SAM: RETHINKING OBJECT HALLUCINATION DETECTION THROUGH VISUAL-EVIDENCE RELIABILITY AND SEMANTIC-UPDATE CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models remain prone to object hallucination, describing objects unsupported by the input image. Existing attention or semantics-based detectors mainly characterize visual support, leaving the relation between selected evidence and subsequent semantic updates less explored. We revisit object hallucination from the joint perspective of visual-evidence reliability and update consistency. Our layer-wise analysis shows that grounded objects exhibit stronger semantically calibrated visual writes, whereas hallucinated objects exhibit higher evidence-mismatch scores over broad layer ranges. *Grounded in these observations, we introduce **S**emantic-**A**ttention and evidence **M**ismatch (**SAM**)*, a dual-signal framework comprising **S**emantic-**A**ttention **R**eliability (**SAR**) and **E**vidence **M**ismatch **S**core (**EMS**). SAR recalibrates source-level attention writes with target-semantic support, while EMS compares the FFN-induced visual preference with target-conditioned attention evidence through a state-calibrated optimal-transport score. A lightweight detector combines their layer-wise trajectories without modifying the LVLM parameters or decoding procedure. Experiments across four LVLMs on three benchmarks covering MS COCO captioning and CLEVR and AMBER question answering show consistent gains over existing detectors, achieving the highest AUROC among the compared methods.

1 INTRODUCTION

Background. Large Vision-Language Models (LVLMs) (Liu et al., 2024b; Bai et al., 2025b; Chen et al., 2025) have demonstrated remarkable capabilities in multimodal content understanding, image captioning, visual question answering, and multimodal instruction following. Nevertheless, they remain prone to object hallucination (Rohrbach et al., 2018; Li et al., 2023; Liu et al., 2024a): generating semantically plausible object descriptions that lack support from the input image. Such errors reduce the factual reliability of generated descriptions and complicate deployment in applications that require image-grounded outputs. Accurately detecting object hallucinations is therefore critical to improving the reliability and safety of LVLMs.

Motivation and gap. Existing detectors estimate visual dependence from attention (Zhang et al., 2025; Jiang et al., 2025b) or visual support from internal predictions and representations (Fieback et al., 2025; Jiang et al., 2025a; Lai et al., 2026). Recent methods also combine attention with cross-modal semantics for patch-level grounding (Nguyen et al., 2026). However, visual selection and target support are distinct: attending to a source does not establish that it supports the generated object, and attention-based explanations require validation (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Brunner et al., 2020). Consistent with this distinction, Wang et al. (2026) found similar visual-attention magnitudes for grounded and hallucinated objects despite differences in target-semantic consistency. Attention indicates which visual sources are routed into the current prediction, while an internal semantic readout indicates whether those sources support the target being generated. We therefore pair each source-level attention write with its target-semantic support to characterize object-specific visual evidence. The attention output is then processed by an FFN. This raises a further question: does the subsequent semantic update remain compatible with the visual evidence formed by attention? This motivates examining the two stages jointly.

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.

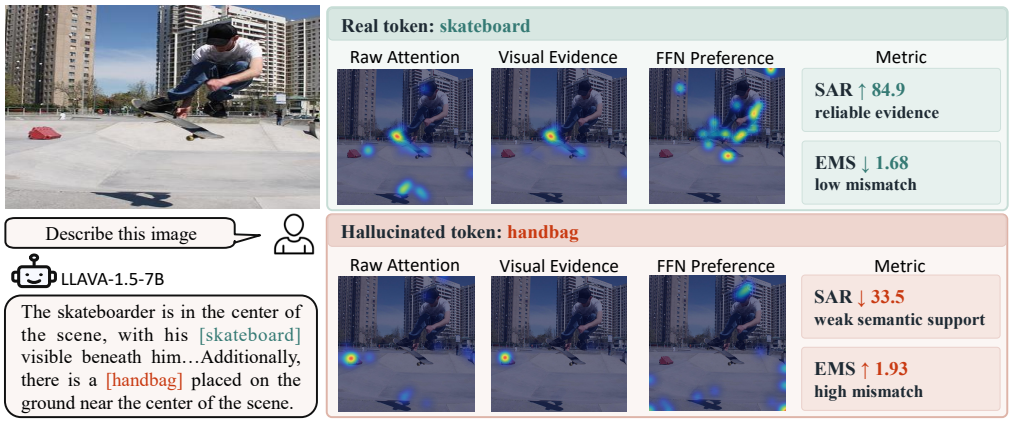


Figure 1: Visualization of a real–hallucinated object pair. Heatmaps show raw attention, visual evidence, and FFN preference. Compared with the hallucinated *handbag*, the real *skateboard* exhibits stronger visual evidence and smaller evidence–preference mismatch, reflected by higher SAR and lower EMS. More examples are provided in Appendix D.

From evidence to the subsequent update. The FFN modifies attention-aggregated states and supports factual recall (Meng et al., 2022; Geva et al., 2023); it can also reshape input contextualization (Kobayashi et al., 2024). Its additive updates can promote semantic concepts in vocabulary space (Geva et al., 2021; 2022), and recent LVLM analysis reports stronger semantic innovation in FFNs than in attention (Xi et al., 2026). Visual support should therefore be examined together with the update that follows it. Fig. 1 illustrates the distinction: the grounded object *skateboard* has stronger semantically calibrated visual writes and lower evidence mismatch than the hallucinated object *handbag*. This example motivates two questions: how reliable is the selected evidence, and how compatible is the subsequent update with it?

Empirical clues. Fig. 3 extends this comparison across decoder depth. Grounded objects generally exhibit higher semantically calibrated write strength, while hallucinated objects tend to have higher mismatch over broad layer ranges. The gaps vary by layer and backbone. These profiles motivate combining evidence reliability and update consistency across layers rather than relying on a single readout position. Complementary layer-averaged distributions are provided in Appendix C.2.

A simple conceptual lens. These observations suggest viewing object generation as two connected processes: *evidence formation* and *evidence-conditioned update*. Attention first selects and writes visual sources, while the model's internal semantic readout measures their support for the current target. We combine the two quantities at each source, using target semantics to recalibrate the magnitude of the actual write and thereby form object-specific visual evidence. The FFN then produces a new semantic update from the aggregated state. Whether its visual directional preference remains compatible with that evidence provides a second, complementary diagnostic of hallucination.

From interpretation to method. We first jointly model the attention write and target-conditioned semantics. At each visual source, we measure both how much information is written to the prediction position and how strongly the source supports the target in the model's vocabulary space. Source-wise semantic recalibration of these quantities yields Semantic-Attention Reliability (SAR). We then read the FFN update as a directional preference over same-layer visual sources and compare that preference with the semantically calibrated visual evidence through optimal transport, obtaining the Evidence Mismatch Score (EMS). The layer-wise trajectories of SAR and EMS are finally combined for object hallucination detection. Fig. 2 summarizes the resulting pipeline.

Our contributions are threefold: **(1) An evidence–update analysis.** We revisit object hallucination through target-specific visual evidence and subsequent semantic updates. Layer-wise analysis reveals stronger semantically calibrated visual writes for grounded objects and larger evidence mismatch for hallucinated objects, motivating the characterization of evidence reliability and update consistency. **(2) A dual-signal detector.** SAM combines Semantic-Attention Reliability (SAR), which recalibrates source-level attention writes with target-semantic support, and Evidence Mismatch Score (EMS), which measures FFN-update consistency with visual evidence through geometric transport and prediction-state calibration. Their layer-wise trajectories are integrated by a lightweight detector without modifying the base LVLM parameters or generation procedure.

(3) Cross-model and cross-task validation. SAM is evaluated across four LVLMs and three benchmarks spanning image captioning and visual question answering, achieving the highest AUROC among the compared methods. Ablations further verify the effectiveness of SAM's key components.

2 Related Work

Large Vision-Language Models (LVLMs). LVLMs extend language models to multimodal understanding by connecting visual encoders and language decoders through lightweight cross-modal interfaces. Representative systems such as LLaVA-1.5 (Liu et al., 2024b), InternVL2.5 (Chen et al., 2025), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a) have advanced instruction following, high-resolution perception, and image understanding. Despite this progress, LVLMs remain prone to object hallucination (Leng et al., 2024; Huang et al., 2024). We evaluate our signals across several LVLM families to test their applicability under different visual-token organizations and decoder implementations.

Object Hallucination Detection. This phenomenon generally refers to an LVLM describing an object that is absent from the input image. Detection methods have increasingly moved from output-only cues to internal model signals. MetaToken uses token-level confidence, uncertainty, and related statistics to identify hallucinated objects (Fieback et al., 2025). ProjectAway applies a Visual Logit Lens to intermediate visual representations and estimates their internal support for a target object (Jiang et al., 2025a). InsLen exploits visual information implicit in instruction-token representations to construct a training-free detection score (Lai et al., 2026). GLSim combines global and local image–text similarities for training-free object hallucination detection (Park & Li, 2025).

A second line of work focuses on visual interaction during generation. SVAR distinguishes grounded from hallucinated objects using visual attention in intermediate layers (Jiang et al., 2025b), while DHCP models attention patterns across layers and heads (Zhang et al., 2025). Recent work further incorporates semantic information into visual evidence assessment. Wang et al. (2026) showed that grounded and hallucinated objects can receive similar visual-attention magnitudes and examined whether highly attended regions semantically support the target. Nguyen et al. (2026) detects hallucinations through patch-level attention concentration and cross-modal semantic alignment.

These methods establish useful visual-support signals, but the relation between evidence formation and the subsequent semantic update remains less explored. In particular, object-specific evidence depends on how visual selection is translated into source-level attention writes and whether those writes support the target being generated. The FFN then produces a subsequent semantic update whose consistency with the formed evidence can be assessed separately. SAM models these two aspects through evidence reliability and update consistency.

3 Preliminaries and Modeling Principles

We first formulate object hallucination detection and introduce the within-layer quantities used by our method. We then state two modeling principles for visual-evidence reliability and the consistency of subsequent semantic updates.

3.1 Problem Setup

Given an image x and a text prompt c, an LVLM converts the image into M visual tokens and feeds them, together with the text tokens, to an autoregressive decoder with L layers. At generation step t, the model predicts y_t from $(x, c, y_{<t})$. We detect object mentions in the generated response and use the first subword of each object span as the target. Appendix B.3 details object-span alignment and causal token localization. The detector determines whether that object is supported by the image using internal quantities available when the target is predicted.

At layer ℓ, let $\{\boldsymbol{v}_m^\ell\}_{m=1}^M \subset \mathbb{R}^d$ denote the pre-layer visual-token states and let $\boldsymbol{q}_t^\ell \in \mathbb{R}^d$ be the pre-layer residual state at the prediction position. This position corresponds to y_{t-1}, or to the last token of the complete prompt at the first generation step. The attention output $\boldsymbol{o}_{A,t}^\ell$ and subsequent FFN update $\boldsymbol{u}_t^\ell$ satisfy

$$\boldsymbol{r}_t^\ell = \boldsymbol{q}_t^\ell + \boldsymbol{o}_{A,t}^\ell, \qquad \boldsymbol{u}_t^\ell = \mathrm{FFN}_\ell\big(N_F^\ell(\boldsymbol{r}_t^\ell)\big)\,, \tag{1}$$

where N_F^ℓ is the native pre-FFN normalization. Residual addition then gives $\boldsymbol{r}_t^\ell + \boldsymbol{u}_t^\ell$.

Let $A_{tm}^{\ell,h}$ be the attention weight from the prediction position to visual token m in head h, and let $\boldsymbol{\nu}_m^{\ell,h}$ be its value vector. The source-resolved attention write is

$$\boldsymbol{a}_{tm}^\ell = \boldsymbol{W}_O^\ell \,\mathrm{Concat}_{h=1}^{H}\left(A_{tm}^{\ell,h}\boldsymbol{\nu}_m^{\ell,h}\right), \tag{2}$$

where H is the number of heads and $\boldsymbol{W}_O^\ell$ is the output projection. We define its magnitude, the cumulative visual-write magnitude, and the head-averaged visual attention by

$$I_{tm}^\ell = \|\boldsymbol{a}_{tm}^\ell\|, \qquad I_t^\ell = \sum_{m=1}^{M} I_{tm}^\ell, \qquad \alpha_{tm}^\ell = \frac{1}{H}\sum_{h=1}^{H} A_{tm}^{\ell,h}. \tag{3}$$

Thus, I_t^ℓ sums source magnitudes, whereas $\sum_m \boldsymbol{a}_{tm}^\ell$ is the visual-source component of the attention output. The full output also includes nonvisual sources and the global output-projection bias. Combining each source across heads and through the output projection before measuring its magnitude follows the norm-based analysis of multi-head attention (Kobayashi et al., 2020; 2021; Ferrando et al., 2022). Summing source magnitudes retains cumulative visual-write strength even when differently directed source vectors partially cancel in their sum.

3.2 Logit Lens

Following vocabulary-space analyses (Dar et al., 2023) and intermediate-layer probing (Belrose et al., 2023), we apply Visual Logit Lens (Jiang et al., 2025a) to read target support from each visual state. Let N_f be the final normalization, $\boldsymbol{W}_U \in \mathbb{R}^{C\times d}$ the vocabulary output matrix, and C the vocabulary size. The target-token readout probability is

$$\beta_{tm}^\ell = \left[\mathrm{softmax}_{\mathrm{vocab}}\left(\boldsymbol{W}_U N_f(\boldsymbol{v}_m^\ell)\right)\right]_{y_t}. \tag{4}$$

The softmax normalizes over the C vocabulary entries. Hence, β_{tm}^ℓ is a scalar response to the fixed target y_t at source m; the source distributions introduced below normalize over visual tokens instead. We use scalar β for vocabulary readout probabilities, bold $\boldsymbol{T}$ for attention-evidence distributions, and bold $\boldsymbol{P}$ for FFN-preference distributions. For a fixed (ℓ, t), we write $I_m = I_{tm}^\ell$, $I = I_t^\ell$, $\alpha_m = \alpha_{tm}^\ell$, and $\beta_m = \beta_{tm}^\ell$, and similarly omit the indices on the residual and visual states. The indices are restored for layer-wise fusion.

3.3 Modeling Principles

Visual selection and target support. Attention determines which sources participate in the current prediction, but object detection must also determine how those sources relate to the target. SaN interprets attention as context-gated selection whose primitive is a gated value (Cai et al., 2026a). IIEU separates a raw response from criterion-dependent importance and connects them through selective recalibration (Cai, 2023). Following this view, we treat the visual write as the information-bearing base quantity and target semantics as an assessment of its source.

Intuition 3.1 (Target-conditioned visual-evidence reliability)**.** *Reliable visual evidence should reflect both how strongly a visual source participates in the prediction and how well that source supports the current target. The two quantities should be aligned at the source level, so stronger target support raises the evidence readout while the original visual write remains available.*

Evidence formation and the subsequent update. The state aggregated by attention is subsequently modified by the FFN (Geva et al., 2023). For object generation, evidence use therefore also concerns the relation between that update and the selected visual sources. We use same-layer visual states as a common reference and compare the FFN update's directional preference with the visual evidence.

Intuition 3.2 (Evidence-conditioned semantic adaptation)**.** *A semantic update may shift preference among nearby visual representations, whereas a shift across substantially different representations should incur a larger penalty. Agreement between the two assignments should also be assessed against the current prediction state, distinguishing agreement on compatible locations from a shared preference for incompatible ones.*

Together, our two modeling principles underpin the construction of complementary readouts for visual-evidence reliability and evidence–update consistency.

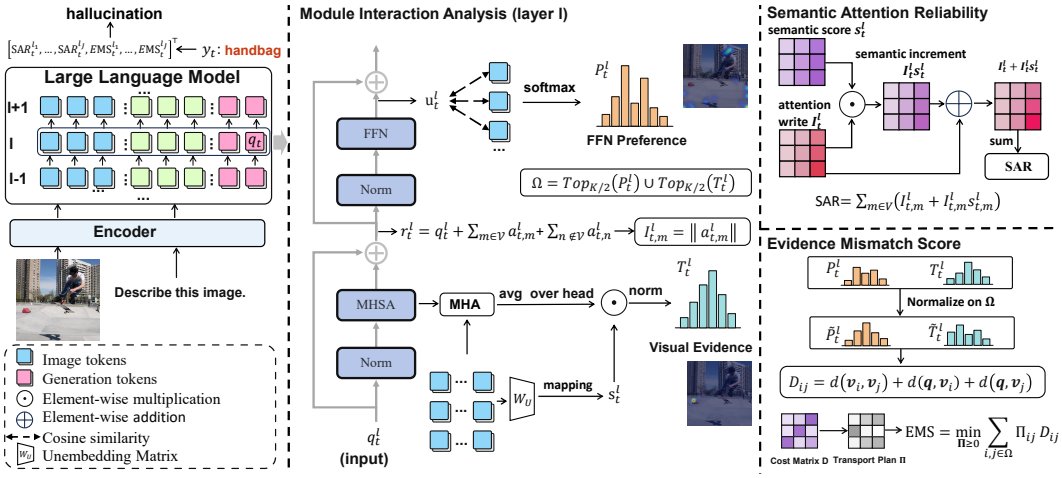


Figure 2: Illustration of our hallucination detection framework. Left: Layer-wise signals are extracted from the LVLM for each selected target token. Middle: Visual Evidence T_t^l and FFN Preference P_t^l characterize the visual support selected by MHSA and the subsequent FFN update, respectively. Right: SAR measures the semantic reliability of visual evidence, while EMS quantifies the mismatch between Visual Evidence and FFN Preference for hallucination detection.

4 METHOD

We instantiate the two modeling principles with SAR and EMS. SAR recalibrates source-resolved visual-write magnitudes with target semantics. EMS compares the FFN-induced visual preference with target-conditioned visual evidence, accounting for both visual geometry and prediction-state compatibility. Fig. 2 shows their shared inputs, separate readouts, and layer-wise fusion. All proofs and supplementary derivations for this section are provided in Appendix A.

4.1 SAR: SEMANTIC-ATTENTION RELIABILITY OF VISUAL EVIDENCE

Following Intuition 3.1, we preserve I_m and use target support to determine a semantic increment:

$$\phi(I_m, \xi_m) = I_m[1 + \varsigma(\xi_m)], \qquad \varsigma(\xi_m) \geq 0. \tag{5}$$

Here ξ_m measures reference-relative target support. This construction separates the base response from its recalibration criterion (Cai, 2023) while retaining the norm information emphasized by AdaShift (Cai, 2024). Using the uniform vocabulary reference $\beta_0 = 1/C$, we set $\xi_m = \log(C\beta_m)$ and choose $s_m = \varsigma(\xi_m) = \text{softplus}(\xi_m) = \log(1 + C\beta_m)$. We use the uniform-vocabulary probability $\beta_0 = 1/C$ as the reference for relative target support. The increment approaches zero as β_m approaches zero. Motivated by FleS's analysis of nearly indistinguishable saturating gates (Cai et al., 2026b), we retain differences between above-reference responses as follows.

Proposition 4.1 (Difference preservation under semantic recalibration)**.** *Let* $1/C \leq \beta_j \leq \beta_i \leq 1$, *where* $C > 1$. *Set* $\xi_k = \log(C\beta_k)$ *for* $k \in \{i, j\}$ *and write* $\sigma(\xi) = (1 + e^{-\xi})^{-1}$. *Then* $0 \leq \sigma(\xi_i) - \sigma(\xi_j) \leq e^{-\xi_j} \log \frac{\beta_i}{\beta_j}$, $\frac{1}{2} \log \frac{\beta_i}{\beta_j} \leq s_i - s_j \leq \log \frac{\beta_i}{\beta_j}$.

Above the vocabulary reference, the semantic increment retains at least half of the relative log-support gap. Its growth is logarithmic in β_m, providing a monotone, nonnegative recalibration without collapsing strong-support differences into a saturating gate.

We aggregate the recalibrated source magnitudes as

$$\text{SAR} = \sum_{m=1}^{M} I_m(1 + s_m) = \underbrace{I}_{\text{base write strength}} + \underbrace{\sum_{m=1}^{M} I_m s_m}_{\text{target-semantic increment}}. \tag{6}$$

The SAR branch in Fig. 2 implements this source-wise recalibration and summation. The base term retains total source strength; the increment favors writes whose sources better support the target. For fixed $\{I_m\}$, stronger semantic support increases SAR.

4.2 EMS: EVIDENCE MISMATCH SCORE FROM THE EVIDENCE

Target-conditioned visual evidence. Following Intuition 3.2, we compare two distributions on the same visual references. The attention-evidence distribution is $T_m = \frac{\alpha_m s_m}{\sum_{n=1}^{M} \alpha_n s_n}, \boldsymbol{T} \in \Delta^{M-1}$, where $\Delta^{M-1} = \{\boldsymbol{w} \in \mathbb{R}^M_{\geq 0} : \sum_m w_m = 1\}$. Unlike SAR, which retains write magnitude, $\boldsymbol{T}$ measures relative attention allocation after semantic calibration. Both share s_m but serve different purposes.

FFN-induced visual preference. For nonzero $\boldsymbol{u}$ and $\boldsymbol{v}_m$, $\cos(\boldsymbol{u}, \boldsymbol{v}_m)$ measures the directional compatibility of the realized update with source m. Minimizing directional cost selects only the best-aligned references. To retain near-compatible alternatives under a unit-mass budget (Cai et al., 2026a), we use the entropy-regularized readout (Blondel et al., 2020)

$$\boldsymbol{P} = \operatorname*{arg\,min}_{\boldsymbol{w} \in \Delta^{M-1}} \left\{ \sum_{m=1}^{M} w_m [1 - \cos(\boldsymbol{u}, \boldsymbol{v}_m)] - \tau \mathcal{H}(\boldsymbol{w}) \right\}, \tag{7}$$

where $\mathcal{H}(\boldsymbol{w}) = -\sum_m w_m \log w_m$ and $\tau > 0$.

Proposition 4.2 (Gibbs visual-preference readout)**.** *The unique minimizer of equation 7 is*

$$P_m = \frac{\exp(\cos(\boldsymbol{u}, \boldsymbol{v}_m)/\tau)}{\sum_{n=1}^{M} \exp(\cos(\boldsymbol{u}, \boldsymbol{v}_n)/\tau)}. \tag{8}$$

The normalization is over visual sources. Accordingly, $\boldsymbol{P}$ records their relative directional compatibility with the FFN update. Its relation to residual readout is characterized below.

Lemma 4.3 (Readout consistency of the FFN visual preference)**.** *For $\widehat{\boldsymbol{v}}_m = \boldsymbol{v}_m / \|\boldsymbol{v}_m\|$, define $g_m(\boldsymbol{z}) = \langle \boldsymbol{z}, \widehat{\boldsymbol{v}}_m \rangle$ and $\Delta g_m = g_m(\boldsymbol{r} + \boldsymbol{u}) - g_m(\boldsymbol{r})$. If $\boldsymbol{u} \neq \boldsymbol{0}$ and $\tau > 0$, then $\log \frac{P_i}{P_j} = \frac{\Delta g_i - \Delta g_j}{\tau \|\boldsymbol{u}\|}$.*

Thus, the relative preference for two visual sources is determined by the difference in their FFN-induced linear-readout increments, scaled by the update norm and the readout temperature.

Evidence–update mismatch. Different visual positions can have similar representations. We therefore compare $\boldsymbol{P}$ and $\boldsymbol{T}$ through the geometry of their visual references (Kusner et al., 2015; Zhao et al., 2019; Zhang et al., 2020). For efficient computation, we retain their joint principal support $\Omega = \mathrm{Top}_{K/2}(\boldsymbol{P}) \cup \mathrm{Top}_{K/2}(\boldsymbol{T}), |\Omega| \leq K$, where K is a positive even support budget and Top_k returns the indices of up to k largest entries. On this support, we renormalize both distributions: $\widetilde{P}_i = \frac{P_i}{\sum_{n \in \Omega} P_n}, \widetilde{T}_i = \frac{T_i}{\sum_{n \in \Omega} T_n}, i \in \Omega$. Retained-mass guarantees for this support restriction are provided in Appendix A.9. We use the pre-attention prediction state $\boldsymbol{q}$ as a shared reference to calibrate the matching cost. For nonzero vectors, let $d(\boldsymbol{a}, \boldsymbol{b}) = 1 - \frac{\boldsymbol{a}^\top \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}$,

$$D_{ij} = \underbrace{d(\boldsymbol{v}_i, \boldsymbol{v}_j)}_{\text{visual geometric deviation}} + \underbrace{d(\boldsymbol{q}, \boldsymbol{v}_i) + d(\boldsymbol{q}, \boldsymbol{v}_j)}_{\text{prediction-state incompatibility}}, \qquad i, j \in \Omega. \tag{9}$$

The first term measures the discrepancy between visual references, while the two endpoint terms assess their compatibility with the current prediction state. The Evidence Mismatch Score is

$$\mathrm{EMS} = \min_{\boldsymbol{\Pi} \geq 0} \sum_{i,j \in \Omega} \Pi_{ij} D_{ij} \qquad \text{s.t.} \quad \boldsymbol{\Pi} \boldsymbol{1} = \widetilde{\boldsymbol{P}}, \qquad \boldsymbol{\Pi}^\top \boldsymbol{1} = \widetilde{\boldsymbol{T}}. \tag{10}$$

Here $\boldsymbol{\Pi} \in \mathbb{R}^{|\Omega| \times |\Omega|}_{\geq 0}$ matches FFN-preference mass to attention-evidence mass. We solve this finite-support problem by network simplex (Peyré & Cuturi, 2019; Flamary et al., 2021), without entropy regularization on $\boldsymbol{\Pi}$. The EMS branch in Fig. 2 follows the sequence: form $\boldsymbol{P}$ and $\boldsymbol{T}$, retain their joint support, and evaluate the calibrated transport score.

The score admits the decomposition

$$\mathrm{EMS} = \mathrm{OT}_d(\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{T}}) + \underbrace{\sum_{i \in \Omega} (\widetilde{P}_i + \widetilde{T}_i) d(\boldsymbol{q}, \boldsymbol{v}_i)}_{R_q}, \tag{11}$$

where OT_d denotes optimal transport with source-to-source cost $d(\boldsymbol{v}_i, \boldsymbol{v}_j)$. The transport term measures the geometric mismatch between FFN preference and attention evidence, while R_q quantifies their respective deviations from the prediction state $\boldsymbol{q}$. The following result characterizes the cost of FFN-preference mass assigned away from the principal evidence.

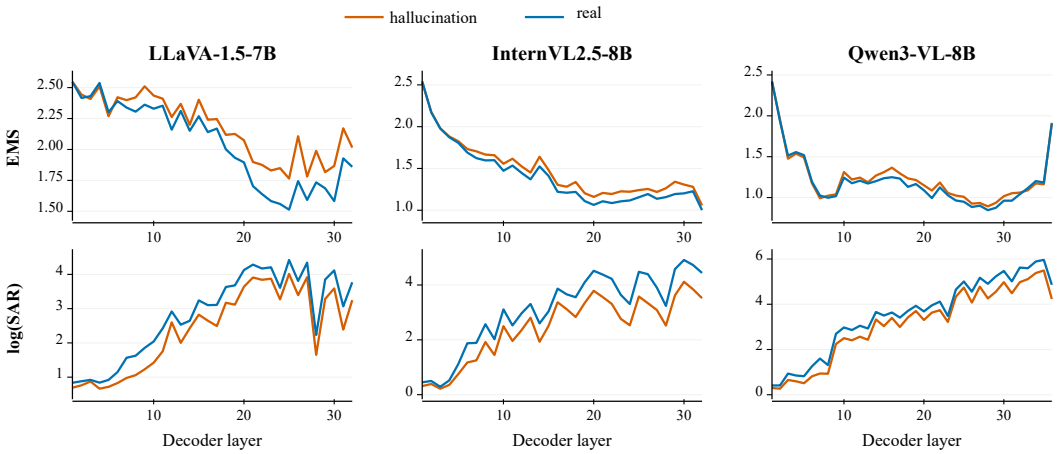


Figure 3: Complementary layer-wise profiles of SAR and EMS. Real objects generally have stronger calibrated visual writes, whereas hallucinated objects tend to exhibit higher mismatch over broad layer ranges. The layer-dependent gaps motivate joint cross-layer detection.

Theorem 4.4 (Geometric lower bound on evidence escape)**.** *Let* $E \subseteq \Omega$ *satisfy* $\widetilde{\boldsymbol{T}}(E) \geq 1-\epsilon$*, with* $0 \leq \epsilon < 1$*. If* $B \subseteq \Omega$ *satisfies* $d(\boldsymbol{v}_i, \boldsymbol{v}_j) \geq \delta > 0$ *for every* $i \in B$ *and* $j \in E$*, then* $\mathrm{EMS} \geq R_q + \delta\big[\widetilde{\boldsymbol{P}}(B) - \epsilon\big]_+$*. Here,* $\widetilde{\boldsymbol{P}}(B) = \sum_{i\in B} \widetilde{P}_i$*,* $\widetilde{\boldsymbol{T}}(E) = \sum_{j\in E} \widetilde{T}_j$*, and* $[z]_+ = \max\{z, 0\}$.

When visual evidence concentrates on E, FFN-preference mass assigned to separated references incurs an unavoidable geometric cost once it exceeds the evidence mass available outside E. This cost depends jointly on the displaced mass and the separation of the visual references.

4.3 Hallucination Detection

We concatenate the layer-wise EMS and SAR scores as

$$\boldsymbol{f}_t = [\mathrm{EMS}_t^1, \ldots, \mathrm{EMS}_t^L; \mathrm{SAR}_t^1, \ldots, \mathrm{SAR}_t^L]^\top \in \mathbb{R}^{2L}. \tag{12}$$

A lightweight MLP maps $\boldsymbol{f}_t$ to the visual-support prediction, preserving the complementary cross-layer patterns of EMS and SAR.

5 Experiments

5.1 Experimental Setup

Models. We evaluate four representative LVLMs: LLaVA-1.5-7B (Liu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL2.5-8B (Chen et al., 2025), and the more recent Qwen3-VL-8B (Bai et al., 2025a). These models differ in visual-token organization, decoder architecture, and attention implementation, allowing us to evaluate the applicability of the proposed signals across LVLM architectures.

Datasets and evaluation protocol. We evaluate hallucination detection on three benchmarks covering complementary settings. On MS COCO (Lin et al., 2014), we sample 4,000 images from val2014, prompt each model with ``Describe this image.'', and label object mentions following CHAIR (Rohrbach et al., 2018). Following prior object-level detection protocols (Jiang et al., 2025b), we retain the first selected occurrence of each canonical object and extract signals at the decoder position predicting its first sub-token (Park & Li, 2025). The same selected targets and causal positions are used for all detectors within a model. On CLEVR (Johnson et al., 2017), we evaluate existence hallucination using object-existence questions. On AMBER (Wang et al., 2024), we further evaluate hallucinations involving object existence, attributes, and relations. For question answering, signals are extracted at the final prompt token that predicts the first answer token. All methods use identical samples, target positions, and data splits. We use image-level 8:1:1 train–validation–test splits with three random seeds and primarily report mean AUC and AUPR.

Table 1: Comparison of token-level object hallucination detection on 4,000 MS COCO images across four LVLMs. The best and second-best results are shown in bold and underlined, respectively.

Method	Venue	LLaVA-1.5-7B		Qwen2.5-VL-7B		InternVL2.5-8B		Qwen3-VL-8B	
		AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
MetaToken	VISAPP'25	$89.86^{\pm 0.01}$	$96.33^{\pm 0.00}$	$84.28^{\pm 0.01}$	$96.98^{\pm 0.00}$	$87.19^{\pm 0.01}$	$96.18^{\pm 0.00}$	$86.95^{\pm 0.00}$	$95.52^{\pm 0.00}$
DHCP	MM'25	$86.68^{\pm 0.10}$	$94.99^{\pm 0.06}$	$82.68^{\pm 0.14}$	$96.55^{\pm 0.05}$	$85.58^{\pm 0.06}$	$95.73^{\pm 0.01}$	$85.64^{\pm 0.06}$	$95.17^{\pm 0.02}$
SVAR	CVPR'25	$\underline{90.58}^{\pm 0.06}$	$\underline{96.72}^{\pm 0.03}$	$\underline{85.20}^{\pm 0.32}$	$\underline{97.08}^{\pm 0.06}$	$\underline{88.13}^{\pm 0.15}$	$\underline{96.66}^{\pm 0.05}$	$\underline{88.13}^{\pm 0.14}$	$\underline{96.21}^{\pm 0.05}$
ProjectAway	ICLR'25	$76.52^{\pm 0.00}$	$89.91^{\pm 0.00}$	$63.13^{\pm 0.00}$	$91.54^{\pm 0.00}$	$75.43^{\pm 0.00}$	$90.73^{\pm 0.00}$	$72.94^{\pm 0.00}$	$88.46^{\pm 0.00}$
TokenGrounding	CVPR'26	$86.77^{\pm 0.00}$	$95.07^{\pm 0.00}$	$80.97^{\pm 0.00}$	$96.46^{\pm 0.00}$	$81.51^{\pm 0.00}$	$94.17^{\pm 0.00}$	$82.01^{\pm 0.00}$	$93.90^{\pm 0.00}$
InsLen	ICML'26	$86.30^{\pm 0.00}$	$95.00^{\pm 0.00}$	$46.31^{\pm 0.00}$	$86.19^{\pm 0.00}$	$80.66^{\pm 0.00}$	$94.02^{\pm 0.00}$	$73.10^{\pm 0.00}$	$90.04^{\pm 0.00}$
SAM (Ours)		$\mathbf{91.52}^{\pm 0.13}$	$\mathbf{97.06}^{\pm 0.05}$	$\mathbf{85.60}^{\pm 0.08}$	$\mathbf{97.26}^{\pm 0.02}$	$\mathbf{88.81}^{\pm 0.17}$	$\mathbf{96.85}^{\pm 0.05}$	$\mathbf{90.31}^{\pm 0.14}$	$\mathbf{96.93}^{\pm 0.04}$

Table 2: Comparison of existence hallucination on CLEVR across four LVLMs.

Method	Venue	LLaVA-1.5-7B		Qwen2.5-VL-7B		InternVL2.5-8B		Qwen3-VL-8B	
		AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
MetaToken	VISAPP'25	$55.53^{\pm 0.00}$	$56.54^{\pm 0.00}$	$\underline{88.58}^{\pm 0.00}$	$99.28^{\pm 0.00}$	$\underline{95.01}^{\pm 0.00}$	$\underline{99.91}^{\pm 0.00}$	$\underline{88.76}^{\pm 0.00}$	$\underline{99.51}^{\pm 0.00}$
DHCP	MM'25	$51.70^{\pm 0.98}$	$50.93^{\pm 1.20}$	$84.42^{\pm 0.13}$	$99.05^{\pm 0.02}$	$94.25^{\pm 0.17}$	$99.89^{\pm 0.00}$	$84.52^{\pm 3.64}$	$99.47^{\pm 0.16}$
SVAR	CVPR'25	$\underline{60.41}^{\pm 0.45}$	$\underline{59.94}^{\pm 0.23}$	$81.98^{\pm 0.23}$	$99.06^{\pm 0.02}$	$83.82^{\pm 0.60}$	$99.67^{\pm 0.02}$	$58.76^{\pm 0.09}$	$97.69^{\pm 0.02}$
ProjectAway	ICLR'25	$51.65^{\pm 0.00}$	$51.45^{\pm 0.00}$	$48.86^{\pm 0.00}$	$95.83^{\pm 0.00}$	$39.01^{\pm 0.00}$	$97.81^{\pm 0.00}$	$44.82^{\pm 0.00}$	$96.85^{\pm 0.00}$
TokenGrounding	CVPR'26	$50.01^{\pm 1.46}$	$49.92^{\pm 1.74}$	$87.42^{\pm 0.38}$	$\underline{99.29}^{\pm 0.02}$	$87.11^{\pm 0.53}$	$99.76^{\pm 0.01}$	$77.23^{\pm 0.12}$	$99.16^{\pm 0.00}$
InsLen	ICML'26	$45.64^{\pm 0.00}$	$48.31^{\pm 0.00}$	$60.30^{\pm 0.00}$	$96.93^{\pm 0.00}$	$88.14^{\pm 0.00}$	$99.77^{\pm 0.00}$	$62.74^{\pm 0.00}$	$98.39^{\pm 0.00}$
SAM (Ours)		$\mathbf{63.66}^{\pm 0.59}$	$\mathbf{64.25}^{\pm 0.47}$	$\mathbf{89.56}^{\pm 0.19}$	$\mathbf{99.35}^{\pm 0.02}$	$\mathbf{96.37}^{\pm 0.08}$	$\mathbf{99.93}^{\pm 0.00}$	$\mathbf{92.58}^{\pm 0.24}$	$\mathbf{99.74}^{\pm 0.01}$

Additional metrics are reported in Appendix C.1. Dataset and implementation details are provided in Appendices B.1 and B.3, respectively.

Implementation details. We fix the transport-support budget to $K = 32$ and the source temperature to $\tau = 0.03$ throughout all experiments. The detector is a lightweight one-hidden-layer MLP with architecture $2L \rightarrow 128 \rightarrow 1$, using ReLU activation and a dropout rate of 0.3 after the hidden layer. All experiments are conducted on two NVIDIA RTX 4090 GPUs. Sensitivity to K and τ is examined in Appendix C.6.

Baselines. We compare with six representative hallucination detectors: MetaToken (Fieback et al., 2025), SVAR (Jiang et al., 2025b), DHCP (Zhang et al., 2025), TokenGrounding (Nguyen et al., 2026), and ProjectAway (Jiang et al., 2025a), and Inslens Lai et al. (2026) For fair comparisons, we follow the original implementation protocols and detector settings of each method. Detailed baseline descriptions and implementation settings are provided in Appendices B.2 and B.3.

5.2 Main Results

Object hallucination detection on COCO captioning. As shown in Tab. 1, our method consistently achieves the best object hallucination detection performance across all four LVLMs on MS COCO. Compared with the strongest baseline SVAR, our method improves both AUROC and AUPR on every model. In particular, on Qwen3-VL-8B, our method outperforms SVAR by 2.18 and 0.72 percentage points in AUROC and AUPR, respectively. Consistent gains are also observed on LLaVA-1.5-7B, Qwen2.5-VL-7B, and InternVL2.5-8B, with AUROC improvements of 0.94, 0.40, and 0.68 percentage points, respectively. These results demonstrate the effectiveness and generalizability of our method across different LVLM architectures.Complete Precision, Recall, F1, and Accuracy results are reported in Appendix C.1.

Hallucination detection on QA benchmarks. As reported in Tabs. 2 and 3, our method consistently delivers strong detection performance under different QA formulations. On CLEVR, SAM achieves the highest AUROC across all four LVLMs, improving over the strongest baseline by 3.25, 0.98, 1.36, and 3.82 percentage points on LLaVA-1.5-7B, Qwen2.5-VL-7B, InternVL2.5-8B, and Qwen3-VL-8B, respectively. On AMBER, it yields the best results for object-existence and relation hallucinations, with improvements of 0.67 and 1.89 percentage points, respectively, while achieving the second-best result on attribute hallucination. These results demonstrate the generalizability of our method across different LVLM architectures, QA formulations, and hallucination categories.

Table 3: Layer ablation and AMBER results. **(Left)** SAR, EMS, and SAM across decoder-layer ranges. **(Right)** AUROC on AMBER using LLaVA-1.5-7B.

Layers	SAR		EMS		SAM	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
1–8	82.29	92.78	75.50	89.89	83.56	93.60
9–18	85.81	94.76	77.38	90.86	86.94	95.28
19–27	81.53	92.78	81.67	92.61	86.00	94.72
28–32	84.25	93.84	83.03	92.55	87.87	95.42
All	**89.98**	**96.48**	**87.34**	**95.24**	**91.52**	**97.06**

Method	Obj.	Attr.	Rel.
MetaToken	$90.72^{\pm 0.00}$	$76.13^{\pm 0.00}$	$63.45^{\pm 0.00}$
DHCP	$94.68^{\pm 0.32}$	$65.20^{\pm 2.63}$	$54.97^{\pm 2.14}$
SVAR	$\underline{98.64}^{\pm 0.26}$	$\mathbf{85.00}^{\pm 0.77}$	$72.92^{\pm 1.43}$
ProjectAway	$96.31^{\pm 0.00}$	$49.80^{\pm 0.00}$	$36.07^{\pm 0.00}$
TokenGrounding	$93.76^{\pm 0.13}$	$61.37^{\pm 0.12}$	$61.79^{\pm 0.33}$
InsLen	$92.16^{\pm 0.00}$	$57.40^{\pm 0.00}$	$\underline{78.16}^{\pm 0.00}$
SAM (Ours)	$\mathbf{99.31}^{\pm 0.08}$	$\underline{81.95}^{\pm 0.30}$	$\mathbf{80.05}^{\pm 0.34}$

Table 4: Ablation and efficiency results on COCO. **(Left)** SAR and EMS variants across two LVLMs. **(Right)** Feature-extraction efficiency and detector complexity. GPU denotes peak memory including LVLM weights, and Dim. the detector input dimension.

Variant	LLaVA-1.5	Qwen3-VL
Signal composition		
SAR only	89.98	89.31
EMS only	87.34	84.62
SAM	**91.52**	**90.31**
Components of full SAM		
w/o s_m in SAR	90.47	89.79
w/o s_m in EMS	91.19	89.98
w/o R_q	91.34	90.03
Union-region JS	91.14	90.13

Method	s/img	ms/obj.	GPU	Dim.	Params
LLaVA-1.5-7B					
SAM (Ours)	0.501	161.8	14.12 GiB	64	8,449
SVAR	0.116	37.3	14.52 GiB	448	111,850
DHCP	0.123	39.8	14.52 GiB	147,456	18,874,754
TokenGrounding	1.065	343.7	14.89 GiB	64	16,897
Qwen3-VL-8B					
SAM (Ours)	0.764	261.6	17.41 GiB	72	9,473
SVAR	0.339	116.0	17.43 GiB	448	111,850
DHCP	0.355	121.5	17.43 GiB	165,888	21,234,050
TokenGrounding	0.770	263.6	17.71 GiB	72	18,945

5.3 Ablation Studies

Layer ablation. Tab. 3(Left) compares four layer ranges with all layers on LLaVA-1.5-7B. SAR peaks at layers 9–18, while EMS improves with depth. SAM consistently outperforms both signals, with larger gains in later layers. All-layer features perform best; SAM exceeds its best individual range by 3.65 AUROC points, supporting cross-layer integration of evidence and mismatch cues.

Component ablation. As shown in Tab. 4(Left), SAR is the stronger standalone signal, yet adding EMS improves AUROC by 1.00–1.54 points across the two models, indicating that FFN-preference mismatch complements visual-evidence strength. Within full SAM, removing semantic calibration s_m from either branch consistently lowers performance, supporting its role in recalibrating visual writes and constructing target-conditioned attention evidence. Removing prediction-state compatibility R_q or using union-region JS also reduces performance. Appendix C.6 further examines sensitivity to the transport-support budget and FFN-preference temperature.

Efficiency analysis. With two scalar signals per layer, SAM uses only 64 and 72 input dimensions and fewer than 10,000 trainable MLP parameters (Tab. 4(Right)), substantially below SVAR and DHCP in both input dimensionality and parameter count. Feature extraction is slower than SVAR and DHCP, but takes roughly half the time of TokenGrounding on LLaVA-1.5-7B and comparable time on Qwen3-VL-8B. Peak GPU memory remains comparable to the baselines.

6 Conclusion

In this work, we revisited object hallucination through visual-evidence reliability and formalized evidence–update mismatch between target-conditioned visual evidence and FFN-induced visual preference. Our layer-wise analysis revealed stronger semantically calibrated visual writes for grounded objects and larger mismatch for hallucinated objects. These observations underpin SAM, an interpretable dual-signal framework that couples source-level semantic recalibration in SAR with geometric transport and prediction-state calibration in EMS. Extensive evaluations across four LVLMs and three benchmarks spanning image captioning and visual question answering demonstrate SAM's effectiveness across architectures and tasks, while targeted ablations support the complementarity of evidence reliability and update mismatch.

AI USE STATEMENT

ChatGPT was used to aid in polishing the writing. Specifically, it was employed to correct grammar, improve readability, and refine the clarity of sentences.

REPRODUCIBILITY STATEMENT

The problem setup, modeling principles, and definitions of SAR and EMS are provided in Sections 3 and 4. The associated assumptions, derivations, and proofs are detailed in Appendix A. Dataset descriptions and baseline protocols appear in Appendices B.1 and B.2. Data sampling and splitting, object labeling, causal token localization, feature extraction, detector training, and hyperparameter selection are specified in Appendix B.3. Additional detection metrics and hyperparameter analyses are reported in Appendices C.1 and C.6.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with fenchel-young losses. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

Sudong Cai. Iieu: Rethinking neural feature activation from decision-making. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5796–5806, October 2023.

Sudong Cai. Adashift: Learning discriminative self-gated neural feature activation with an adaptive shift factor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5947–5956, June 2024.

Sudong Cai, Shuai Yuan, Bingzhi Chen, Rui Mao, and Bing Wang. Selection-as-nonlinearity: Bridging attention and activation via a joint game-decision lens for interpretable, discriminative visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, June 2026a.

Sudong Cai, Shuyuan Zheng, Bingzhi Chen, Shuai Yuan, Chuan Xiao, Jianbin Qin, and Bing WANG. Toward principled flexible scaling for self-gated neural activation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=XGODWn7HeJ.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893/.

Javier Ferrando, Gerard I Gállego, and Marta R Costa-Jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8698–8714, 2022.

Laura Fieback, Jakob Spiegelberg, and Hanno Gottschalk. MetaToken: Detecting hallucination in image descriptions by meta classification. In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: VISAPP*, pp. 126–137. INSTICC, SciTePress, 2025. ISBN 978-989-758-728-3. doi: 10.5220/0013165700003912.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing (EMNLP)*, pp. 30–45, 2022.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235, 2023.

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, June 2024.

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.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=94kQgWXojH.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 25004–25014, June 2025b.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4547–4568, 2021.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Analyzing feed-forward blocks in transformers through the lens of attention maps. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=mYWsyTuiRp.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kusnerb15.html.

Runhe Lai, Xinhua Lu, Yanqi Wu, Jinlun Ye, Weijiang Yu, and Ruixuan Wang. Instruction lens score: Your instruction contributes a powerful object hallucination detector for multimodal large language models. In *Forty-third International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=kCJ9kaebgd.

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, June 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, pp. 740–755. Springer, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40235–40244, June 2026.

Seongheon Park and Sharon Li. Glsim: Detecting object hallucinations in lvlms via global-local similarity. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 37747–37776. Curran Associates, Inc., 2025. doi: 10.52202/085713-1265. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3640a1997a4c9571cea9db2c82e1fc35-Paper-Conference.pdf.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 02 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

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

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, January 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. URL https://arxiv.org/abs/2311.07397.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25315–25325, June 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.

Gongli Xi, Ye Tian, Mengyu Yang, Huahui Yi, Liang Lin, Xiaoshuai Hao, Kun Wang, and Wendong Wang. Large vision-language models get lost in attention. In *Forty-third International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=pt7x2tJm4L.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 3555–3564, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755118. URL https://doi.org/10.1145/3746027.3755118.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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. 563–578, 2019.

A Supplementary Derivations and Proofs

In this appendix, we provide the derivations and proofs supporting § 4. We first examine semantic recalibration and attention evidence, then characterize the FFN visual preference, and finally establish the geometric and approximation properties of EMS. The analysis inherits the notation of the main paper. We fix a target and a decoder layer and omit their indices. Throughout, $M \geq 1$ is the number of visual sources, $C > 1$ is the vocabulary size, and $\widehat{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|$ for a nonzero vector $\boldsymbol{x}$. Entropy and KL terms use the convention $0 \log 0 = 0$.

A.1 Proof of Proposition 4.1

In the main paper, Proposition 4.1 characterizes how semantic recalibration preserves differences between above-reference target responses. Let $s_m = \log(1 + C\beta_m)$ and, for $\beta_m > 0$, let $\xi_m = \log(C\beta_m)$. Thus $s_m = s(\xi_m)$, where $s(z) = \log(1 + e^z)$.

Retrospect. For ease of reference, we restate the proposition below.

Proposition A.1 (Difference preservation under semantic recalibration)**.** *Let* $1/C \leq \beta_j \leq \beta_i \leq 1$ *and* $\sigma(z) = (1 + e^{-z})^{-1}$*. Then*

$$0 \leq \sigma(\xi_i) - \sigma(\xi_j) \leq e^{-\xi_j} \log \frac{\beta_i}{\beta_j} . \tag{13}$$

Moreover, the semantic increments satisfy

$$\frac{1}{2} \log \frac{\beta_i}{\beta_j} \leq s_i - s_j \leq \log \frac{\beta_i}{\beta_j} . \tag{14}$$

Proof. By the assumed ordering of the probabilities, $0 \leq \xi_j \leq \xi_i \leq \log C$ and $\xi_i - \xi_j = \log(\beta_i/\beta_j)$. On this interval, $\sigma'(z) = e^{-z}/(1 + e^{-z})^2 \leq e^{-\xi_j}$. Hence, by the fundamental theorem of calculus,

$$0 \leq \sigma(\xi_i) - \sigma(\xi_j) = \int_{\xi_j}^{\xi_i} \sigma'(z)\, \mathrm{d}z \leq e^{-\xi_j} (\xi_i - \xi_j) . \tag{15}$$

This gives Eq. equation 13. Similarly, $s'(z) = \sigma(z)$, and $1/2 \leq \sigma(z) \leq C/(1 + C)$ on the same interval. Therefore,

$$\frac{1}{2}(\xi_i - \xi_j) \leq s_i - s_j = \int_{\xi_j}^{\xi_i} \sigma(z)\, \mathrm{d}z \leq \frac{C}{1 + C}(\xi_i - \xi_j) . \tag{16}$$

Substituting the log-probability ratio yields the tighter bound

$$\frac{1}{2} \log \frac{\beta_i}{\beta_j} \leq s_i - s_j \leq \frac{C}{1 + C} \log \frac{\beta_i}{\beta_j} , \tag{17}$$

which also establishes Eq. equation 14. This completes the proof. □

The result quantifies the retained difference in relative log support. Above the vocabulary reference $1/C$, the semantic increment preserves at least half of this difference. This property follows the motivation of maintaining discriminative recalibration for strong responses in FleS (Cai et al., 2026b). For a fixed C, its range is $0 \leq s_m \leq \log(1 + C)$.

A.2 Anchor Preservation and Source-wise Semantic Alignment

We further examine how semantic support modifies the source-write magnitudes. Let $I_m \geq 0$ denote the magnitude of source m and $I = \sum_m I_m$. The SAR construction is

$$\mathrm{SAR} = \sum_{m=1}^{M} I_m(1 + s_m) = I + \sum_{m=1}^{M} I_m s_m . \tag{18}$$

Proposition A.2 (Magnitude preservation and semantic monotonicity). *For $I_m \geq 0$ and $\beta_m \in [0,1]$,*

$$I \leq \mathrm{SAR} \leq [1 + \log(1+C)]I\,. \tag{19}$$

For fixed write magnitudes, SAR is nondecreasing in each β_m, and strictly increasing whenever $I_m > 0$. For $I > 0$, it also satisfies

$$\mathrm{SAR} = I\left(1 + \sum_m \omega_m s_m\right), \qquad \omega_m = I_m/I, \quad \sum_m \omega_m = 1\,. \tag{20}$$

Proof. The range $0 \leq s_m \leq \log(1+C)$ follows from $\beta_m \in [0,1]$. Multiplication by $I_m \geq 0$ and summation give Eq. equation 19. The partial derivatives are

$$\frac{\partial \mathrm{SAR}}{\partial \beta_m} = \frac{C I_m}{1 + C\beta_m}, \qquad \frac{\partial \mathrm{SAR}}{\partial I_m} = 1 + s_m\,. \tag{21}$$

The first derivative gives the claimed semantic monotonicity. For $I > 0$, substituting $I_m = I\omega_m$ into Eq. equation 18 gives Eq. equation 20. For $I = 0$, all nonnegative I_m are zero and hence $\mathrm{SAR} = 0$. This completes the proof. □

For $I > 0$, ω_m is the share of cumulative write magnitude contributed by source m. Eq. 20 therefore expresses SAR as cumulative magnitude recalibrated by its write-weighted semantic support. This preserves the response anchor underlying the recalibration perspectives of IIEU and AdaShift (Cai, 2023; 2024).

Proposition A.3 (Source-alignment contribution). *Let $\bar{I} = M^{-1}\sum_m I_m$, $\bar{s} = M^{-1}\sum_m s_m$, and $\mathrm{Cov}_{\mathcal{V}}(I,s) = M^{-1}\sum_m (I_m - \bar{I})(s_m - \bar{s})$. Then*

$$\mathrm{SAR} = I(1+\bar{s}) + M\,\mathrm{Cov}_{\mathcal{V}}(I,s)\,. \tag{22}$$

If $I_i \geq I_j$ and $s_i \geq s_j$, assigning s_i to I_i and s_j to I_j increases SAR relative to the exchanged assignment by

$$\Delta\,\mathrm{SAR} = (I_i - I_j)(s_i - s_j) \geq 0\,. \tag{23}$$

Proof. Expanding the covariance gives

$$\begin{aligned} M\,\mathrm{Cov}_{\mathcal{V}}(I,s) &= \sum_m I_m s_m - M\bar{I}\bar{s} \\ &= \sum_m I_m s_m - I\bar{s}\,. \end{aligned} \tag{24}$$

Combining this identity with Eq. equation 18 proves Eq. equation 22. For the two-source exchange, the base magnitudes and all other source terms are fixed. Thus

$$\begin{aligned} \Delta\,\mathrm{SAR} &= (I_i s_i + I_j s_j) - (I_i s_j + I_j s_i) \\ &= (I_i - I_j)(s_i - s_j) \geq 0\,, \end{aligned} \tag{25}$$

where the last inequality follows from the assumed orderings. This completes the proof. □

For a fixed target and layer, the covariance measures the source-wise correspondence between write magnitude and semantic support. The exchange identity further characterizes how aligning stronger writes with stronger support increases the semantic increment.

Corollary A.4 (Stability of SAR in log-support coordinates). *Let $\boldsymbol{I} = (I_m)$ and $\boldsymbol{I}' = (I'_m)$ have nonnegative entries. For $\beta_m, \beta'_m \in (0,1]$, define $\xi_m = \log(C\beta_m)$ and $\xi'_m = \log(C\beta'_m)$, using the same vocabulary size and source index set. Then*

$$\begin{aligned} &|\,\mathrm{SAR}(\boldsymbol{I},\boldsymbol{\xi}) - \mathrm{SAR}(\boldsymbol{I}',\boldsymbol{\xi}')| \\ &\quad\leq [1+\log(1+C)]\sum_m |I_m - I'_m| + \sum_m I'_m |\xi_m - \xi'_m|\,. \end{aligned} \tag{26}$$

Proof. Since $s'(z) = \sigma(z) \in (0, 1)$, the mean value theorem gives $|s(\xi_m) - s(\xi'_m)| \le |\xi_m - \xi'_m|$. Adding and subtracting $\sum_m I'_m[1 + s(\xi_m)]$, we obtain

$$\begin{aligned} &|\operatorname{SAR}(\boldsymbol{I}, \boldsymbol{\xi}) - \operatorname{SAR}(\boldsymbol{I}', \boldsymbol{\xi}')| \\ &\quad \le \sum_m |I_m - I'_m|[1 + s(\xi_m)] + \sum_m I'_m |s(\xi_m) - s(\xi'_m)| \\ &\quad \le [1 + \log(1 + C)] \sum_m |I_m - I'_m| + \sum_m I'_m |\xi_m - \xi'_m| \,. \end{aligned} \tag{27}$$

Here the last inequality uses $\xi_m \le \log C$. This completes the proof. □

A.3 Semantic Reweighting of Attention Evidence

The evidence distribution combines the realized attention allocation with source-wise semantic support. For $\alpha_m \ge 0$ and $s_m \ge 0$, we write

$$Z_T = \sum_{m=1}^{M} \alpha_m s_m > 0, \qquad T_m = \frac{\alpha_m s_m}{Z_T} \,. \tag{28}$$

Proposition A.5 (Support and odds of the evidence distribution)**.** *The vector $\boldsymbol{T}$ in Eq. equation 28 belongs to Δ^{M-1}, and $T_m > 0$ exactly when $\alpha_m s_m > 0$. For two positive-mass sources,*

$$\frac{T_i}{T_j} = \frac{\alpha_i}{\alpha_j} \frac{s_i}{s_j} \,. \tag{29}$$

Multiplying all attention weights, or all semantic scores, by a common positive factor preserves $\boldsymbol{T}$.

Proof. Every entry is nonnegative and $\sum_m T_m = Z_T^{-1} \sum_m \alpha_m s_m = 1$. Since $Z_T > 0$, each entry has the support of its numerator. Dividing two positive entries cancels Z_T and gives Eq. equation 29. A common positive scaling cancels between numerator and denominator. This completes the proof. □

Variational characterization. Let $\mathcal{J} = \{m : \alpha_m > 0\}$ and suppose $s_m > 0$ on $\mathcal{J}$. Define $\bar{\alpha}_m = \alpha_m / \sum_{n \in \mathcal{J}} \alpha_n$, and let $\Delta(\mathcal{J})$ be the probability simplex supported on $\mathcal{J}$. Then

$$\boldsymbol{T} = \operatorname*{arg\,max}_{\boldsymbol{w} \in \Delta(\mathcal{J})} \left\{ \sum_{m \in \mathcal{J}} w_m \log s_m - \operatorname{KL}(\boldsymbol{w} \| \bar{\boldsymbol{\alpha}}) \right\}. \tag{30}$$

Indeed, put $Z_s = \sum_{m \in \mathcal{J}} \bar{\alpha}_m s_m$. Substituting $T_m = \bar{\alpha}_m s_m / Z_s$ into the objective yields

$$\begin{aligned} \sum_{m \in \mathcal{J}} w_m \log s_m - \operatorname{KL}(\boldsymbol{w} \| \bar{\boldsymbol{\alpha}}) &= \sum_{m \in \mathcal{J}} w_m \log \frac{\bar{\alpha}_m s_m}{w_m} \\ &= \log Z_s - \operatorname{KL}(\boldsymbol{w} \| \boldsymbol{T}) \,. \end{aligned} \tag{31}$$

KL divergence is nonnegative and vanishes exactly at $\boldsymbol{w} = \boldsymbol{T}$, proving the unique maximizer. This variational form characterizes semantic reweighting of the realized attention allocation. SAR aggregates the source-write magnitudes $I_m(1 + s_m)$, while $\boldsymbol{T}$ records the normalized allocation $\alpha_m s_m / Z_T$.

A.4 Proof of Proposition 4.2

We next derive the visual preference induced by a realized FFN update. For fixed nonzero $\boldsymbol{u}$ and $\boldsymbol{v}_m$, write $r_m = \cos(\boldsymbol{u}, \boldsymbol{v}_m)$. With $\tau > 0$, the entropy-regularized objective is

$$F_\tau(\boldsymbol{w}) = \sum_{m=1}^{M} w_m (1 - r_m) + \tau \sum_{m=1}^{M} w_m \log w_m, \qquad \boldsymbol{w} \in \Delta^{M-1}, \tag{32}$$

where $0 \log 0 = 0$. The following result specializes the variational characterization of exponential-family distributions to this readout (Wainwright & Jordan, 2008).

Retrospect. We restate Proposition 4.2 below.

Proposition A.6 (Gibbs visual-preference readout)**.** *The unique minimizer of Eq. equation 32 is*

$$P_m = \frac{e^{r_m/\tau}}{Z_\tau}, \qquad Z_\tau = \sum_{n=1}^{M} e^{r_n/\tau}\,. \tag{33}$$

Proof. The vector in Eq. equation 33 is a strictly positive probability distribution. For any $\boldsymbol{w} \in \Delta^{M-1}$, substituting $\log P_m = r_m/\tau - \log Z_\tau$ gives

$$\begin{aligned} \tau\,\mathrm{KL}(\boldsymbol{w}\|\boldsymbol{P}) &= \tau \sum_m w_m \log w_m - \sum_m w_m r_m + \tau \log Z_\tau \\ &= F_\tau(\boldsymbol{w}) - 1 + \tau \log Z_\tau\,. \end{aligned} \tag{34}$$

Consequently,

$$F_\tau(\boldsymbol{w}) = 1 - \tau \log Z_\tau + \tau\,\mathrm{KL}(\boldsymbol{w}\|\boldsymbol{P})\,. \tag{35}$$

Since $\tau > 0$ and KL divergence vanishes exactly at $\boldsymbol{w} = \boldsymbol{P}$, this vector uniquely attains the minimum $1 - \tau \log Z_\tau$. This completes the proof. □

Entropy regularization yields a directional preference with positive mass on every visual source and greater mass on better-aligned references. EMS subsequently evaluates the transport cost between this preference and the attention evidence.

A.5 Readout Consistency and Temperature-dependent Preference

We relate the Gibbs preference to linear readouts of the additive FFN update and then examine the role of temperature. Throughout this subsection, $\boldsymbol{u}$ and all visual references are nonzero, and $\tau > 0$.

Retrospect. The following lemma restates Lemma 4.3.

Lemma A.7 (Readout consistency of the FFN visual preference)**.** *Let* $g_m(\boldsymbol{z}) = \langle \boldsymbol{z}, \widehat{\boldsymbol{v}}_m \rangle$ *and* $\Delta g_m = g_m(\boldsymbol{r} + \boldsymbol{u}) - g_m(\boldsymbol{r})$*. Then*

$$\log \frac{P_i}{P_j} = \frac{\Delta g_i - \Delta g_j}{\tau \|\boldsymbol{u}\|}\,. \tag{36}$$

Proof. The linearity of g_m gives

$$\begin{aligned} \Delta g_m &= \langle \boldsymbol{r} + \boldsymbol{u}, \widehat{\boldsymbol{v}}_m \rangle - \langle \boldsymbol{r}, \widehat{\boldsymbol{v}}_m \rangle \\ &= \langle \boldsymbol{u}, \widehat{\boldsymbol{v}}_m \rangle = \|\boldsymbol{u}\| r_m\,. \end{aligned} \tag{37}$$

The common normalizer in Eq. equation 33 cancels in the ratio, yielding

$$\log \frac{P_i}{P_j} = \frac{r_i - r_j}{\tau} = \frac{\Delta g_i - \Delta g_j}{\tau \|\boldsymbol{u}\|}\,. \tag{38}$$

This completes the proof. □

The identity connects relative visual preference to the difference between FFN-induced linear-readout increments. For fixed nonzero update and reference vectors, positive rescaling preserves the cosine scores and hence $\boldsymbol{P}$.

Proposition A.8 (Temperature, concentration, and entropy)**.** *For fixed finite scores* r_m*, let* $\mathcal{A} = \{m : r_m = \max_n r_n\}$*. The Gibbs preference satisfies*

$$\lim_{\tau \downarrow 0} P_m(\tau) = \begin{cases} 1/|\mathcal{A}|, & m \in \mathcal{A}, \\ 0, & m \notin \mathcal{A}, \end{cases} \qquad \lim_{\tau \to \infty} P_m(\tau) = \frac{1}{M}\,. \tag{39}$$

Furthermore,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}_{\boldsymbol{P}}[r] = -\frac{\mathrm{Var}_{\boldsymbol{P}}(r)}{\tau^2}, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{H}(\boldsymbol{P}) = \frac{\mathrm{Var}_{\boldsymbol{P}}(r)}{\tau^3} \geq 0\,. \tag{40}$$

Proof. Let $r_\star = \max_m r_m$. Dividing the numerator and denominator of Eq. equation 33 by $e^{r_\star/\tau}$ gives

$$P_m(\tau) = \frac{e^{(r_m - r_\star)/\tau}}{\sum_n e^{(r_n - r_\star)/\tau}}\,. \tag{41}$$

As $\tau \downarrow 0$, each maximizing term tends to one and each remaining term tends to zero. As $\tau \to \infty$, all exponential terms tend to one. These observations establish both limits.

For the derivatives, direct differentiation gives

$$\frac{\mathrm{d}P_m}{\mathrm{d}\tau} = -\frac{P_m}{\tau^2}\big(r_m - \mathbb{E}_{\boldsymbol{P}}[r]\big)\,. \tag{42}$$

Multiplying by r_m and summing proves the expectation derivative in Eq. equation 40. Also, $\mathcal{H}(\boldsymbol{P}) = \log Z_\tau - \mathbb{E}_{\boldsymbol{P}}[r]/\tau$ and $(\mathrm{d}/\mathrm{d}\tau)\log Z_\tau = -\mathbb{E}_{\boldsymbol{P}}[r]/\tau^2$. Therefore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{H}(\boldsymbol{P}) &= -\frac{\mathbb{E}_{\boldsymbol{P}}[r]}{\tau^2} - \frac{1}{\tau}\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}_{\boldsymbol{P}}[r] + \frac{\mathbb{E}_{\boldsymbol{P}}[r]}{\tau^2} \\ &= \frac{\mathrm{Var}_{\boldsymbol{P}}(r)}{\tau^3}\,.\end{aligned} \tag{43}$$

This completes the proof. □

Proposition A.9 (Sensitivity to directional-score perturbations)**.** *For $\tau > 0$ and finite score vectors $\boldsymbol{a}, \boldsymbol{b}$ on the same M sources, define $\boldsymbol{P}(\boldsymbol{a}) = \mathrm{softmax}(\boldsymbol{a}/\tau)$. Then*

$$\|\boldsymbol{P}(\boldsymbol{a}) - \boldsymbol{P}(\boldsymbol{b})\|_1 \le \min\left\{2, \frac{\|\boldsymbol{a} - \boldsymbol{b}\|_\infty}{\tau}\right\}. \tag{44}$$

Proof. Let $\boldsymbol{h} = \boldsymbol{b} - \boldsymbol{a}$ and $\boldsymbol{a}(t) = \boldsymbol{a} + t\boldsymbol{h}$ for $t \in [0,1]$. Along this segment, differentiating the softmax gives

$$\frac{\mathrm{d}P_m(\boldsymbol{a}(t))}{\mathrm{d}t} = \frac{P_m}{\tau}\left(h_m - \sum_n P_n h_n\right). \tag{45}$$

Here each expectation is taken with respect to $\boldsymbol{P}(\boldsymbol{a}(t))$. By Cauchy–Schwarz and $\mathrm{Var}_{\boldsymbol{P}}(h) \le \mathbb{E}_{\boldsymbol{P}}[h^2] \le \|\boldsymbol{h}\|_\infty^2$,

$$\begin{aligned}\sum_m \left|\frac{\mathrm{d}P_m}{\mathrm{d}t}\right| &= \frac{1}{\tau}\mathbb{E}_{\boldsymbol{P}}|h - \mathbb{E}_{\boldsymbol{P}}[h]| \\ &\le \frac{\sqrt{\mathrm{Var}_{\boldsymbol{P}}(h)}}{\tau} \le \frac{\|\boldsymbol{h}\|_\infty}{\tau}\,.\end{aligned} \tag{46}$$

Integration over $t \in [0,1]$ yields the second bound in Eq. equation 44. The ℓ_1 distance between two probability distributions is at most two, giving the first bound. This completes the proof. □

The temperature derivative characterizes the entropy increase of the preference distribution. The perturbation inequality supplies a uniform sensitivity bound proportional to $1/\tau$.

A.6 Existence and Decomposition of the EMS Optimum

We now examine the state-calibrated transport score. Fix a finite nonempty support Ω, nonzero visual references $\{\boldsymbol{v}_i\}_{i\in\Omega}$, and a nonzero prediction state $\boldsymbol{q}$. For nonzero vectors, define

$$\begin{aligned}d(\boldsymbol{x}, \boldsymbol{y}) &= 1 - \widehat{\boldsymbol{x}}^{\mathsf{T}}\widehat{\boldsymbol{y}}, \\ G_{ij} = d(\boldsymbol{v}_i, \boldsymbol{v}_j), &\qquad c_i = d(\boldsymbol{q}, \boldsymbol{v}_i), \qquad D_{ij} = G_{ij} + c_i + c_j\,.\end{aligned} \tag{47}$$

For probability marginals $\boldsymbol{p}, \boldsymbol{t}$ on Ω, let

$$\mathcal{U}(\boldsymbol{p}, \boldsymbol{t}) = \big\{\boldsymbol{\Pi} \ge 0 : \boldsymbol{\Pi}\mathbf{1} = \boldsymbol{p}, \quad \boldsymbol{\Pi}^{\mathsf{T}}\mathbf{1} = \boldsymbol{t}\big\}\,. \tag{48}$$

We write $\mathrm{OT}_d(\boldsymbol{p}, \boldsymbol{t})$ for the minimum with cost G, and $\mathrm{EMS}(\boldsymbol{p}, \boldsymbol{t}; \boldsymbol{q})$ for the minimum with cost D. For the proposed detector, $\boldsymbol{p} = \widetilde{\boldsymbol{P}}$ and $\boldsymbol{t} = \widetilde{\boldsymbol{T}}$. The following results use these fixed-reference settings.

Proposition A.10 (State calibration preserves the optimal matching)**.** *Both transport minima exist, and*

$$\mathrm{EMS}(\boldsymbol{p}, \boldsymbol{t}; \boldsymbol{q}) = \mathrm{OT}_d(\boldsymbol{p}, \boldsymbol{t}) + R_q, \qquad R_q = \sum_{i \in \Omega} (p_i + t_i) c_i \,. \tag{49}$$

For fixed marginals and support, D and G have the same set of minimizing couplings.

Proof. The feasible set contains $\boldsymbol{p}\boldsymbol{t}^\mathsf{T}$ and is closed. Its entries are nonnegative and sum to one, so it is also bounded. Compactness and continuity of the finite linear objectives ensure that both minima are attained. For any feasible $\boldsymbol{\Pi}$, the marginal constraints give

$$\begin{aligned} \sum_{i,j \in \Omega} \Pi_{ij} D_{ij} &= \sum_{i,j \in \Omega} \Pi_{ij} G_{ij} + \sum_i c_i \sum_j \Pi_{ij} + \sum_j c_j \sum_i \Pi_{ij} \\ &= \sum_{i,j \in \Omega} \Pi_{ij} G_{ij} + \sum_i p_i c_i + \sum_j t_j c_j \\ &= \sum_{i,j \in \Omega} \Pi_{ij} G_{ij} + R_q \,. \end{aligned} \tag{50}$$

The term R_q is constant over the feasible set. Minimizing gives Eq. equation 49, and addition of this constant preserves the set of minimizing couplings. This completes the proof. □

Dual interpretation. The geometric transport dual is (Peyré & Cuturi, 2019)

$$\mathrm{OT}_d(\boldsymbol{p}, \boldsymbol{t}) = \max_{\boldsymbol{\lambda}, \boldsymbol{\eta}} \left\{ \sum_i p_i \lambda_i + \sum_j t_j \eta_j : \lambda_i + \eta_j \le G_{ij} \right\}. \tag{51}$$

The transformation $\lambda_i^D = \lambda_i + c_i$ and $\eta_j^D = \eta_j + c_j$ is a bijection between the feasible dual pairs for G and D. It increases every dual objective by R_q. Thus, the primal and dual formulations give the same state-calibration term for fixed marginals and support.

A.7 Range, Zero Sets, and Reference Geometry of EMS

The cost in Eq. equation 47 couples the directions of two visual references with the prediction state. We use this shared geometry to characterize the range and agreement properties of EMS.

Proposition A.11 (Range and agreement properties)**.** *Under the fixed-reference settings above,*

$$0 \le D_{ij} \le \frac{9}{2}, \qquad 0 \le \mathrm{EMS} \le \frac{9}{2} \,. \tag{52}$$

EMS is symmetric in its probability arguments. For equal arguments,

$$\mathrm{EMS}(\boldsymbol{p}, \boldsymbol{p}; \boldsymbol{q}) = 2 \sum_i p_i d(\boldsymbol{q}, \boldsymbol{v}_i) \,. \tag{53}$$

Proof. Let $\boldsymbol{a} = \widehat{\boldsymbol{q}}$, $\boldsymbol{b} = \widehat{\boldsymbol{v}}_i$, and $\boldsymbol{c} = \widehat{\boldsymbol{v}}_j$. Expanding their squared sum and using their unit norms gives

$$\begin{aligned} D_{ij} &= 3 - \boldsymbol{a}^\mathsf{T}\boldsymbol{b} - \boldsymbol{a}^\mathsf{T}\boldsymbol{c} - \boldsymbol{b}^\mathsf{T}\boldsymbol{c} \\ &= \frac{9 - \|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|^2}{2} \,. \end{aligned} \tag{54}$$

The triangle inequality implies $0 \le \|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|^2 \le 9$, which proves the cost range. Every feasible transport objective is an average of these entries with total mass one, so its minimum has the same range.

Since $D_{ij} = D_{ji}$, transposition maps a feasible coupling for $(\boldsymbol{p}, \boldsymbol{t})$ to a coupling for $(\boldsymbol{t}, \boldsymbol{p})$ with the same cost, proving symmetry. When the marginals agree, the diagonal coupling $\boldsymbol{\Pi} = \mathrm{diag}(\boldsymbol{p})$ has zero geometric cost and is optimal for G. Eq. 49 then gives Eq. equation 53. This completes the proof. □

Even identical preference and evidence distributions receive a positive EMS when their positive-mass visual references are misaligned with the prediction state. The next result characterizes the corresponding zero sets through reference directions.

Proposition A.12 (Zero geometric cost and reference directions)**.** *Partition Ω into equivalence classes $\{\mathcal{C}_k\}$ under $i \sim j$ when $\widehat{\boldsymbol{v}}_i = \widehat{\boldsymbol{v}}_j$. Then*

$$\mathrm{OT}_d(\boldsymbol{p}, \boldsymbol{t}) = 0 \quad \Longleftrightarrow \quad \boldsymbol{p}(\mathcal{C}_k) = \boldsymbol{t}(\mathcal{C}_k) \quad \textit{for every } k\,. \tag{55}$$

Moreover, $\mathrm{EMS} = 0$ *exactly when every source with $p_i + t_i > 0$ satisfies $\widehat{\boldsymbol{v}}_i = \widehat{\boldsymbol{q}}$.*

Proof. For normalized references, $G_{ij} = \frac{1}{2}\|\widehat{\boldsymbol{v}}_i - \widehat{\boldsymbol{v}}_j\|^2$, so zero-cost pairs belong to the same equivalence class. An attained zero-cost optimum therefore transports all mass within classes. Summing its marginals over a class gives equal class masses.

Conversely, let $a_k = \boldsymbol{p}(\mathcal{C}_k) = \boldsymbol{t}(\mathcal{C}_k)$. Define the coupling

$$\Pi_{ij} = \begin{cases} p_i t_j / a_k, & i, j \in \mathcal{C}_k,\ a_k > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

Each positive-mass block has the prescribed row and column marginals; zero-mass classes contain only zero marginal entries. This coupling is feasible and has zero geometric cost.

Finally, both terms in Eq. equation 49 are nonnegative. A zero EMS forces $c_i = 0$ whenever $p_i + t_i > 0$, or equivalently $\widehat{\boldsymbol{v}}_i = \widehat{\boldsymbol{q}}$. Under this condition, all positive-mass references have the same direction, and both the geometric and endpoint costs vanish. This completes the proof. □

Proposition A.13 (Barycentric lower bound and nearby-reference matching)**.** *Define $\boldsymbol{\mu}_p = \sum_i p_i \widehat{\boldsymbol{v}}_i$ and $\boldsymbol{\mu}_t = \sum_i t_i \widehat{\boldsymbol{v}}_i$. Then*

$$\mathrm{EMS} \geq \frac{1}{2}\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\|^2 + 2 - \widehat{\boldsymbol{q}}^{\mathsf{T}}(\boldsymbol{\mu}_p + \boldsymbol{\mu}_t)\,. \tag{57}$$

If a partition $\{\mathcal{C}_k\}$ has equal class masses under $\boldsymbol{p}$ and $\boldsymbol{t}$, and $G_{ij} \leq \eta$ within every class, then $\mathrm{OT}_d(\boldsymbol{p}, \boldsymbol{t}) \leq \eta$.

Proof. For any feasible coupling, Jensen's inequality gives

$$\begin{aligned} \sum_{i,j} \Pi_{ij} G_{ij} &= \frac{1}{2} \sum_{i,j} \Pi_{ij} \|\widehat{\boldsymbol{v}}_i - \widehat{\boldsymbol{v}}_j\|^2 \\ &\geq \frac{1}{2} \left\| \sum_{i,j} \Pi_{ij} (\widehat{\boldsymbol{v}}_i - \widehat{\boldsymbol{v}}_j) \right\|^2 \\ &= \frac{1}{2} \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\|^2\,. \end{aligned} \tag{58}$$

The endpoint contribution is $R_q = 2 - \widehat{\boldsymbol{q}}^{\mathsf{T}}(\boldsymbol{\mu}_p + \boldsymbol{\mu}_t)$. Adding it and minimizing prove Eq. equation 57.

For the upper bound, apply the blockwise coupling in Eq. equation 56 to the stated partition. Each positive-mass pair costs at most η, and the total transported mass is one. Hence its objective, and therefore the optimum, is at most η. This completes the proof. □

Both bounds use distances between hidden representations to quantify reference similarity. Together they characterize displacement between weighted reference means and redistribution within nearby-reference groups.

A.8 Proof of Theorem 4.4

Theorem 4.4 relates geometric mismatch to preference mass assigned to references separated from the principal attention evidence. We use the shorthand $\boldsymbol{p}(B) = \sum_{i \in B} p_i$ and $\boldsymbol{t}(E) = \sum_{j \in E} t_j$ on the renormalized support Ω.

Retrospect. We restate the theorem using the notation introduced above.

Theorem A.14 (Geometric lower bound on evidence escape)**.** *Let $E \subseteq \Omega$ satisfy $\boldsymbol{t}(E) \geq 1 - \epsilon$, where $0 \leq \epsilon < 1$. Suppose $B \subseteq \Omega$ satisfies $G_{ij} \geq \delta > 0$ for every $i \in B$ and $j \in E$. Then*

$$\mathrm{EMS} \geq R_q + \delta[\boldsymbol{p}(B) - \epsilon]_+ , \tag{59}$$

where $[z]_+ = \max\{z, 0\}$.

Proof. For any feasible coupling, let $\Pi(B, E) = \sum_{i \in B, j \in E} \Pi_{ij}$ and $E^c = \Omega \setminus E$. The marginal constraints give

$$\begin{aligned} \Pi(B, E) &= \boldsymbol{p}(B) - \Pi(B, E^c) \\ &\geq \boldsymbol{p}(B) - \boldsymbol{t}(E^c) = \boldsymbol{p}(B) + \boldsymbol{t}(E) - 1 . \end{aligned} \tag{60}$$

Combining nonnegativity with $\boldsymbol{t}(E) \geq 1 - \epsilon$ yields

$$\Pi(B, E) \geq [\boldsymbol{p}(B) + \boldsymbol{t}(E) - 1]_+ \geq [\boldsymbol{p}(B) - \epsilon]_+ . \tag{61}$$

Each pair in $B \times E$ has geometric cost at least δ. Since all geometric costs are nonnegative,

$$\begin{aligned} \sum_{i,j \in \Omega} \Pi_{ij} G_{ij} &\geq \sum_{i \in B, j \in E} \Pi_{ij} G_{ij} \\ &\geq \delta \Pi(B, E) \geq \delta[\boldsymbol{p}(B) - \epsilon]_+ . \end{aligned} \tag{62}$$

Minimization over feasible couplings and addition of R_q establish Eq. equation 59. This completes the proof. □

Further discussion. Using the first bound in Eq. equation 61 gives the sharper expression

$$\mathrm{EMS} \geq R_q + \delta[\boldsymbol{p}(B) + \boldsymbol{t}(E) - 1]_+ . \tag{63}$$

The separation condition implies $B \cap E = \varnothing$, since $G_{ii} = 0$. For $\boldsymbol{p}(B) \leq \epsilon$, the bound in Eq. equation 59 reduces to R_q; for $\epsilon = 0$, all preference mass in B contributes at least δ per unit to the geometric cost.

The mass bound is attained by a two-reference construction. Take $G_{12} = G_{21} = \delta \in (0, 2]$, $E = \{1\}$, $B = \{2\}$, $\boldsymbol{p} = (1 - b, b)$, and $\boldsymbol{t} = (1 - \epsilon, \epsilon)$, with $0 \leq \epsilon \leq b \leq 1$ and $\epsilon < 1$. The coupling

$$\boldsymbol{\Pi} = \begin{pmatrix} 1 - b & 0 \\ b - \epsilon & \epsilon \end{pmatrix} \tag{64}$$

has the required marginals and geometric cost $\delta(b - \epsilon)$. This equals the lower bound. For any nonzero $\boldsymbol{q}$, the endpoint decomposition therefore gives $\mathrm{EMS} = R_q + \delta(b - \epsilon)$, establishing tightness.

A.9 Joint-support Restriction and Retained-mass Guarantees

We analyze the joint-support construction with a positive even budget K. Let $k = \min\{K/2, M\}$ and define

$$\Omega = \mathrm{Top}_k(\boldsymbol{P}) \cup \mathrm{Top}_k(\boldsymbol{T}), \qquad \eta_P = \boldsymbol{P}(\Omega), \quad \eta_T = \boldsymbol{T}(\Omega) . \tag{65}$$

Each top set contains exactly k indices, with a fixed rule for ties. The restricted distributions are $p_i = P_i/\eta_P$ and $t_i = T_i/\eta_T$ for $i \in \Omega$.

Proposition A.15 (Joint top-support coverage and exact restriction)**.** *The joint support satisfies*

$$k \leq |\Omega| \leq \min\{2k, M\}, \qquad \eta_P \geq \frac{k}{M}, \quad \eta_T \geq \frac{k}{M} . \tag{66}$$

Its renormalized distributions are

$$p_i = \frac{e^{r_i/\tau}}{\sum_{j \in \Omega} e^{r_j/\tau}}, \qquad t_i = \frac{\alpha_i s_i}{\sum_{j \in \Omega} \alpha_j s_j}, \qquad i \in \Omega . \tag{67}$$

Let $\boldsymbol{p}^\Omega$ and $\boldsymbol{t}^\Omega$ be their zero extensions to all M sources. Then

$$\|\boldsymbol{P} - \boldsymbol{p}^\Omega\|_1 = 2(1 - \eta_P), \qquad \|\boldsymbol{T} - \boldsymbol{t}^\Omega\|_1 = 2(1 - \eta_T) . \tag{68}$$

Proof. Each top set has k elements, which gives the stated cardinality bounds. For a sorted probability vector $a_1 \geq \cdots \geq a_M$,

$$1 = \sum_{i=1}^{M} a_i \leq \sum_{i=1}^{k} a_i + (M-k)a_k \leq \frac{M}{k} \sum_{i=1}^{k} a_i \,. \tag{69}$$

Hence its top-k mass is at least k/M. Each distribution contributes its own top set to Ω, so both retained masses have this lower bound and are positive. Substituting the definitions of P_i and T_i into $p_i = P_i/\eta_P$ and $t_i = T_i/\eta_T$ cancels their full-support normalizers, giving Eq. equation 67.

Within Ω, $p_i^{\Omega} = P_i/\eta_P \geq P_i$. Therefore,

$$\begin{aligned} \|\boldsymbol{P} - \boldsymbol{p}^{\Omega}\|_1 &= \sum_{i \in \Omega} \left(\frac{P_i}{\eta_P} - P_i \right) + \sum_{i \notin \Omega} P_i \\ &= (1-\eta_P) + (1-\eta_P) = 2(1-\eta_P) \,. \end{aligned} \tag{70}$$

The same calculation with $\boldsymbol{T}$ proves the second identity. This completes the proof. □

For $K = 32$, Ω is the union of two top-16 sets and contains at most 32 sources. Their overlap determines its actual size. The coverage bound is distribution-free; the following approximation result uses the actual retained masses.

Lemma A.16 (Stability of bounded-cost discrete transport)**.** *For a fixed finite cost matrix $0 \leq C_{ij} \leq B_C$ and probability marginals on the same finite index sets,*

$$\begin{aligned} &|\operatorname{OT}_C(\boldsymbol{p}, \boldsymbol{t}) - \operatorname{OT}_C(\boldsymbol{p}', \boldsymbol{t}')| \\ &\qquad \leq B_C \big[\operatorname{TV}(\boldsymbol{p}, \boldsymbol{p}') + \operatorname{TV}(\boldsymbol{t}, \boldsymbol{t}') \big], \end{aligned} \tag{71}$$

where $\operatorname{TV}(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_1$ and OT_C denotes the minimum with cost C.

Proof. Couple $\boldsymbol{p}$ and $\boldsymbol{p}'$ by placing $\min\{p_i, p_i'\}$ on each diagonal entry and coupling any positive residual mass by its normalized outer product. The residual marginal supports are disjoint. When the marginals agree, the diagonal coupling completes the construction. Applying the same construction to $\boldsymbol{t}, \boldsymbol{t}'$ gives pairs (X, X') and (Y, Y') with

$$\Pr(X \neq X') = \operatorname{TV}(\boldsymbol{p}, \boldsymbol{p}'), \qquad \Pr(Y \neq Y') = \operatorname{TV}(\boldsymbol{t}, \boldsymbol{t}') \,. \tag{72}$$

Start with an optimal pair (X, Y) for $(\boldsymbol{p}, \boldsymbol{t})$. Use these marginal couplings to draw X' conditional on X and Y' conditional on Y, independently given (X, Y). The resulting (X', Y') is feasible for $(\boldsymbol{p}', \boldsymbol{t}')$.

Write $\delta_p = \operatorname{TV}(\boldsymbol{p}, \boldsymbol{p}')$ and $\delta_t = \operatorname{TV}(\boldsymbol{t}, \boldsymbol{t}')$. The cost changes only when at least one index changes. Since all entries lie in $[0, B_C]$, the union bound gives

$$\begin{aligned} \operatorname{OT}_C(\boldsymbol{p}', \boldsymbol{t}') &\leq \mathbb{E}[C_{X'Y'}] \\ &\leq \mathbb{E}[C_{XY}] + B_C \Pr(X \neq X' \text{ or } Y \neq Y') \\ &\leq \operatorname{OT}_C(\boldsymbol{p}, \boldsymbol{t}) + B_C(\delta_p + \delta_t) \,. \end{aligned} \tag{73}$$

Exchanging the two marginal pairs proves the reverse inequality. This completes the proof. □

Corollary A.17 (Retained-mass bound for the EMS approximation)**.** *Fix all nonzero visual references and the nonzero state $\boldsymbol{q}$. Let $\mathrm{EMS}_{\mathrm{full}}$ use $(\boldsymbol{P}, \boldsymbol{T})$ on all M sources, and let EMS_{Ω} use their renormalized restrictions. Then*

$$|\,\mathrm{EMS}_{\mathrm{full}} - \mathrm{EMS}_{\Omega}\,| \leq \min\left\{ \frac{9}{2}, 4\big[(1-\eta_P) + (1-\eta_T)\big] \right\}. \tag{74}$$

Proof. The zero-extended marginals have zero mass on every omitted source, so their feasible couplings are supported on $\Omega \times \Omega$. Their full-matrix transport optimum thus equals the restricted optimum. Let $\varepsilon_{\Omega} = (1-\eta_P) + (1-\eta_T)$. Applying Lemma A.16 to $G_{ij} \in [0, 2]$ and using Eq. equation 68 gives

$$|\operatorname{OT}_d(\boldsymbol{P}, \boldsymbol{T}) - \operatorname{OT}_d(\boldsymbol{p}^{\Omega}, \boldsymbol{t}^{\Omega})| \leq 2\varepsilon_{\Omega} \,. \tag{75}$$

For probability vectors $\boldsymbol{a}, \boldsymbol{b}$, the signed mass $a_i - b_i$ has zero sum and equal positive and negative total magnitudes, each equal to $\mathrm{TV}(\boldsymbol{a}, \boldsymbol{b})$. Since $0 \le c_i \le 2$, it follows that

$$\left| \sum_i (a_i - b_i) c_i \right| \le 2\,\mathrm{TV}(\boldsymbol{a}, \boldsymbol{b})\,. \tag{76}$$

Applying this bound to both endpoint expectations gives $|R_q^{\mathrm{full}} - R_q^{\Omega}| \le 2\varepsilon_\Omega$. Consequently, Eq. equation 49 yields

$$\begin{aligned} |\,\mathrm{EMS}_{\mathrm{full}} - \mathrm{EMS}_{\Omega}\,| &\le 2\varepsilon_\Omega + 2\varepsilon_\Omega \\ &= 4\big[(1-\eta_P) + (1-\eta_T)\big]\,. \end{aligned} \tag{77}$$

Both EMS values belong to $[0, 9/2]$, which supplies the complementary bound $9/2$. This completes the proof. □

The result quantifies the approximation error through the omitted probability masses of the two source distributions. Increasing their retained masses tightens the resulting error guarantee.

B Experimental Details

B.1 Datasets

MS COCO (Lin et al., 2014) is a large-scale scene-understanding dataset of complex everyday images. Each image is annotated with instance-level segmentations over 80 object categories and five human-written reference captions, and typical scenes contain several objects in context rather than a single centered subject. Because the annotated category set is closed and instance-complete, the objects genuinely present in an image are known, which is what makes caption-level hallucination labeling possible: CHAIR (Rohrbach et al., 2018) compares the objects a model mentions against this annotated set, using a curated synonym list to map caption words to the 80 categories. COCO is the standard testbed for object hallucination in captioning, and its dense multi-object scenes are exactly the condition under which LVLMs are known to hallucinate co-occurring objects.

CLEVR (Johnson et al., 2017) is a synthetic diagnostic benchmark of rendered 3D scenes containing simple primitives—cubes, spheres, and cylinders in two sizes, two materials, and eight colors—with roughly one million questions over 100,000 images. Every question is generated from a functional program executed against the scene graph, so answers are exact and free of annotation noise; existence questions correspond to programs whose terminal function is `exist`. For our purposes CLEVR is a controlled complement to COCO: scene content is fully specified, no language prior links the objects, and visual grounding is the only route to a correct answer.

AMBER (Wang et al., 2024) is an LLM-free multi-dimensional hallucination benchmark of 1,004 images with fine-grained human annotations of objects, attributes, and inter-object relations. Its discriminative track contains 14,216 `Yes`/`No` questions spanning three query types—object existence, attribute, and relation—evaluated directly against the annotations without any LLM judge.

How each dataset is sampled, prompted, labeled, and split in our pipeline is specified in Appendix B.3.

B.2 Baselines

All baselines are evaluated under the same data and token-level evaluation protocol as our method. On MS COCO, they consume the identical generated captions, first-sub-token object positions, CHAIR-derived labels, and image-level 8:1:1 splits under the same three random seeds. On CLEVR and AMBER, all methods use the same question–answer instances, causal prediction positions, and data splits. For fair comparisons, we preserve the original feature construction and detector or scoring rule of each baseline rather than applying a shared detection head. Below, we restate their defining quantities in our notation. Let $A_{tm}^{\ell,h}$ denote the attention weight from prediction position t to visual token v_m at layer ℓ and head h, $\boldsymbol{h}_\ell(\cdot)$ the layer-ℓ hidden state, $\boldsymbol{W}_U$ the vocabulary output matrix, M the number of visual tokens, and H_ℓ the number of attention heads at layer ℓ.

MetaToken. MetaToken (Fieback et al., 2025) assesses hallucination using a meta-classifier built on token-level generation statistics. Specifically, it constructs a $10{+}H_L$-dimensional feature vector from the relative position and occurrence count of the object word, the per-head mean absolute visual attention $\frac{1}{M}\sum_m |A^{L,h}_{tm}|$ at the final decoder layer, and statistics of the generating distribution. The latter include the object log-probability $\sum_{i=s}^{e} \log p_i(y_i)$ over the object span $[s, e]$, a length-normalized sequence score, the vocabulary-wise variance of $\log p_s(\cdot)$, and

$$H_s = -\frac{1}{\log|\mathcal{C}|}\sum_{c\in\mathcal{C}} p_s(c)\log p_s(c), \qquad \mathrm{VR}_s = 1 - p_{(1)}, \qquad \mathrm{PM}_s = (1-p_{(1)}) + p_{(2)}, \tag{78}$$

where $p_{(1)}$ and $p_{(2)}$ are the two largest next-token probabilities, corresponding to normalized predictive entropy, variation ratio, and probability margin, respectively.

SVAR. SVAR (Jiang et al., 2025b) assesses hallucination by measuring how strongly a generated token attends to visual information. For layer ℓ and head h, its summed visual attention ratio is defined as

$$r_{\ell,h}(t) = \sum_{m\in\mathcal{V}} A^{\ell,h}_{tm}, \tag{79}$$

which measures the fraction of the attention row allocated to visual tokens. SVAR observes that this signal exhibits stronger hallucination–real separation in intermediate decoder layers. Following the original implementation, we extract $\{r_{\ell,h}(t)\}$ over the recommended intermediate-layer window and flatten the resulting matrix as the detection feature.

DHCP. DHCP (Zhang et al., 2025) detects hallucination directly from the cross-modal attention pattern associated with the generated token. Specifically, it averages visual attention across heads at each decoder layer and concatenates the resulting maps across layers:

$$\boldsymbol{d}_t = \Big[\tfrac{1}{H_1}\textstyle\sum_{h=1}^{H_1} A^{1,h}_{t,:};\ \ldots;\ \tfrac{1}{H_L}\sum_{h=1}^{H_L} A^{L,h}_{t,:}\Big], \tag{80}$$

which is then fed to its lightweight MLP detector. To accommodate LVLMs with different or dynamic visual-token grids, we resize each per-layer attention map to a fixed spatial grid using mass-preserving interpolation, with the same adaptation applied to all evaluated models.

TokenGrounding. TokenGrounding (Nguyen et al., 2026) characterizes hallucination through attention dispersion and cross-modal grounding consistency. Its Attention Dispersion Score measures whether visual attention fails to form a coherent spatial focus. With the top 10% of patches treated as foreground and grouped into spatially connected regions, it computes

$$\mathrm{ADS}_\ell = \big(1 - m^{\mathrm{fg}}_\ell\big)\cdot\frac{H^{\mathrm{bg}}_\ell}{\log M}, \tag{81}$$

where m^{fg}_ℓ is the attention mass captured by the foreground regions and H^{bg}_ℓ is the entropy of the renormalized background distribution. Its Cross-modal Grounding Consistency further measures whether the generated-token representation agrees with relevant visual representations:

$$\mathrm{CGC}_\ell = \frac{1}{k}\sum_{m\in\mathrm{TopK}_k} \cos\big(\boldsymbol{h}_\ell(q_t), \boldsymbol{h}_\ell(v_m)\big), \tag{82}$$

where $k{=}5$. Following the original method, the layer-wise ADS and CGC signals over the selected intermediate layers are used to construct the detection feature.

ProjectAway. ProjectAway (Jiang et al., 2025a) assesses hallucination by probing the object confidence encoded in intermediate visual representations through the logit lens. It projects visual hidden states at different decoder layers into the vocabulary space and defines

$$\mathrm{IC}(o) = \max_{\ell}\ \max_{m\in\mathcal{V}}\ \max_{w\in\mathrm{sub}(o)} \big[\mathrm{softmax}\big(\boldsymbol{h}_\ell(v_m)\boldsymbol{W}_U\big)\big]_w, \tag{83}$$

i.e., the largest probability assigned by any visual token at any decoder layer to any sub-token w of the object word o. The hallucination score is $1 - \mathrm{IC}(o)$. Different from our semantic recalibration, ProjectAway directly uses the vocabulary-level confidence as the detection signal, whereas our semantic score is used to recalibrate the corresponding attention evidence.

InsLen. InsLen (Lai et al., 2026) is a training-free detector that exploits visual information implicitly encoded in instruction-token representations. Let $\boldsymbol{z}_j$ denote the representation of the j-th instruction token and $\boldsymbol{h}_t$ the representation of the target object token. InsLen first measures the maximum target confidence carried by the instruction tokens and uses it to calibrate local visual evidence:

$$S_{\text{cafe}}(y_t) = \max_j \left[\text{softmax}\left(\boldsymbol{W}_U \boldsymbol{z}_j / \tau_{\text{Ins}}\right)\right]_{y_t}, \qquad S_{\text{cls}}(y_t) = S_{\text{cafe}}(y_t) \frac{1}{K_{\text{Ins}}} \sum_{k=1}^{K_{\text{Ins}}} \cos(\boldsymbol{h}_t, \boldsymbol{v}_k), \tag{84}$$

where τ_{Ins} is the temperature and $\{\boldsymbol{v}_k\}_{k=1}^{K_{\text{Ins}}}$ are the selected visual representations. InsLen further aggregates the m instruction representations assigning the highest probabilities to y_t as $\bar{\boldsymbol{z}} = \frac{1}{m}\sum_{n=1}^{m} \hat{\boldsymbol{z}}_n$, and measures their consistency with the generated object:

$$\begin{aligned} S_{\text{con}}(y_t) &= \alpha - \frac{\|\boldsymbol{h}_t - \bar{\boldsymbol{z}}\|}{\|\boldsymbol{h}_t\|}, \\ S_{\text{ccs}}(y_t) &= S_{\text{con}}(y_t) \frac{1}{m} \sum_{n=1}^{m} \left[\text{softmax}\left(\boldsymbol{W}_U \hat{\boldsymbol{z}}_n / \tau_{\text{Ins}}\right)\right]_{y_t}, \\ S_{\text{Ins}}(y_t) &= \omega S_{\text{cls}}(y_t) + (1-\omega) S_{\text{ccs}}(y_t). \end{aligned} \tag{85}$$

Here S_{ccs} measures context consistency and $\omega \in [0, 1]$ balances the two components. We follow the original implementation and directly use S_{Ins} for hallucination detection without an additional classifier.

B.3 Implementation Details

LVLM inference settings. All LVLMs are evaluated using greedy decoding. We generate at most 512 new tokens for MS COCO captioning and 8 new tokens for CLEVR and AMBER question answering.

Detector and hyperparameter selection. For each LVLM, we train an independent lightweight hallucination detector while keeping the backbone model frozen. The detector takes the layer-wise SAR and EMS signals as input. Feature normalization statistics are estimated exclusively from the training split and fixed thereafter. We employ a single-hidden-layer MLP and search over the hidden width and initial learning rate:

$$d_{\text{hid}} \in \{64, 128, 256\}, \qquad \eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}. \tag{86}$$

We adopt $d_{\text{hid}} = 128$ for the reported experiments, as it achieves the best validation performance among the candidate widths. The hidden layer is followed by ReLU and dropout with a rate of 0.3, and the linear layers are initialized with Kaiming initialization. We optimize the detector using Adam with a weight decay of 10^{-5}, a batch size of 256, and binary cross-entropy loss for at most 100 epochs. The learning rate is scheduled by `ReduceLROnPlateau` with a factor of 0.5 and a patience of 10 epochs. For each backbone and random seed, the initial learning rate and checkpoint are selected according to validation performance.

Training-based baseline detectors. For training-based baselines, we retain the classifier design and optimization protocol of the corresponding methods while using our unified data splits and target positions.

MetaToken (Fieback et al., 2025) evaluates Logistic Regression (LR) and Gradient Boosting (GB) meta-classifiers. Its LR classifier uses the `saga` solver with `max_iter`$= 1000$, tol $= 10^{-3}$, and `random_state`$= 0$; GB also uses `random_state`$= 0$ with the remaining scikit-learn defaults.

SVAR (Jiang et al., 2025b) employs a two-layer MLP trained with Adam $(\beta_1, \beta_2) = (0.9, 0.999)$ for 200 epochs. For each candidate intermediate-layer range, it searches

$$d_{\text{hid}} \in \{64, 128, 256, 512\}, \qquad \eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}, \tag{87}$$

and selects the corresponding detector configuration from the search.

DHCP (Zhang et al., 2025) uses a two-layer MLP that projects the flattened cross-modal attention feature to a 128-dimensional hidden representation and then to two output classes. The detector is

optimized using Adam with an initial learning rate of 10^{-3} and a batch size of 1024. Following its class-imbalance treatment, training samples are weighted inversely proportional to their class frequencies.

TokenGrounding (Nguyen et al., 2026) evaluates XGBoost (XGB), Random Forest (RF), and MLP classifiers, with hyperparameters selected by grid search:

$$\begin{aligned} \text{XGB}: &\quad d_{\max} \in \{4, 6, 8\}, \quad \eta \in \{0.1, 0.05\}, \quad N_{\text{est}} \in \{100, 200, 500\}, \\ \text{RF}: &\quad d_{\max} \in \{\text{None}, 10, 20\}, \quad N_{\text{tree}} \in \{200, 400, 600\}, \\ \text{MLP}: &\quad d_{\text{hid}} \in \{64, 128, 256\}, \quad \eta \in \{10^{-2}, 10^{-3}\}, \quad \text{opt} \in \{\text{Adam}, \text{SGD}\}. \end{aligned} \tag{88}$$

Data protocol. For MS COCO (Lin et al., 2014), we randomly sample $4{,}000$ images from `val2014` and reuse the identical image set for all evaluated LVLMs. Each model is prompted with `"Describe this image."` and generates at most 512 new tokens. Following CHAIR (Rohrbach et al., 2018), object mentions in each generated caption are matched against the annotated object set. A mentioned object contained in the annotation is labeled as grounded, and otherwise as hallucinated. For an object word consisting of multiple sub-tokens, we take the prediction position of its first sub-token as the detection target.

For CLEVR (Johnson et al., 2017), we construct **CLEVR-Exist-9K** with $9{,}000$ object-existence questions covering $8{,}629$ unique images. For AMBER (Wang et al., 2024), we use $14{,}216$ questions over $1{,}004$ images, including $4{,}924$ existence, $7{,}628$ attribute, and $1{,}664$ relation questions. For VQA samples, internal signals are extracted at the causal position predicting the first answer token, and the detection label is determined by whether the generated answer matches the ground-truth answer. We perform image-level $8{:}1{:}1$ train–validation–test splits with three random seeds for all datasets, assigning all samples associated with the same physical image to the same split.

Object-mention labeling. For MS COCO captioning, we follow the standard CHAIR synonym list, lemmatization, and compound-word handling to map generated object mentions to the 80 COCO object categories. Each matched mention is associated with its character span in the generated caption and the corresponding indices in the generated response tokens. For each canonical object, we retain its first occurrence in the caption as one detection sample. The aligned first-occurrence samples are subsequently used for feature extraction.

Exact token localization. We align the generated response token IDs with their character offsets in the decoded response. For tokenizers providing reliable offset mappings, we directly use the returned offsets; for SentencePiece-based tokenizers, the token intervals are recovered through progressive decoding. Let i denote the response index of the first sub-token of a target object. We construct the model input as

$$\text{Prompt} \oplus \text{Response}_{<i}, \tag{89}$$

and read the internal states at the final causal position, corresponding to the prediction state that generates the target sub-token at index i. Character-span alignment therefore provides the response index for each detection sample. The same resolved target positions are shared by all compared methods.

Feature extraction. For each target token, we extract the internal quantities required by SAR and EMS from all decoder layers in a single forward pass. For SAR, we first recover the attention write from each visual source to the current prediction state and measure its intensity by

$$I_{tm}^{\ell} = \left\| \boldsymbol{a}_{tm}^{\ell} \right\|. \tag{90}$$

The target semantics of visual source m are read through the Logit Lens defined in the main paper. Given its layer-ℓ representation $\boldsymbol{v}_m^{\ell}$, we apply the model-specific final normalization N_f and vocabulary output matrix $\boldsymbol{W}_U$:

$$\beta_{tm}^{\ell} = \left[\text{softmax}_{\text{vocab}} \left(\boldsymbol{W}_U N_f(\boldsymbol{v}_m^{\ell}) \right) \right]_{y_t}. \tag{91}$$

The resulting β_{tm}^{ℓ} is transformed into the semantic score s_{tm}^{ℓ} and combined with I_{tm}^{ℓ} to obtain SAR as defined in the main paper.

For EMS, we construct the target-conditioned attention-evidence distribution $\boldsymbol{T}$ and the FFN-induced visual-preference distribution $\boldsymbol{P}$. The attention allocation is obtained by averaging visual

Table 5: Classification results (%) on LLaVA-1.5-7B and Qwen2.5-VL-7B.

	LLaVA-1.5-7B				Qwen2.5-VL-7B			
Method	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑
MetaToken	87.12	92.02	89.50	83.93	<u>90.30</u>	96.51	<u>93.30</u>	<u>87.95</u>
SVAR	<u>88.43</u>	90.95	<u>89.67</u>	<u>84.40</u>	**91.42**	93.85	92.61	86.98
ProjectAway	77.11	**96.77**	85.83	76.21	86.96	**100.00**	93.03	86.96
TokenGrounding	82.98	<u>92.67</u>	87.55	80.38	88.22	98.49	93.07	87.25
InsLen	83.43	92.65	87.80	80.83	86.95	<u>99.93</u>	92.99	86.90
Ours	**91.61**	88.63	**90.09**	**85.18**	89.24	98.53	**93.65**	**88.22**

Table 6: Classification results (%) on InternVL2.5-8B and Qwen3-VL-8B.

	InternVL2.5-8B				Qwen3-VL-8B			
Method	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑
MetaToken	87.24	94.73	<u>90.83</u>	<u>84.76</u>	86.61	93.45	89.90	83.61
SVAR	**89.87**	90.57	90.22	84.35	<u>88.34</u>	91.94	<u>90.10</u>	<u>84.23</u>
ProjectAway	82.70	<u>96.91</u>	89.24	81.39	78.28	<u>99.68</u>	87.69	78.16
TokenGrounding	84.11	95.04	89.24	81.74	82.22	95.65	88.43	80.45
InsLen	80.48	**99.20**	88.87	80.20	78.07	**100.00**	87.69	78.07
Ours	<u>88.66</u>	93.33	**90.93**	**85.11**	**89.40**	93.73	**91.50**	**86.34**

attention over heads and is calibrated by s_{tm}^{ℓ}. The Gibbs temperature for the FFN preference is fixed to $\tau = 0.03$. We set the joint principal-support budget to $K = 32$. The Top-$K/2$ visual sources are selected independently from $\boldsymbol{P}$ and $\boldsymbol{T}$, and their index union forms the support Ω. The transport cost is constructed on Ω, and the resulting optimal transport problem is solved using the network simplex algorithm.

Layer-wise detection features. SAR and EMS are extracted from all decoder layers. For an LVLM with L decoder layers, the final detector input is

$$\boldsymbol{f}_t = \left[\mathrm{EMS}_t^1, \ldots, \mathrm{EMS}_t^L; \mathrm{SAR}_t^1, \ldots, \mathrm{SAR}_t^L\right]^\top \in \mathbb{R}^{2L}. \tag{92}$$

Accordingly, the feature dimensions are 64, 56, 64, and 72 for LLaVA-1.5-7B, Qwen2.5-VL-7B, InternVL2.5-8B, and Qwen3-VL-8B, respectively.

Computational setup. All LVLM parameters remain frozen during feature extraction, and only the lightweight hallucination detector is optimized. All experiments are conducted using two NVIDIA RTX 4090 GPUs.

C Complementary Quantitative Studies

C.1 Additional Detection Metrics on COCO

Tables 5–7 provide the complete Precision, Recall, F1, and Accuracy results. Precision, Recall, and F1 are computed for the real class, while Accuracy denotes the overall classification accuracy. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

Our method achieves the highest F1 and Accuracy on all four models, with an average F1 of 91.54% and an average Accuracy of 86.21%. Several baselines obtain higher Recall at the cost of lower Precision, whereas our method maintains a better Precision–Recall balance and shows consistent performance across different LVLMs.

C.2 Distribution Analysis of Layer-Averaged EMS and SAR

Fig. 4 visualizes the distributions of EMS and SAR after averaging each feature across decoder layers. Across all four LVLMs, hallucinated samples exhibit larger EMS values, reflecting stronger dis-

Table 7: Average classification results (%) across four LVLMs.

Method	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑
MetaToken	87.82	94.18	<u>90.88</u>	<u>85.06</u>
SVAR	<u>89.52</u>	91.83	90.65	84.99
ProjectAway	81.26	**98.34**	88.95	80.68
TokenGrounding	84.38	95.46	89.57	82.46
InsLen	82.23	<u>97.95</u>	89.34	81.50
Ours	**89.73**	93.56	**91.54**	**86.21**

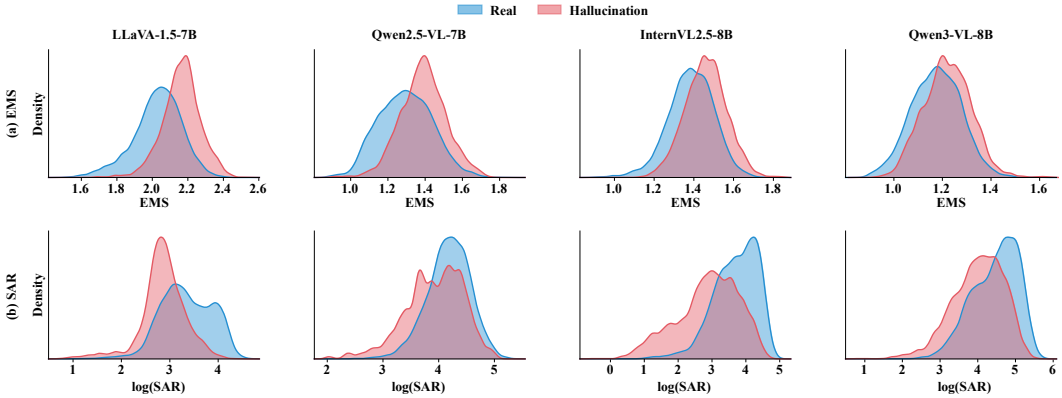


Figure 4: Distributions of layer-averaged EMS and $\log(\mathrm{SAR})$ across four LVLMs. Hallucinated samples exhibit larger EMS, whereas real samples exhibit larger $\log(\mathrm{SAR})$, revealing complementary distributional patterns between evidence–update mismatch and attention-evidence reliability.

agreement between attention evidence and the subsequent FFN preference. In contrast, real samples exhibit larger $\log(\mathrm{SAR})$, indicating stronger semantic support from the attended visual evidence. The two features therefore capture complementary aspects of the prediction process: EMS measures evidence–update mismatch, while SAR characterizes the reliability of attention evidence.

C.3 Supplementary Analysis of Visual WRITE and Semantic Support

Figures 5 and 6 further compare the target-related visual semantic support and visual WRITE intensity between real and hallucinated tokens. We obtain two key observations.

First, real tokens exhibit clearly stronger target-related visual semantic support in middle-to-late layers. At early decoder layers, both real and hallucinated tokens receive relatively weak semantic support from visual representations. As the representations evolve through the decoder, the visual support for real tokens increases substantially and becomes clearly separated from that for hallucinated tokens. This trend is particularly evident on LLaVA-1.5-7B, InternVL2.5-8B, and Qwen3-VL-8B, and is also observed in the final layers of Qwen2.5-VL-7B. These results indicate that grounded generation is accompanied by progressively strengthened visual semantics that are consistent with the current prediction y_t.

Second, visual WRITE intensity shows a weaker and less stable real–hallucination distinction. Across most models and layers, real tokens generally have larger I_t^l, indicating that the visual WRITE itself carries useful grounding information. However, the two groups exhibit highly similar layer-wise variations with substantial overlap. For example, the WRITE intensities of real and hallucinated tokens evolve almost synchronously on InternVL2.5-8B and Qwen3-VL-8B. On LLaVA-1.5-7B, their values further approach each other in several late layers and occasionally exhibit a reversed relation.

These observations indicate that a stronger visual WRITE does not necessarily correspond to more reliable visual evidence. The WRITE intensity primarily reflects the magnitude of visual infor-

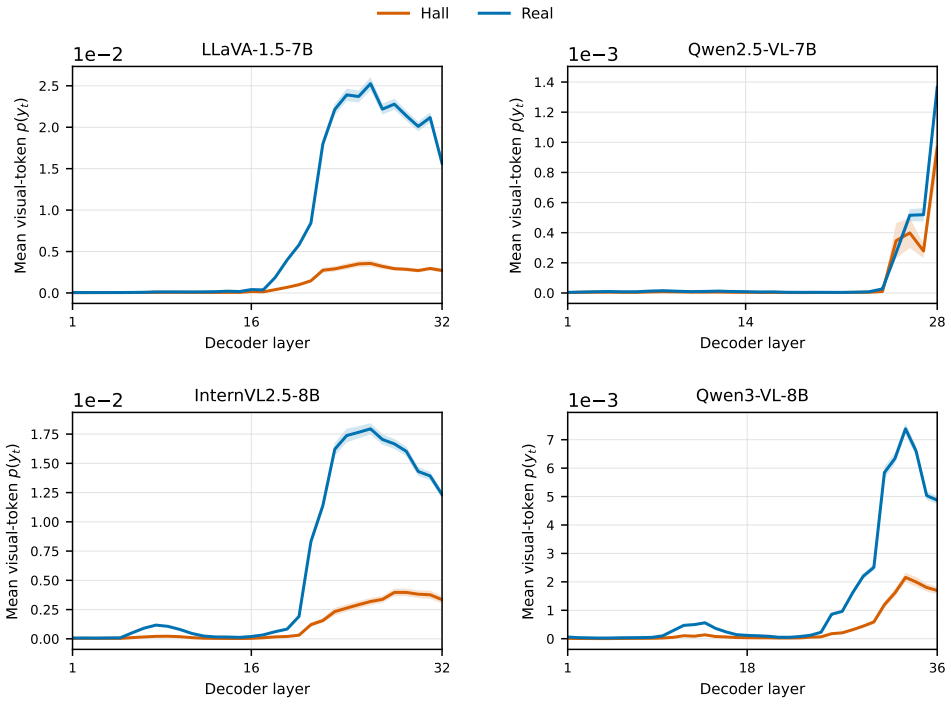


Figure 5: **Layer-wise visual semantic support for real and hallucinated tokens.** Real tokens receive clearly stronger target-related visual support in middle-to-late decoder layers, while the semantic support of hallucinated tokens remains relatively weak.

mation written into the current prediction state, but does not determine whether such information is semantically consistent with y_t. Consequently, raw WRITE intensity may conflate evidence strength with target relevance.

Together with the clearer semantic separation in Fig. 5, these findings further support the semantic recalibration adopted by SAR in Sec. 4.1: visual WRITE provides the evidence intensity, whereas target-conditioned semantics further identifies the visual evidence relevant to the current prediction.

C.4 Target-Semantic Recalibration in SAR

We investigate the contribution of target-semantic recalibration by replacing SAR with the original visual-write magnitude $I = \sum_{m\in\mathcal{V}} I_m$, where $I_m = \|\boldsymbol{a}_m\|$, while keeping EMS and the detector unchanged. Table 8 reports the comparative results. SAR improves AUROC by 1.05, 0.86, 1.34, and 0.52 percentage points on LLaVA-1.5, Qwen2.5-VL, InternVL2.5, and Qwen3-VL, respectively. AUPR and F1 also improve on LLaVA-1.5, InternVL2.5, and Qwen3-VL. These findings indicate that target-semantic support provides discriminative information beyond write magnitude. This supports Intuition 3.1: visual evidence should be assessed jointly by how strongly each source is written and how well it supports the target. The source-wise semantic increment in SAR implements this joint assessment.

C.5 Target Construction for EMS

Target variants. We examine the visual weighting, residual contribution, and normalization of the EMS evidence target, keeping SAR, the FFN-preference distribution, the transport cost, and the detector fixed. Let $A_m = \alpha_m / \sum_{n\in\mathcal{V}} \alpha_n$ be the visually normalized mean-head attention, with s_m defined in the main text. For a nonnegative score vector x, define

$$\mathcal{N}(x)_m = \frac{x_m}{\sum_{n\in\mathcal{V}} x_n}, \qquad \mathrm{SM}(x)_m = \frac{\exp(x_m)}{\sum_{n\in\mathcal{V}} \exp(x_n)}. \tag{93}$$

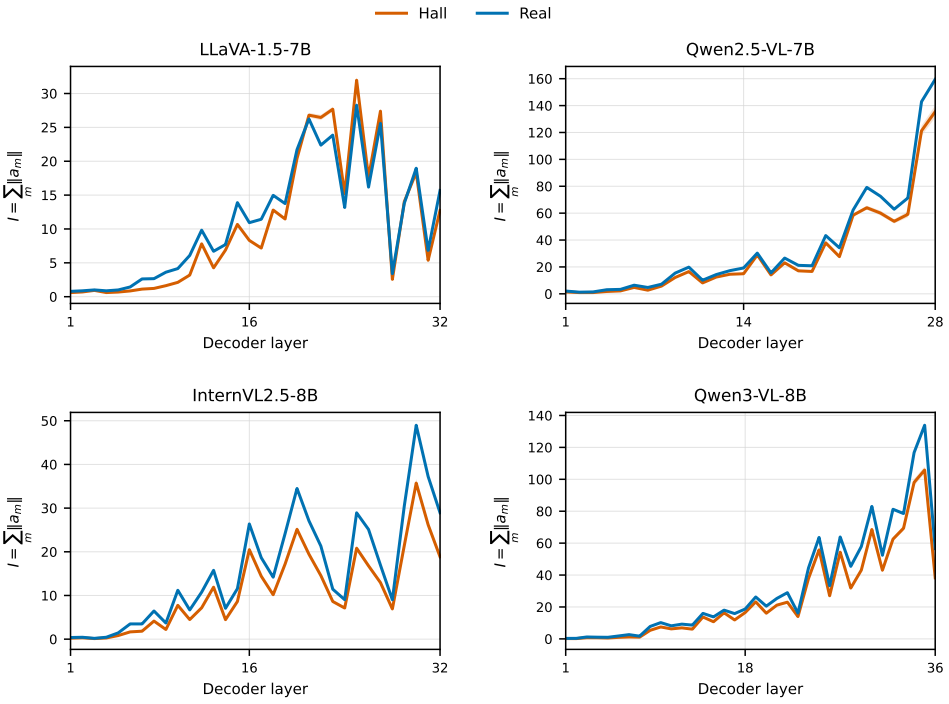


Figure 6: **Layer-wise visual WRITE intensity I_t^l for real and hallucinated tokens.** Real tokens generally exhibit stronger visual WRITE, while the two groups share highly similar layer-wise variations with considerable overlap and occasional local reversals.

Table 8: Ablation of target-semantic recalibration in SAR. Replacing SAR with I removes its semantic increment while retaining the same EMS.

Features	LLaVA-1.5			Qwen2.5-VL		
	AUC↑	AUPR↑	F1↑	AUC↑	AUPR↑	F1↑
I + EMS	90.47	96.58	89.15	84.74	**97.27**	92.78
SAM (Ours)	**91.52**	**97.06**	**90.09**	**85.60**	97.26	**93.65**
Features	InternVL2.5			Qwen3-VL		
	AUC↑	AUPR↑	F1↑	AUC↑	AUPR↑	F1↑
I + EMS	87.47	96.41	90.32	89.79	96.57	90.40
SAM (Ours)	**88.81**	**96.85**	**90.94**	**90.31**	**96.93**	**91.47**

All metrics are reported in percent. Bold marks the best result for each model and metric.

We use a uniform distribution for $\mathcal{N}(\mathbf{0})$. Here, Is denotes $(I_m s_m)_{m\in\mathcal{V}}$, and similarly for the other score vectors. The default target satisfies $\mathcal{N}(As) = \mathcal{N}(\alpha s)$. Softmax acts directly on the token scores at target temperature one. All targets are constructed over all visual tokens before the same joint-support restriction, with $K = 32$ and source temperature $\tau = 0.03$.

Visual weighting. Table 9 shows that replacing A_m with I_m reduces AUC on all four models by 0.03–0.36 percentage points. AUPR and F1 also decrease on three models, with the largest F1 drop of 0.92 points on Qwen2.5-VL. These results support using attention to describe relative source selection in EMS, while retaining write magnitude in SAR's cumulative evidence readout.

Residual contribution. Adding the unit residual to $\mathcal{N}(Is)$ reduces AUC and AUPR on three models. On Qwen2.5-VL, AUC, AUPR, and F1 decrease by 0.77, 0.17, and 0.43 points, respectively.

Table 9: Ablation of the evidence target in EMS. The variants change the visual weighting, residual contribution, and normalization of the target scores, with SAR fixed.

EMS target	LLaVA-1.5			Qwen2.5-VL		
	AUC↑	AUPR↑	F1↑	AUC↑	AUPR↑	F1↑
$\mathcal{N}(As)$ (Ours)	**91.52**	97.06	90.09	**85.60**	**97.26**	**93.65**
$\mathcal{N}(Is)$	91.47	**97.08**	90.01	85.24	97.24	92.73
$\mathcal{N}\big(I(1+s)\big)$	91.20	96.96	89.98	84.47	97.07	92.30
$\mathrm{SM}(As)$	91.46	97.03	**90.16**	85.47	97.25	92.70
$\mathrm{SM}\big(I(1+s)\big)$	91.10	96.94	89.75	84.41	97.01	92.77
EMS target	InternVL2.5			Qwen3-VL		
	AUC↑	AUPR↑	F1↑	AUC↑	AUPR↑	F1↑
$\mathcal{N}(As)$ (Ours)	**88.81**	**96.85**	**90.94**	**90.31**	96.93	**91.47**
$\mathcal{N}(Is)$	88.57	96.77	90.75	90.28	96.90	91.43
$\mathcal{N}\big(I(1+s)\big)$	88.53	96.76	90.79	90.30	96.91	91.31
$\mathrm{SM}(As)$	88.47	96.80	90.68	90.26	**96.94**	91.30
$\mathrm{SM}\big(I(1+s)\big)$	87.65	96.54	90.41	89.35	96.70	90.59

All metrics are reported in percent. Bold marks the best result for each model and metric.

For $I > 0$, the residual target before support restriction can be written as

$$\mathcal{N}\big(I(1+s)\big)_m = \frac{I}{\mathrm{SAR}}\frac{I_m}{I} + \frac{\mathrm{SAR}-I}{\mathrm{SAR}}\mathcal{N}(Is)_m. \tag{94}$$

Thus, the residual mixes a write-only distribution into the target-modulated evidence. These results support distinguishing cumulative evidence strength from relative evidence allocation: SAR retains the original write in its cumulative readout, whereas EMS uses target-modulated scores rather than normalized SAR contributions as its evidence reference.

Score normalization. For $A_m s_m$, sum normalization yields higher AUC than softmax on all four models, with gains of 0.06, 0.13, 0.34, and 0.05 points on LLaVA-1.5, Qwen2.5-VL, InternVL2.5, and Qwen3-VL, respectively. It also yields higher F1 on Qwen2.5-VL, InternVL2.5, and Qwen3-VL, by 0.95, 0.26, and 0.17 points, respectively. For $I_m(1+s_m)$, replacing sum normalization with softmax reduces AUC and AUPR on all four models. Sum normalization preserves score proportions and is invariant to common positive rescaling, retaining the relative strengths established by attention selection and target-semantic support. Overall, $\boldsymbol{T} = \mathcal{N}(As)$ achieves the highest AUC among the evaluated targets on all four models, supporting the proportional attention–semantic assignment as the evidence reference for FFN-update comparison.

C.6 Hyperparameter Robustness

Setup. We study the sensitivity of EMS to two key hyperparameters: the transport-support budget K and the FFN visual-preference temperature τ. Here, K controls the size of the joint support used for optimal transport, while τ controls the concentration of the FFN-induced visual-preference distribution. We consider $K \in \{16, 32, 64, 128\}$ and $\tau \in \{0.02, 0.03, \ldots, 0.08\}$, while keeping all other settings fixed. We use $K = 32$ and $\tau = 0.03$ as the shared default configuration for all four LVLMs.

Results. Tables 10 and 11 report AUROC / F1 under different combinations of K and τ. We make three observations. First, EMS remains stable over a broad range of configurations. Across the full parameter grid, the maximum AUROC variations are only 0.20, 0.64, 0.19, and 0.66 percentage points on LLaVA-1.5-7B, Qwen2.5-VL-7B, InternVL2.5-8B, and Qwen3-VL-8B, respectively; the corresponding F1 variations are at most 0.60, 0.56, 0.45, and 0.62 points. Second, increasing K brings only modest and model-dependent changes. At $\tau = 0.03$, increasing K from 32 to 128 changes the four-model average AUROC from 89.06% to 89.13% and F1 from 91.36% to 91.33%, while increasing the maximum number of pairwise transport-cost entries from $32^2 = 1{,}024$ to $128^2 = 16{,}384$, i.e., by $16\times$. Third, different LVLMs exhibit slightly different preferences for τ, yet

Table 10: Hyperparameter sensitivity of EMS on LLaVA-1.5-7B and Qwen2.5-VL-7B. Each entry reports AUROC / F1 (%). The shaded entry denotes the shared default configuration $(K, \tau) = (32, 0.03)$.

LLaVA-1.5-7B							
$K \backslash \tau$	0.02	0.03	0.04	0.05	0.06	0.07	0.08
16	91.49 / 90.22	91.48 / 90.00	91.47 / 90.10	91.51 / 90.33	91.45 / 90.31	91.44 / 90.33	91.39 / 90.28
32	91.53 / 90.13	91.52 / 90.09	91.52 / 89.98	91.47 / 89.95	91.48 / 90.00	91.47 / 89.97	91.45 / 89.92
64	91.50 / 89.87	91.59 / 89.77	91.52 / 89.78	91.50 / 89.88	91.49 / 89.89	91.51 / 89.90	91.47 / 89.84
128	91.49 / 89.98	91.48 / 89.73	91.53 / 89.78	91.53 / 89.88	91.54 / 89.96	91.52 / 90.04	91.53 / 90.03
Qwen2.5-VL-7B							
$K \backslash \tau$	0.02	0.03	0.04	0.05	0.06	0.07	0.08
16	85.75 / 92.94	85.82 / 92.96	85.74 / 93.25	85.61 / 93.31	85.52 / 93.33	85.48 / 93.23	85.42 / 93.21
32	85.63 / 93.14	85.60 / 93.65	85.47 / 93.29	85.38 / 93.37	85.35 / 93.35	85.25 / 93.20	85.18 / 92.98
64	85.68 / 93.10	85.62 / 93.09	85.68 / 93.38	85.49 / 93.10	85.51 / 93.16	85.52 / 92.92	85.41 / 92.82
128	85.57 / 93.10	85.56 / 93.16	85.48 / 93.27	85.39 / 93.07	85.49 / 93.12	85.58 / 93.01	85.64 / 93.03

Table 11: Hyperparameter sensitivity of EMS on InternVL2.5-8B and Qwen3-VL-8B, together with the four-model average. Each entry reports AUROC / F1 (%). The shaded entry denotes the shared default configuration $(K, \tau) = (32, 0.03)$.

InternVL2.5-8B							
$K \backslash \tau$	0.02	0.03	0.04	0.05	0.06	0.07	0.08
16	88.77 / 90.90	88.74 / 90.82	88.78 / 91.10	88.70 / 91.01	88.70 / 91.12	88.66 / 91.09	88.66 / 90.88
32	88.85 / 90.94	88.81 / 90.94	88.78 / 90.91	88.77 / 91.12	88.78 / 91.03	88.74 / 91.07	88.80 / 91.23
64	88.84 / 91.03	88.79 / 91.12	88.72 / 90.97	88.75 / 90.82	88.71 / 90.93	88.72 / 90.86	88.70 / 90.81
128	88.83 / 90.98	88.81 / 90.98	88.78 / 91.00	88.78 / 91.06	88.85 / 90.90	88.85 / 91.13	88.80 / 90.78
Qwen3-VL-8B							
$K \backslash \tau$	0.02	0.03	0.04	0.05	0.06	0.07	0.08
16	90.21 / 91.47	90.13 / 91.42	90.11 / 91.25	90.09 / 91.22	90.05 / 91.13	90.03 / 91.28	90.02 / 91.21
32	90.34 / 91.62	90.31 / 91.47	90.26 / 91.24	90.28 / 91.15	90.34 / 91.15	90.32 / 91.12	90.26 / 91.00
64	90.42 / 91.48	90.42 / 91.27	90.48 / 91.26	90.47 / 91.24	90.50 / 91.13	90.48 / 91.09	90.49 / 91.10
128	90.52 / 91.52	90.68 / 91.43	90.62 / 91.47	90.66 / 91.48	90.67 / 91.47	90.61 / 91.41	90.60 / 91.33
Four-model Average							
$K \backslash \tau$	0.02	0.03	0.04	0.05	0.06	0.07	0.08
16	89.06 / 91.38	89.04 / 91.30	89.02 / 91.42	88.98 / 91.47	88.93 / 91.47	88.90 / 91.48	88.87 / 91.40
32	89.09 / 91.46	89.06 / 91.36	89.01 / 91.36	88.98 / 91.40	88.99 / 91.38	88.94 / 91.34	88.92 / 91.28
64	89.11 / 91.37	89.10 / 91.31	89.10 / 91.35	89.05 / 91.26	89.05 / 91.28	89.06 / 91.19	89.02 / 91.14
128	89.10 / 91.40	89.13 / 91.33	89.10 / 91.38	89.09 / 91.37	89.14 / 91.36	89.14 / 91.40	89.14 / 91.29

the overall performance varies only mildly within the examined range of 0.02–0.08. Taken together, these results indicate that EMS is insensitive to K and τ within reasonable ranges. Accordingly, we set $K = 32$ and $\tau = 0.03$ by default, which achieves a favorable balance between stable detection performance and compact transport support.

C.7 Cross-Layer Dynamics of SAR and EMS

Layer-range ablation. Tab. 12 reports the comparative results of SAR, EMS, and their combination over different decoder-layer ranges, where we have three major observations. (1) SAR maintains clear discriminability across different layer ranges, with Layers 9–18 achieving the strongest performance among the individual ranges; further aggregating all decoder layers substantially improves AUC, AUPR, and F1. This indicates that the visual-evidence cues captured by SAR are

Table 12: Ablation study of SAR, EMS, and their combination over different decoder-layer ranges.

Layer Range	SAR			EMS			SAM		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
Layers 1–8	82.29	92.78	86.54	75.50	89.89	85.37	83.56	93.60	86.80
Layers 9–18	85.81	94.76	87.34	77.38	90.86	85.74	86.94	95.28	87.56
Layers 19–27	81.53	92.78	86.50	81.67	92.61	87.47	86.00	94.72	88.07
Layers 28–32	84.25	93.84	87.25	83.03	92.55	87.92	87.87	95.42	88.63
All layers	**89.98**	**96.48**	**89.24**	**87.34**	**95.24**	**88.34**	**91.52**	**97.06**	**90.09**

broadly distributed across decoder layers rather than concentrated within a local stage. (2) EMS exhibits an evident depth-enhanced discriminative trend, especially with AUC and F1 progressively increasing from early to late layers. This indicates that the mismatch between visual evidence and FFN preference becomes increasingly discriminative as higher-level representations are formed. (3) SAR+EMS consistently outperforms the corresponding individual branches across all layer ranges and achieves the best overall results when all layers are used. This phenomenon further indicates the complementary cross-layer properties of SAR and EMS: the former provides broadly distributed visual-evidence cues, while the latter exhibits stronger discriminability in deeper representations. Integrating the full decoder hierarchy therefore enables more effective use of their complementary information.

D Qualitative Analysis

We further inspect the spatial behaviors of raw attention (A), the semantically calibrated visual target (T), and the FFN-derived source preference (P) for representative object mentions. Each example shows the input image, the corresponding response excerpt, and the spatial responses from layers 15, 18, 21, 24, and 27. Each heatmap retains the top-16 visual patches and is independently normalized for visualization.

D.1 Individual Object Mentions

We further present individual object mentions across different LVLMs. For real tokens, T generally concentrates on the corresponding visible objects, with P preserving similar spatial support. For hallucinated tokens, the responses instead concentrate on regions without a clear visual counterpart for the generated object, while P often exhibits a shifted spatial preference. These examples visualize the different spatial evolution of visual evidence for real and hallucinated predictions through the decoder.

D.2 Paired Object Mentions

We compare real and hallucinated object mentions generated from the same image and response. The paired visualization highlights the differences in where T concentrates visual evidence and whether the subsequent preference P preserves or shifts this spatial support.



Figure 7: Visualization of real object mentions in LLaVA-1.5-7B.



Figure 8: Visualization of hallucinated object mentions in LLaVA-1.5-7B.

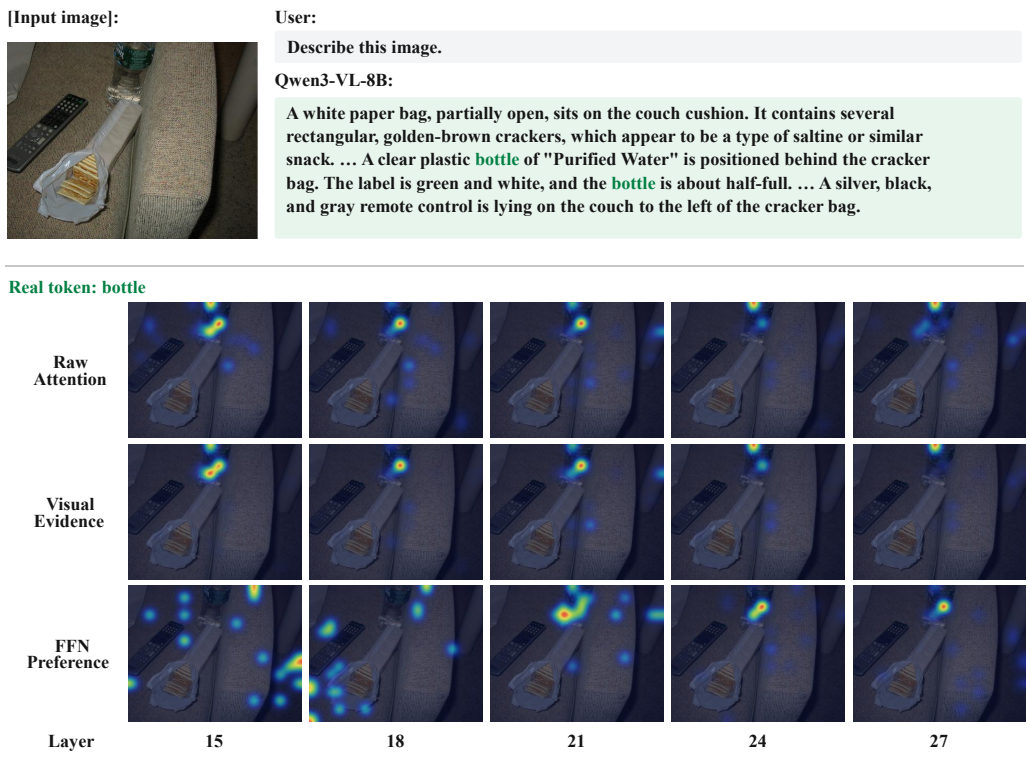


Figure 9: Visualization of real object mentions in QWEN3-VL-8B.

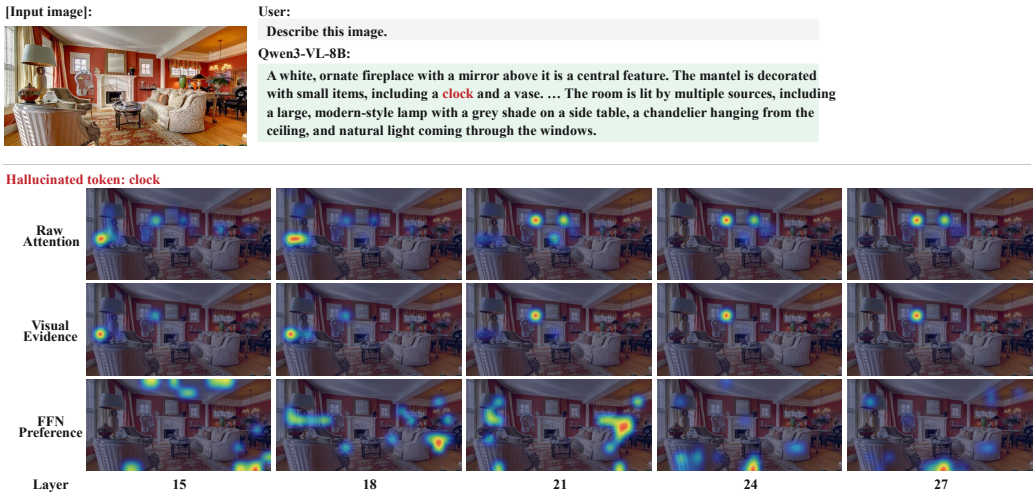


Figure 10: Visualization of hallucinated object mentions in QWEN3-VL-8B.

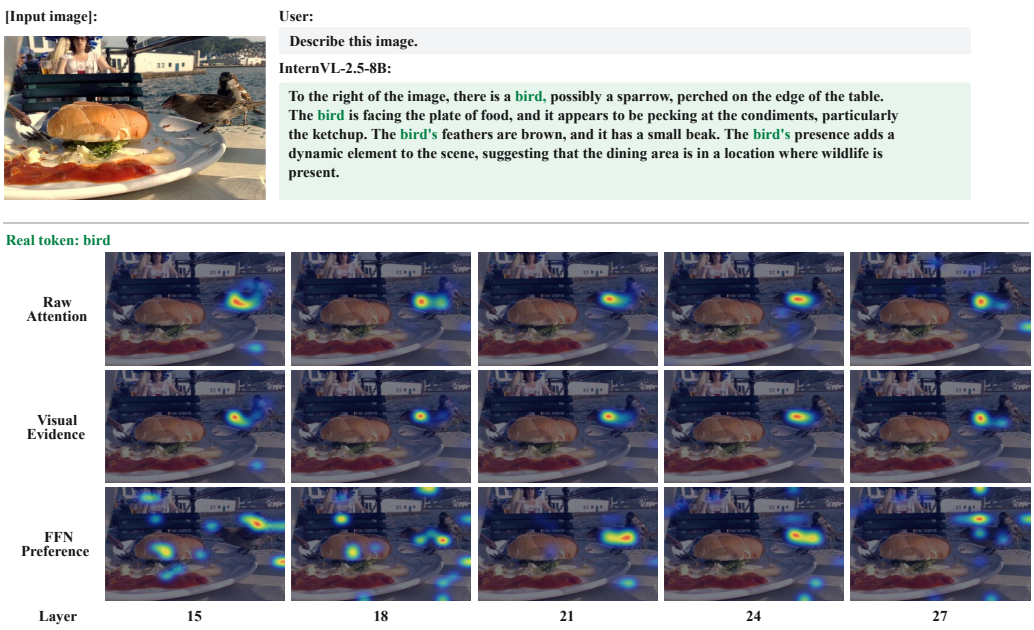


Figure 11: Visualization of real object mentions in InternVL2.5-8B.

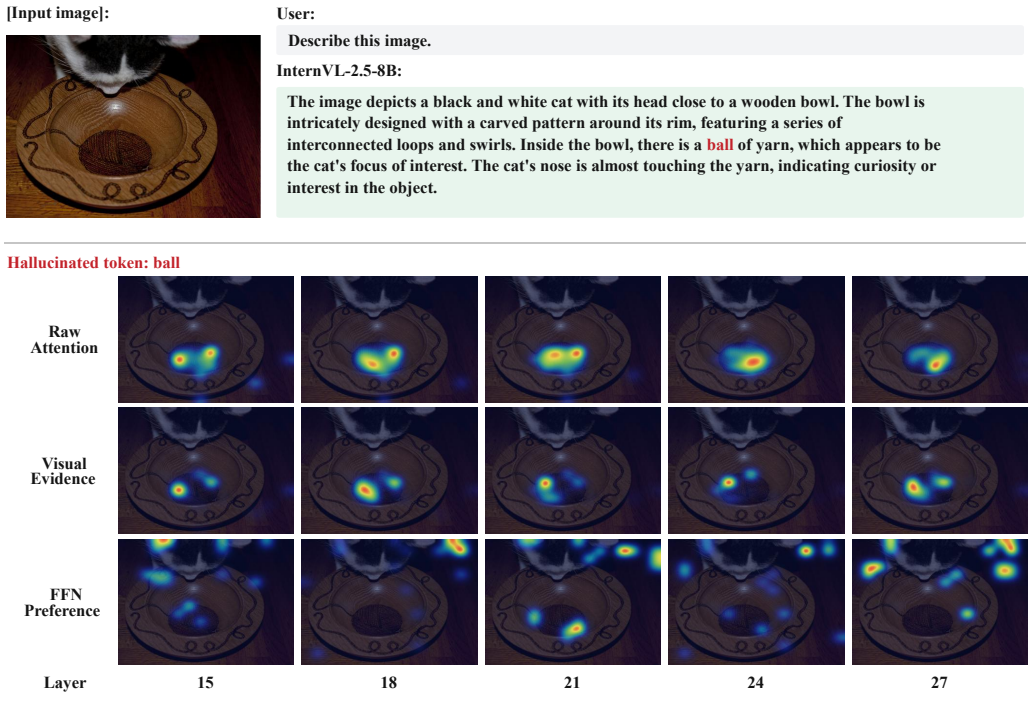


Figure 12: Visualization of hallucinated object mentions in InternVL2.5-8B.

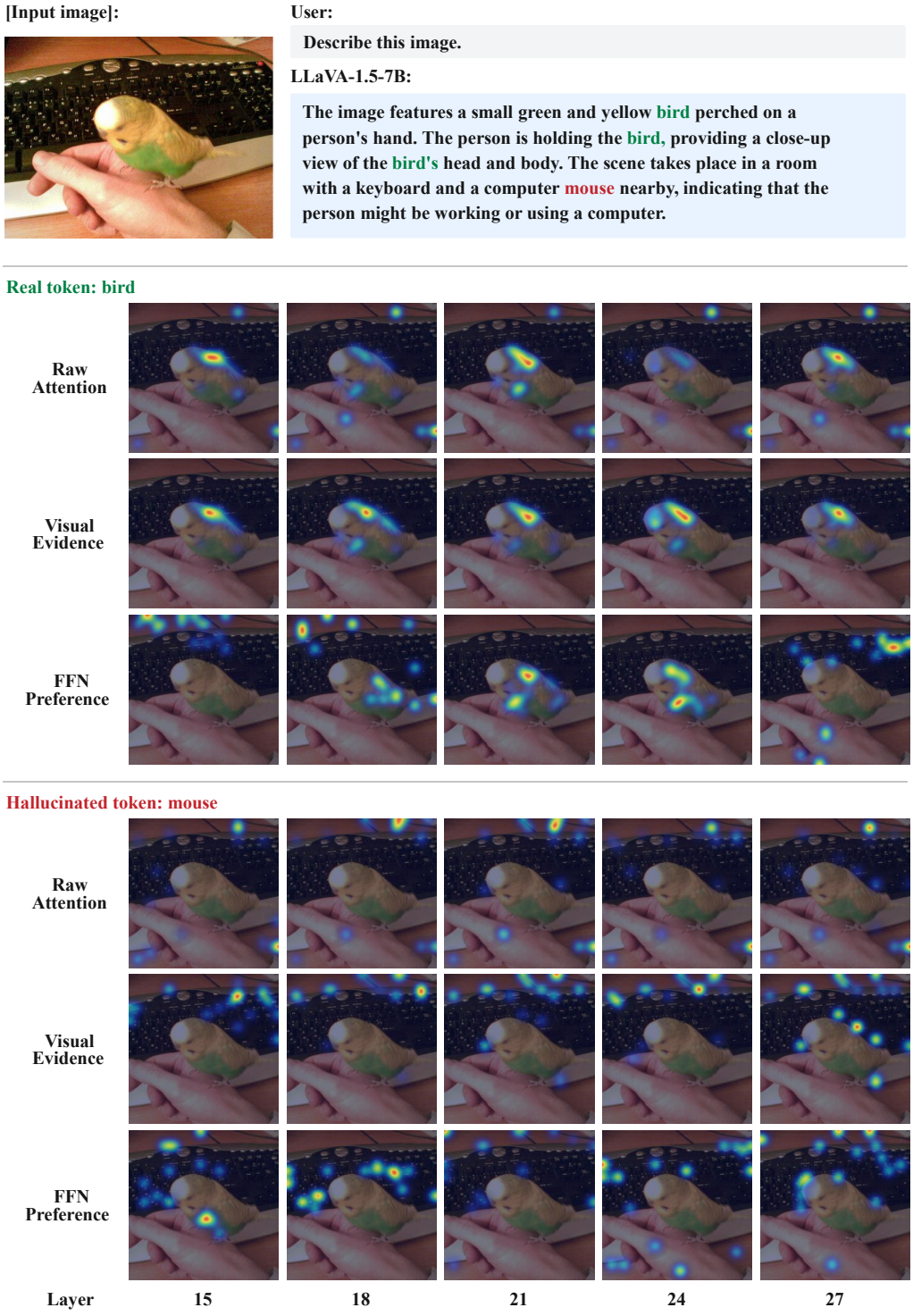


Figure 13: Visualization of paired object mentions in LLaVA-1.5-7B.

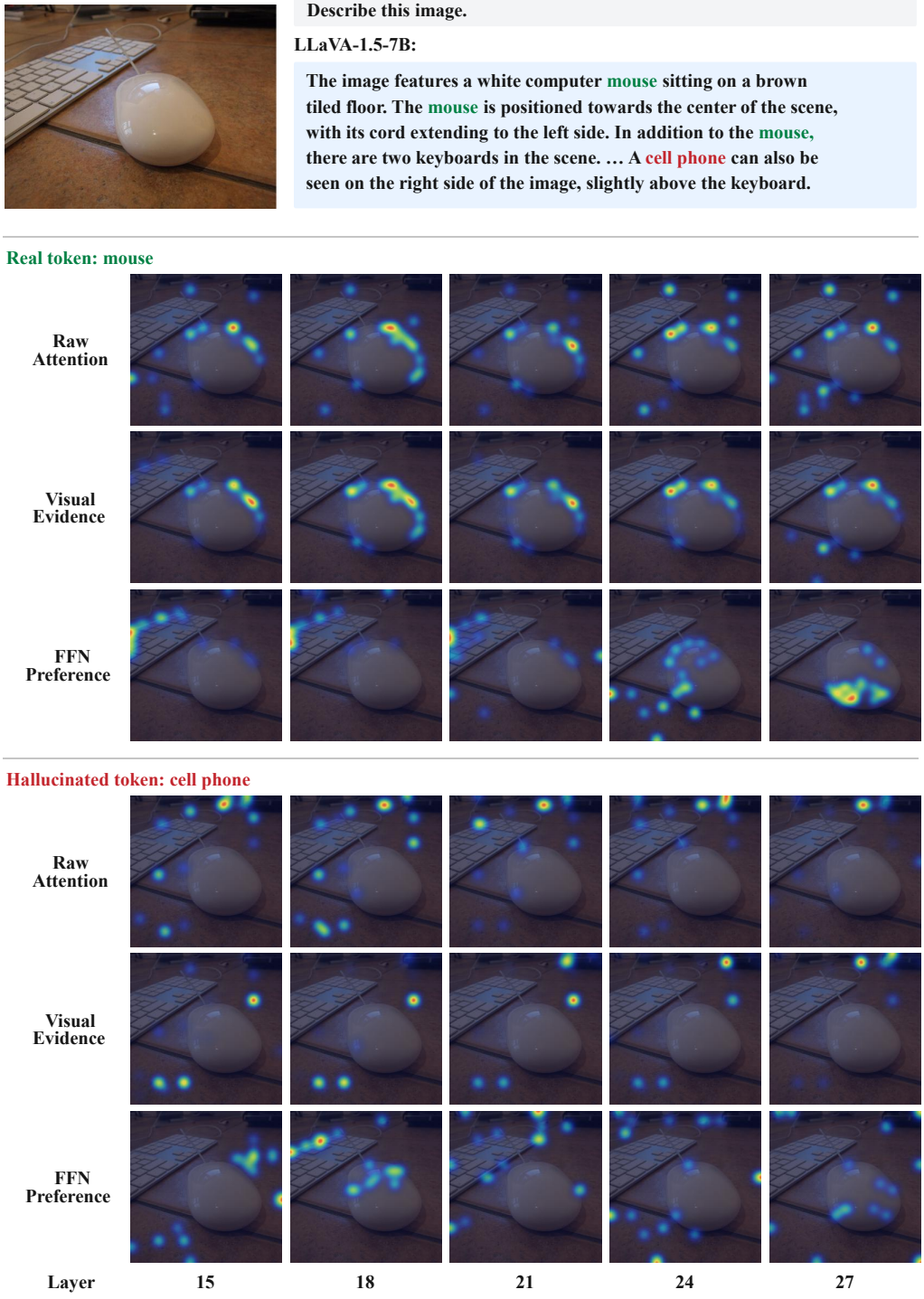


Figure 14: Visualization of paired object mentions in LLaVA-1.5-7B.

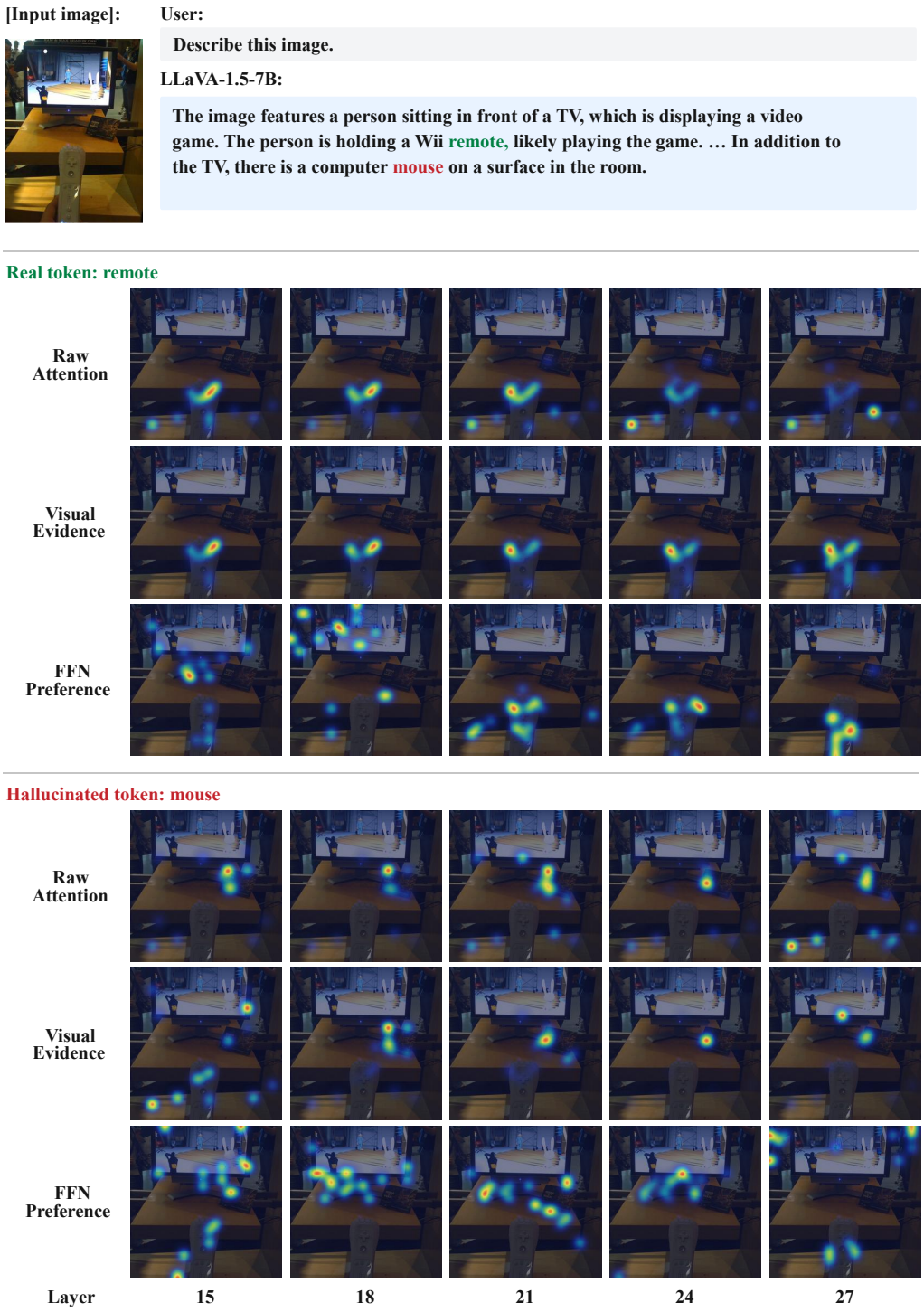


Figure 15: Visualization of paired object mentions in LLaVA-1.5-7B.

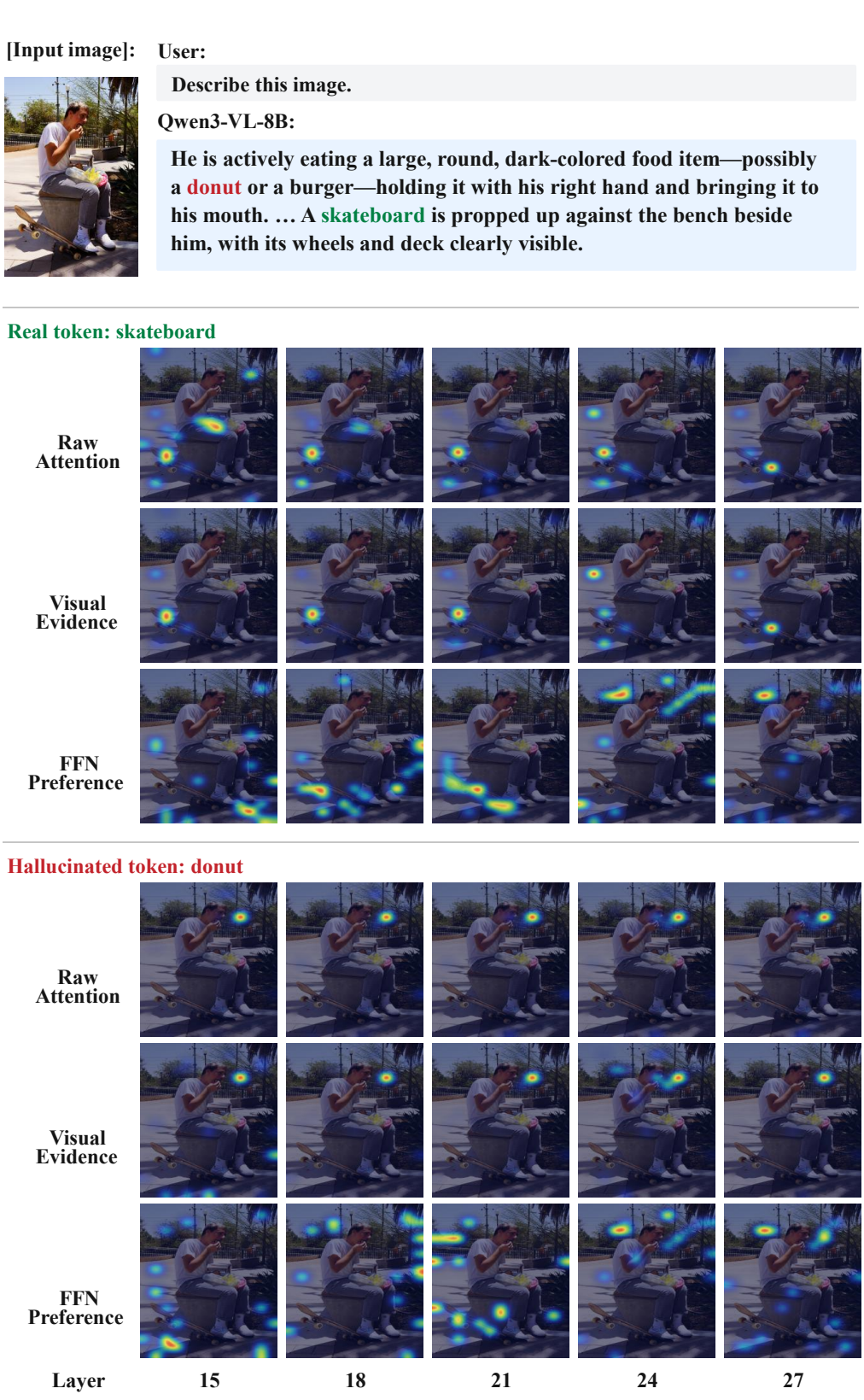


Figure 16: Visualization of paired object mentions in QWEN3-VL-8B.

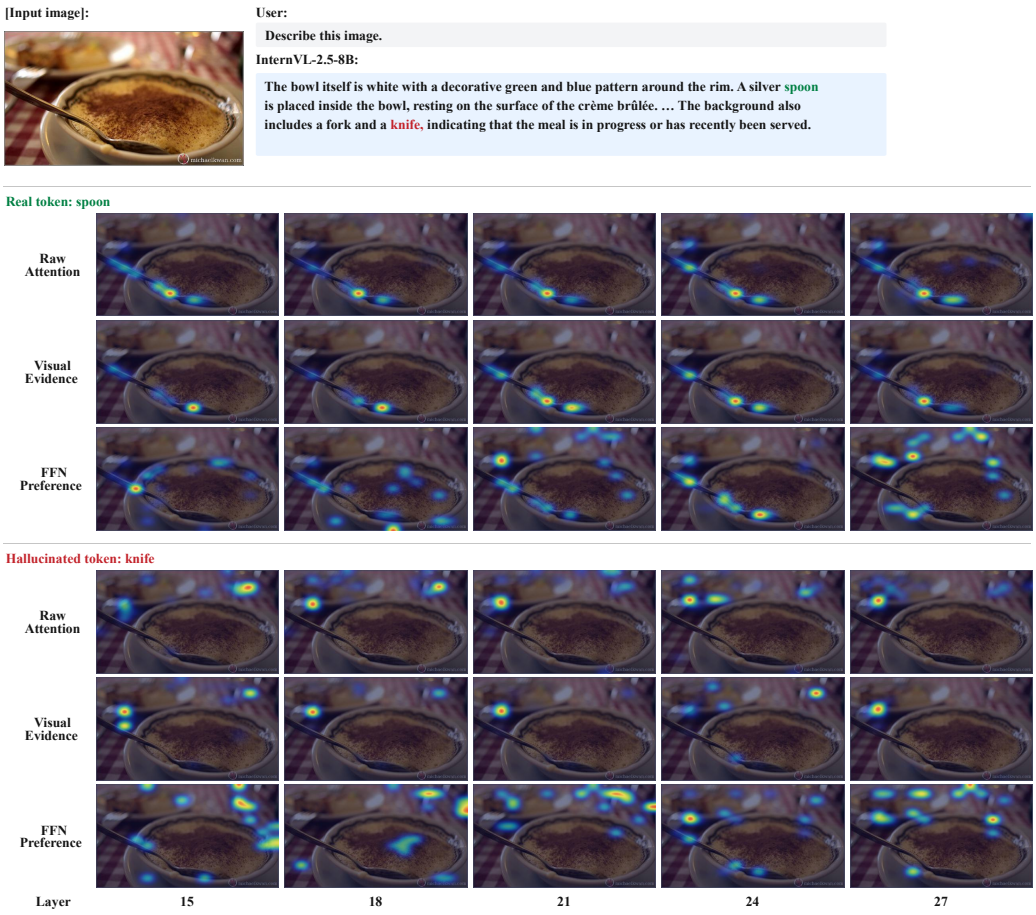


Figure 17: Visualization of paired object mentions in InternVL2.5-8B.