ATTENDING YET HALLUCINATING: INHIBITORY EVIDENCE DEFICITS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) combine visual understanding with language generation, yet hallucinated responses continue to limit their reliability. Recent studies show that these errors persist even when attention concentrates on relevant visual regions. This raises a question: **why does attending to the right location still lead to an incorrect answer?** We investigate how visual tokens within these regions affect competing answer candidates. Our analysis reveals **Competitive Evidence Imbalance**: (1) Visual tokens within relevant regions provide conflicting evidence, with a substantial share favoring competing answers over the correct one. (2) Tokens favoring the correct answer support it either directly or indirectly by suppressing competing answers. We call the latter **Inhibitory Support Tokens**, whose average count and regional proportion are markedly lower in hallucinated answers despite substantial remaining direct support. Motivated by this finding, we introduce Suppression-prioritized Evidence for Answer Recalibration (SPEAR), a training-free inference method. SPEAR proposes a relevant region from attention, prioritizes inhibitory support tokens while retaining direct support and opposing effects, and aggregates their signed responses to calibrate candidate probabilities under a bounded distributional change. Experiments show improved object hallucination mitigation, compositional discrimination, open-ended answer quality, and aggregate visual perception. These results highlight the value of using visual evidence to suppress competing answers as well as support plausible ones.[1]

1 INTRODUCTION

Large Vision-Language Models (LVLMs) perform strongly in image captioning, visual question answering, and multimodal reasoning (Radford et al., 2021; Li et al., 2023a; Liu et al., 2023). Yet they still hallucinate objects, attributes, and relations absent from the image (Rohrbach et al., 2018; Bai et al., 2025b). These errors undermine deployment reliability when visual outputs guide downstream decisions, including on resource-limited edge devices (Yao et al., 2025).

Mitigation uses visual alignment, preference learning, and inference-time visual or distributional adjustments (Sun et al., 2024; Leng et al., 2024; Huang et al., 2024). Region magnification and selective reinforcement strengthen the use of relevant visual content (Mao et al., 2026; Zhu et al., 2026; Ren et al., 2026). Yet recent studies show that hallucinations persist even when attention concentrates on answer-relevant regions (Zhang et al., 2025; 2026b). Why does attending to the correct image region still produce an incorrect answer?

We investigate how tokens within that region influence the final choice, requiring a measure of each token's effect on candidates and their relative advantage. Attention weights, their value mixture, and gradients alone do not specify this influence. We propose Γ, which measures a candidate-centered score's local response as the current read moves toward a visual token. The response difference $\Delta\Gamma$ measures the resulting change in two candidates' relative advantage. On annotated TextVQA regions, emphasis sometimes reduces discrimination despite a positive target response. Further analysis reveals competing token effects within the same region: a substantial share favors an incorrect competitor (Figure 1). Four roles separate direct support, stronger rival suppression, greater rival gains under shared support, and other rival-favoring effects. Tokens with stronger rival

[1] Code is available for review at `https://anonymous.4open.science/r/SPEAR-9EBA/`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

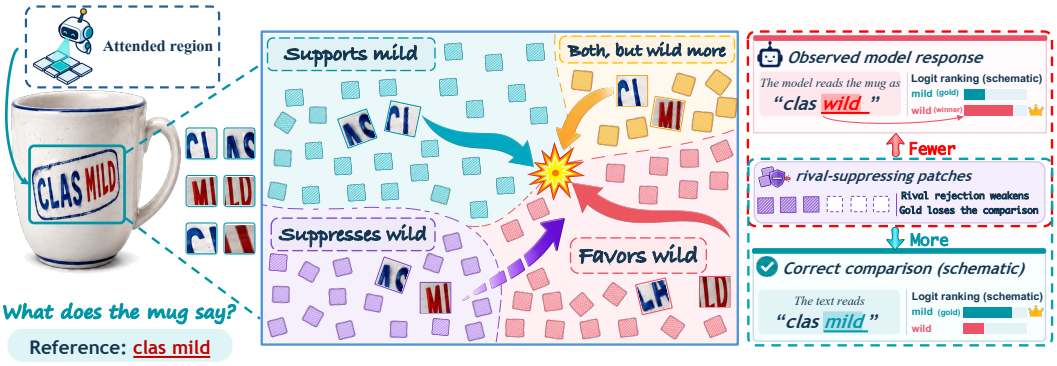


Figure 1: **Attending to the relevant region does not ensure a correct answer.** Reading *"clas wild"* for *"clas mild"* illustrates conflicting evidence within the attended region. The schematic shows weaker rival suppression leaving the correct candidate behind despite direct support.

suppression indirectly support the correct answer: we call them **inhibitory support tokens**. Our experiments show that inhibitory support tokens decline in mean count and regional proportion in incorrect answers across three models, despite substantial direct support. We interpret this decline as weaker rival suppression that erodes the correct answer's advantage and contributes to hallucination. We call this conflict and diminished inhibitory support **competitive evidence imbalance**.

To address this imbalance, we propose Suppression-prioritized Evidence for Answer Recalibration (SPEAR) for training-free inference. Attention proposes a compact region for response computation. SPEAR prioritizes suppression alongside direct support for both candidates. Signed responses guide bounded probability calibration, connecting rival suppression to answer selection. Benchmarks show gains in hallucination mitigation, compositional discrimination, answer quality, and aggregate perception. We summarize our primary contributions as follows:

- We propose interpretable candidate-specific responses that connect token effects to relative answer preference and explain the net effect of regional visual emphasis.
- We find competitive evidence imbalance in three models: regional tokens favor rivals; inhibitory support counts and proportions fall in errors despite substantial direct support.
- We propose SPEAR for suppression-prioritized, constrained candidate calibration and evaluate it across hallucination and multimodal capability benchmarks.

2 Analysis and Motivation

2.1 Hallucinations Despite Relevant Regional Attention

We analyze TextVQA, a visual question-answering dataset requiring image text recognition (Singh et al., 2019), using its manually annotated answer-region extension (Zhang et al., 2025).[2] For Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision-1.5-4B, we extract layer/head attention at answer positions on 1,000 fixed records. Mean attention per token compares reference region $\mathcal{R}$ with ten equal-sized background sets; head averages reveal layerwise concentration (Appendix D).

Across all three models, Figure 2(a) shows greater attention density in the annotated region than in the background for correct and incorrect answers across broad depth ranges, reaching tens of times in middle-to-late layers. Figure 2(b) shows an incorrect clock answer with attention on the annotated region; the accompanying group-mean head maps reinforce this concentration at later depths. These errors motivate tracing token effects within the attended region.

2.2 Candidate-specific Visual Responses

Explaining these errors requires tracing each regional token's effect on candidates. Attention weights determine how value vectors are mixed; weights alone do not measure the evidence carried by

[2] https://huggingface.co/datasets/jrzhang/TextVQA_GT_bbox

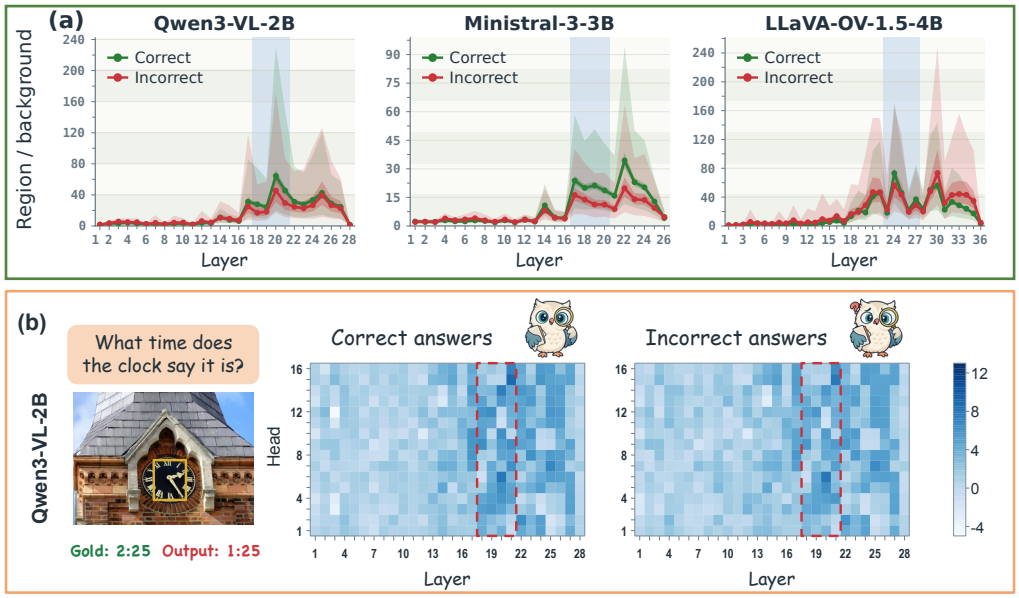


Figure 2: **Incorrect answers retain regional attention.** (a) Regional/background geometric mean ratios; light/dark bands show event interquartile ranges/95% confidence intervals of means. Green/red: correct/incorrect; blue: layer windows. (b) Clock error and Qwen cohort means: $\log_2$ head density ratios (shared scale). Red: 18–21; gold: reference box; OV: OneVision (Appendix A).

those vectors. The resulting attention-head output summarizes the read content without separating each token's effect on candidate ranking. Candidate gradients measure sensitivity to this aggregate output; signed token attribution requires token-specific directions. Prior work uses these signals to analyze hallucination and guide inference (Huang et al., 2024; Zhang et al., 2026c). Combining token directions from the current read with candidate sensitivity links patches to answer preferences.

The response measures a candidate-centered score's local change as the head output moves toward a visual token. At a fixed generation step, $\mathcal{C}$ is the natural top-K candidate set and z_c candidate c's logit. $K = 8$ balances gradient cost and coverage, including compatible reference tokens in 96.0%–98.3% of valid events across four models; larger K improves coverage (Appendix A.2). Freeze weights π_c at the unmodified state $t = 0$ and center score m_c:

$$\pi_c = \mathrm{stopgrad}\left(\frac{\exp z_c(0)}{\sum_{k\in\mathcal{C}} \exp z_k(0)}\right), \qquad m_c(t) = z_c(t) - \sum_{k\in\mathcal{C}} \pi_k z_k(t) \tag{1}$$

For layer ℓ and head h, let $\mathcal{T}$ index all visible positions, $A_{h,i}^{(\ell)}$ be attention weights, and $\boldsymbol{v}_{h,i}^{(\ell)} \in \mathbb{R}^{d_h}$ be value vectors of head dimension d_h. The complete read and direction toward visual token j are

$$\boldsymbol{o}_h^{(\ell)} = \sum_{i\in\mathcal{T}} A_{h,i}^{(\ell)} \boldsymbol{v}_{h,i}^{(\ell)}, \qquad \boldsymbol{d}_{h,j}^{(\ell)} = \boldsymbol{v}_{h,j}^{(\ell)} - \boldsymbol{o}_h^{(\ell)} \tag{2}$$

Along $\boldsymbol{o}_h^{(\ell)}(t) = \boldsymbol{o}_h^{(\ell)} + t\boldsymbol{d}_{h,j}^{(\ell)}$, the chain rule gives the local response through downstream computation, with ∇ taken with respect to $\boldsymbol{o}_h^{(\ell)}$:

$$\Gamma_{h,j}^{(\ell)}(c) = \left.\frac{dm_c(t)}{dt}\right|_{t=0} = \left\langle \nabla m_c, \boldsymbol{d}_{h,j}^{(\ell)} \right\rangle = \left\langle \nabla z_c - \sum_{k\in\mathcal{C}} \pi_k \nabla z_k, \boldsymbol{d}_{h,j}^{(\ell)} \right\rangle \tag{3}$$

Gradients are evaluated at $t = 0$, where Γ also equals the conditional log-probability derivative $d\log q_c(t)/dt$ for $q_c(t) = \exp z_c(t)/\sum_{k\in\mathcal{C}} \exp z_k(t)$ (Equation 28). Thus Γ connects a token and the current read to candidate score/probability changes. For reference answer g and competitor r, subtracting responses cancels the shared candidate center:

$$\begin{aligned} \Delta\Gamma_{h,j}^{(\ell)}(g,r) &= \Gamma_{h,j}^{(\ell)}(g) - \Gamma_{h,j}^{(\ell)}(r) = \left\langle \nabla(z_g - z_r), \boldsymbol{d}_{h,j}^{(\ell)} \right\rangle \\ &= \left.\frac{d[z_g(t) - z_r(t)]}{dt}\right|_{t=0} = \left.\frac{d\log[p_g(t)/p_r(t)]}{dt}\right|_{t=0} \end{aligned} \tag{4}$$

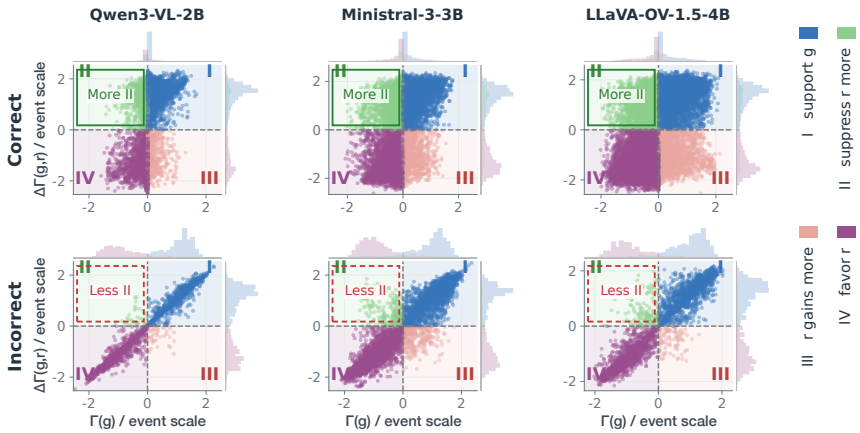


Figure 3: **Candidate competition within relevant regions.** Columns/rows: backbones/correctness; colors: roles. Green solid/red dashed frames mark inhibitory support (II). Event-scaled responses use inverse hyperbolic sine displays, raw ticks, and event-equal marginals (Appendix A).

Here $p_c(t)$ is the full-vocabulary probability. Joint signs of $\Gamma(g)$ and $\Delta\Gamma(g, r)$ distinguish the candidate's own response from comparative gain. Appendix F gives complete proofs, starting with Proposition F.1; Appendix D defines reference probes outside $\mathcal{C}$.

2.3 Visual Emphasis and Evidence Competition

Effects of Visual Emphasis. We test whether greater attention to visual targets within the correct region improves discrimination. We redistribute attention toward a target at fixed regional mass and outside attention (Appendix A.3). The local net margin response is $J_{\text{net}} = J_{\text{target}} - J_{\text{base}}$: both target and original-mixture responses include the head's regional mass (Proposition F.3).

Appendix Figure 12 shows positive targets with negative net responses across three target rules and four depths in all three models. Three finite intervention strengths decrease this group's mean reference/competitor margin and increase the positive-net group's (Appendix Figure 13). We interpret this loss as a **disruption of regional evidence balance**: redistributing weight toward shared or less discriminative content may displace tokens that better separate candidates. Replacing more discriminative evidence with a supportive target weakens discrimination, motivating token-level analysis.

Competitive Evidence Imbalance. We analyze regional competition with Γ and $\Delta\Gamma(g, r)$, summing signed head responses and averaging layers. In 69%–80% of events, favorable and unfavorable comparisons each occupy at least 5% of regional tokens. Joint signs define four roles (Table 1). Roles I and II favor the correct answer through direct support and stronger rival suppression, respectively. Role III raises both scores with greater gains for the rival; role IV also favors the rival. $|\Delta\Gamma(g, r)| \leq \delta$ marks neutral comparisons.

Table 1: **Diagnostic token roles.** g: correct candidate; r: rival. Neutral tolerance: $\delta = 10^{-7}$.

Role	$\Gamma(g)$	$\Delta\Gamma(g, r)$
I: Direct support	$> \delta$	$> \delta$
II: Inhibitory support	$\leq \delta$	$> \delta$
III: Rival gains more	$> \delta$	$< -\delta$
IV: Rival favored	$\leq \delta$	$< -\delta$

Figure 3 reveals the second feature: **inhibitory support tokens** (role II) decline in incorrect answers. Mean count and regional proportion fall from 4.30 and 13.19% to 0.40 and 0.62% in Qwen, as also observed in Ministral and LLaVA (Appendix A). Role II suppresses r more: $\Gamma(r) < \Gamma(g) \leq 0$ away from tolerance. Qwen errors have 59.60% direct support; role III declines in all models (Table 9).

Substantial regional evidence favors competitors, while inhibitory support declines in count and proportion in errors. Supporting tokens act through direct score gains or stronger rival suppression. We interpret this decline as weaker constraints on rivals, reducing the correct answer's advantage despite substantial direct support. SPEAR prioritizes suppression and retains opposing effects, weighing challenger support against evidence for the original answer.

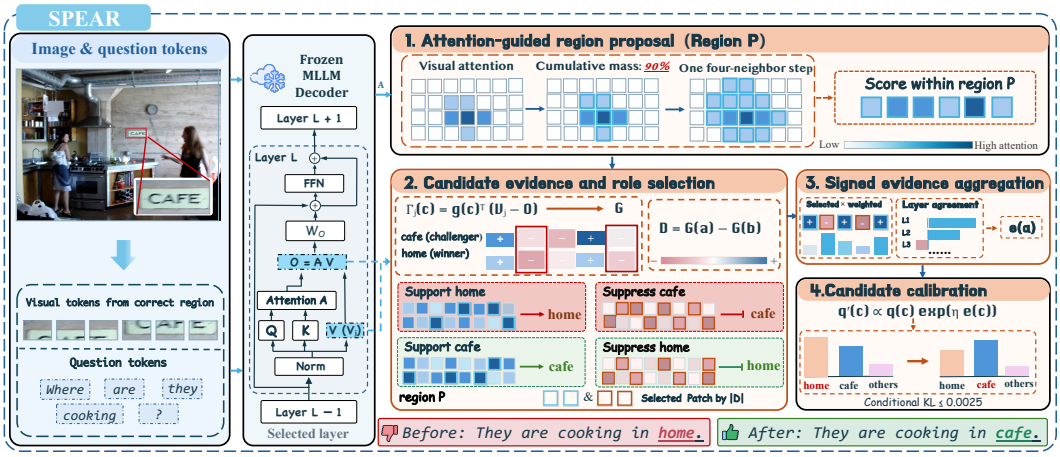


Figure 4: **Overview of SPEAR.** A frozen multimodal model supplies attention and candidate responses for region proposal, suppression-prioritized selection, signed aggregation, and constrained calibration. Example outputs and score bars are schematic.

3 The SPEAR Approach

SPEAR uses signed visual evidence to address competitive evidence imbalance (Figure 4). Attention localizes response evaluation to reduce patchwise computation and storage. For image $\boldsymbol{I}$, question $\boldsymbol{x}$, and prefix $\widehat{\boldsymbol{y}}_{<t}$, fixed model θ returns $\boldsymbol{p}'_t = f_{\theta,\mathrm{SPEAR}}(\boldsymbol{I}, \boldsymbol{x}, \widehat{\boldsymbol{y}}_{<t})$, the calibrated distribution.

3.1 Attention-based Region Proposal

To reduce token-response computation, we first locate an answer-relevant region and evaluate visual-token responses there. Figure 2 shows that middle-to-late attention concentrates on answer-relevant regions. We use this attention to identify their visual tokens without region annotations, averaging over layer window $\mathcal{L}$ and its H heads: $\bar{A}_j = (|\mathcal{L}|H)^{-1} \sum_{\ell\in\mathcal{L}} \sum_h A^{(\ell)}_{h,j}$.

To test localization, we map selected tokens to TextVQA images and measure annotated-region coverage and overall visual-token retention (Appendix Figure 7). The shortest descending-attention prefix containing 90% of visual attention mass forms $\mathcal{P}_0$; one four-neighbor expansion yields $\mathcal{P}$ (Equation 10). On 908 Qwen events, the proposal covers 97.29% of reference tokens while retaining 49.00% of all visual tokens, focusing responses on relevant tokens and context.

3.2 Candidate Evidence Construction

Suppression-prioritized Selection. Within $\mathcal{P}$, we compare each challenger a in the top-8 set $\mathcal{C}$ with original winner $b = \arg\max_{c\in\mathcal{C}} z_c$. Centering on $\mathcal{C}$ avoids binary support/suppression collapse (Equation 1; Proposition F.2). We sum signed head responses, scale by the shared floored median token-wise root-mean-square response $s^{(\ell)} > 0$ (Appendix F.4), and average layers:

$$G^{(\ell)}_j(c) = \frac{\sum_h \Gamma^{(\ell)}_{h,j}(c)}{s^{(\ell)}}, \qquad G_j(c) = |\mathcal{L}|^{-1} \sum_{\ell\in\mathcal{L}} G^{(\ell)}_j(c), \quad D_j(a,b) = G_j(a) - G_j(b) \tag{5}$$

D compares normalized responses; $\Delta\Gamma$ compares raw responses. For threshold $\delta > 0$, tokens with $D_j(a,b) > \delta$ support a if $G_j(a) > \delta$, or suppress b otherwise. For $D_j(a,b) < -\delta$, testing $G_j(b)$ separates support for b from suppression of a, extending inhibitory support to either candidate.

To address diminished inhibitory support, suppression roles receive twice the support retention rate. This gives rival constraints greater representation alongside direct support on both sides. Within each role, we retain the largest $|D_j(a,b)|$ values at rates $(f, \min(2f,1)), f, \min(2f,1))$, with $f = 0.05$. Each nonempty role contributes $\lceil f_i n_i \rceil$ tokens to $\mathcal{S}_a$ without a global cap. Neutral/tie rules appear in Appendix E; exchange symmetry and rounding bounds in Propositions F.8–F.9.

Signed Evidence Aggregation. The selected effects retain their signs, including evidence favoring b. Normalize attention over the entire proposal, $\alpha_j = \bar{A}_j / \sum_{k\in\mathcal{P}} \bar{A}_k$. The layer evidence $\xi_a^{(\ell)}$, its mean $\bar{\xi}_a$, agreement factor ρ_a, and final evidence e_a are

$$\begin{aligned} \xi_a^{(\ell)} &= \sum_{j\in\mathcal{S}_a} \alpha_j [G_j^{(\ell)}(a) - G_j^{(\ell)}(b)], \qquad \bar{\xi}_a = |\mathcal{L}|^{-1} \sum_{\ell\in\mathcal{L}} \xi_a^{(\ell)} \\ \rho_a &= \frac{|\bar{\xi}_a|}{|\mathcal{L}|^{-1} \sum_{\ell\in\mathcal{L}} |\xi_a^{(\ell)}| + \epsilon_\rho}, \qquad e_a = \rho_a \bar{\xi}_a, \qquad e_b = 0 \end{aligned} \tag{6}$$

The stabilizer $\epsilon_\rho > 0$ handles vanishing evidence; cross-layer cancellation reduces ρ_a. Shared weights preserve subset attention mass. Suppression priority changes selection, without multiplying evidence or renormalizing roles. Challenger-specific sets jointly calibrate the distribution.

3.3 Constrained Candidate Calibration

Calibration uses $\boldsymbol{e} = (e_c)_{c\in\mathcal{C}}$ when a challenger has positive evidence; otherwise it retains the original distribution. Let p_v be vocabulary probability, $P_\mathcal{C} = \sum_{c\in\mathcal{C}} p_c$ the candidate mass, and $q_c = p_c / P_\mathcal{C}$ the normalized candidate probability. For $\eta \geq 0$, update

$$q'_c = \frac{q_c \exp(\eta e_c)}{\sum_{k\in\mathcal{C}} q_k \exp(\eta e_k)}, \qquad p'_v = \begin{cases} P_\mathcal{C} q'_v, & v \in \mathcal{C} \\ p_v, & v \notin \mathcal{C} \end{cases} \tag{7}$$

The update preserves the candidate mass and every noncandidate probability. With $e_b = 0$, the exact change in a challenger's log odds is $\log(p'_a/p'_b) - \log(p_a/p_b) = \eta e_a$.

To retain contextual linguistic preferences, we bound distributional change using $\eta \leq \eta_{\max} = 1$ and the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(\boldsymbol{q}' \| \boldsymbol{q}) \leq \kappa = 0.0025$. Bisection enforces the budget through monotonicity (Proposition F.11). The decoder uses $\boldsymbol{p}'$; attention, values, and the key-value cache stay intact. Tokens alter later prefixes. Appendices E and F give the algorithm and proofs.

4 Experiments

4.1 Experimental Setup

Base Models and Baselines. We evaluate SPEAR on seven backbones: Qwen3-VL-2B/4B, Ministral-3-3B/8B, LLaVA-OneVision-1.5-4B/8B, and Qwen3.5-2B (Table 3). On Qwen3-VL-2B-Instruct, we compare against greedy, beam search, nucleus sampling, VCD, OPERA, HALC, IM-CCD, DeCo, LVLMs-Saliency, and HulluEdit, using their respective decoding operations (Appendix B.2). Appendix Tables 17 and 11 list SPEAR's parameters and benchmark sample counts.

Evaluation Benchmarks and Metrics. We use Polling-based Object Probing Evaluation (POPE) on Common Objects in Context (COCO) and GQA images (Li et al., 2023b; Lin et al., 2014; Hudson & Manning, 2019), Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018), AMBER (Wang et al., 2024), the Multimodal Large Language Model Evaluation (MME) benchmark (Fu et al., 2025), Attribution, Relation, and Order (ARO) evaluation (Yuksekgonul et al., 2023), and MMHal-Bench (Sun et al., 2024). We report POPE accuracy/F1, CHAIR sentence/instance rates with up to 256 new tokens, and benchmark-specific capability scores (Appendix B). MMHal-Bench responses are scored by GPT-5.6 Terra through the OpenRouter API.[3]

4.2 Hallucination Mitigation and Answer Quality

Table 2 shows that SPEAR exceeds greedy in POPE accuracy on every COCO/GQA subset and leads most settings, supporting reliable judgments under random, popular, and adversarial distractors. It leads or ties for best F1 across COCO subsets and leads GQA random/popular. On CHAIR, SPEAR achieves the lowest instance-level rate and the second-lowest sentence-level rate after HALC. Candidate recalibration reduces unsupported object mentions and improves object verification.

[3] `https://openrouter.ai/`

Table 2: **Comparison on POPE-COCO, POPE-GQA, and CHAIR.** CHAIR: up to 256 new tokens; C_S/C_I: percentages of captions with hallucinated objects/hallucinated object mentions (lower is better). Parentheses: percentage-point changes from greedy. Best: **bold**; second-best: underlined.

	POPE-COCO						POPE-GQA						CHAIR	
	Random		Popular		Adversarial		Random		Popular		Adversarial			
Method	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	C_S ↓	C_I ↓
Greedy	85.67	86.63	81.11	83.06	76.18	79.52	87.51	92.06	83.53	89.92	80.86	86.53	39.80	12.18
Beam search	85.77	86.75	81.21	83.18	76.28	79.64	87.52	92.07	83.54	89.93	80.87	87.54	39.80	11.97
Nucleus sampling	85.80	86.80	80.90	83.00	76.03	79.56	87.29	91.94	83.29	89.68	80.63	88.40	40.60	12.91
VCD Leng et al. (2024)	89.53	89.40	87.57	87.56	85.67	85.83	90.81	93.81	86.30	91.04	82.30	88.81	38.40	8.58
OPERA Huang et al. (2024)	89.28	88.85	87.07	86.80	85.80	85.47	90.71	93.61	85.66	90.66	81.80	88.46	34.40	6.99
HALC Chen et al. (2024)	89.45	88.85	87.23	86.78	85.83	84.72	90.66	93.57	85.64	90.63	81.78	88.43	**22.80**	6.80
IMCCD Li et al. (2025a)	90.73	**91.21**	87.63	87.33	85.67	**86.92**	**92.82**	94.23	86.43	91.15	82.51	**90.25**	40.80	8.87
DeCo Wang et al. (2025)	89.26	89.32	86.93	86.88	85.47	85.53	91.19	93.95	86.06	90.94	82.06	89.29	36.80	6.61
LVLMs-Saliency Zhang et al. (2026c)	89.65	89.36	86.97	86.61	85.97	85.55	90.04	93.15	85.18	90.30	81.34	88.17	37.20	8.52
HulluEdit Lin et al. (2026)	89.92	88.95	87.57	86.81	86.03	85.70	90.87	93.59	85.81	90.62	81.88	88.43	34.80	6.80
SPEAR (Ours)	**91.77** (+6.10)	**91.21** (+4.58)	**88.93** (+7.82)	**88.54** (+5.48)	**87.13** (+10.95)	**86.92** (+7.40)	92.81 (+5.30)	**95.44** (+3.38)	**87.57** (+4.04)	**92.38** (+2.46)	**83.73** (+2.87)	88.73 (+2.20)	30.80 (-9.00)	**5.52** (-6.66)

Table 3: **Cross-model results.** POPE-COCO: adversarial. Blue rows/bold values: SPEAR; green: changes from base. C_S/C_I: sentence/instance hallucination (%); OV: OneVision.

	POPE-COCO		CHAIR		MME			AMBER
Model	Acc↑	F1↑	C_S ↓	C_I ↓	Total↑	Perception↑	Cognition↑	Score↑
Qwen3-VL-2B	76.18	79.52	39.80	12.18	1333.75	1009.11	324.64	86.00
+ SPEAR	**87.13** (+10.95)	**86.92** (+7.40)	**30.80** (-9.00)	**5.52** (-6.66)	**2034.51** (+700.76)	**1526.30** (+517.19)	**508.21** (+183.57)	**92.25** (+6.25)
Qwen3-VL-4B	83.93	83.45	37.20	9.02	1555.66	1129.48	426.18	87.40
+ SPEAR	**89.00** (+5.07)	**88.50** (+5.05)	**29.40** (-7.80)	**5.42** (-3.60)	**2365.55** (+809.89)	**1715.91** (+586.43)	**649.64** (+223.46)	**93.00** (+5.60)
Ministral-3-3B	80.40	78.40	36.80	9.03	1084.17	863.33	220.84	84.40
+ SPEAR	**85.37** (+4.97)	**83.24** (+4.84)	**31.40** (-5.40)	**6.20** (-2.83)	**1426.68** (+342.51)	**1134.54** (+271.21)	**292.14** (+71.30)	**88.50** (+4.10)
Ministral-3-8B	83.47	81.41	32.80	7.99	1397.71	1030.98	366.73	86.60
+ SPEAR	**87.40** (+3.93)	**86.32** (+4.91)	**28.20** (-4.60)	**5.20** (-2.79)	**1748.46** (+350.75)	**1289.89** (+258.91)	**458.57** (+91.84)	**91.85** (+5.25)
LLaVA-OV-1.5-4B	83.63	83.00	32.00	7.77	1690.62	1209.66	480.96	87.30
+ SPEAR	**88.07** (+4.44)	**87.42** (+4.42)	**26.20** (-5.80)	**5.06** (-2.71)	**2143.86** (+453.24)	**1520.65** (+310.99)	**623.21** (+142.25)	**91.85** (+4.55)
LLaVA-OV-1.5-8B	84.33	83.79	26.20	5.99	1768.14	1277.30	490.84	89.00
+ SPEAR	**88.43** (+4.10)	**87.93** (+4.14)	**21.80** (-4.40)	**4.42** (-1.57)	**2242.81** (+474.67)	**1628.17** (+350.87)	**614.64** (+123.80)	**93.05** (+4.05)
Qwen3.5-2B	83.60	82.78	42.40	9.91	1640.72	1214.45	426.27	86.30
+ SPEAR	**87.90** (+4.30)	**87.02** (+4.24)	**33.60** (-8.80)	**6.44** (-3.47)	**2175.78** (+535.06)	**1612.92** (+398.47)	**562.86** (+136.59)	**90.80** (+4.50)

SPEAR improves all eight metrics across seven 2B–8B backbones (Table 3). Adversarial POPE accuracy gains 3.93–10.95 percentage points, with lower CHAIR rates throughout. MME improves perception and cognition; AMBER gains 4.05–6.25 points.

4.3 Multimodal Capability and Compositional Reasoning

SPEAR leads MME total and perception. Gains over greedy include existence, scene recognition, translation, and code reasoning (Figure 5(a)). Cognition gains support broader visual understanding.

SPEAR improves all ARO accuracies over greedy and leads relation/attribute discrimination, with gains in word-order judgments. It leads AMBER score and generative CHAIR and ties best discriminative accuracy/F1. Cover measures retained objects.

SPEAR improves every MMHal-Bench category over greedy and leads average, attribute, comparison, environment, and holistic scores (Figure 5(b)). Hallucination rates fall as open-ended answer quality improves. Appendix Table 12 compares both metrics.

4.4 Component Ablations

Table 5 tests suppression priority and signed evidence on two backbones. Equal-rate selection matches SPEAR's count per candidate pair and within-role $|D|$ ranking (Appendix E.2). SPEAR

Table 4: **Comparison on MME, ARO, and AMBER.** Best: **bold**; second-best: underlined.

	MME			ARO				AMBER				
				Visual Genome		COCO	Flickr	Overall	Generative		Discriminative	
Method	Total↑	Percep.↑	Cogn.↑	Rel.↑	Attr.↑	Order↑	Order↑	Score↑	CHAIR↓	Cover↑	Acc.↑	F1↑
Greedy	1333.75	1009.11	324.64	71.44	79.15	85.96	87.22	86.00	12.70	77.70	81.70	84.80
Beam search	1320.15	998.76	321.39	73.54	80.58	86.00	87.25	86.20	12.30	**78.70**	81.60	84.70
Nucleus sampling	1336.11	1011.23	324.88	72.86	78.79	85.90	87.18	85.90	12.30	76.40	81.10	84.30
VCD Leng et al. (2024)	1981.03	1455.31	525.72	80.11	80.14	69.24	75.52	90.15	8.70	73.20	85.50	89.00
OPERA Huang et al. (2024)	1992.57	1473.99	518.58	86.72	93.86	**98.55**	**98.10**	91.65	6.80	72.80	**87.00**	**90.10**
HALC Chen et al. (2024)	1991.46	1474.74	516.72	79.44	85.62	65.04	68.08	91.00	7.70	73.10	86.70	89.70
IMCCD Li et al. (2025a)	1958.03	1442.46	515.57	81.93	78.48	79.23	83.86	90.45	8.20	73.50	86.30	89.10
DeCo Wang et al. (2025)	1976.29	1467.36	508.93	77.44	81.91	63.55	69.72	91.65	6.50	74.10	86.70	89.80
LVLMs-Saliency Zhang et al. (2026c)	1955.60	1420.95	**534.65**	68.56	92.01	96.92	98.06	90.25	8.80	70.30	86.00	89.30
HulluEdit Lin et al. (2026)	1976.99	1471.64	505.35	73.56	94.10	97.15	97.50	91.65	6.80	73.60	**87.00**	**90.10**
SPEAR (Ours)	**2034.51**	**1526.30**	508.21	**88.96**	**94.13**	96.20	97.06	**92.25**	**5.60**	73.60	**87.00**	**90.10**

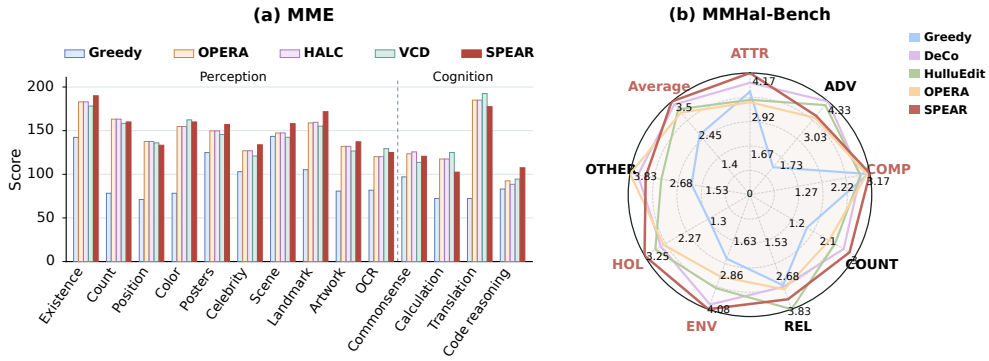


Figure 5: **MME and MMHal-Bench results on Qwen3-VL-2B.** (a) MME scores; dashes separate perception/cognition. (b) MMHal category/average scores; radial axes span zero to each all-method maximum, with raw ticks. Red: SPEAR's best scores, including ties. Panels show greedy, SPEAR, and the top three named baselines by aggregate score.

Table 5: **Component ablations.** POPE-COCO adversarial F1 and CHAIR sentence/instance rates (%). OV: OneVision. SPEAR is shaded and bold; † matches its selected-token count.

	Qwen3-VL-2B			LLaVA-OV-1.5-4B		
	CHAIR		POPE	CHAIR		POPE
Variant	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑
Greedy	39.80	12.18	79.52	32.00	7.77	83.00
Equal-rate†	32.60	7.57	85.73	27.80	5.63	87.41
w/o suppression	35.60	8.54	82.93	29.80	8.63	85.63
Positive only	34.20	7.16	84.49	28.40	6.65	86.41
SPEAR (Ours)	**30.80**	**5.52**	**86.92**	**26.20**	**5.06**	**87.42**

Table 6: **Hyperparameter sensitivity.** CHAIR-500 rates (%). Left: support/suppression retention $f/2f$; right: KL budget κ. Each scan varies one parameter; minima are shaded and bold.

Qwen3-VL-2B						
Retention scan				KL-budget scan		
Support f	Suppress. $2f$	$C_S \downarrow$	$C_I \downarrow$	κ	$C_S \downarrow$	$C_I \downarrow$
0.05	**0.10**	**30.80**	**5.52**	0.0005	34.20	7.77
0.10	0.20	33.20	6.10	0.0010	32.60	6.33
0.15	0.30	31.40	5.70	**0.0025**	**30.80**	**5.52**
0.20	0.40	34.20	7.02	0.0050	31.40	5.92
0.30	0.60	33.60	6.38	0.0100	35.20	7.70
0.40	0.80	35.20	8.79	0.0200	35.60	8.63
0.50	1.00	34.60	7.81	0.0500	33.80	6.97

leads POPE F1 and CHAIR on both models. Against equal-rate selection, instance-level hallucination falls from 7.57% to 5.52% on Qwen3-VL-2B and from 5.63% to 5.06% on LLaVA-OneVision-1.5-4B: prioritizing suppression improves the comparison at matched counts.

Removing suppression raises instance-level hallucination to 8.54% and 8.63%, consistent with weakened competitor constraints. Positive-only selection also underperforms: it retains challenger-favoring tokens and discards the reverse direction. Both directions let SPEAR weigh corrections against evidence for the original answer, supporting signed aggregation with suppression priority.

4.5 Hyperparameter Sensitivity

Table 6 reports two seven-point scans on CHAIR-500. Retention scans support fraction f with suppression $\min(2f, 1)$. Rates reach a minimum at $f = 0.05$ and fluctuate upward at larger fractions. Larger $|D|$-ranked quotas admit weaker comparisons, favoring compact selection.

Calibration rates are lowest at $\kappa = 0.0025$. Smaller budgets restrict supported challengers' probability shifts; larger budgets amplify those shifts from the same evidence. Higher rates at both ends favor moderate recalibration. Retention selects responses; the budget controls probability shifts.

4.6 Region Localization and Inference Cost

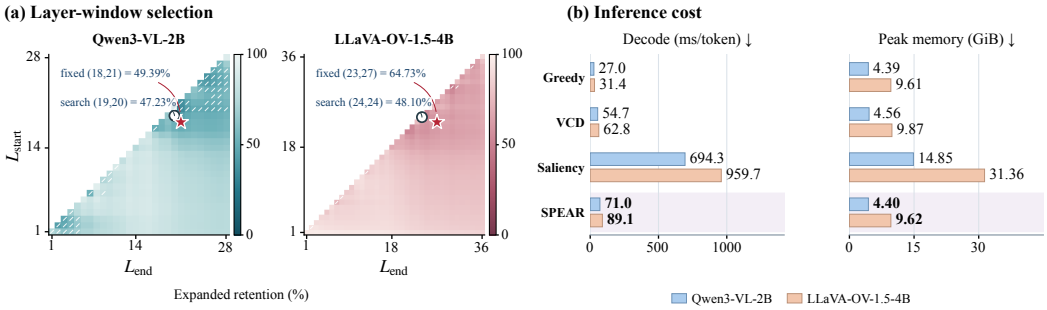


Figure 6: **Layer-window selection and inference cost.** (a) Development-set retention (%) by one-based layer windows. Slashes: below 95% reference coverage in either correctness group; stars/circles: fixed/search windows. (b) Mean decode time and peak allocated memory for 256-token outputs. Saliency: full LVLMs-Saliency; shading: SPEAR. Full metrics: Table 14.

Figure 6(a) minimizes token retention over contiguous windows with at least 95% reference coverage in each correctness group. The narrow search windows and fixed study windows both meet this criterion. SPEAR costs 2.63/2.84 times greedy on Qwen/LLaVA, about one tenth of full LVLMs-Saliency, with memory close to greedy. Similar proposal-variant costs point to shared forward/gradient computation as the main cost; neighbor expansion retains context at little extra cost. Appendix B.3 gives window choices and complete measurements.

5 Related Work

Preference learning aligns responses with factual visual feedback (Sun et al., 2024). Inference-time methods modify decoding through contrastive predictions under different visual conditions (Leng et al., 2024; Chen et al., 2024) or penalties on concentrated attention (Huang et al., 2024). Subspace editing separates visual evidence from conflicting priors in hidden states and selectively suppresses the latter (Lin et al., 2026). SPEAR instead calibrates candidate probabilities using signed token responses that distinguish direct support from stronger rival suppression.

Attention-guided visual interventions use internal model signals to enhance small visual details (Zhang et al., 2025). Adaptive reinforcement selects aligned visual patches (Zhu et al., 2026), while layer-wise interventions adjust image and text attention (Ren et al., 2026). Saliency combines attention and gradients to measure token influence and guide selection or context refocusing (Zhang et al., 2026c). SPEAR examines candidate competition within attended regions, prioritizing suppressive evidence while retaining support for both candidates.

6 Conclusion

In this paper, we propose SPEAR for training-free hallucination mitigation. SPEAR localizes visual evidence and aggregates signed responses, prioritizing rival suppression alongside direct support. KL-bounded calibration adjusts candidate probabilities while preserving candidate mass. Experiments demonstrate SPEAR's effectiveness in mitigating hallucinations.

AI USE STATEMENT

Artificial intelligence (AI) assistance contributes to manuscript drafting, LaTeX editing, mathematical derivation review, and figure composition from existing experimental records. It also supports bibliographic and formatting checks. Mechanistic measurements come from archived observations, and benchmark results come from the supplied experimental records. The authors are responsible for verifying the scientific claims and submission materials.

REPRODUCIBILITY STATEMENT

Code is available for review at `https://anonymous.4open.science/r/SPEAR-9EBA/`. The paper includes the analysis protocol (Appendix D), mathematical derivations (Appendix F), algorithm specification (Appendix E), and additional mechanistic statistics (Appendix A). Appendix B describes the evaluation metrics, baseline methods, ablations, and efficiency protocol. Table 17 lists the default token-selection, calibration, and inference parameters.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. arXiv preprint arXiv:2509.23661v3, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631v2, 2025a.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of Multimodal Large Language Models: A Survey. arXiv preprint arXiv:2404.18930, 2025b.

Minghui Chen, Chenxu Yang, Hengjie Zhu, Dayan Wu, Qingyi Si, and Zheng Lin. Online Self-Calibration Against Hallucination in Vision-Language Models. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 989–997, 2026.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7824–7846, 2024.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating Hallucinations in Large Vision-Language Models via Causal Route Gating. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-Sparsify: Attribution Guided Sparse Strategy for Visual Hallucination Mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18251–18260, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024.

Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Yubo Jiang, Yitong An, Xin Yang, Abudukelimu Wuerkaixi, Xuxin Cheng, Fengying Xie, Zhiguo Jiang, Cao Liu, Ke Zeng, and Haopeng Zhang. Breaking the Illusion: When Positive Meets Negative in Multimodal Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4210–4220, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025.

Mincheol Kwon, Minseung Lee, Seonga Choi, Miso Choi, Kyeongjin Oh, Hyunyoung Lee, Cheonyoung Park, Yongho Song, Seunghyun Park, and Jinkyu Kim. Focus, Don't Prune: Identifying Instruction-Relevant Regions for Information-Rich Image Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31900–31909, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating Hallucination for Large Vision Language Model by Inter-Modality Correlation Calibration Decoding. arXiv preprint arXiv:2501.01926v2, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023a.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, 2025b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models Via Visual Information Steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 35799–35819, 2025c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. HulluEdit: Single-Pass Evidence-Consistent Subspace Editing for Mitigating Hallucinations in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11086–11095, 2026.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. arXiv preprint arXiv:2601.08584v1, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Shunqi Mao, Chaoyi Zhang, and Weidong Cai. Through the Magnifying Glass: Adaptive Perception Magnification for Hallucination-Free VLM Decoding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44480–44501, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Hairui Ren, Zixuan Wang, Yibo Yang, He Zhao, Fan Tang, Dandan Guo, and Yi Chang. Imitating the Truth: Attention-aware Truth-Guided Enhancement for Hallucination Mitigation in Large Vision-Language Models. In *International Conference on Learning Representations*, volume 2026, pp. 43994–44020, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Tripti Shukla and Zsolt Kira. SAGE: Sink-Aware Grounded Decoding for Multimodal Hallucination Mitigation. arXiv preprint arXiv:2603.27898, 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3319–3328, 2017.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing Far and Clearly: Mitigating Hallucinations in MLLMs with Attention Causal Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26147–26159, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. arXiv preprint arXiv:2311.07397v2, 2024.

Shan Wang, Maying Shen, Nadine Chang, Chuong Nguyen, Hongdong Li, and Jose M. Alvarez. Mitigating Multimodal Hallucinations via Gradient-based Self-Reflection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32539–32549, 2026.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. REVIS: Sparse Latent Steering to Mitigate Object Hallucination in Large Vision-Language Models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yutong Xie, Zhenglin Hua, Ran Wang, Wing W. Y. Ng, Xizhao Wang, and Yuheng Jia. Finding the Correct Visual Evidence Without Forgetting: Mitigating Hallucination in LVLMs via Inter-Layer Visual Attention Discrepancy. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient GPT-4V level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509, 2025.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic Multimodal Activation Steering for Hallucination Mitigation in Large Vision-Language Models. In *International Conference on Learning Representations*, volume 2026, pp. 2092–2109, 2026.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-Grounded Dual-Path Attention Intervention for Object Hallucination Mitigation in LVLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):36021–36029, 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-Time Intervention for Mitigating Hallucination in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25293–25303, 2026a.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

RuiPeng Zhang, Zhihao Li, Haozhang Yuan, C.L.Philip Chen, and Tong Zhang. P^2-DPO: Grounding Hallucination in Perceptual Processing via Calibration Direct Preference Optimization. In *International Conference on Learning Representations*, volume 2026, pp. 56520–56546, 2026b.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. Hallucination Begins Where Saliency Drops. In *International Conference on Learning Representations*, volume 2026, pp. 15062–15083, 2026c.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look Carefully: Adaptive Visual Reinforcements in Multimodal Large Language Models for Hallucination Mitigation. In *International Conference on Learning Representations*, volume 2026, pp. 81235–81256, 2026.

A Additional Mechanistic Results

A.1 Attention Localization and Proposal Coverage

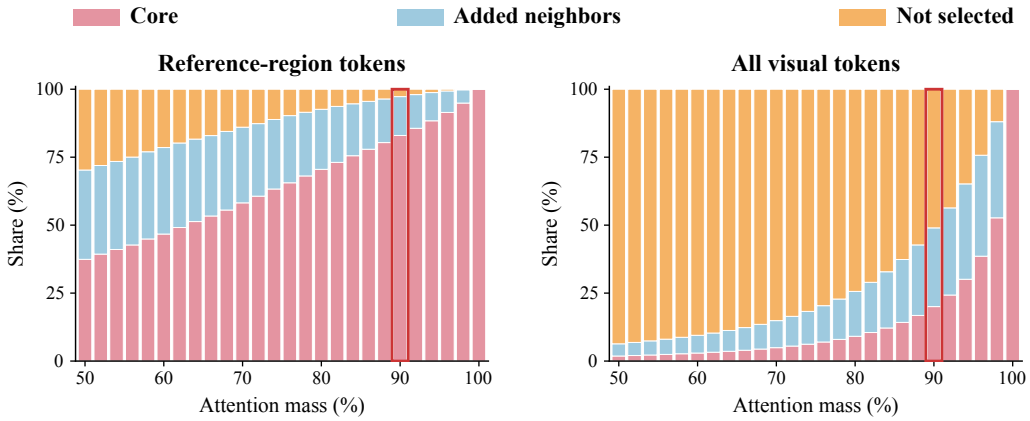


Figure 7: **Attention narrows response computation.** Coverage/retention for 908 Qwen3-VL-2B events. Stacks: core, added neighbors, unselected tokens; 2% attention-mass steps; red: 90%.

Figure 2(a) shows regional attention throughout the decoder. Ratios exceed one across broad depth ranges in both correctness groups; architecture-specific profiles motivate different layer windows. Among incorrect events, regional attention exceeds matched background attention in 97.73%, 97.66%, and 98.80% of cases for Qwen, Ministral, and LLaVA, with geometric mean ratios of 8.31, 5.15, and 11.95, respectively.

Figure 8 presents two correct and two incorrect answers per model. The Qwen clock example from Figure 2(b) predicts *"1:25"* for a clock annotated as *"2:25"*, while the coin example correctly reads *"2"*. Both place their maximum attention inside the reference region. The gallery shows attention and expanded proposals covering relevant content in both groups.

At 90% attention mass, the full 908-event core recalls 82.98% of reference tokens and retains 20.04% of visual tokens. Expansion recalls 97.29% with a 95% event-bootstrap interval of [96.81, 97.74]%, and retains 49.00% with interval [47.97, 50.05]%. The expanded correct/incorrect recall is 97.36%/96.60%, and retention is 48.91%/49.85%. Full reference coverage occurs in 76.76% of events. The proposal retains reference-region tokens and neighboring visual context. Figure 7 compares reference-token coverage and full-image retention across thresholds of 50%–100%.

Figures 9 and 10 provide paired head maps and full-cohort coverage curves for the three models. The descriptive windows average all query heads over layers 18 to 21 for Qwen, 17 to 20 for Ministral, and 23 to 27 for LLaVA, matching the middle-to-late ranges in Figure 2(a). Head maps report mean $\log_2$ regional/background attention-density ratios.

At 90% mass, expanded correct/incorrect recall is 98.47%/97.87% in Ministral and 99.08%/98.94% in LLaVA. The retained visual-token fractions are 75.14%/72.40% and 64.68%/62.97%, respectively. Both models achieve high reference coverage with larger proposals than Qwen.



Figure 8: **Localized attention in correct and incorrect answers.** Two correct cases precede two incorrect cases per model. Red shows image-wise normalized attention within proposals; gold boxes locate complete reference regions. Answers: reference → output; OV: OneVision.

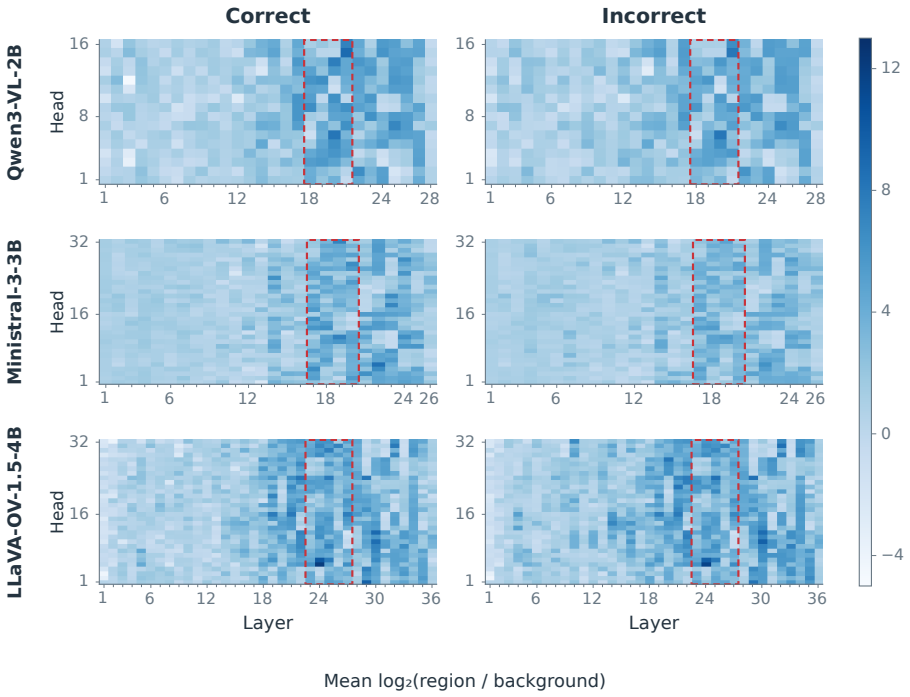


Figure 9: Head-level regional attention in correct (left) and incorrect (right) events. Rows show Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision (OV)-1.5-4B. The shared vertical scale gives mean $\log_2$ regional/background density ratios; all observed values lie within its range. Red dashed frames mark layers 18 to 21, 17 to 20, and 23 to 27, respectively.

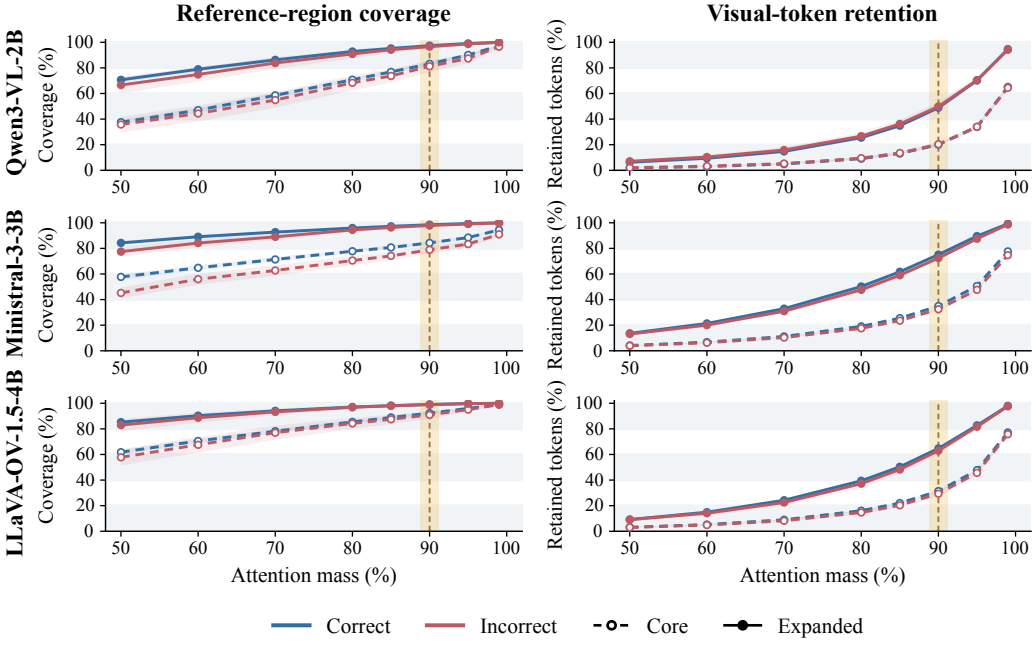


Figure 10: Full-cohort coverage and retention by correctness group. Rows show Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision (OV)-1.5-4B. Blue/rose curves show correct/incorrect events; dashed/solid curves show core/expanded proposals. Bands give event means plus or minus 1.96 standard errors. The shaded vertical band marks 90% attention mass.

A.2 Candidate Budget and Reference-token Coverage

Figure 11 measures whether the natural top-K set contains at least one token compatible with the reference continuation. We evaluate Qwen3-VL-2B, Ministral-3-3B, LLaVA-OneVision-1.5-4B, and Qwen3.5-2B on TextVQA. Means and 95% image-bootstrap intervals summarize coverage across valid image-question pairs evaluated by each model.

At $K = 8$, reference-token coverage reaches 96.0%–98.3% across the four models. Larger candidate sets increase coverage and require more candidate-gradient evaluations. The top-8 budget retains broad reference-token coverage with eight candidate responses per step.

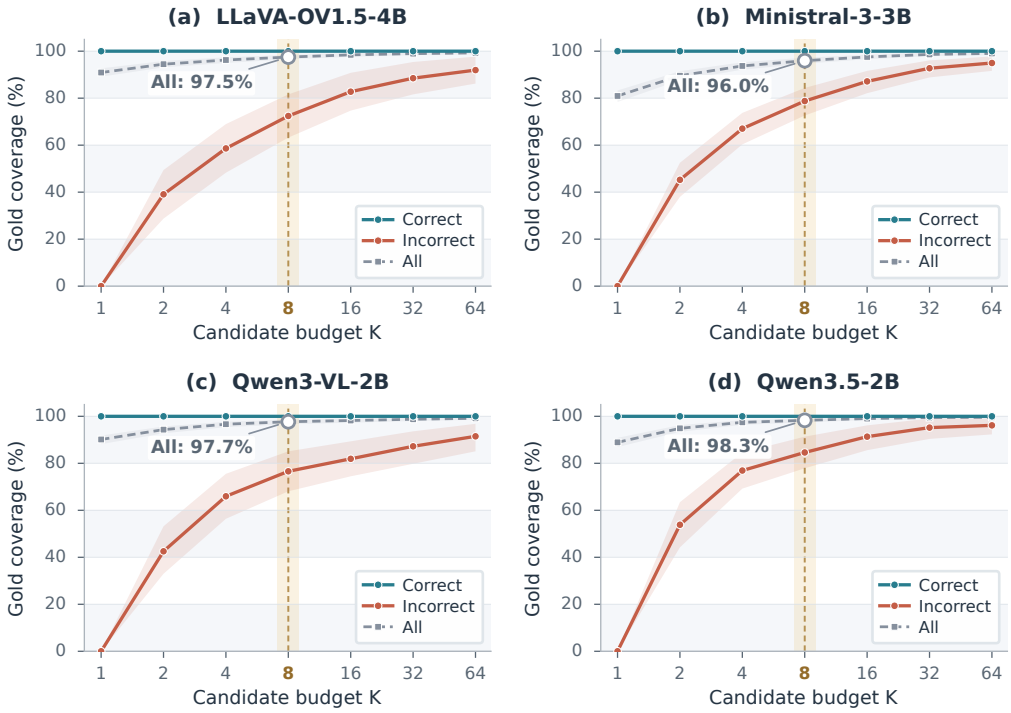


Figure 11: **Reference-token coverage versus candidate budget.** Coverage tests whether any compatible reference continuation belongs to the natural top-K set, for $K \in \{1, 2, 4, 8, 16, 32, 64\}$. Curves separate correct, incorrect, and all valid events; shaded bands give 95% image-bootstrap intervals. The highlighted $K = 8$ annotations report all-event coverage. OV denotes OneVision.

A.3 Regional Emphasis Across Targets and Strengths

We test whether emphasizing useful content within the correct region improves discrimination. At a fixed layer, let $\mu^{(\ell)}_{\mathcal{R},h} = \sum_{j\in\mathcal{R}} A^{(\ell)}_{h,j}$ be the regional attention mass and $w^{(\ell)}_{h,j} = A^{(\ell)}_{h,j}/\mu^{(\ell)}_{\mathcal{R},h}$ its conditional distribution. We redistribute attention toward a target $u^{(\ell)}_{h,j}$ with strength ϵ:

$$A'^{(\ell)}_{h,j} = A^{(\ell)}_{h,j} + \epsilon\mu^{(\ell)}_{\mathcal{R},h}(u^{(\ell)}_{h,j} - w^{(\ell)}_{h,j}), \qquad j \in \mathcal{R} \tag{8}$$

Regional mass and attention outside $\mathcal{R}$ remain fixed. The net margin response, $J_{\text{net}} = J_{\text{target}} - J_{\text{base}}$, subtracts the original mixture response from the target response, both weighted by each head's regional mass (Proposition F.3).

Figure 12 compares all three backbones using the same target rules and depths.

Figure 12 applies $\operatorname{asinh}(x/0.005)$ to both scatter coordinates and retains raw-value ticks. Figure 13 measures full-vocabulary softmax probability for the reference token. Directions are averaged within each event before computing group means and normal 95% intervals.

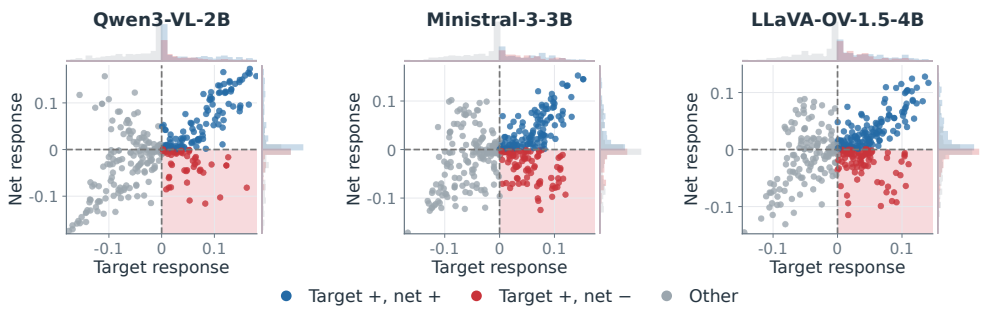


Figure 12: **Visual emphasis across three models.** Panels show Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision-1.5-4B. Each panel shows target versus net margin response for 480 directions, with inverse hyperbolic sine displays and raw-value ticks. Red points identify positive target responses with negative net responses.

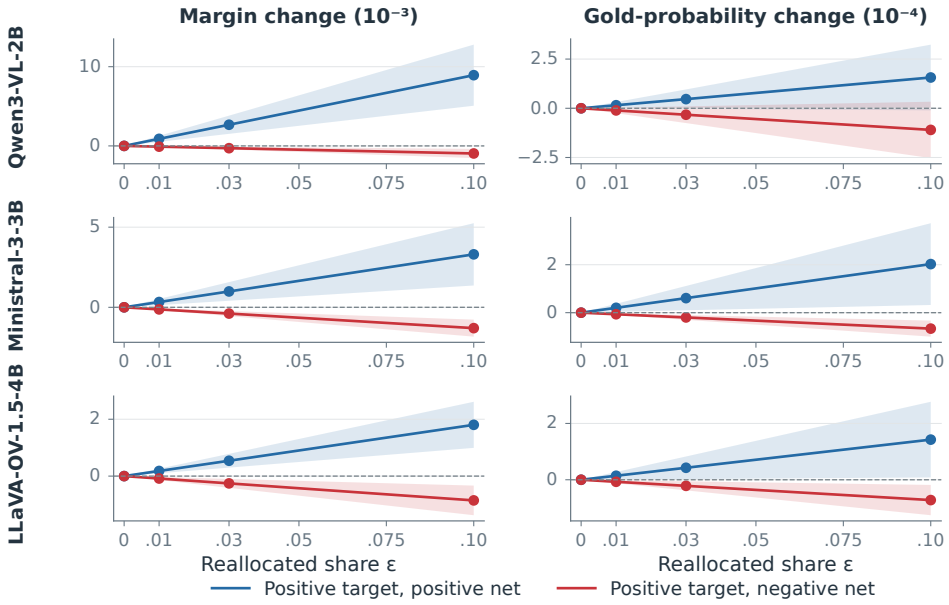


Figure 13: Measured finite effects of regional emphasis. Rows show Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision (OV)-1.5-4B. Columns report changes in the reference/competitor margin and full-vocabulary reference-token probability. Baseline derivatives fix the two positive-target groups. Curves average directions within events, then events, with 95% intervals; zero strength is the identity.

Table 7 reports negative-net prevalence among positive-target directions. The effect occurs under high-attention, random, and uniform targets in both correctness groups.

Table 8 reports actual changes at all three strengths. Subgroups use the original derivative signs, and each event has equal weight after averaging its directions. Negative mean margins support the emphasis interpretation in Section 2.3.

A.4 Candidate Roles, Coexistence, and Sensitivity

Table 9 gives the complete four-role proportions for Figure 3. Direct support remains frequent in both groups, whereas inhibitory support tokens (role II) occupy a smaller token fraction in incorrect events across all three models. The coexistence rates establish competing effects within individual events.

Table 7: Negative-net directions among positive-target directions, as numerator/denominator. The target rules use the top 20% of regional attention, an equally sized random subset, or the full uniform region. OV denotes OneVision.

Model	Group	High attention	Random	Uniform
Qwen3-VL-2B	Correct	15/45	13/46	20/38
Qwen3-VL-2B	Incorrect	15/42	13/43	13/43
Ministral-3-3B	Correct	14/41	20/43	28/41
Ministral-3-3B	Incorrect	15/44	17/41	16/37
LLaVA-OV-1.5-4B	Correct	15/52	17/51	29/48
LLaVA-OV-1.5-4B	Incorrect	17/40	21/45	22/46

Table 8: Actual finite margin change for positive-target, negative-net directions at all strengths. Values and 95% event-bootstrap intervals are in units of 10^{-3}. Counts distinguish contributing events from directions; OV denotes OneVision.

Model	Group	ϵ	Events	Dirs.	Mean Δ	95% interval
Qwen3-VL-2B	Correct	0.01	16	48	-0.075	[-0.173, -0.016]
		0.03	16	48	-0.229	[-0.522, -0.054]
		0.10	16	48	-0.771	[-1.755, -0.184]
Qwen3-VL-2B	Incorrect	0.01	14	41	-0.117	[-0.195, -0.048]
		0.03	14	41	-0.354	[-0.586, -0.149]
		0.10	14	41	-1.165	[-1.941, -0.487]
Ministral-3-3B	Correct	0.01	19	62	-0.134	[-0.237, -0.066]
		0.03	19	62	-0.402	[-0.712, -0.195]
		0.10	19	62	-1.340	[-2.375, -0.646]
Ministral-3-3B	Incorrect	0.01	18	48	-0.125	[-0.182, -0.069]
		0.03	18	48	-0.376	[-0.548, -0.211]
		0.10	18	48	-1.255	[-1.830, -0.703]
LLaVA-OV-1.5-4B	Correct	0.01	19	61	-0.074	[-0.122, -0.038]
		0.03	19	61	-0.224	[-0.366, -0.115]
		0.10	19	61	-0.745	[-1.219, -0.385]
LLaVA-OV-1.5-4B	Incorrect	0.01	15	60	-0.099	[-0.218, -0.026]
		0.03	15	60	-0.297	[-0.657, -0.080]
		0.10	15	60	-0.993	[-2.191, -0.271]

Table 9: Four diagnostic roles within the reference region (percent) and mean role-II token count. Each event has equal weight. Coexistence requires favorable and unfavorable tokens each to occupy at least 5% of the region. OV denotes OneVision.

Model	Group	I	II	III	IV	Coexist.	Mean II
Qwen3-VL-2B	Correct	57.14	13.19	6.23	23.45	69.15	4.30
Qwen3-VL-2B	Incorrect	59.60	0.62	1.91	37.87	69.32	0.40
Ministral-3-3B	Correct	43.75	13.76	8.04	34.45	78.97	5.17
Ministral-3-3B	Incorrect	47.63	4.17	4.57	43.62	80.12	1.33
LLaVA-OV-1.5-4B	Correct	41.27	13.69	9.69	35.35	74.67	5.75
LLaVA-OV-1.5-4B	Incorrect	56.37	3.95	2.69	36.99	74.70	1.70

For each event, raw head responses are summed and then averaged over all layers before assigning roles with threshold 10^{-7}. The display divides both coordinates by $\max\{\sqrt{\text{mean}_{j\in\mathcal{R}}[\Gamma_j(g)^2 + \Delta\Gamma_j(g,r)^2]}, 10^{-12}\}$ and applies $\text{asinh}(x/0.2)$. The scatter retains one point per regional token; each event contributes unit total weight to each marginal histogram. Tokens with $|\Delta\Gamma| \leq 10^{-7}$ form a neutral category.

Table 10 excludes candidate pairs that differ only in whitespace or letter case and also reports a larger-threshold check. Within-region coexistence remains high after alias exclusion, and the larger threshold preserves the combined favorable proportions. Both support the competition in Figure 3.

Table 10: Surface-alias and threshold checks. Removed counts exclude candidate pairs equal after trimming whitespace and case folding. Favorable and coexistence percentages use this subset; the last column uses all events with the larger 10^{-5} threshold. OV denotes OneVision.

Model	Group	Removed	Favorable	Coexist.	10^{-5} fav.
Qwen3-VL-2B	Correct	23	69.72	69.76	70.32
Qwen3-VL-2B	Incorrect	1	60.72	68.97	60.22
Ministral-3-3B	Correct	20	57.14	78.80	57.51
Ministral-3-3B	Incorrect	8	52.14	79.14	51.80
LLaVA-OV-1.5-4B	Correct	98	55.75	73.35	54.96
LLaVA-OV-1.5-4B	Incorrect	3	59.99	73.75	60.32

A.5 Spatial Attention and Candidate-response Comparisons

Figures 14 and 15 compare four Qwen examples using the original image, question attention, answer attention, and two signed candidate-response maps. Question and answer attention are each normalized over visual tokens and share a square-root color scale. The source labels D and q denote unit-read and attention-weighted value-scaling margin responses; the latter differs from SPEAR's conditional probability q_c. Responses sum heads and average layers. Each response map is divided by its maximum absolute value; each pair shares the symmetric range $[-1, 1]$. Their spatial patterns show how attended content affects the reference candidate relative to its rival.

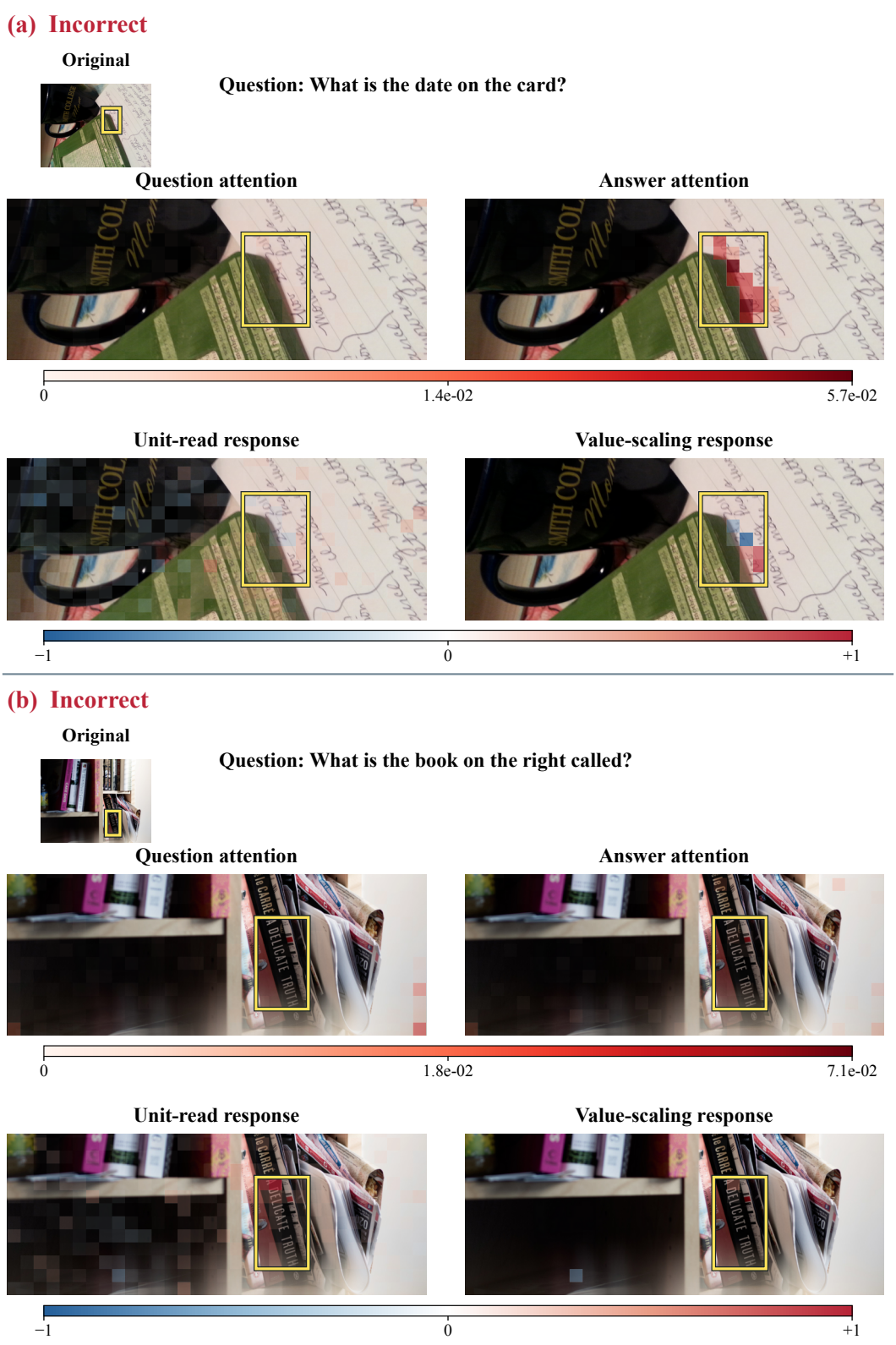


Figure 14: **Spatial evidence in two incorrect answers.** Top: visual-token-normalized attention. Bottom: signed responses divided by each map's maximum absolute value. Each row shares a scale. Gold boxes locate reference regions; red/blue responses favor the reference/rival.

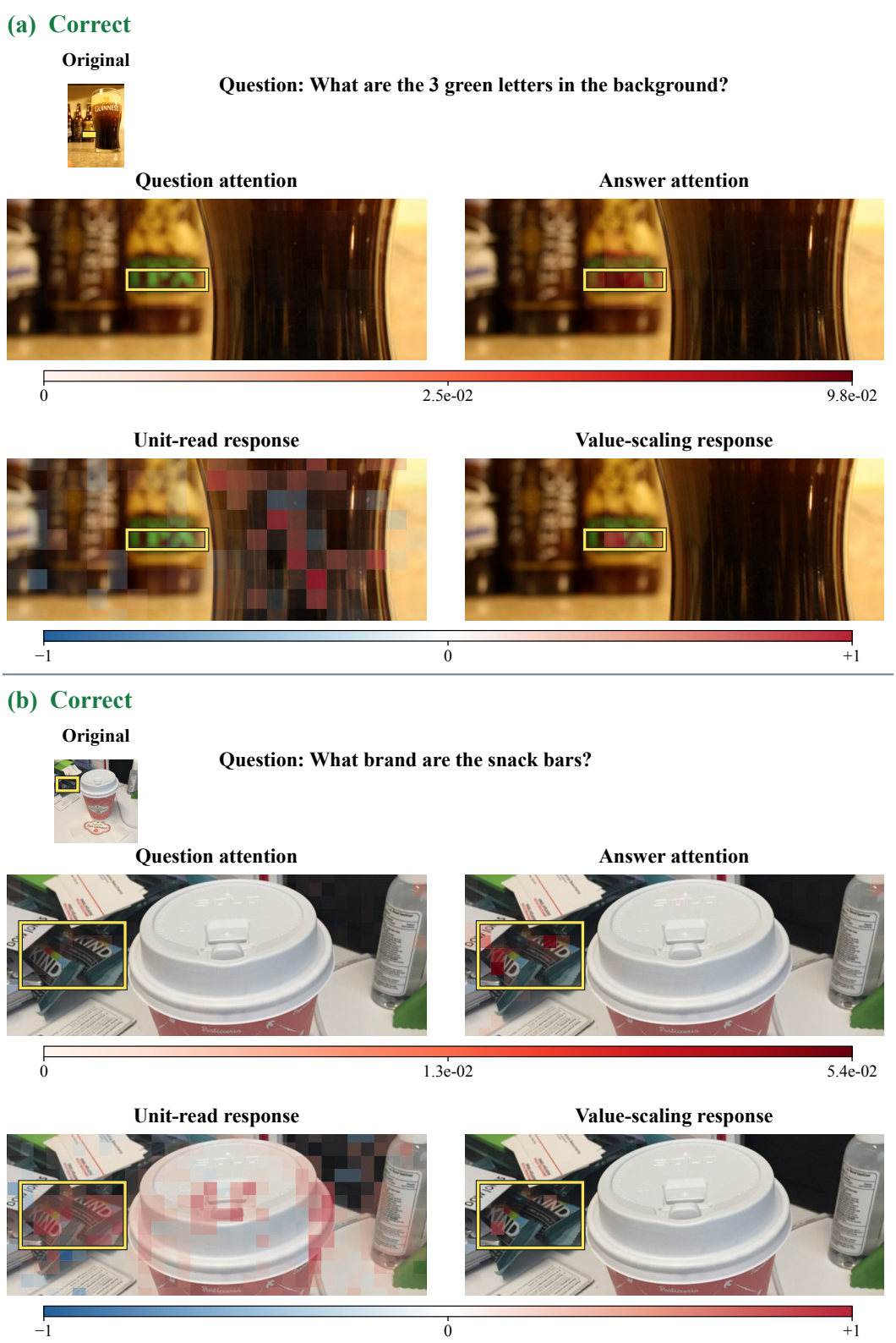


Figure 15: **Spatial evidence in two correct answers.** Layout and color conventions follow Figure 14. Each row shares a color scale. Response maps are normalized by their maximum absolute values. Gold boxes locate reference regions; headers give the questions.

B Evaluation Details and Full Results

The transfer evaluation compares Qwen3-VL-2B/4B, Ministral-3-3B/8B, LLaVA-OneVision-1.5-4B/8B, and Qwen3.5-2B on POPE-COCO, CHAIR, MME, and AMBER. Shared selection and calibration parameters stay fixed, with layer windows mapped to each architecture. Table 3 tests model size within families and architecture transfer across families. We report adversarial POPE-COCO accuracy/F1, two CHAIR rates, three MME scores, and AMBER score for each base/SPEAR pair.

B.1 Datasets and Metric Definitions

Table 11: **Benchmark sample counts.** Image and instance counts used in evaluation.

Benchmark	# Images	# Instances
POPE-COCO	500	9,000
POPE-GQA	500	23,025
CHAIR-500	500	500
MMHal-Bench	96	96
AMBER	1,004	15,220
ARO	12,603	82,695
MME	1,187	2,374

Table 11 lists the image and instance counts for each benchmark. Across its three subsets, POPE-COCO contains 4,500 positive and 4,500 negative questions. POPE-GQA contains 18,525 positive and 4,500 negative questions across the same three settings.

Table 4 compares visual capabilities: MME's ten perception and four cognition tasks, ARO's relation/attribute and word-order tests, and AMBER's open and discriminative judgments. AMBER averages $100 - \mathrm{CHAIR}$ and discriminative F1; Cover measures object coverage (Wang et al., 2024). ARO and AMBER use percentage scales.

POPE measures yes/no object existence on COCO and GQA images, with random, popular, and adversarial negative-object subsets. Its primary comparison uses accuracy and F1 for each negative-object subset; the complete record also retains precision, recall, and yes-response ratio. CHAIR measures object hallucination in generated captions: the sentence-level rate is the fraction of captions with at least one hallucinated object, and the instance-level rate is the fraction of mentioned object instances that are hallucinated. CHAIR evaluation uses a maximum of 256 new tokens.

AMBER separates generative hallucination, hallucinated-answer frequency, and object coverage from discriminative accuracy and F1 across existence, attributes, and relations. Its overall score is the arithmetic mean of $100 - \mathrm{CHAIR}$ and discriminative F1, both on a percentage scale. MME reports total, perception, and cognition scores separately. ARO tests Visual Genome attribution/relation and COCO/Flickr30k caption word-order discrimination. MMHal-Bench uses GPT-5.6 Terra through the OpenRouter API to score answers and hallucination rates with a shared prompt.

Table 12 gives all reported MMHal-Bench aggregate results, supplementing the five-method category visualization. Method selection for each panel uses the corresponding aggregate score. Radar spokes divide scores by the per-metric maximum across all reported methods, preserving a common zero origin. Red axis labels identify SPEAR's best results, including ties.

B.2 Baseline Methods

We compare against ten baselines: greedy decoding, beam search, nucleus sampling, visual contrastive decoding (VCD) (Leng et al., 2024), over-trust penalty and retrospection-allocation (OPERA) (Huang et al., 2024), object hallucination reduction via adaptive focal-contrast decoding (HALC) (Chen et al., 2024), inter-modality correlation calibration decoding (IMCCD) (Li et al., 2025a), dynamic correction decoding (DeCo) (Wang et al., 2025), LVLMs-Saliency (Zhang et al., 2026c), and HulluEdit (Lin et al., 2026).

VCD contrasts the image with a perturbed visual condition; OPERA combines over-trust penalty with retrospection-allocation; HALC uses adaptive focal-contrast decoding. DeCo uses

Table 12: MMHal-Bench aggregate results on Qwen3-VL-2B. The average answer score is on the original 0-to-6 scale; hallucination rate is a percentage. Colors and ranking follow Table 2. Parentheses give changes from greedy in each displayed unit.

MMHal-Bench		
Method	Average score↑	Hallucination rate↓
Greedy	2.25	85.44
Beam search	2.07	87.50
Nucleus sampling	2.26	84.38
VCD	3.00	59.38
OPERA	3.08	54.17
HALC	2.43	**46.88**
IMCCD	2.93	60.42
DeCo	<u>3.39</u>	<u>51.04</u>
LVLMs-Saliency	2.52	67.71
HulluEdit	3.21	56.25
SPEAR (Ours)	**3.50** (↑1.25)	52.08 (↓33.36)

Table 13: **Layer-window recheck on TextVQA.** Coverage is the percentage of reference-region tokens retained; retention is the percentage of all visual tokens selected. Both use the expanded proposal and event means. Layer indices are one-based.

Model	Window	Layers	Coverage (%)	Retention (%)
Qwen3-VL-2B	Search choice	19–20	96.33	47.43
	Fixed study	18–21	96.91	49.08
	All layers	1–28	98.95	76.57
LLaVA-OV-1.5-4B	Search choice	24	97.53	48.22
	Fixed study	23–27	98.87	65.31
	All layers	1–36	99.61	87.53

intermediate-layer predictions for dynamic correction. LVLMs-Saliency uses saliency-guided rejection sampling and local context refocusing. HulluEdit performs evidence-consistent subspace editing. IMCCD uses its 2025 second preprint version.

B.3 Ablations, Layer Selection, and Efficiency

The five conditions in Table 5 use the same candidate-response and calibration definitions. Equal-rate selection redistributes the full method's selected-token count across nonempty roles using the allocation in Appendix E.2. The suppression-removal control deletes both suppression roles from the full selection. Positive-evidence-only selection retains tokens with positive mean $D(a, b)$ for natural challenger a against original winner b. Both removal controls keep the original proposal-attention normalization. Table 5 reports POPE-COCO adversarial F1 and CHAIR rates for all five conditions on both models. Table 6 reports seven-point CHAIR-500 scans with suppression retention $2f$.

Layer-window selection requires at least 95% mean expanded reference-token recall in each development correctness group, then minimizes mean token retention. On the recheck split, the selected Qwen/LLaVA windows retain 47.43%/48.22% of visual tokens while covering 96.33%/97.53% of reference tokens. Table 13 compares selected, fixed-study, and full-layer windows.

Efficiency compares greedy, VCD, full LVLMs-Saliency (SGRS+LocoRE), and SPEAR on both backbones. Qwen also includes full-image and core-only proposal variants. Runs use the same 100 COCO images, batch size one, BF16 precision, and three rounds with exactly 256 output tokens per request. Timing includes image encoding, prefill, candidate gradients, and recalibration.

Table 14 reports first-token latency, mean decoding time, total generation time, and peak allocated memory, including model weights. Latencies are averaged over rounds and images; memory reports the maximum per configuration. Similar memory peaks across the full-image, core-only, and expanded-proposal variants reflect their shared model and cache state.

Table 14: **Complete inference-cost measurements.** First-token latency, decoding time per output token, fixed 256-token generation time, and peak allocated memory. Lower values are better for all four metrics. SPEAR rows use the ablation highlight color.

Method	First token (ms)	Decode (ms/token)	Generation (s)	Peak memory (GiB)
Qwen3-VL-2B				
Greedy	58.022	26.955	6.932	4.391
VCD	116.376	54.674	14.058	4.563
LVLMs-Saliency	645.476	694.332	177.700	14.852
SPEAR (Ours)	**136.328**	**70.992**	**18.239**	**4.400**
Full-image proposal	135.157	71.182	18.287	4.400
Core-only proposal	132.801	68.987	17.724	4.400
LLaVA-OneVision-1.5-4B				
Greedy	82.666	31.425	8.096	9.612
VCD	163.807	62.791	16.175	9.866
LVLMs-Saliency	844.541	959.739	245.578	31.359
SPEAR (Ours)	**171.897**	**89.125**	**22.899**	**9.619**

B.4 Additional Related Literature

Attribution studies distinguish reading weights, content, and coordinate sensitivity as explanatory quantities (Sundararajan et al., 2017; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019).

Complementary approaches use online self-calibration (Chen et al., 2026), visual evidence from specialist models (Li et al., 2025b), or intermediate-layer prediction correction (Chuang et al., 2024; Wang et al., 2025). Attention intervention also addresses repetition, overconfidence, and visual attention bias (Jiang et al., 2025; Tang et al., 2025; Xie et al., 2026). Other methods identify tokens for retention (Kwon et al., 2026), steer representations at selected layers (Li et al., 2025c; Dang et al., 2026; Wu et al., 2026), or adapt intervention timing and context (Zhang et al., 2026a; Yin et al., 2026). These routes motivate evaluation of both grounding and preservation of answer content.

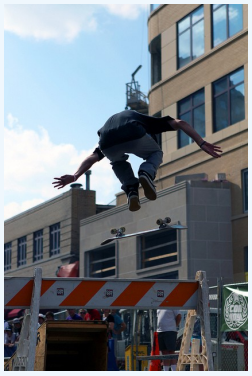
Gradient-based approaches examine candidate sensitivity and visual dependence (Wang et al., 2026; Cheng et al., 2026); spatial agreement and modality balance inform attention adjustment (Shukla & Kira, 2026; Yu et al., 2026). Positive and negative decoding paths contrast reinforced and degraded visual conditions (Jiang et al., 2026). SPEAR defines support and suppression through the signed response of individual candidates to a specified local visual direction. The signed comparison is retained even for evidence favoring the original winner. The matched-count ablation isolates the preference for suppression from an increase in the number of selected tokens.

C Captioning Case Studies

Figure 16 shows three Qwen3-VL-2B captioning cases. SPEAR describes the airborne skateboard, the swirl's visible color, and the patterned head and feet of the toy. The highlighted corrections address spatial relations and attribute assignment.

Figure 17 shows the corresponding LLaVA-OneVision-1.5-4B cases. SPEAR removes the invented umbrella and specific wine-label claim. It also drops the whiteboard-on-desk description while retaining the staircase and whiteboard.

(a)



Question

Please describe this image in detail.

Baseline (Qwen3-VL-2B)

This is a dynamic action photograph capturing a skateboarder in mid-air during a trick, set against an urban backdrop. The image is taken from a low angle, emphasizing the height and motion of the skateboarder. The central figure is a young man, dressed in a black t-shirt and grey pants, performing a jump. He is airborne, with his body extended horizontally, arms outstretched for balance. His skateboard is clearly visible, with its wheels and deck in the air, indicating he has just launched off a surface. The skateboarder's feet are positioned on the board, and his body is angled slightly forward. The setting is an outdoor urban environment. In the background, there are modern buildings with large windows and a mix of architectural styles. The sky is bright blue with scattered white clouds, suggesting it is a sunny day. The scene is framed by a temporary barrier, likely a safety or event barrier, which has orange and white diagonal stripes.

SPEAR (Ours)

This is a dynamic action photograph capturing a skateboarder in mid-air during a trick, set against an urban backdrop. The image is taken from a low angle, emphasizing the height and motion of the jump. The central subject is a young man, dressed in a black t-shirt and grey pants, who is airborne. His body is in a dynamic, athletic pose, with his arms outstretched for balance. He is wearing black shoes and is captured in the moment just after landing or during the peak of his jump. His skateboard is clearly visible, with its wheels and deck in the air, indicating the trick is in progress. The setting is an outdoor urban environment. In the background, there are modern buildings with large windows and a mix of architectural styles. The sky is bright blue with scattered white clouds, suggesting it is a sunny day. The scene is framed by a temporary obstacle, likely a barrier or a ramp, which is painted with orange and white diagonal stripes. This structure is positioned in the foreground, adding a sense of depth and context to the action.

(b)



Question

Please describe this image in detail.

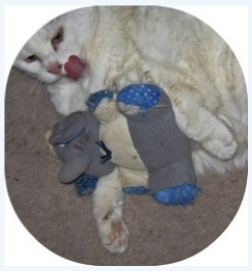
Baseline (Qwen3-VL-2B)

This is a close-up, slightly blurred photograph of a young man with dark, shoulder-length hair and glasses, captured in a moment of quiet contemplation. He is wearing a black t-shirt and is holding a white lollipop with a chocolate swirl, which he is about to take a bite of. The lollipop is on a white stick, and the chocolate swirl is visible on the top of the candy. The man is looking slightly to his right, and his expression is neutral, almost pensive. The setting appears to be a modern, minimalist interior, possibly a studio or a home office. In the background, to the left, there is a white staircase with a simple metal railing. A small easel stands near the stairs, holding a white sheet of paper with a sketch of a person's face. The walls are plain white, and the lighting is soft and diffused, creating a calm and somewhat ethereal atmosphere.

SPEAR (Ours)

This is a close-up, slightly overexposed photograph of a young man with dark, shoulder-length hair and glasses, who is in the process of eating a lollipop. He is positioned on the right side of the frame, looking towards the camera with a neutral expression. The man is wearing a black t-shirt and is holding a lollipop in his right hand. The lollipop has a white, possibly chocolate, base with a pink or red swirl on top, and it is being held up to his mouth. The background is a bright, minimalist interior space, likely a modern apartment or studio, with white walls and a staircase visible on the left. The staircase has a simple metal railing and white steps. In the background, there is a small easel with a sketch on it, which appears to be a drawing of a person in a suit. The overall lighting is bright and somewhat harsh, creating a high-contrast image with a noticeable vignette effect, where the corners are darker than the center.

(c)



Question

Please describe this image in detail.

Baseline (Qwen3-VL-2B)

This is a close-up, circular photograph of a white cat lying on a carpeted floor. The cat is the central subject, and it is positioned in a relaxed, reclining posture. Its fur is predominantly white, with some slight cream or light yellow undertones, and it has a soft, fluffy texture. The cat's eyes are a pale yellow, and its nose is pink. It is licking its own pink tongue, which is extended out of its mouth. The cat is holding a small, damaged stuffed animal, which appears to be an elephant. The elephant is made of fabric and is in a state of disrepair, with its body covered in patches of blue and white polka-dotted fabric, and its legs are a faded grey. The stuffed animal is also covered in blue fabric with white polka dots, and its feet are blue. The cat is lying on its side, with its front paws stretched out in front of it. The carpet is a light brown or beige color, and the overall lighting is soft and even.

SPEAR (Ours)

This is a close-up, circular photograph of a white cat lying on a carpeted floor. The cat is the central subject, and it is positioned in a relaxed, reclining posture. Its fur is predominantly white, with some slight cream or light yellow undertones, and it has a soft, fluffy texture. The cat's eyes are a pale yellow, and its nose is pink. It is licking its own pink tongue, which is extended out of its mouth, and its pink tongue is visible. The cat is holding a small, damaged stuffed elephant toy in its paws. The toy is made of fabric and has a grey body with blue polka-dotted fabric on its head and feet. The toy appears to be worn and torn, with visible fraying and holes. The cat is lying on a light brown or beige carpet, and the image is framed with a circular border, giving it a vignette effect. The overall mood of the image is peaceful and endearing.

Figure 16: **Captioning cases with Qwen3-VL-2B.** Top to bottom: skateboard position, swirl appearance, and toy fabric. Red marks unsupported baseline statements; green marks corresponding image-grounded SPEAR descriptions.

(a)



Question

Please describe this image in detail.

Baseline (LLaVA-OV-1.5-4B)

The image captures a scene of everyday life, set against the backdrop of a multi-story building. The building, constructed of white tiles, stands tall with its windows reflecting the world outside. In the foreground, a clothesline stretches across the frame, holding a variety of items. A pink shirt and a blue shirt hang side by side, their colors vibrant against the white tiles. A white umbrella, perhaps a symbol of protection against the elements, is also part of this ensemble. Adding a touch of whimsy to the scene is a brown teddy bear, suspended from the clothesline. The teddy bear, with its soft and cuddly appearance, seems to be enjoying its high vantage point, overlooking the world from its perch. The image is taken from a low angle, looking up at the building and the clothesline.

SPEAR (Ours)

The image captures a scene of everyday life, set against the backdrop of a multi-story building with a tiled facade. The building's exterior is adorned with a series of windows, each framed by a white railing. In the foreground, a clothesline stretches across the scene, holding a variety of items. A teddy bear, its soft fur a stark contrast to the hard tiles of the building, hangs from the line. It's a playful touch, adding a sense of whimsy to the otherwise mundane setting. Next to the teddy bear, a pink shirt and a blue jacket are also hanging out to dry. Their vibrant colors stand out against the muted tones of the building and the clothesline. The image is taken from a low angle, looking up at the building and the clothesline. This perspective gives a sense of scale and grandeur to the building, while also highlighting the small details like the teddy bear and the clothes on the line.

(b)



Question

Please describe this image in detail.

Baseline (LLaVA-OV-1.5-4B)

The image is a photograph of a dining scene featuring a large, round pizza with various toppings, including what appears to be bacon, onions, and possibly some greens, such as spinach or kale. The pizza is placed on a metal stand, which elevates it above the table. To the left of the pizza, there is a person's arm visible, holding a plate with a slice of pizza on it. The table is set with multiple wine glasses, some of which are empty, and there are two bottles of wine, one of which is a green bottle with a label that is not clearly visible, and the other is a dark bottle with a red label that reads "Château Lafite Rothschild." There is also a white plate with a piece of pizza on it, and a glass of red wine. The setting appears to be a restaurant, as indicated by the presence of other tables and chairs in the background, and the overall ambiance suggests a casual dining experience.

SPEAR (Ours)

The image is a photograph of a dining scene featuring a large, round pizza with various toppings, including what appears to be bacon, onions, and possibly some greens, such as spinach or kale. The pizza is placed on a metal stand in the center of a wooden table. To the left of the pizza, there is a white plate with a small piece of pizza on it, and to the right, there is a glass of red wine. In the background, there are additional wine bottles and glasses, suggesting a social or celebratory setting. The lighting in the image is warm, and the focus is on the pizza, making it the central element of the photograph. There are no visible texts or logos that provide additional context or information.

(c)



Question

Please describe this image in detail.

Baseline (LLaVA-OV-1.5-4B)

In the image, a person is seated in a room with a modern aesthetic. The individual, who appears to be a young adult, is wearing glasses and a black top. They are holding a lollipop with a pink wrapper, which they are about to take a bite from. The person's hair is black and short.The room has a minimalist design with white walls and a staircase with black railings on the left side. On the right, there is a white desk with a chair, and on the desk, there are a few items including what looks like a whiteboard with some writing on it. The lighting in the room is soft and natural, suggesting daytime.The image has a slightly blurred quality, giving it a candid and informal feel. There are no visible texts or discernible brands in the image. The style of the photograph is casual and seems to capture a moment of everyday life.

SPEAR (Ours)

In the image, a person is seated indoors, holding a lollipop close to their mouth. The individual appears to be in a room with a staircase in the background, and there is a whiteboard with some drawings or notes on it. The person is wearing glasses and has short hair. The room has a modern and minimalistic design, with white walls and a light-colored floor. The lighting in the room is soft and diffused, creating a calm atmosphere. There are no visible texts or discernible brands in the image. The style of the photograph is candid and informal, capturing a moment of leisure or relaxation.

Figure 17: **Captioning cases with LLaVA-OneVision-1.5-4B.** Top to bottom: clothesline objects, wine-label details, and whiteboard location. Red and green highlights follow Figure 16.

D NOTATION AND ANALYSIS PROTOCOL

We examine regional attention in Qwen3-VL-2B, Ministral-3-3B, and LLaVA-OneVision-1.5-4B (Bai et al., 2025a; Liu et al., 2026; An et al., 2025) on 1,000 TextVQA image-question pairs (Singh et al., 2019). The manually annotated answer-relevant boxes released by Zhang et al. (2025) define a reference token region $\mathcal{R}$; its source link appears in Section 2.1. We compare attention per token in the reference region and ten equal-sized background sets across layers and heads.

D.1 NOTATION AND REFERENCE QUANTITIES

Table 15 collects the quantities introduced in Sections 2 and 3. Sets use calligraphic letters, vectors use bold lowercase letters, and layer indices appear in parenthesized superscripts. A visual token denotes a spatial token after the model's image preprocessing and visual merging operations.

Table 15: Notation used in the response analysis and SPEAR. The diagnostic reference region and online proposal have different roles.

Symbol	Definition
$\mathcal{V}, \mathcal{J}$	Language vocabulary and spatial visual-token indices.
$\boldsymbol{z}, \boldsymbol{p}$	Next-token logits and full-vocabulary probabilities.
$\mathcal{C}, \mathcal{C}_{\text{probe}}$	Natural candidate set and expanded offline probe set.
π_c, m_c	Frozen candidate weight and centered score for candidate c.
$\boldsymbol{o}_h^{(\ell)}, \boldsymbol{v}_{h,j}^{(\ell)}$	Current-query head output and visual-token value.
$\Gamma, \Delta\Gamma$	Raw local candidate response and pairwise margin response.
$G_j^{(\ell)}, G_j, D_j$	Normalized layer response, mean response, and pair comparison.
$\mathcal{R}, \mathcal{U}, \mathcal{P}$	Annotated region, matched background, and online proposal.
$\mu_{\mathcal{R},h}^{(\ell)}, \boldsymbol{w}, \boldsymbol{u}$	Regional attention mass, original weights, and target weights.
$\alpha_j, \mathcal{S}_a$	Full-proposal attention weight and selected token set.
$\xi_a^{(\ell)}, \rho_a, e_a$	Layer evidence, agreement factor, and signed final evidence.
$P_{\mathcal{C}}, \boldsymbol{q}, \boldsymbol{q}'$	Candidate mass, original and updated conditional distributions.
η, κ	Calibration step and conditional divergence budget.

The diagnostic pair (g, r) uses a reference answer and a designated competitor to explain observed decisions. The inference pair (a, b) compares a natural challenger with the model's original highest-scoring candidate. The inference procedure receives neither the diagnostic answer nor its annotated region. The raw difference $\Delta\Gamma$ and normalized algorithmic score D are distinct quantities.

D.2 DATA AND MANUALLY ANNOTATED REFERENCE REGIONS

TextVQA supplies the image-question pairs and human answers (Singh et al., 2019). We use the regional extension released by Zhang et al. (2025), which contains 4,370 single-region questions from TextVQA validation. The human-annotated boxes define our reference regions; the dataset is available at `https://huggingface.co/datasets/jrzhang/TextVQA_GT_bbox`.

Each source box in (x, y, w, h) coordinates is clipped to the image and mapped to visual tokens whose spatial centers fall inside it. Empty mappings are excluded. Ten background sets $\mathcal{U}$ contain the same number of tokens as $\mathcal{R}$ and exclude known annotated boxes and border positions. Table 16 records the resulting model-specific population and exclusions.

D.3 ANSWER-LEVEL RESPONSE ANALYSIS

We measure candidate responses at the extracted answer position using the model's generated prefix. Reference and competing tokens define the diagnostic pair (g, r). Correct and incorrect groups follow the original answers, identified by automatic extraction.

Responses are computed in single precision (FP32), with the natural top-8 set fixed from BF16 generation. The center uses this candidate set, and diagnostic probes evaluate both g and r. Appendix A reports regional responses and the surface-alias check across the three models.

Table 16: Analysis population and exclusion counts from the fixed 1,000-record manifest. Missing events or regions denote unavailable valid extraction artifacts. The finite-probe subset contains 20 correct and 20 incorrect events for each model.

Quantity	Qwen	Ministral	LLaVA
Correct events	820	737	841
Incorrect events	88	171	83
Missing valid event or region	46	61	42
Insufficient matched background	26	22	23
No token center inside reference box	17	9	11
Missing verified single-precision cache	3	0	0
Finite-probe subset	40	40	40

D.4 Regional Probes and Statistical Aggregation

Finite probes use 20 correct and 20 incorrect events per model, each with at least two regional tokens. Targets assign uniform weights to the highest-attention 20% of tokens, an equally sized random subset, or the full reference region. Each event is probed at relative layer depths $0.25, 0.50, 0.75, 1.00$ and strengths $\epsilon \in \{0.01, 0.03, 0.10\}$, yielding 480 directions per model.

The target and original margin responses, at one probed layer, are

$$\begin{aligned} J_{\text{target}} &= \sum_h \mu^{(\ell)}_{\mathcal{R},h} \sum_{j\in\mathcal{R}} u^{(\ell)}_{h,j} \Delta\Gamma^{(\ell)}_{h,j}(g,r) \\ J_{\text{base}} &= \sum_h \mu^{(\ell)}_{\mathcal{R},h} \sum_{j\in\mathcal{R}} w^{(\ell)}_{h,j} \Delta\Gamma^{(\ell)}_{h,j}(g,r), \qquad J_{\text{net}} = J_{\text{target}} - J_{\text{base}} \end{aligned} \tag{9}$$

Directions with $J_{\text{target}} > 10^{-7}$ and $J_{\text{net}} < -10^{-7}$ define the subgroup in Figure 12(a). Membership uses unmodified-state derivatives. Actual change is the replayed difference in $z_g - z_r$.

Attention ratios first average heads and the ten matched backgrounds, then take the base-two logarithm of regional/background density at each layer. Averaging these logarithms over layers and events and exponentiating gives the reported geometric mean. Candidate-role statistics first sum heads with signs, average layers, and classify each regional token. Each event receives equal weight in the reported role proportions.

Figure 2 bootstraps event log ratios and exponentiates the endpoints. Figure 12(a) first averages all eligible directions within each event and then bootstraps those event means. Both figures use 4,000 bootstrap resamples. Layer and proposal curves use pointwise mean plus or minus 1.96 standard errors across events. Event-level resampling preserves dependence among token responses.

E Algorithm and Implementation Details

E.1 One Generation Step and Default Configuration

For layer window $\mathcal{L}$, H heads per layer, and visual index set $\mathcal{J}$, mean attention is $\bar{A}_j = (|\mathcal{L}|H)^{-1} \sum_{\ell\in\mathcal{L}} \sum_{h=1}^{H} A^{(\ell)}_{h,j}$. Sort $j_1, \ldots, j_N$ by descending $\bar{A}_j$, with $N = |\mathcal{J}|$. The core is the shortest prefix $\mathcal{P}_0 = \{j_1, \ldots, j_{n_0}\}$ satisfying

$$n_0 = \min\left\{ n : \sum_{i=1}^{n} \bar{A}_{j_i} \geq \tau_A \sum_{j\in\mathcal{J}} \bar{A}_j \right\}, \qquad \mathcal{P} = \mathcal{P}_0 \cup \mathcal{N}_4(\mathcal{P}_0) \tag{10}$$

Here $\tau_A = 0.90$ thresholds cumulative visual attention, and $\mathcal{N}_4$ adds one step of four-neighbor expansion. Fixed spatial identifiers break attention ties.

Algorithm 1 specifies a next-token decision at a fixed prefix. Table 17 records the default design. The candidate gradients include the output projection and downstream computation. The layer window controls which head outputs receive probes, while the model's forward computation remains complete.

Algorithm 1 SPEAR at one generation step

Require: Image $\boldsymbol{I}$, question $\boldsymbol{x}$, generated prefix $\widehat{\boldsymbol{y}}_{<t}$, fixed model θ
Ensure: Next token $\widehat{y}_t$ and updated generated prefix

```
Run the original forward pass and retain logits, attention, head outputs, and values
Construct natural C = TopK(z) and b = arg max_{c∈C} z_c, with K = 8
Form P from cumulative attention and spatial expansion (Equation 10)
if P is empty or its attention mass is zero then
    Decode from the original distribution and return
end if
Freeze π_c and compute Γ^(ℓ)_{h,j}(c) for j ∈ P
Compute signed head sums, common layer scales, G, and D (Equation 5)
Set e_b ← 0
for each a ∈ C \ {b} do
    Classify the four mirrored support/suppression roles with threshold δ
    Select the top ⌈f_i n_i⌉ in each role by |D|, with (f_1, f_2, f_3, f_4) = (0.05, 0.10, 0.05, 0.10)
    Form S_a and compute signed e_a (Equation 6)
end for
if max_{a≠b} e_a > 0 then
    Set η ← η_max and form q_η
    if D_KL(q_η ‖ q) > κ then
        Bisect [0, η_max] and retain the largest tested feasible step
    end if
    Construct p′ using Equation 7
else
    Set p′ ← p
end if
Decode ŷ_t with the original decoding rule, release the graph, and extend the prefix
```

Table 17: **SPEAR hyperparameters and inference configuration.** Qwen3-VL-2B-Instruct defaults used in the reported evaluations. Retention is applied separately to each role.

Component	Symbol	Value or configuration
Probed layers	$\mathcal{L}$	18–21, one-based; all 16 heads.
Proposal mass	τ_A	0.90, then one four-neighbor step.
Candidates	K	8, selected from current logits.
Support retention	f	0.05 for each support role.
Suppression retention	$\min(2f, 1)$	0.10 for each suppression role.
Response threshold	δ	10^{-7}.
Scale floor	ϵ_s	10^{-7}.
Agreement stabilizer	ϵ_ρ	10^{-12}.
Maximum update	$\eta_{\max}$	1.
Conditional KL budget	κ	0.0025.
CHAIR generation limit		256 new tokens.
Model parameters	θ	Frozen; no training or parameter update.

For the default rates, exact integer quotas are $\lfloor (n_i + 19)/20 \rfloor$ for support and $\lfloor (n_i + 9)/10 \rfloor$ for suppression. Tokens with near-zero pair differences are excluded, empty roles have zero quota, and small nonempty roles retain at least one token. No global quota truncates the union, and no extra factor multiplies the selected suppression values.

Diagnostic analysis uses a shared 20% retention rate for each role. The archived implementation computes full-image responses before filtering. SPEAR assigns 5% to support roles and 10% to suppression roles, with proposal-first contractions to reduce patchwise computation.

E.2 Equal-rate Control with Matched Selection Count

For each candidate pair, let $N_a = |\mathcal{S}_a|$ be the total selected by SPEAR, let n_i be the four role capacities, and set $l_i = 1[n_i > 0]$. The control chooses continuous quotas $k_i^* = \min(n_i, \max(l_i, t_0 n_i))$ with $\sum_i k_i^* = N_a$, where t_0 is a shared retention fraction. Such quotas exist because $\sum_i l_i \le N_a \le$

$\sum_i n_i$. Exact rational arithmetic fixes any proportional quota below one to one and recomputes t_0 from the remaining count and capacity until all lower bounds hold.

Take $\lfloor k_i^* \rfloor$ and assign the remaining units to the largest fractional remainders among nonfull groups. A tied remainder is resolved by the identifier of the physical group's next unselected token in its $|D|$ ordering. The control preserves the exact total count, capacity bounds, one token per nonempty group, and exchange symmetry. It changes allocation across roles while retaining within-role ranking, proposal weights, and downstream calibration.

E.3 Model Adapters, Numerical Handling, and Cost

The Qwen layer rule uses original attention averaged over layers 18 to 21 and all heads. Spatial expansion operates on the patch-coordinate grid and excludes padding and separators. Other architectures require explicit mappings for spatial tokens, query heads, and value heads. Grouped-query attention repeats each value head only for its associated query heads. Qwen3.5 patchwise analysis uses its full-attention layers. Table 3 compares the evaluated backbones and model sizes.

Zero proposal mass, an empty candidate comparison, or absence of positive challenger evidence returns the original distribution. Calibration uses log-sum-exp arithmetic, and nonfinite responses invalidate a run. End-of-sequence and length termination follow the original decoder. Bisection returns the feasible lower endpoint, keeping the update within the KL budget.

For N visual tokens, restricting directional inner products to $|\mathcal{P}|$ positions reduces their spatial dimension from N to $|\mathcal{P}|$. The candidate gradient computations still traverse the original network. End-to-end timing includes both gradient computation and the restricted patchwise contractions. Each generation step computes responses at its current prefix.

Recentering already centered responses with the same frozen weights is an identity (Proposition F.2). A different center preserves raw pair differences but changes individual signs and possibly the normalization scale. SPEAR fixes the candidate set and center weights at the current decision state.

F Mathematical Properties of the Response Scores and Calibration

This appendix establishes the local meaning of the response score, the constraints imposed by regional attention redistribution, and exact properties of the output update. The analysis fixes the current model state and characterizes candidate responses, evidence selection, and probability changes under the stated operations.

F.1 Notation and assumptions

We fix an image, a prompt, a generated prefix, and its unmodified forward pass. Let $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{V}|}$ be the next-token logits and $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$. The candidate set $\mathcal{C} \subseteq \mathcal{V}$ is selected from this pass and is held fixed during each local probe. In the current configuration, $\mathcal{C}$ is the natural top-8 set. Define

$$\pi_c = \frac{\exp z_c}{\sum_{k \in \mathcal{C}} \exp z_k}, \qquad m_c = z_c - \sum_{k \in \mathcal{C}} \pi_k z_k, \qquad c \in \mathcal{C} \tag{11}$$

where $\boldsymbol{\pi}$ is evaluated at the unmodified pass and treated as a constant when differentiating m_c. This convention also freezes the candidate selection operation; no derivative through top-K is taken.

For a layer ℓ and a head h, let $\boldsymbol{o}_h^{(\ell)} \in \mathbb{R}^{d_h}$ be the attention output for the current query, and let $\boldsymbol{v}_{h,j}^{(\ell)} \in \mathbb{R}^{d_h}$ be the value vector at visual token j. The output $\boldsymbol{o}_h^{(\ell)}$ includes the contributions of all tokens visible to that head, including nonvisual tokens. The response is

$$\Gamma_{h,j}^{(\ell)}(c) = \left\langle \nabla_{\boldsymbol{o}_h^{(\ell)}} m_c, \boldsymbol{v}_{h,j}^{(\ell)} - \boldsymbol{o}_h^{(\ell)} \right\rangle \tag{12}$$

The gradient includes the output projection and the downstream network. We assume differentiability at the probed state, and hold the intervention direction and value vectors fixed at that state. Probability results below assume finite logits. Equivalently, the analysis restricts them to the positive support of a distribution without introducing probability on a zero-mass candidate.

F.2 LOCAL RESPONSE AND CANDIDATE COMPARISON

Proposition F.1 (Local derivative and conditional log probability)**.** *Fix a layer, a head, and a visual token, and suppress their indices. Consider the head-output probe*

$$\boldsymbol{o}(t) = \boldsymbol{o} + t(\boldsymbol{v}_j - \boldsymbol{o}) \tag{13}$$

Let $z_c(t)$ be the resulting downstream logits and let $q_c(t) = \exp z_c(t) / \sum_{k\in\mathcal{C}} \exp z_k(t)$ on the fixed set $\mathcal{C}$. Then, at $t = 0$,

$$\begin{aligned} \Gamma_j(c) &= \left.\frac{dm_c(t)}{dt}\right|_0 = \left.\frac{d\log q_c(t)}{dt}\right|_0, \\ \Gamma_j(a) - \Gamma_j(b) &= \left.\frac{d(z_a(t) - z_b(t))}{dt}\right|_0 = \left.\frac{d\log(p_a(t)/p_b(t))}{dt}\right|_0 \end{aligned} \tag{14}$$

Moreover, $\sum_{c\in\mathcal{C}} \pi_c \Gamma_j(c) = 0$.

Proof. Write $u_c = dz_c(t)/dt|_0$. Since $\boldsymbol{\pi}$ is fixed, the chain rule gives

$$\left.\frac{dm_c(t)}{dt}\right|_0 = u_c - \sum_{k\in\mathcal{C}} \pi_k u_k = \langle \nabla_{\boldsymbol{o}} m_c, \boldsymbol{v}_j - \boldsymbol{o}\rangle = \Gamma_j(c) \tag{15}$$

Independently, differentiation of the log-sum-exp in $\log q_c(t)$ yields

$$\left.\frac{d\log q_c(t)}{dt}\right|_0 = u_c - \sum_{k\in\mathcal{C}} q_k(0) u_k \tag{16}$$

Since $\boldsymbol{q}(0) = \boldsymbol{\pi}$, this proves the first claim. Subtracting the expressions for a and b cancels their common center and gives $u_a - u_b$. The identity $\log(p_a(t)/p_b(t)) = z_a(t) - z_b(t)$ holds for the full-vocabulary softmax, so it has the same derivative. Finally,

$$\sum_{c\in\mathcal{C}} \pi_c \Gamma_j(c) = \sum_c \pi_c u_c - \left(\sum_c \pi_c\right)\left(\sum_k \pi_k u_k\right) = 0 \tag{17}$$

□

The equality with a conditional log-probability derivative is a statement at the unmodified state. In general, $m_c(t)$ and $\log q_c(t)$ are different functions at finite t. Likewise, an individual $\Gamma_j(c)$ need not equal the derivative of the raw logit z_c or of the full-vocabulary log probability $\log p_c$. Only the pair difference in Equation 14 eliminates the choice of a common center.

Diagnostic tokens outside the natural candidate set. The archived mechanism analyses probe tokens in a larger set $\mathcal{C}_{\text{probe}} \supseteq \mathcal{C}$, including a gold token outside the natural top-K set. For any such token, define $m_c = z_c - \sum_{k\in\mathcal{C}} \pi_k z_k$ using the same center over $\mathcal{C}$. Equation 15 and the local derivative of m_c remain valid, and subtracting two probed responses still gives the logit-margin derivative, whether or not either token belongs to $\mathcal{C}$. However, $\exp z_c / \sum_{k\in\mathcal{C}} \exp z_k$ is not a candidate conditional probability when $c \notin \mathcal{C}$; the conditional log-probability identity in Proposition F.1 is restricted to $c \in \mathcal{C}$. The weighted-centering identity likewise sums over $\mathcal{C}$, not over the expanded probe set. Online calibration uses the natural set $\mathcal{C}$ only and does not receive or append a gold token.

Freezing the center is essential to the individual-score interpretation. Differentiating the weights in $\sum_k \pi_k(t) z_k(t)$ as well adds the additional term

$$\sum_k z_k(0)\pi_k'(0) = \mathrm{Cov}_{k\sim\boldsymbol{\pi}}\big(z_k(0), u_k\big) \tag{18}$$

Differentiating the center subtracts the same extra quantity from every candidate response. The pair differences remain unchanged, but the individual signs and their conditional log-probability interpretation change.

Proposition F.2 (Recentring and the two-candidate degeneracy)**.** *For a fixed probability vector $\boldsymbol{\pi}$, define $T_{\boldsymbol{\pi}}(\boldsymbol{r})_c = r_c - \sum_k \pi_k r_k$. Then $T_{\boldsymbol{\pi}}$ preserves all pair differences, satisfies $T_{\boldsymbol{\pi}}(T_{\boldsymbol{\pi}}(\boldsymbol{r})) = T_{\boldsymbol{\pi}}(\boldsymbol{r})$, and is invariant to a common offset in $\boldsymbol{r}$. If $\mathcal{C} = \{a, b\}$, any centered response satisfies*

$$\Gamma_j(a) = \pi_b \Delta\Gamma_j(a,b), \qquad \Gamma_j(b) = -\pi_a \Delta\Gamma_j(a,b) \tag{19}$$

where $\Delta\Gamma_j(a,b) = \Gamma_j(a) - \Gamma_j(b)$.

Proof. Subtracting the same scalar from every component preserves pair differences. The centered vector has $\boldsymbol{\pi}$-weighted mean zero, so a second centering changes nothing. For a scalar c_0, the center of $\boldsymbol{r} + c_0\mathbf{1}$ is the center of $\boldsymbol{r}$ plus c_0, which proves offset invariance. For two candidates, Equation 15 becomes

$$\Gamma_j(a) = u_a - (\pi_a u_a + \pi_b u_b) = \pi_b(u_a - u_b) \tag{20}$$

and similarly $\Gamma_j(b) = -\pi_a(u_a - u_b)$. Equation 14 identifies $u_a - u_b$ with $\Delta\Gamma_j(a,b)$. □

With positive candidate probabilities and a zero sign threshold, Equation 19 excludes both $\Gamma_j(a) \le 0, \Delta\Gamma_j(a,b) > 0$ and $\Gamma_j(a) > 0, \Delta\Gamma_j(a,b) < 0$. A candidate set with at least three elements avoids this forced binary relationship. A nonzero numerical threshold places small positive responses in a threshold-near category, even when the exact binary identity holds.

Recentring already centered responses is an identity when the same $\boldsymbol{\pi}$ is used. Recentring cached responses using a different $\boldsymbol{\pi}$ still preserves raw pair differences, but may change individual signs and any subsequent scale computed from those centered responses.

F.3 Mass-preserving regional redistribution

The probe in Equation 13 moves a head output toward one value vector. A regional redistribution is a different operation: it keeps the total attention assigned to a region fixed and changes its internal allocation. The following result specifies the relevant derivative.

Proposition F.3 (Target-minus-original response)**.** *For one head, let $\mathcal{R}$ be a fixed set of visual tokens with original attention mass $\mu_{\mathcal{R}} = \sum_{j\in\mathcal{R}} A_j > 0$, and let $w_j = A_j/\mu_{\mathcal{R}}$. Let $\boldsymbol{u}$ be any probability distribution on $\mathcal{R}$. Define*

$$A_j(t) = \begin{cases} \mu_{\mathcal{R}}\big((1-t)w_j + tu_j\big), & j \in \mathcal{R}, \\ A_j, & j \notin \mathcal{R} \end{cases} \qquad 0 \le t \le 1 \tag{21}$$

This path preserves nonnegativity, the region mass, and the total attention mass. Its local response is

$$\begin{aligned} \left.\frac{dm_c(t)}{dt}\right|_0 &= \mu_{\mathcal{R}} \left[\sum_{j\in\mathcal{R}} u_j \Gamma_j(c) - \sum_{j\in\mathcal{R}} w_j \Gamma_j(c)\right] \\ &= \mu_{\mathcal{R}} \sum_{j\in\mathcal{R}} u_j \widetilde{\Gamma}_j(c), \qquad \widetilde{\Gamma}_j(c) = \Gamma_j(c) - \sum_{k\in\mathcal{R}} w_k \Gamma_k(c) \end{aligned} \tag{22}$$

If the original region mass is zero, a redistribution preserving that mass is a no-op.

Proof. The regional weights are a convex combination of two distributions, so their sum is one and each weight is nonnegative. Region-exterior weights are unchanged. Holding value vectors fixed at the probed state gives

$$\left.\frac{d\boldsymbol{o}(t)}{dt}\right|_0 = \mu_{\mathcal{R}} \sum_{j\in\mathcal{R}} (u_j - w_j)\boldsymbol{v}_j \tag{23}$$

Since $\sum_{j\in\mathcal{R}}(u_j - w_j) = 0$, subtracting $\boldsymbol{o}$ from every value in this sum does not change it. Therefore the chain rule gives

$$\begin{aligned} \left.\frac{dm_c(t)}{dt}\right|_0 &= \mu_{\mathcal{R}} \sum_{j\in\mathcal{R}} (u_j - w_j)\langle \nabla_{\boldsymbol{o}} m_c, \boldsymbol{v}_j - \boldsymbol{o}\rangle \\ &= \mu_{\mathcal{R}} \sum_{j\in\mathcal{R}} (u_j - w_j)\Gamma_j(c) \end{aligned} \tag{24}$$

Expanding the difference and using $\sum_j u_j = 1$ gives both expressions in Equation 22. If $\mu_{\mathcal{R}} = 0$, nonnegative regional attention weights summing to zero must all remain zero. □

For simultaneous probes of multiple heads in a single layer, with explicit head strengths $0 \le \beta_h \le 1$, linearity of the directional derivative gives

$$\left.\frac{dm_c(t)}{dt}\right|_0 = \sum_h \beta_h \mu^{(\ell)}_{\mathcal{R},h} \sum_{j\in\mathcal{R}} (u^{(\ell)}_{h,j} - w^{(\ell)}_{h,j}) \Gamma^{(\ell)}_{h,j}(c) \tag{25}$$

Each head contribution is weighted by its original regional mass and retains its sign. A weighted average of separate layer probes summarizes those probes; interpreting it as one joint multilayer intervention additionally requires an explicit intervention on the corresponding computation graph. SPEAR uses these probes to construct evidence for output calibration.

Corollary F.4 (Positive responses do not guarantee a beneficial redistribution)**.** *Even if a target regional distribution assigns mass only to tokens with positive $\Gamma_j(c)$, the derivative in Equation 22 is negative in some cases.*

Proof. Consider two regional responses $(4, 10)$, original weights $(0.2, 0.8)$, and target weights $(0.3, 0.7)$. Both responses are positive, but the original weighted response is 8.8 and the target weighted response is 8.2. The derivative is therefore $-0.6\mu_{\mathcal{R}} < 0$. □

This algebraic example isolates the role of the original mixture: a positive target improves the score when its response exceeds that of the mixture it replaces.

Proposition F.5 (Finite-step remainder)**.** *Let φ denote a centered candidate score as a function of the intervened head-output vector, and let $\boldsymbol{d}$ be a fixed intervention direction. If φ is twice differentiable on $\{\boldsymbol{o} + t\boldsymbol{d} : 0 \le t \le \tau\}$ and its Hessian has operator norm at most M there, then for $0 \le t \le \tau$,*

$$|\varphi(\boldsymbol{o} + t\boldsymbol{d}) - \varphi(\boldsymbol{o}) - t\nabla\varphi(\boldsymbol{o})^\top \boldsymbol{d}| \le \frac{M}{2} t^2 \|\boldsymbol{d}\|_2^2 \tag{26}$$

Proof. Apply the integral form of Taylor's theorem to $g(t) = \varphi(\boldsymbol{o} + t\boldsymbol{d})$:

$$g(t) - g(0) - tg'(0) = \int_0^t (t - s) g''(s)\, ds \tag{27}$$

Since $g''(s) = \boldsymbol{d}^\top \nabla^2 \varphi(\boldsymbol{o} + s\boldsymbol{d})\boldsymbol{d}$, its absolute value is at most $M\|\boldsymbol{d}\|_2^2$. Integrating $(t - s)$ gives the stated bound. □

The remainder controls the difference between the local prediction and the finite intervention. The mechanism probes therefore report both intervention strength and measured downstream logit changes.

Conditional-probability interpretation. All gradients are evaluated at $t = 0$. For the conditional candidate probability $q_c(t) = \exp z_c(t) / \sum_{k\in\mathcal{C}} \exp z_k(t)$, differentiability yields

$$\begin{aligned} m_c(t) &= m_c(0) + t\Gamma^{(\ell)}_{h,j}(c) + o(t) \\ \left.\frac{d \log q_c(t)}{dt}\right|_{t=0} &= z_c'(0) - \sum_{k\in\mathcal{C}} \pi_k z_k'(0) = \Gamma^{(\ell)}_{h,j}(c), \qquad c \in \mathcal{C} \end{aligned} \tag{28}$$

F.4 Signed aggregation and common normalization

For a common proposal $\mathcal{P}$, define

$$\begin{aligned} X^{(\ell)}_j(c) &= \sum_h \Gamma^{(\ell)}_{h,j}(c), \\ \sigma^{(\ell)} &= \mathrm{median}_{j\in\mathcal{P}} \sqrt{\frac{1}{K} \sum_{c\in\mathcal{C}} X^{(\ell)}_j(c)^2}, \qquad s^{(\ell)} = \max(\sigma^{(\ell)}, \epsilon_s), \\ G^{(\ell)}_j(c) &= X^{(\ell)}_j(c)/s^{(\ell)}, \qquad G_j(c) = \frac{1}{|\mathcal{L}|} \sum_{\ell\in\mathcal{L}} G^{(\ell)}_j(c), \qquad D_j(a, b) = G_j(a) - G_j(b) \end{aligned} \tag{29}$$

Here $\epsilon_s > 0$ is fixed. We use $\Delta\Gamma$ for a raw response difference and D for the normalized algorithmic comparison.

Proposition F.6 (Properties of a common positive scale)**.** *A positive scale common to all candidates and tokens of a layer preserves the signs of $X_j^{(\ell)}(c)$ and of its pair differences. It also preserves the identity $\sum_c \pi_c G_j^{(\ell)}(c) = 0$. If all $X_j^{(\ell)}(c)$ in a layer are multiplied by $t^{(\ell)} > 0$, and $\sigma^{(\ell)} \geq \epsilon_s$ and $t^{(\ell)}\sigma^{(\ell)} \geq \epsilon_s$, then all normalized scores in that layer are unchanged.*

Proof. Division by a positive scalar preserves signs, and the same scalar factors out of every candidate difference and the $\boldsymbol{\pi}$-weighted sum. Root mean square and median are positively homogeneous, so the transformed statistic is $t^{(\ell)}\sigma^{(\ell)}$. Under the stated assumptions the denominator scales by the same factor $t^{(\ell)}$ as the numerator, and their ratio is unchanged. □

Scale invariance holds in the regime specified above. Near the fixed floor, its value affects the normalized magnitude. Normalization also changes the relative weighting of layers, so the aggregate sign can differ from that of the unnormalized sum. The resulting G and D serve as normalized scores for selection and calibration.

Signed head sums preserve cancellation: responses g and $-g$ sum to zero. Absolute-value aggregation yields $2|g|$, measuring magnitude and discarding cancellation.

F.5 Evidence weights and cross-layer agreement

To expose exchange symmetry, we use pair subscripts: $\mathcal{S}_{ab} = \mathcal{S}_a$, $\xi_{ab}^{(\ell)} = \xi_a^{(\ell)}$, and $e_{ab} = e_a$ when b is the fixed winner. For a candidate pair (a, b) and its selected set $\mathcal{S}_{ab} \subseteq \mathcal{P}$, let

$$\alpha_j = \frac{\bar{A}_j}{\sum_{k\in\mathcal{P}} \bar{A}_k}, \qquad \xi_{ab}^{(\ell)} = \sum_{j\in\mathcal{S}_{ab}} \alpha_j \big(G_j^{(\ell)}(a) - G_j^{(\ell)}(b)\big) \tag{30}$$

The denominator for α is the entire proposal, irrespective of the pair or the selected role. A zero-attention proposal yields no update.

Proposition F.7 (Shared-weight and agreement bounds)**.** *Suppose $\alpha_j \geq 0$ and $\sum_{j\in\mathcal{P}} \alpha_j = 1$. Set $a_{\mathcal{S}} = \sum_{j\in\mathcal{S}_{ab}} \alpha_j$. For a nonempty $\mathcal{S}_{ab}$,*

$$|\xi_{ab}^{(\ell)}| \leq a_{\mathcal{S}} \max_{j\in\mathcal{S}_{ab}} |G_j^{(\ell)}(a) - G_j^{(\ell)}(b)|, \qquad 0 \leq a_{\mathcal{S}} \leq 1 \tag{31}$$

For any finite layer evidence vector, define its mean $\bar{\xi}_{ab}$, mean absolute magnitude ν_{ab}, agreement ρ_{ab}, and net evidence e_{ab} by

$$\bar{\xi}_{ab} = \frac{1}{|\mathcal{L}|}\sum_{\ell} \xi_{ab}^{(\ell)}, \quad \nu_{ab} = \frac{1}{|\mathcal{L}|}\sum_{\ell} |\xi_{ab}^{(\ell)}|, \quad \rho_{ab} = \frac{|\bar{\xi}_{ab}|}{\nu_{ab} + \epsilon_\rho}, \quad e_{ab} = \rho_{ab}\bar{\xi}_{ab} \tag{32}$$

where $\epsilon_\rho > 0$. Then $0 \leq \rho_{ab} < 1$ and $|e_{ab}| \leq |\bar{\xi}_{ab}| \leq \nu_{ab}$. If $\bar{\xi}_{ab} \neq 0$, e_{ab} has its sign; if $\bar{\xi}_{ab} = 0$, then $e_{ab} = 0$. Negating every layer contribution leaves ρ_{ab} unchanged and negates e_{ab}.

Proof. The triangle inequality gives

$$|\xi_{ab}^{(\ell)}| \leq \sum_{j\in\mathcal{S}_{ab}} \alpha_j |G_j^{(\ell)}(a) - G_j^{(\ell)}(b)| \tag{33}$$

which is at most the maximum magnitude times $a_{\mathcal{S}}$. For an empty selected set the evidence is zero. Another application of the triangle inequality gives $|\bar{\xi}_{ab}| \leq \nu_{ab}$, hence $\rho_{ab} \leq \nu_{ab}/(\nu_{ab} + \epsilon_\rho) < 1$ when $\nu_{ab} > 0$ and $\rho_{ab} = 0$ when $\nu_{ab} = 0$. Multiplying $\bar{\xi}_{ab}$ by the nonnegative factor ρ_{ab} proves the magnitude and sign statements. Negation changes $\bar{\xi}_{ab}$ to $-\bar{\xi}_{ab}$ and leaves both $|\bar{\xi}_{ab}|$ and ν_{ab} unchanged. □

If the idealized expression omits ϵ_ρ and $\nu_{ab} > 0$, then $\rho_{ab} = 1$ exactly when all nonzero layer contributions have the same sign, by the equality condition for the scalar triangle inequality. With the implemented positive stabilizer, the ratio is instead strictly less than one. Renormalizing α on the selected set removes the factor $a_{\mathcal{S}}$ from Equation 31; a subset carrying arbitrarily small original proposal attention then receives unit total weight. Shared proposal weights preserve the attention mass carried by the selected subset.

F.6 CANDIDATE EXCHANGE AND ROLE QUOTAS

With a common positive threshold δ, define four disjoint roles:

$$\begin{aligned}
\mathcal{T}_1(a,b) &= \{j \in \mathcal{P} : D_j(a,b) > \delta,\ G_j(a) > \delta\},\\
\mathcal{T}_2(a,b) &= \{j \in \mathcal{P} : D_j(a,b) > \delta,\ G_j(a) \le \delta\},\\
\mathcal{T}_3(a,b) &= \{j \in \mathcal{P} : D_j(a,b) < -\delta,\ G_j(b) > \delta\},\\
\mathcal{T}_4(a,b) &= \{j \in \mathcal{P} : D_j(a,b) < -\delta,\ G_j(b) \le \delta\}
\end{aligned} \tag{34}$$

These represent support for a, suppression of b, support for b, and suppression of a, respectively. In $\mathcal{T}_2$, $G_j(b) = G_j(a) - D_j(a,b) < 0$; in $\mathcal{T}_4$, the analogous inequality gives $G_j(a) < 0$. The condition $G_j(a) \le \delta$ includes small positive values up to δ.

For $0 < f \le 1$, define the mirrored retention rates

$$(f_1, f_2, f_3, f_4) = (f, \min(2f, 1), f, \min(2f, 1)) \tag{35}$$

For each role with n_r tokens, select $k_r = \lceil f_r n_r \rceil$ if $n_r > 0$, and zero otherwise. Tokens are ordered by descending $|D_j(a,b)|$, with ties resolved by a fixed ascending token identifier. The selected union is $\mathcal{S}_{ab}$. SPEAR uses $f = 0.05$, giving rates $(0.05, 0.10, 0.05, 0.10)$. These rates change only selection; the shared weights α_j and signed evidence are not multiplied by the retention-rate ratio, and no global quota truncates the union. Appendix E gives the algorithm and its configuration.

Proposition F.8 (Pairwise exchange antisymmetry)**.** *Keep $\mathcal{C}$, its center, $\mathcal{P}$, G, and α fixed. Under the role-wise selection rule above, exchanging a and b yields*

$$\mathcal{S}_{ba} = \mathcal{S}_{ab}, \qquad \xi_{ba}^{(\ell)} = -\xi_{ab}^{(\ell)}, \qquad \rho_{ba} = \rho_{ab}, \qquad e_{ba} = -e_{ab} \tag{36}$$

Proof. The comparison changes sign: $D_j(b,a) = -D_j(a,b)$. Direct substitution in Equation 34 gives $\mathcal{T}_1(b,a) = \mathcal{T}_3(a,b)$ and $\mathcal{T}_2(b,a) = \mathcal{T}_4(a,b)$, with the other two identities following by exchanging the pairs. Thus each physical role group has the same size after relabeling. Since $f_1 = f_3$ and $f_2 = f_4$, it also has the same retention rate and integer quota. Its absolute comparison scores and token-identifier tie breaks are unchanged, so the same tokens are selected. For each selected token the layer comparison changes sign, while its weight is unchanged. Equation 30 therefore changes sign. Proposition F.7 gives the last two conclusions. □

This proposition concerns one fixed candidate pair. In the complete rule, the unmodified winner b is the reference and $e_a = e_{ab}$, $e_b = 0$. Different challengers may select different token sets. Pairwise antisymmetry therefore does not imply the cycle identity $e_{ab} + e_{bc} = e_{ac}$ or invariance of the entire rule to replacing its winner reference.

Proposition F.9 (Rounding with role-specific retention rates)**.** *Let M be the number of nonempty roles and let $N_{\text{valid}} = \sum_r n_r$. Define $Q = \sum_r f_r n_r$, with $0 < f_r \le 1$ for each nonempty role. If $N_{\text{valid}} > 0$, the role-wise rule satisfies*

$$Q \le |\mathcal{S}_{ab}| < Q + M \le Q + 4, \qquad M \le |\mathcal{S}_{ab}| \le N_{\text{valid}} \tag{37}$$

At the default rates, writing $N_{\text{sup}} = n_1 + n_3$ and $N_{\text{suppress}} = n_2 + n_4$ gives $Q = 0.05N_{\text{sup}} + 0.10N_{\text{suppress}}$. More generally, Proposition F.8 continues to hold for an integer quota rule if its quotas transform under $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4) \mapsto (\mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_1, \mathcal{T}_2)$ by the same permutation, and within-role ordering remains unchanged.

Proof. For each nonempty role, $f_r n_r \le \lceil f_r n_r \rceil < f_r n_r + 1$. Summing over the disjoint roles gives the first bound. Since $0 < f_r \le 1$ and n_r is a positive integer, $1 \le \lceil f_r n_r \rceil \le n_r$, which gives the lower and capacity bounds. For the generalized statement, permutation-equivariant quotas assign the same integer count to each physical group after exchange. The proof of Proposition F.8 then applies unchanged. □

Thus the default rule does not impose an exact overall retention percentage. Before rounding, its overall rate is $0.05 + 0.05N_{\text{suppress}}/N_{\text{valid}}$; small roles may retain a much larger fraction after rounding. At $f \ge 0.5$ the suppression rates saturate at one, and at $f = 1$ every valid token is retained.

Equal-rate control with a matched total count. To test whether suppression preference matters beyond selecting more tokens, let $N = |\mathcal{S}_{ab}|$ under Equation 35 and assign continuous control quotas

$$q_r^{\mathrm{ctrl}} = \min\{n_r, \max\{\mathbf{1}[n_r > 0], tn_r\}\}, \qquad \sum_r q_r^{\mathrm{ctrl}} = N \tag{38}$$

Such quotas exist because $M \le N \le N_{\mathrm{valid}}$ and the sum is continuous and nondecreasing in t, from M at $t = 0$ to N_{valid} at $t = 1$. The vector of quotas is unique even if a boundary value admits several choices of t. For integer capacities and total, it is computed using rational arithmetic: fix quotas below one to one, recompute t from the residual count divided by the remaining total capacity, and repeat. No upper saturation occurs before $t = 1$.

Take the floor of each control quota and assign the remaining units to the largest exact fractional remainders among nonfull roles. Break a tied remainder by the token identifier of the physical group's next unselected token in its $|D|$ ordering. The allocation yields total N, respects capacities, and retains at least one token per nonempty role. The continuous quotas, their remainders, and the physical tie keys move with the role permutation; the generalized statement of Proposition F.9 therefore preserves pairwise exchange antisymmetry. A fixed role-index tie break does not provide that guarantee. The control uses the same within-role ranking and downstream evidence rule, so it changes the allocation of a matched count rather than the total count itself.

F.7 Candidate-mass-preserving output calibration

Let $P_{\mathcal{C}} = \sum_{c\in\mathcal{C}} p_c > 0$ and $q_c = p_c/P_{\mathcal{C}}$. For a fixed finite evidence vector $\boldsymbol{e}$ and a nonnegative step η, define

$$Z(\eta) = \sum_{c\in\mathcal{C}} q_c \exp(\eta e_c), \qquad q_{\eta,c} = \frac{q_c \exp(\eta e_c)}{Z(\eta)}, \qquad p'_v = \begin{cases} P_{\mathcal{C}} q_{\eta,v}, & v \in \mathcal{C}, \\ p_v, & v \notin \mathcal{C} \end{cases} \tag{39}$$

The method uses $e_b = 0$ at the original winner. It returns the unmodified distribution when no challenger has positive evidence; the identities below concern an applied update, including its zero-step limit.

Proposition F.10 (Preserved mass and exact odds)**.** *The update in Equation 39 is a valid probability distribution, preserves the total candidate mass $P_{\mathcal{C}}$, and leaves each noncandidate probability unchanged. For any $a, b \in \mathcal{C}$,*

$$\log \frac{p'_a}{p'_b} - \log \frac{p_a}{p_b} = \eta(e_a - e_b) \tag{40}$$

An equivalent full-vocabulary logit update is

$$z'_v = \begin{cases} z_v + \eta e_v - \log Z(\eta), & v \in \mathcal{C}, \\ z_v, & v \notin \mathcal{C} \end{cases} \tag{41}$$

Adding a common constant to all entries of $\boldsymbol{e}$ changes neither $\boldsymbol{q}_\eta$ nor $\boldsymbol{p}'$.

Proof. All candidate weights in Equation 39 are nonnegative and sum to one after division by $Z(\eta)$. They therefore contribute exactly $P_{\mathcal{C}}$ to $\boldsymbol{p}'$, while the unchanged complement contributes $1 - P_{\mathcal{C}}$. Taking the ratio of two candidate probabilities cancels both $P_{\mathcal{C}}$ and $Z(\eta)$, proving Equation 40. Let $Z_{\mathcal{C}}^{(0)} = \sum_{c\in\mathcal{C}} \exp z_c$, so that $q_c = \exp z_c / Z_{\mathcal{C}}^{(0)}$. Then

$$\sum_{c\in\mathcal{C}} \exp z'_c = \frac{\sum_{c\in\mathcal{C}} \exp z_c \exp(\eta e_c)}{Z(\eta)} = Z_{\mathcal{C}}^{(0)} \tag{42}$$

The full-vocabulary softmax denominator is unchanged, which proves the equivalent logit expression. Finally, for any scalar c_0, replacing e_c by $e_c + c_0$ multiplies both the numerator and $Z(\eta)$ by $\exp(\eta c_0)$. □

With $e_b = 0$, a positive e_a increases the odds of a relative to b. The winner's absolute probability $p'_b = p_b/Z(\eta)$ decreases exactly when $Z(\eta) > 1$. Other signed evidence values influence this normalizer. In particular, fixing $e_b = 0$ fixes a reference score, not a probability.

F.8 KL MONOTONICITY AND THE STEP-SIZE CONSTRAINT

The implemented conditional constraint has direction $D_{\mathrm{KL}}(\boldsymbol{q}_\eta\|\boldsymbol{q})$, with the updated distribution in the first argument. All logarithms in this appendix are natural.

Proposition F.11 (Monotonic KL along the exponential update)**.** *Let $\psi(\eta) = \log Z(\eta)$ and $\mathcal{K}(\eta) = D_{\mathrm{KL}}(\boldsymbol{q}_\eta\|\boldsymbol{q})$ for a fixed evidence vector. Then*

$$\begin{aligned} \psi'(\eta) &= \mathbb{E}_{c\sim\boldsymbol{q}_\eta}[e_c], \qquad \psi''(\eta) = \mathrm{Var}_{c\sim\boldsymbol{q}_\eta}(e_c), \\ \mathcal{K}(\eta) &= \eta\psi'(\eta) - \psi(\eta), \quad \mathcal{K}'(\eta) = \eta\,\mathrm{Var}_{c\sim\boldsymbol{q}_\eta}(e_c) \geq 0 \quad (\eta \geq 0) \end{aligned} \tag{43}$$

If $\boldsymbol{e}$ is nonconstant on the positive support of $\boldsymbol{q}$, then $\mathcal{K}$ is strictly increasing for $\eta > 0$.

Proof. Differentiating the finite positive sum Z gives

$$\psi'(\eta) = \frac{\sum_c q_c e_c \exp(\eta e_c)}{Z(\eta)} = \mathbb{E}_{c\sim\boldsymbol{q}_\eta}[e_c] \tag{44}$$

A second derivative yields $\mathbb{E}_{c\sim\boldsymbol{q}_\eta}[e_c^2] - (\mathbb{E}_{c\sim\boldsymbol{q}_\eta}[e_c])^2$. From $\log(q_{\eta,c}/q_c) = \eta e_c - \psi(\eta)$,

$$\mathcal{K}(\eta) = \sum_c q_{\eta,c}\big(\eta e_c - \psi(\eta)\big) = \eta\psi'(\eta) - \psi(\eta) \tag{45}$$

Differentiating cancels the first-derivative terms and leaves $\mathcal{K}'(\eta) = \eta\psi''(\eta)$. At a finite step, the update preserves the positive support of $\boldsymbol{q}$. Hence a nonconstant $\boldsymbol{e}$ on that support has strictly positive variance, proving strict monotonicity for $\eta > 0$. □

Since $\mathcal{K}(0) = 0$ and $\mathcal{K}$ is continuous and nondecreasing, a valid step under budget κ is chosen within $[0, \eta_{\max}]$: use $\eta_{\max}$ if it satisfies the budget, and otherwise bisect this interval. The procedure does not increase η beyond its cap to exhaust the budget. Constant evidence produces no distribution change, and a zero budget permits the zero-step update. The decision to require a positive challenger is a separate algorithmic trigger, not a consequence of the KL constraint.

Proposition F.12 (Variational characterization)**.** *For $\eta > 0$, $\boldsymbol{q}_\eta$ uniquely maximizes*

$$\langle \boldsymbol{r}, \boldsymbol{e}\rangle - \frac{1}{\eta} D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{q}) \tag{46}$$

over distributions $\boldsymbol{r}$ supported on the positive support of $\boldsymbol{q}$.

Proof. The log-ratio identity used above gives

$$D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{q}_\eta) = D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{q}) - \eta\langle \boldsymbol{r}, \boldsymbol{e}\rangle + \psi(\eta) \tag{47}$$

Rearrangement expresses Equation 46 as $\psi(\eta)/\eta - D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{q}_\eta)/\eta$. Nonnegativity of KL, with equality only when the distributions agree, proves the claim. □

For the supplied evidence vector, the objective balances expected evidence against departure from the original candidate distribution.

F.9 DISTRIBUTIONAL BOUNDS AND THEIR LIMITS

We use total variation $\mathrm{TV}(\boldsymbol{r}, \boldsymbol{s}) = \frac{1}{2}\sum_i |r_i - s_i|$.

Proposition F.13 (Full-vocabulary divergence and total variation)**.** *Under Equation 39,*

$$\begin{aligned} D_{\mathrm{KL}}(\boldsymbol{p}'\|\boldsymbol{p}) &= P_{\mathcal{C}} D_{\mathrm{KL}}(\boldsymbol{q}_\eta\|\boldsymbol{q}), \\ \mathrm{TV}(\boldsymbol{p}', \boldsymbol{p}) &= P_{\mathcal{C}}\,\mathrm{TV}(\boldsymbol{q}_\eta, \boldsymbol{q}) \end{aligned} \tag{48}$$

Consequently, $D_{\mathrm{KL}}(\boldsymbol{q}_\eta\|\boldsymbol{q}) \leq \kappa$ implies

$$D_{\mathrm{KL}}(\boldsymbol{p}'\|\boldsymbol{p}) \leq P_{\mathcal{C}}\kappa, \qquad \mathrm{TV}(\boldsymbol{p}', \boldsymbol{p}) \leq P_{\mathcal{C}}\sqrt{\kappa/2} \tag{49}$$

Proof. The complement of $\mathcal{C}$ contributes zero to both the KL and the absolute probability difference. Within $\mathcal{C}$, the ratio $p'_c/p_c = q_{\eta,c}/q_c$ and the difference $p'_c - p_c = P_{\mathcal{C}}(q_{\eta,c} - q_c)$ give Equation 48.

For completeness, the needed form of Pinsker's inequality follows by aggregating an arbitrary pair $(\boldsymbol{r}, \boldsymbol{s})$ onto the event $\mathcal{A} = \{i : r_i \geq s_i\}$. Put $x = \sum_{i\in\mathcal{A}} r_i$ and $y = \sum_{i\in\mathcal{A}} s_i$, so that $x - y = \mathrm{TV}(\boldsymbol{r}, \boldsymbol{s})$. The log-sum inequality on $\mathcal{A}$ and its complement gives

$$D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{s}) \geq d_{\mathrm{bin}}(x\|y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \tag{50}$$

The log-sum inequality itself follows from Jensen's inequality for the convex function $u \log u$: on $\mathcal{A}$, use normalized weights s_i/y and arguments r_i/s_i, and multiply by y; apply the same argument to the complement. For fixed $0 < y < 1$, the binary divergence has value and first derivative zero at $x = y$, and

$$\frac{\partial^2}{\partial x^2} d_{\mathrm{bin}}(x\|y) = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)} \geq 4 \tag{51}$$

Integrating this lower bound twice gives $d_{\mathrm{bin}}(x\|y) \geq 2(x-y)^2$; endpoint cases follow by limits. Thus $\mathrm{TV}(\boldsymbol{r}, \boldsymbol{s}) \leq \sqrt{D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{s})/2}$. Apply this to $(\boldsymbol{q}_\eta, \boldsymbol{q})$ and use Equation 48. □

Proposition F.14 (Odds bounds and the threshold for changing the winner)**.** *Let* $e_{\min} = \min_{c\in\mathcal{C}} e_c$, $e_{\max} = \max_{c\in\mathcal{C}} e_c$, *and* $0 \leq \eta \leq \eta_{\max}$. *Then, for* $a, b, c \in \mathcal{C}$,

$$\begin{aligned} \left| \log \frac{p'_a}{p'_b} - \log \frac{p_a}{p_b} \right| &\leq \eta_{\max}(e_{\max} - e_{\min}), \\ |\log(p'_c/p_c)| &\leq \eta_{\max}(e_{\max} - e_{\min}) \end{aligned} \tag{52}$$

If b *is the original winner and* $e_b = 0$, *then* $p'_a > p'_b$ *exactly when*

$$\eta e_a > \log(p_b/p_a) \tag{53}$$

If $\mathcal{C}$ *is a natural top-*K *set, a noncandidate cannot strictly exceed all candidates after a mass-preserving update.*

Proof. The first bound follows from Equation 40 and $|e_a - e_b| \leq e_{\max} - e_{\min}$. Since $\boldsymbol{q}$ sums to one,

$$\exp(\eta e_{\min}) \leq Z(\eta) \leq \exp(\eta e_{\max}) \tag{54}$$

Combining these inequalities with $\log(p'_c/p_c) = \eta e_c - \log Z(\eta)$ proves the second bound. Taking the sign of the updated log ratio in Equation 40 gives the exact threshold. Finally, a natural top-K set satisfies $\min_{c\in\mathcal{C}} p_c \geq \max_{v\notin\mathcal{C}} p_v$. Candidate mass preservation implies

$$\max_{c\in\mathcal{C}} p'_c \geq P_{\mathcal{C}}/K \geq \min_{c\in\mathcal{C}} p_c \geq \max_{v\notin\mathcal{C}} p'_v \tag{55}$$

which proves the final claim; boundary ties depend on the decoder's tie rule. □

In particular, the original winner stays strictly ahead of all other candidates if, for every $a \neq b$,

$$\log(p_b/p_a) > \eta_{\max} \max(e_a, 0) \tag{56}$$

Without such a margin condition, an arbitrarily small distributional change may change a greedy token. For example, take $\boldsymbol{q} = (1/2 + \varepsilon, 1/2 - \varepsilon)$ and $\boldsymbol{r} = (1/2 - \varepsilon, 1/2 + \varepsilon)$. Their winners differ, whereas

$$D_{\mathrm{KL}}(\boldsymbol{r}\|\boldsymbol{q}) = 2\varepsilon \log \frac{1/2 + \varepsilon}{1/2 - \varepsilon} \longrightarrow 0 \qquad \text{as } \varepsilon \longrightarrow 0 \tag{57}$$

The pair also has a representation as an exponential update with the first candidate's evidence zero and positive evidence for the second.

The constraints characterize the next-token distribution at a fixed prefix. A changed token alters later prefixes; extending the bound to a sequence requires conditional policies at those prefixes and the chain rule for divergence. Benchmark evaluation measures how these local updates affect complete answers.