confirming the exact computation. The ranking diverges materially; the diagnostic (Table 14) does not.

Backbone	AMBER format	Dom. tail	%
LLaVA-7B	generative	over-conc.	51.5
LLaVA-7B	discriminative	over-spread	98.3
Qwen-7B	generative	over-conc.	66.6
Qwen-7B	discriminative	over-spread	81.2

Table 16: AMBER diagnostic, $n \approx 300$ per cell, same images across the two formats. The task-conditional split reproduces *within* a single benchmark and tracks the CHAIR (53/66%) and POPE (95/88%) shares of Table 1.

Track	Backbone	Base	SBC
Generative (% captions w/ halluc. ↓)			
	LLaVA-7B	13.0	13.2
	Qwen-7B	7.1	7.1
Existence, all-negative, 4,924q (Acc ↑)			
	LLaVA-7B	92.3	82.5
	Qwen-7B	95.4	93.5
Attribute, balanced, 7,628q (F1 ↑)			
	LLaVA-7B	78.2	79.5
	Qwen-7B	82.2	83.7

Table 17: AMBER decoding, both backbones, same shared configuration as Tables 5–8. Generative: response-level hallucination rate (% of captions mentioning $\geq$1 of AMBER's annotated absent objects, CHAIR_S-style; our scorer), on which SBC matches baseline. The existence split is an *all-negative* probe set (all 4,924 answers are "no") and reproduces the PMI break mode of App. O: flips to a wrong answer are PMI-routed (LLaVA $504/504$, Qwen $93/96$). On the balanced attribute subset (50/50 yes/no) SBC improves F1 on both backbones.

Backbone	CHAIR *over-conc. %*	POPE *over-spread %*
LLaVA-1.5-7B	53.2	94.9
Qwen2.5-VL-7B	66.3	88.4
Qwen2-VL-7B	54.5	86.1
LLaVA-1.5-13B (8-bit)	50.8	94.6
InternVL3-8B	64.6	93.6

Table 18: Dominant tail and share across five backbones from three families (LLaVA, Qwen, InternVL). Qwen2-VL-7B and LLaVA-1.5-13B are diagnostic-only; the 13B model is 8-bit, which confounds the attention-based actuator. InternVL3-8B uses an InternViT encoder and is also decoded against greedy (Table 19).

	CHAIR			POPE F1				MME
Method	C_S↓	C_I↓	Rec↑	rand	pop	adv	avg	score
Baseline	27.8	7.58	70.1	92.5	89.9	86.6	89.70	693.3
SBC	26.6	7.13	70.1	93.1	90.3	86.2	89.86	683.3

Table 19: InternVL3-8B decoding, greedy baseline vs. SBC. SBC uses the three-condition router and the shared configuration of Table 5 (b=1.8, δ=0.5, top-k=2) with no re-tuning; each image is one 448×448 view (dynamic tiling disabled), as in the diagnostic; bf16, single run. CHAIR: 1000 images; POPE: 3000q per split, average over the three splits; MME: hallucination-subset score (240q). Paired bootstrap 95% CIs of SBC−baseline include 0 for CHAIR_I, POPE avg F1 and MME (App. H.1). VCD, OPERA and AIF were not run on this backbone.

max K	4	6 (def.)	8
POPE-random F1 (1000q)	88.70	88.82	88.82
CHAIR_I (300 img)	13.93	13.19	13.15

Table 20: Segment-count ablation (LLaVA-1.5-7B, all other hyperparameters at defaults). POPE F1 is flat and the H tails are unchanged (dominant share at K=4/8: CHAIR over-concentration $52.9/53.5\%$, POPE over-spread $94.6/94.5\%$, n=297 CHAIR / 598 POPE; K=6: $53/95\%$, Table 1)—expected, since H is normalised by $\log K$. The K=6 CHAIR cell closely matches the full 1000-image run (13.19 vs. 13.14, Table 5); the K=4 offset comes from coarser segment ranking inside MSB, i.e., from the actuator, not the diagnostic.

Quantity (LLaVA-7B)	**M2F**	**SAM**
CHAIR_I (MSB, b=1.5)	13.69	13.35
CHAIR_I (MSB-sent, b=1.5)	13.45	13.51
POPE over-spread share	95%	98.5%
CHAIR dominant tail	over-conc.	over-spread
its share	53%	54.0%
PMI fired on CHAIR	0/300	0/300

Table 21: Segmenter swap: Mask2Former (M2F) vs. class-agnostic SAM (Kirillov et al., 2023). End metrics and the router are unaffected; the CHAIR H distribution is not. SAM cuts objects into parts, so a caption grounded in one object spreads over several segments and the CHAIR tail moves to over-spread; the low-H tail is thus defined relative to an object-aligned partition. CHAIR rows use the matched development setting b=1.5, hence the offset from Table 5. The M2F-MSB and both SAM CHAIR_I cells come from development runs whose per-sample outputs were not retained; only their summary values are available.

Variant	**CHAIR s/img**	**POPE s/q**
SBC (eager, latency harness)	5.05	1.65
SBC (lazy)	5.66	1.11
attention-ranked	4.85	0.80
Baseline greedy	3.01	0.23

Table 22: Provisional run-harness wall clock for the deferred (lazy) and attention-ranked variants, not comparable digit-for-digit with the latency-harness rows of Table 4; the harness-matched measurement is in Table 4. The full POPE runs' own timings (adversarial split, run harness) give $1.70/1.14/0.81$ s for eager/lazy/attention-ranked; the relative reductions (33%, 52%) agree across all three measurements.

	Len	D-1	D-2	Rep-4	CLIP
LLaVA-1.5-7B					
Baseline	90.4	0.589	0.901	0.0079	0.723
VCD	93.4	0.627	0.924	0.0036	0.720
OPERA	88.5	0.605	0.905	0.0065	0.718
AIF	86.8	0.599	0.908	0.0065	0.716
MSB-only	90.5	0.588	0.899	0.0087	0.724
PMI-only	94.0	0.609	0.913	0.0057	0.722
SBC	90.5	0.588	0.899	0.0088	0.724
Qwen2.5-VL-7B					
Baseline	159.4	0.397	0.628	0.304	0.713
VCD	143.3	0.543	0.824	0.115	0.699
OPERA	113.8	0.495	0.754	0.175	0.709
AIF	149.3	0.403	0.634	0.297	0.715
MSB-only	161.4	0.402	0.641	0.286	0.715
PMI-only	150.5	0.491	0.744	0.199	0.700
SBC	160.7	0.404	0.644	0.282	0.715

Table 23: Reference-free quality proxies on the full 1000-caption CHAIR outputs. *Len*: average caption length in words. *D-1/D-2*: distinct-1/2 (diversity). *Rep-4*: 4-gram repetition rate (degeneration). *CLIP*: sentence-level CLIPScore with CLIP ViT-B/32 (semantic consistency). On LLaVA every proxy is within noise of baseline; on Qwen every proxy is flat-to-better, including *less* degeneration ($0.304 \rightarrow 0.282$). VCD (which samples, App. A) and PMI-only raise diversity but lower CLIPScore on Qwen. AIF on LLaVA (-3.6 words, alongside the -6.6pt recall of Table 5) and OPERA on Qwen (-29%) come with markedly shorter captions.

τ_{mid}	random	popular	adv.
0.30	88.2/88.4	85.5/86.1	79.0/81.1
0.40	88.2/88.4	85.5/86.1	79.0/81.1
0.50	88.2/88.4	85.5/86.1	79.0/81.1
0.60	88.2/88.3	85.5/86.0	79.0/81.0
0.70	88.1/88.3	85.4/86.0	78.9/81.0

Table 24: Router threshold ablation (POPE, LLaVA-1.5-7B, Acc/F1) at default $b = 1.8$. All cells lie within ± 0.5pt of each other within each split. Cells come from an earlier router without the margin guard (recall $\approx 90\%$) and therefore differ from the main-table δ=0.5 numbers (Tables 6/7); the conclusion is unchanged because the δ guard is orthogonal to the τ decision (App. M).

b	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$	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 25: Boost factor ablation (SBC, CHAIR, LLaVA-1.5-7B) at $\tau_{\text{mid}} = 0.5$ (δ is immaterial: PMI never fires on CHAIR). $b = 1.8$ 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 26: Joint $\tau \times b$ sweep on POPE (LLaVA-1.5-7B). At fixed b, all τ cells lie within ± 0.3pt of each other; the router and the MSB boost do not interact: each one's optimum is unchanged across the other's settings. b=1.5 rows are 1000q runs and b=1.8 rows 3000q runs, so the offset between the two b levels mixes b with sample size. The τ=0.50, b=1.8 cell is the default in Tables 5–6; cells here come from the earlier router without the margin guard and therefore differ from the main-table δ=0.5 values (Tables 6/7). The relative ordering across τ, b is unchanged.

Backbone	δ	Acc	F1	Rec	MME
LLaVA-7B	0.0†	88.90	89.09	90.6	586.7
LLaVA-7B	0.1	87.30	85.90	77.4	635.0
LLaVA-7B	0.3	88.70	87.73	80.8	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	90.40	89.50	81.8	668.3
Qwen2.5-VL	0.5	91.00	90.28	83.6	663.3

Table 27: Image-margin threshold δ sweep. POPE columns are on the first 1000 questions of the random split (taken from the 3000q run where only that exists) and MME on the hallucination subset; numbers thus differ from main Tables 6/8 which average the three POPE splits over 3000q each. Setting $\delta = 0$ disables the margin guard. †LLaVA $\delta = 0$ is from an earlier router version with the guard off (no $\delta = 0$ run of the final router exists for LLaVA).

Fill	n	H mean	H med.	Hist.	OS %
mean	298	0.941	0.998	2/8/9/8/271	90.9
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 28: 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 near-identical means, the same median, and over-spread shares within ± 0.5pt: the mass-tail diagnostic does not depend on the fill choice.

Router 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 29: Router-condition substitution ablation on LLaVA-1.5-7B POPE random. The SBC (ref) row is the full 3000q run (89.13/88.59/84.40/93.23); the six variant rows below it are the 1000q ablation subsample, so they are comparable with each other but not digit-for-digit with the reference row. All variants hold the actuators (MSB, PMI) and the margin guard δ=0.5 fixed; only the entropy and blank-agreement conditions vary. `attn_h` = replace H with last-layer attention entropy; `logit_h` = replace with top-100 logit entropy; `task_pmi` / `task_msb` = oracle task-label router (always PMI candidate / always MSB); `no_h` = drop the $H \geq \tau$ check, keep blank-agreement + 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) `task_pmi` applies the same rule as `no_h` on binary QA (identical outputs), and both match SBC because $H \geq \tau$ already holds for 88–95% of POPE responses (Table 1); (iii) `attn_h` tracks H *on average*, even though the two signals are per-sample uncorrelated (App. P, r=−0.016). Within POPE the `pmi_guard` row looks best, but see Table 30 for its cross-task cost.

Backbone	Variant	$C_S\downarrow$	$C_I\downarrow$	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 30: Cross-task `pmi_guard` ablation: drop both $H \geq \tau$ and blank-agreement, keep only δ, on CHAIR (1000 images) and MME (score column). Benign on POPE (Table 29), it inflates CHAIR_S by 2.5pt (LLaVA) and 4.2pt (Qwen) on captioning—the cost of removing the blank-agreement check.