GLSim: Detecting Object Hallucinations in LVLMs via Global-Local Similarity

Seongheon Park [1] **Yixuan Li** [1]

Abstract

Object hallucination in large vision-language models presents a significant challenge to their safe deployment in real-world applications. Recent works have proposed object-level hallucination scores to estimate the likelihood of object hallucination; however, these methods typically adopt either a global or local perspective in isolation, which may limit detection reliability. In this paper, we introduce GLSim, a novel training-free object hallucination detection framework that leverages complementary global and local embedding similarity signals between image and text modalities, enabling more accurate and reliable hallucination detection in diverse scenarios. We comprehensively benchmark existing object hallucination detection methods and demonstrate that GLSim achieves superior detection performance, outperforming competitive baselines by a significant margin.

1. Introduction

Large Vision-Language Models (LVLMs) (Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023; Chen et al., 2024c; Ye et al., 2024; Wang et al., 2024; Bai et al., 2023; Lu et al., 2024) have made striking advances in understanding real-world visual data, enabling systems that can describe images, answer visual questions, and follow multi-modal instructions with fluency and creativity (Liu et al., 2023; Liang et al., 2024). Yet beneath this surface of impressive capability lies a critical vulnerability—object hallucinations (OH)—where the model generates plausible-sounding mentions of objects that are not present in the image (Rohrbach et al., 2018). An example is illustrated in Figure 1, where the LVLM describes a "dining table" in a birthday party scene, even though the image contains no such object. These hallucinations can undermine user trust, and are particularly concerning in high-stakes domains including medical imaging (Li et al., 2023a), autonomous navigation (Cui et al., 2024), and accessibility applications (Yang et al., 2024). Detecting such hallucinations is thus essential for safe and reliable deployment of LVLMs, and has become an increasingly active area of research (Bai et al., 2024).

Existing approaches to object hallucination detection often rely on external knowledge sources, such as human-annotated ground truth annotations (Rohrbach et al., 2018; Li et al., 2023b; Fu et al., 2023; Wang et al., 2023a; Chen et al., 2024b; Petryk et al., 2024). Others prompt or fine-tune external large language or vision-language models as judge to detect hallucinations (Jing et al., 2024; Sun et al., 2024; Liu et al., 2024a; Guan et al., 2024; Gunjal et al., 2024; Chen et al., 2024a). However, these approaches face practical limitations: ground-truth references are often unavailable in real-world scenarios, and external LLMs are prone to hallucinate themselves, thereby limiting reliability. This highlights the need for a lightweight, model-internal approach that can detect hallucinations without supervision or auxiliary models.

In this paper, we propose an object-level hallucination scoring function that operates without relying on external sources, leveraging the similarity between image and text modalities within the latent embedding space of LVLMs. We introduce **G**lobal-**L**ocal **Sim**ilarity (**GLSim**) score, a method that unify two complementary perspectives: a global similarity score, which captures how well an object semantically fits the overall scene, and a local grounding score, which checks whether any specific region in the image actually supports the object's presence. This fusion addresses a key shortcoming of prior approaches that rely on one perspective in isolation (Zhou et al., 2024; Jiang et al., 2025b;a; Phukan et al., 2025). For instance, a global-only method may wrongly consider a "dining table" plausible in a birthday party scene (Figure 1), simply because such contextual associations are common in pretraining data—even if no table is visually present. On the other hand, local-only approaches may struggle when a hallucinated object is visually similar to real objects in the scene, as in Figure 2, where a model hallucinates a "handbag" due to confusion with a leather seat of a motorcycle. By integrating both global and local signals, our method can ask not only "does this object belong contextually to the scene?" but also "is there

[1]Department of Computer Sciences, University of Wisconsin-Madison. Contact: Seongheon Park <seongheon_park@cs.wisc.edu>. Correspondence to: Yixuan Li <sharonli@cs.wisc.edu>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

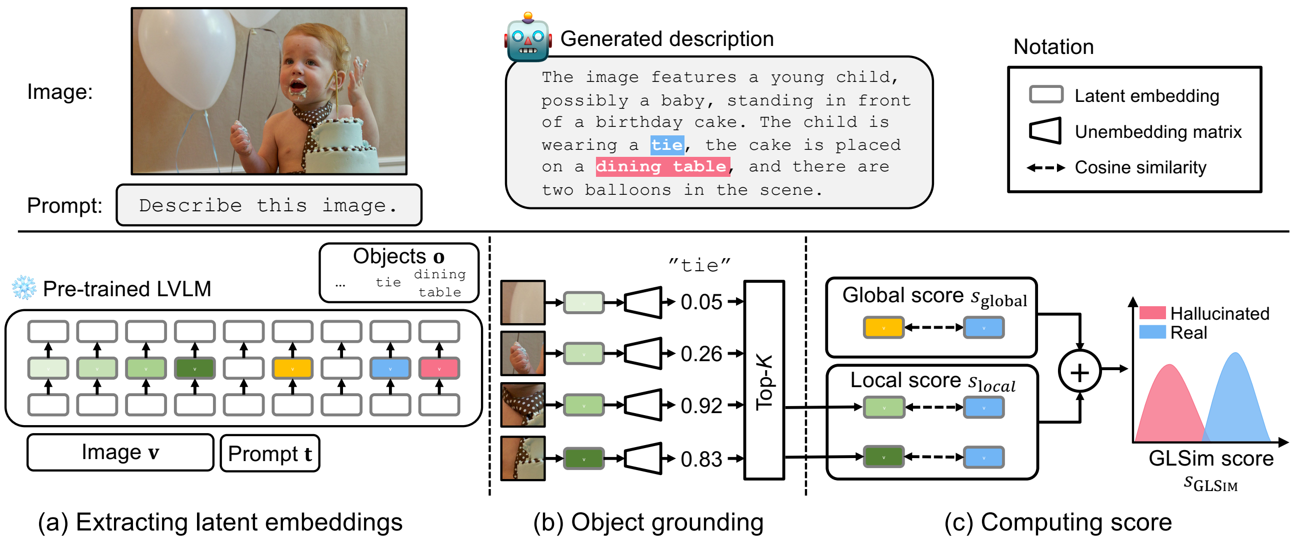


Figure 1. **Overall framework**. (a) We detect object-level hallucinations by leveraging latent embedding similarity. (b) For each object, the most relevant image regions are identified via unembedding from latent image representations. (c) The final GLSIM score is computed as a weighted combination of local (Section 4.2) and global (Section 4.3) signals, capturing both scene-level plausibility and spatial alignment, enhancing object hallucination detection accuracy.

concrete visual evidence for it?", resulting in more accurate, well-rounded, and interpretable hallucination detection across diverse scenarios.

As illustrated in Figure 1, GLSIM works by evaluating each object mention along two axes. First, the global score measures the similarity between the object token's embedding and the overall scene embedding—captured by the final token of the multimodal instruction prompt (highlighted in yellow). Next, we compute a local similarity score that checks for spatial grounding. Specifically, we identify the top image patches most relevant to the object using an adapted Logit Lens technique (nostalgebraist, 2020), then assess whether these regions provide strong visual evidence for the object using the average similarity between the object token's embedding and the top-K image token embeddings (highlighted in green). By combining these two complementary signals, GLSIM produces a holistic score that reflects both contextual fit and visual grounding—effectively distinguishing real objects from hallucinations.

We extensively evaluate GLSIM across multiple benchmark datasets and LVLMs, including LLaVA-1.5 (Li et al., 2023a), MiniGPT-4 (Zhu et al., 2023), and Shikra (Chen et al., 2023), demonstrating strong generalization and *state-of-the-art* performance in detecting object hallucinations. On both MSCOCO and Objects365 datasets, GLSIM consistently outperforms the latest baselines, including Internal Confidence (Jiang et al., 2025a) and attention-based grounding scores (Jiang et al., 2025b), achieving up to a **12.7%** improvement in AUROC. Ablation studies confirm the complementary roles of the global and local components: removing either degrades performance, while their combination yields the most reliable detection. Qualitative results further illustrate how GLSIM accurately flags subtle hallucinations, making it a practical tool for real-world deployment.

Our key contributions are summarized as follows:

1. We propose GLSIM, a novel object hallucination detection method that combines global and local similarity scores between latent embeddings. To the best of our knowledge, this is the first work to demonstrate their complementary effectiveness for the OH detection task.

2. We provide a comprehensive benchmarking of existing OH detection methods, addressing an important gap that has been overlooked in prior work.

3. We demonstrate the superior performance of GLSIM through extensive experiments, conduct in-depth ablations to analyze the contributions of each component and design choice, and verify the generalizability of our method across various LVLMs and datasets.

2. Related Works

Object Hallucination Detection in LVLMs. Object hallucination (OH) refers to the phenomenon where LVLMs generate textual descriptions that include *non-existent objects* in the image—a critical but underexplored problem in LVLMs with direct implications for reliable decision-making. Such hallucinations can stem from factors including statistical biases in training data (Delétang et al., 2024), strong language model prior (Liu et al., 2023), or visual information loss (Favero et al., 2024). Recent studies have focused on evaluating and detecting OH by leveraging ground-truth annotations (Rohrbach et al., 2018; Li et al., 2023b; Fu et al., 2023; Wang et al., 2023a; Chen et al., 2024b; Petryk et al., 2024). For instance, CHAIR (Rohrbach et al., 2018) suggests utilizing the discrete ratio of objects presented in the answer relative to a ground-truth object list to identify OH. Another line of work evaluates OH using external LLMs or LVLMs (Jing et al., 2024; Sun et al., 2024; Liu et al., 2024a; Guan et al., 2024; Gunjal et al., 2024; Chen et al., 2024a). For instance, GAIVE (Liu et al., 2024a) leverages a stronger LVLM (*e.g.*, GPT-4 (Achiam et al., 2023)) as a teacher to assess the responses of a student model, while HaLEM (Wang et al., 2023b) fine-tunes an LLM (*e.g.*, LLaMA (Touvron et al., 2023)) to score LVLM generations. While effective, these methods are resource-intensive and often lack transparency.

Several recent works have proposed object-level hallucination scores that estimate OH likelihood without requiring an external judge model or additional training. For instance, LURE (Zhou et al., 2024) utilizes the negative log-likelihood (NLL) of the object token generation probability; Internal Confidence (IC) (Jiang et al., 2025a) computes the maximum probability of the object token across all image hidden states (nostalgebraist, 2020); and Summed Visual Attention Ratio (SVAR) (Jiang et al., 2025b) leverages attention weights assigned to image tokens with respect to the object token. While promising, these methods primarily target hallucination mitigation and often fall short in detection performance: these methods typically leverage either global (*e.g.*, NLL, SVAR) or localized (*e.g.*, IC) signals in isolation and thus fail to capture the nuanced interplay between the overall semantic context and fine-grained visual grounding. Moreover, NLL often fail since LVLMs tend to favor linguistic fluency over factual accuracy (Radford et al., 2019); IC does not fully capture contextual information from the generated text; and SVAR can be biased toward previously generated text tokens (Liu et al., 2024c) and vulnerable to attention sink effects (Kang et al., 2025). Different from prior works, *we introduce the first object hallucination detection method that explicitly integrates both global and local signals—unifying localized attribution with holistic semantic alignment between the image and generated text*. We benchmark our approach against existing object-level hallucination detection methods across diverse settings to offer a comprehensive comparison in this space. Further related works are provided in Appendix C.2.

3. Problem Setup

Large Vision-Language Models for text generation typically consist of three main components: a vision encoder (*e.g.*, CLIP (Radford et al., 2021)) which extracts visual features, a multi-modal connector (*e.g.*, MLP) that projects these visual features into the language space, and an autoregressive language model that generates text conditioned on the projected visual and prompt embeddings.

Given an input image, the vision encoder processes it into a set of patch-level visual embeddings, commonly referred to as visual tokens. These tokens are then projected into the language model's embedding space through the multi-modal connector, resulting in a sequence of N visual embeddings: $\mathbf{v} = \{v_1, \dots, v_N\} \in \mathbb{R}^{N\times d}$, where each v_i corresponds to a transformed visual token of dimension d. On the language side, the input text prompt (e.g., "Describe this image in detail.") is tokenized and embedded into a sequence of language embeddings: $\mathbf{t} = \{t_1, \dots, t_L\} \in \mathbb{R}^{L\times d}$, where L is the prompt length. These two modalities—the projected visual tokens $\mathbf{v}$ and the textual embeddings $\mathbf{t}$—are concatenated and passed as the input sequence to the language model. The language model then generates a sequence of output tokens: $\mathbf{y} = \{y_1, \dots, y_M\}$, where each $y_i \in \mathcal{V}$ is drawn from a vocabulary space and M is the output length.

Object hallucination detection. In this work, we focus on detecting ***object existence hallucination*** in LVLMs—cases where the model generates text that references objects not present in the image (Li et al., 2023b; Zhai et al., 2023; Bai et al., 2024). This represents the most fundamental and critical form of errors affecting model reliability. We provide the formal task definition below.

Definition 3.1 (**Object Hallucination Detector**). *Let $\mathbf{x} = (\mathbf{v}, \mathbf{t})$ denote the input to the LVLM, and $\mathbf{y} = \{y_1, \dots, y_M\}$ be the sequence of generated tokens from the model. From $\mathbf{y}$, we extract a set of object mentions $\mathbf{o} = \{o_1, \dots, o_{n_h+n_r}\} \subset \mathcal{O}$, where n_h and n_r denote the number of hallucinated and real objects, respectively. The task of object hallucination detection is to design a scoring function $s : \mathcal{O} \times \mathcal{X} \to [0, 1]$, where $s(o, \mathbf{x})$ quantifies the likelihood that object $o \in \mathcal{O}$ is present in the input $\mathbf{x} \in \mathcal{X}$. Here $\mathcal{O}$ and $\mathcal{X}$ denote the space of objects and input, respectively. Based on this score, we define the object hallucination detector:*

$$G(o, \mathbf{x}) = \begin{cases} 1, & \text{if } s(o, \mathbf{x}) \geq \tau \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\tau \in [0, 1]$ is a decision threshold. Here, $G(o, \mathbf{x}) = 1$ indicates that object o is real (i.e., occurs in the image), while $G(o, \mathbf{x}) = 0$ indicates a hallucinated object.

4. Method

Overview. In this section, we propose an object-level hallucination scoring function that operates without relying on external sources, leveraging the similarity between image and text modalities within the latent embedding space of

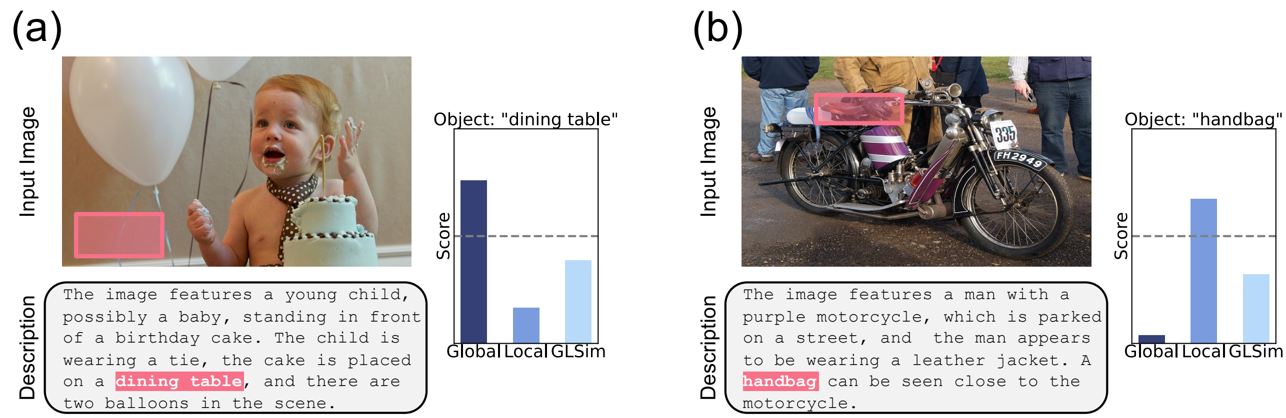


Figure 2. **Qualitative evidence.** In the generated descriptions, hallucinated objects are highlighted in red. The localized image regions are shaded with the same color as their corresponding objects. The gray line shows a threshold value τ. If an object's score is lower than the threshold τ, we consider it a hallucination. In (a), the local score successfully compensates for the failure of the global score, while in (b), the global score offsets the limitations of the local score.

LVLMs. We introduce **G**lobal-**L**ocal **Sim**ilarity (GLSIM) score, a method that leverages both global and local similarity measures, and discuss how these complementary signals contribute to effective object hallucination detection.

4.1. Motivation: Both Local and Global Signals Matter

Object hallucination in LVLMs often arises when models generate plausible-sounding descriptions that are not visually grounded. But detecting such hallucinations is challenging: they can stem from subtle biases, background patterns, or statistical co-occurrence in training data (Li et al., 2023b; Gong et al., 2024; Zhou et al., 2024). Critically, relying on a single perspective—either a global similarity or a local region-level score—is often not enough to reliably catch them. In particular, global similarity quantifies *how semantically related the object is to the image as a whole*. It captures holistic alignment between the object mention and the overall scene, and is useful for assessing whether the object "makes sense" in context. In contrast, local similarity measures *how well the object is visually grounded in a specific region*. It focuses on fine-grained evidence aligned with spatial areas most relevant to the object, helping verify whether it is actually present.

Qualitative evidence. Figure 2 illustrates how each signal alone can be insufficient. In panel (a), the LVLM-generated description includes a "`dining table`", yet no table is present in the image. A global similarity score fails to flag this hallucination—likely because the overall scene (*e.g.*, birthday cake, party setting) frequently co-occurs with tables in training data, leading to a high false-positive signal. In contrast, a local score that focuses on the visual region associated with "`dining table`" correctly assigns a low similarity, reflecting the absence of meaningful grounding in that region. In contrast, panel (b) shows a failure case for local similarity. The model hallucinates a "`handbag`," and while the global similarity correctly captures that the handbag is not semantically compatible with the overall scene, the local score becomes unreliable—likely due to a visually similar object in the image (*i.e.*, leather seat of the motorcycle).

These examples underscore the inherent limitations of using either signal in isolation. Global similarity can be overly influenced by high-level contextual associations, leading to false positives when hallucinated objects are contextually plausible within the scene but not visually present. On the other hand, local similarity is sensitive to spatial precision, but can misfire when localization is noisy or there are visually similar objects. As a result, each signal captures only a partial view of the grounding problem. To overcome this, we propose a unified approach, **G**lobal-**L**ocal **Sim**ilarity (GLSIM), that leverages the complementary strengths of both perspectives and offers more accurate and reliable detection of object hallucinations across a diverse range of visual-textual contexts. In the next subsections, we introduce the score definition in detail—explaining how we design global and local similarity for each object mention, and how they are integrated into a single decision score for hallucination detection. More qualitative results are presented in Appendix B.1.

4.2. Object Grounding via Local Similarity

A key component of our approach is the computation of the local similarity score, which captures how well an object mention is visually grounded in a specific region of the image. Unlike global similarity, which reflects scene-level plausibility, the local score focuses on verifying the presence of the object at the spatial level. The main challenge lies in identifying the most relevant region for each object mention—without relying on external annotations or bounding boxes.

Unsupervised object grounding. We leverage an unsupervised approach that leverages internal representations of the LVLM itself, to ground whether a predicted object token o is hallucinated or not. Given the LVLM input $\mathbf{x} = (\mathbf{v}, \mathbf{t})$, where $\mathbf{v} = \{v_1, \dots, v_N\}$ are the visual tokens and $\mathbf{t}$ are the prompt embeddings, we extract the hidden representations $h_l(v_i) \in \mathbb{R}^d$ of each visual token v_i at decoder layer l. To project these representations into the vocabulary space, we can leverage Visual Logit Lens (VLL) as:

$$\mathrm{VLL}_l(v_i) = h_l(v_i) \cdot W_U,$$

where $W_U \in \mathbb{R}^{d\times|\mathcal{V}|}$ is the unembedding layer matrix. Unlike the original Logit Lens (nostalgebraist, 2020), which operates solely in language models, our approach adapts it to a multimodal setting to attribute generated object mentions to relevant visual tokens. We apply a softmax and extract the predicted probability for the target object token o: $\mathrm{softmax}(\mathrm{VLL}_l(v_i))[o]$, probability quantifies how likely a visual token v_i is to predict the object word o, offering a model-internal signal of relevance between the image patch and object token. Importantly, we select the Top-K image patches with the highest probabilities as the localized regions corresponding to the object o:

$$\mathcal{I}(o) = \mathrm{TopK}_{v_i\in\mathbf{v}} \left(\{\mathrm{softmax}(\mathrm{VLL}_l(v_i))[o]\}\right). \quad (2)$$

We visualize object grounding results in Section 5.3 and Appendix B.2.

Local similarity score. Based on the localized regions $\mathcal{I}(o)$, we compute average cosine similarity between each localized image embedding and object embedding:

$$s_{\text{local}}(o, \mathbf{x}) = \frac{1}{K} \sum_{v_i\in\mathcal{I}(o)} \mathrm{sim}(h_l(v_i), h_{l'}(o)), \quad (3)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, and l' is the decoder layer used to represent the text embedding at the position of the object word. The score should be higher for real objects and relatively lower for hallucinated objects.

4.3. Scene-Level Grounding via Global Similarity

While the local similarity score focuses on spatially grounding an object in specific image regions, it alone may be insufficient—especially in cases where localization is ambiguous. To complement this, we introduce a *global similarity score* that measures scene-level semantic coherence between an object mention and the entire image. This can be useful for identifying out-of-context hallucinations (*e.g.*, referencing a “handbag” in a motorcycle scene).

Global similarity score. We compute the global similarity as the cosine similarity between the embedding of the object/text token and the embedding of the final token in the instruction prompt. The final instruction token often encodes a condensed summary of the model’s understanding of both image and prompt context. By comparing the object token to this representation, the global score quantifies how well the object semantically aligns with the overall scene. This allows the model to down-weight mentions that may be contextually implausible, even if they are locally aligned with some visual region.

Formally, given an object mention o and LVLM input $\mathbf{x} = (\mathbf{v}, \mathbf{t})$, let $h_{l'}(o) \in \mathbb{R}^d$ be the object token representation at layer l', and let $h_l(\mathbf{v}, \mathbf{t}) \in \mathbb{R}^d$ be the hidden representation of the last visual-text prompt token at layer l. The global similarity score is then defined as:

$$s_{\text{global}}(o, \mathbf{x}) = \mathrm{sim}\left(h_l(\mathbf{v}, \mathbf{t}), h_{l'}(o)\right), \quad (4)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity.

Global-Local Similarity (GLSIM) score. To fully leverage the complementary strengths of both grounding signals, we define the final hallucination detection score as a weighted combination of local and global similarity. Specifically, we define the GLSIM score as:

$$s_{\text{GLSIM}}(o, \mathbf{x}) = w \cdot s_{\text{global}}(o, \mathbf{x}) + (1 - w) \cdot s_{\text{local}}(o, \mathbf{x}), \quad (5)$$

where $w \in [0, 1]$ is a hyperparameter controlling the balance between local evidence and global context. This fused score captures both spatial alignment and scene-level plausibility, enabling more accurate detection of hallucinated objects. In practice, we find that a moderate value of w (e.g., 0.6) yields consistently strong performance across diverse scenarios (see Section 5.3). Based on the scoring function, the object hallucination detector is $G(o, \mathbf{x}) = \mathbb{1}\{s_{\text{GLSIM}}(o, \mathbf{x}) \geq \tau\}$, where 1 indicates a real object and 0 indicates a hallucinated object.

5. Experiments

5.1. Setup

Datasets and models. We utilize the MSCOCO dataset (Lin et al., 2014), which is widely adopted as the primary evaluation benchmark in numerous LVLM object hallucination studies and contains 80 object classes. In addition, we employ the Objects365 dataset (Shao et al., 2019), which offers a more diverse set of images and a larger category set comprising 365 object classes, along with denser object annotations per image. For evaluation, we randomly sample 5,000 images each from the validation sets of MSCOCO and Objects365. We conduct experiments on three representative LVLMs: LLaVA-1.5 (Li et al.,

Table 1. **Main results**. Comparison with competitive object hallucination detection methods on different datasets. For our method, the mean and standard deviation are computed across three different random seeds. All values are percentages, and the best results are shown in **bold**.

Dataset	Method		LLaVA-1.5-7B		LLaVA-1.5-13B		MiniGPT-4		Shikra	
			AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
MSCOCO	NLL	ICLR'24	63.7	84.9	63.1	86.1	59.4	81.2	60.4	82.1
	Entropy	ICLR'21	64.0	85.0	63.2	86.3	60.6	83.2	62.9	84.0
	Internal Conf.	ICLR'25	72.9	89.3	71.0	90.0	75.7	93.0	69.1	88.5
	SVAR	CVPR'25	74.7	91.2	75.2	92.9	83.6	95.9	70.7	89.1
	Contextual Lens♠	NACCL'25	75.4	90.7	78.7	92.8	84.9	96.2	69.5	87.6
	GLSIM (Ours)		**83.7**$^{\pm 0.3}$	**94.2**$^{\pm 0.2}$	**84.8**$^{\pm 0.5}$	**95.8**$^{\pm 0.2}$	**87.0**$^{\pm 0.4}$	**97.0**$^{\pm 0.1}$	**83.0**$^{\pm 0.7}$	**94.9**$^{\pm 0.3}$
Objects365	NLL	ICLR'24	62.9	60.8	59.4	61.0	56.7	70.4	58.9	64.8
	Entropy	ICLR'21	63.3	60.9	59.1	60.4	57.3	70.7	60.7	67.6
	Internal Conf.	ICLR'25	68.7	67.4	65.5	70.0	68.5	75.0	64.4	72.5
	SVAR	CVPR'25	64.9	66.6	63.5	68.2	71.0	79.4	60.6	68.3
	Contextual Lens♠	NACCL'25	63.2	62.6	62.1	65.6	70.2	77.8	59.6	67.0
	GLSIM (Ours)		**72.6**$^{\pm 0.5}$	**74.6**$^{\pm 0.4}$	**70.4**$^{\pm 0.8}$	**74.0**$^{\pm 0.6}$	**74.8**$^{\pm 0.6}$	**82.4**$^{\pm 0.7}$	**69.7**$^{\pm 1.0}$	**75.9**$^{\pm 0.9}$

2023a), MiniGPT-4 (Zhu et al., 2023), and Shikra (Chen et al., 2023). For LLaVA-1.5, we evaluate both 7B and 13B model variants to study scalability. Implementation details are provided in Appendix A. We evaluate on three additional LVLMs—InstructBLIP (Dai et al., 2023), LLaVA-NeXT-7B (Liu et al., 2024b), and Cambrian-1-8B (Tong et al., 2024), in Appendix 5.3.

Evaluation. We formulate the object hallucination detection problem as an object-level binary classification task, where a positive sample is a real object and a negative sample is a hallucinated object. We extract objects from the generated descriptions and perform exact string matching against the ground-truth object classes of each image and their synonyms, following CHAIR (Rohrbach et al., 2018). To evaluate OH detection performance, we report: (1) the area under the receiver operating characteristic curve (AUROC), and (2) the area under the precision-recall curve (AUPR), both of which are threshold-independent metrics widely used for binary classification tasks.

Baselines. We compare our approach against a comprehensive set of baselines, categorized as follows: (1) *Token probability*-based approaches—Negative Log-Likelihood (NLL) (Zhou et al., 2024) and Entropy (Malinin & Gales, 2021); (2) *Logit Lens probability*-based approach—Internal Confidence (Jiang et al., 2025a); (3) *Attention*-based approach—Summed Visual Attention Ratio (SVAR) (Jiang et al., 2025b); and (4) *Embedding similarity*-based approach—Contextual Lens♠ (Phukan et al., 2025). To ensure a fair comparison, we evaluate all baselines on identical test sets using the default experimental configurations provided in their respective papers. As Contextual Lens♠ was originally proposed for sentence-level hallucination detection, we adapt it for object-level hallucination detection. Further details of these baselines are discussed in Appendix C.

5.2. Main results

As shown in Table 1, we compare our method, GLSIM, with competitive object hallucination detection methods, including the latest ones published in 2025. GLSIM consistently outperforms existing state-of-the-art approaches across different models and datasets by a significant margin. Specifically, on the MSCOCO dataset with LLaVA-1.5-7B, GLSIM outperforms SVAR by **9.0%** AUROC, and achieves an **8.3%** AUROC improvement over Contextual Lens, an embedding similarity-based baseline. Unlike Contextual Lens, which relies on the maximum cosine similarity between text embeddings and all image embeddings, GLSIM integrates global and local signals, resulting in more robust detection performance. Notably, our method also demonstrates strong performance on Shikra, achieving a **12.7%** improvement in AUROC on the MSCOCO dataset compared to SVAR. Given that Shikra is trained with a focus on region-level inputs and understanding, this result suggests that our method is effective in models with strong spatial alignment capabilities.

Comparison with Internal Confidence. Recently, the Internal Confidence (IC) method (Jiang et al., 2025a) was proposed to detect hallucinations using visual logit lens probabilities. Our approach differs from IC in three key ways. First, IC directly uses the maximum probability from the visual logit lens across all image patches and layers, which can be overconfident—assigning high scores to hallucinated objects (see Figure 3). In contrast, we compute the semantic *similarity in representation space* between the object token embedding and the Top-K visual tokens, yielding a more reliable and semantically meaningful signal. For hallucinated objects, this leads to alignment with semantically irrelevant regions, resulting in lower similarity scores, thereby enabling more reliable object hallucination

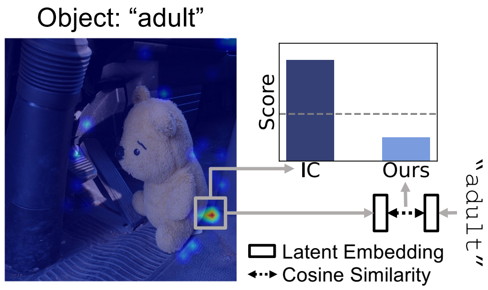


Figure 3. Internal Confidence (IC) can assign high confidence to incorrect regions for hallucinated objects. Our local score (Section 4.2) mitigates this by cross-modal embedding similarity.

Table 2. Comparison of global and local scores.

	Method	Metric	LLaVA	Shikra
Global	NLL	TP	63.7	60.4
	Entropy	TP	64.0	62.9
	SVAR	AT	74.7	70.7
	s_{global} (Eq. (4))	ES	79.3	78.9
Local	Internal Conf.	LLP	72.9	69.1
	Contextual Lens♠	ES	75.4	69.5
	s_{local} (Top-1)	ES	76.5	73.1
	s_{local} (Top-K)	ES	78.8	76.8
G & L	s_{GLSIM} (Top-1)	ES	82.0	81.0
	s_{GLSIM} (Top-K)	ES	**83.7**	**83.0**

Table 3. Object grounding methods.

Score	Grd. Method	LLaVA	Shikra
s_{global}	-	79.3	79.8
s_{local}	Attention	66.3	65.0
	Cosine Sim.	76.2	70.1
	Logit Lens	78.8	76.8
s_{GLSIM}	Attention	79.4	80.0
	Cosine Sim.	80.7	80.9
	Logit Lens	83.7	82.0

detection. Second, IC considers only the most probable patch, while we aggregate over the Top-K most relevant patches. As shown in our ablation study (Section 5.3), using multiple visual regions improves performance by capturing spatially distributed evidence and reducing sensitivity to local noise. Third, IC is purely local in nature, whereas our framework also integrates a global similarity score that captures object-scene coherence at the image level. Together, these advantages enable our method to outperform IC by a substantial margin of **10.8%** AUROC.

5.3. Ablation Studies

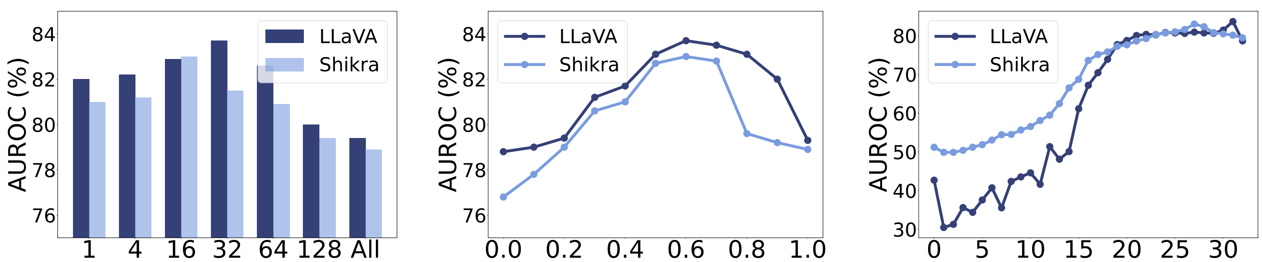
In this section, we provide various in-depth analysis of each component of our method. All experiments are conducted using LLaVA-1.5-7B and Shikra on the MSCOCO dataset, and results are reported in terms of AUROC (%). Further ablation studies are provided in Appendix D.

Analysis of global and local scores. We systematically compare global and local scores on the MSCOCO dataset, as shown in Table 2. **Finding 1: Embedding similarity is an effective scoring metric.** Embedding similarity (ES)-based methods consistently outperform other scoring functions, with performance improvements ranging from 1.1% to 22.6%. In contrast, token probability (TP)-based approaches are optimized for linguistic fluency rather than object existence accuracy; attention weight (AT)-based methods often fail to align with causal attributions (Jain & Wallace, 2019); and Logit Lens probability (LLP) methods tend to exhibit overconfidence. By directly capturing the semantic alignment between image and text modalities, embedding similarity provides a more reliable signal for OH detection. **Finding 2: Object grounding improves OH detection.** Among local methods, our approach leverages grounded objects in the image and computes embedding similarity directly with those object representations, achieving 0.3% to 11% improvements over baselines. This enables fine-grained alignment, unlike Internal Confidence and Contextual Lens methods, which rely only on the maximum token probability or cosine similarity score. **Finding 3: Combining global and local scores further improves performance.** By combining global (s_{global}) and local (s_{local}) similarity scores, we observe additional gains of 2.7% in Top-1 and 4.4% in Top-K for the LLaVA model. This demonstrates that our scoring function design in Equation (5) effectively integrates the complementary strengths of both global and local signals.

Comparison of object grounding methods. We explore several design choices for the patch selection for object grounding in Section 4.2, with results summarized in Table 3. Specifically, we vary the metric used for Top-K ($K = 32$) patch selection, comparing (1) attention weights, (2) cosine similarity, and (3) our method (visual logit lens). Our method outperforms attention weights by 12.5% and cosine similarity by 2.6% in local score evaluation. When combining global and local scores, our method achieves gains of 4.3% over attention weights and 3.0% over cosine similarity. We further visualize the Top-K patch scores for each metric in Figure 4. From the visualization, we observe that high attention weights tend to be assigned to irrelevant regions (Kang et al., 2025); cosine similarity better localizes object regions but still assigns spuriously high scores to background areas. In contrast, ours accurately highlights object regions, leading to more reliable patch selection for grounding.

(a) Object: “car” (b) Object: “surfboard”

Figure 4. Object grounding results with LLaVA. Ground-truth bounding boxes are shown in red.



(a) Number of patches K (b) Weighting parameter w (c) Text embedding layer l'

Figure 5. (a) Effect of the number of selected image patches K; (b) effect of the weighting parameter w in Equation (5); and (c) effect of the text embedding layer index l'.

How do the selected number of patches K affect the performance? We analyze the impact of varying the number of selected patches K in Equation (2) on object hallucination detection performance in Figure 5(a). Performance improves with increasing K up to K=32 for LLaVA and K=16 for Shikra, after which it degrades. This trend suggests that a small K may fail to capture sufficient object information, while a large K introduces irrelevant regions, adding noise. Given that LLaVA processes 576 image tokens, the optimal K roughly corresponds to 6% of total image tokens. This highlights the importance of choosing K relative to input resolution for effective OH detection.

How does the weighting parameter w affect performance? In Figure 5(b), we examine the effect of the weighting parameter w in Equation (5) on OH detection performance. Performance increases with w up to 0.6, after which it declines. Smaller values of w place greater emphasis on the local score, while larger values prioritize the global score. We find that moderate values consistently yield the best results across models, suggesting that global and local signals are complementarily informative—where the global score captures scene-level semantics, and the local score captures fine-grained, spatial-level semantics. These results support our design choice of combining both components through a balanced weighting scheme, effectively enhancing overall performance.

How does the text embedding layer index affect performance? We examine how the choice of text embedding layer l' influences overall performance when computing embedding similarity in Equation (3) and Equation (4). We fix the image embedding layer l to the 32nd layer for LLaVA and the 30th layer for Shikra, as specified in Table 6. As shown in Figure 5(c), the best performance is achieved at the 31st layer for LLaVA and the 27th layer for Shikra. Performance improves with later layers, suggesting that semantic representations are progressively refined in later layers. However, it slightly drops afterward, which supports the observation from (Skean et al., 2025) that the optimal layer for downstream tasks may not necessarily be the final layer. These findings indicate that later-intermediate layers are particularly effective for object hallucination detection. The complete performance matrix over all (l, l') layer pairs is provided in Appendix E.

Design choices for global and local scores. We ablate several key design choices for each scoring function in Table 5. For the global score (s_{global}), we compare (1) similarity with the last image token embedding, (2) average similarity across all image tokens, and (3) similarity with the last instruction token. The last instruction token performs best, outperforming the average similarity by 8%, highlighting its strength in capturing scene-level semantics. For the local score (s_{local}), we compare (1) a Logit Lens probability-weighted average of local similarities among top-K patches and (2) a non-weighted average as in Equation (3), where the latter works slightly better. Finally, combining global and local scores improves performance for both variants of the global score. This confirms that the two signals are complementary and supports the design of our

Table 4. Performance on additional models evaluated on MSCOCO.

Method		InstructBLIP		LLaVA-NeXT		Cambrian-1	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
NLL	ICLR'24	65.1	82.3	56.1	84.6	50.1	76.3
Entropy	ICLR'21	65.6	82.6	57.5	84.8	50.2	76.4
Internal Conf.	ICLR'25	81.9	91.6	77.8	95.8	65.4	96.0
SVAR	CVPR'25	78.4	90.6	76.9	95.6	60.5	90.8
Contextual Lens♠	NACCL'25	83.0	92.8	70.1	93.7	63.4	94.8
GLSIM (Ours)		**85.0**	**94.3**	**81.4**	**97.6**	**79.7**	**98.0**

Table 5. Design choices for global and local scoring functions.

Score	Method	LLaVA	Shikra
Global	Last image token	65.9	56.2
	Average image token	71.3	66.7
	Last instruction token	79.3	79.8
Local	Weighted average	75.8	73.0
	Non-weighted average	78.8	76.8
GLSIM	Average & Eq. (3)	79.2	77.0
	Last inst. & Eq. (3)	83.7	82.0

scoring function.

Results for additional models. We further evaluate our approach on three additional large vision-language models—InstructBLIP (Dai et al., 2023), LLaVA-NeXT-7B (Liu et al., 2024b), and Cambrian-1-8B (Tong et al., 2024)—using the MSCOCO dataset. As shown in Table 4, our method consistently outperforms baselines, achieving a 2.0% AUROC improvement on InstructBLIP, 4.2% on LLaVA-NeXT, and a notable 14.3% gain on Cambrian-1. These results demonstrate the robustness of GLSIM across varying model architectures and scales.

6. Conclusion

In this paper, we propose GLSIM, a novel training-free framework for object hallucination detection, which exploits the complementary strengths of global scene-level semantics and fine-grained spatial alignment by leveraging embedding similarity. Empirical results demonstrate that GLSIM achieves superior performance across diverse families of LVLMs and two representative datasets. Our in-depth quantitative and qualitative ablations provide further insights into understanding the effectiveness of GLSIM. We hope our work will inspire future research on OH detection from diverse perspectives.

Limitations and future work. Our analysis in this paper focuses on object existence hallucinations, as annotations and benchmarks for attribute and relation hallucinations are currently limited. Nonetheless, it would be interesting to investigate further the grounding ability of the Logit Lens technique for attributes and relationships to quantify local similarity beyond object presence. Moreover, leveraging accurate OH detection from our method to guide model editing or prediction refinement presents a promising future direction for mitigating object hallucinations.

Acknowledgement

We gratefully acknowledge Changdae Oh and Hyeong Kyu Choi for their valuable comments on the draft. Seongheon Park and Yixuan Li are supported in part by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS2331669, Alfred P. Sloan Fellowship, and Schmidt Sciences Foundation.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. In *EMNLP Findings*, 2023.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., and Shou, M. Z. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023.

Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. Mllmas-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024a.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D., and Chai, J. Multi-object hallucination in vision language models. In *NeurIPS*, 2024b.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., et al. A survey on multimodal large language models for autonomous driving. In *WACV*, 2024.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., et al. Language modeling is compression. In *ICLR*, 2024.

Du, X., Xiao, C., and Li, S. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *NeurIPS*, 2024.

Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., and Soatto, S. Multi-modal hallucination control by visual information grounding. In *CVPR*, 2024.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Gong, X., Ming, T., Wang, X., and Wei, Z. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*, 2024.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. In *AAAI*, 2024.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025.

Jain, S. and Wallace, B. C. Attention is not explanation. In *NACCL*, 2019.

Jiang, N., Kachinthaya, A., Petryk, S., and Gandelsman, Y. Interpreting and editing vision-language representations to mitigate hallucinations. In *ICLR*, 2025a.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., and Yang, X. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025b.

Jing, L., Li, R., Chen, Y., and Du, X. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *EMNLP Findings*, 2024.

Kang, S., Kim, J., Kim, J., and Hwang, S. J. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023a.

Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., and Karray, F. Reference-free hallucination detection for large vision-language models. In *EMNLP Findings*, 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., and Liu, K. A survey of multimodel large language models. *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. In *TMLR*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, 2024b.

Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*, 2024c.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, 2023.

nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong.

Park, S., Du, X., Yeh, M.-H., Wang, H., and Li, Y. How to steer llm latents for hallucination detection? In *ICML*, 2025.

Paszke, A. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Petryk, S., Chan, D. M., Kachinthaya, A., Zou, H., Canny, J., Gonzalez, J. E., and Darrell, T. Aloha: A new measure for hallucination in captioning models. In *NACCL*, 2024.

Phukan, A., Morj, H. K., Saxena, A., Goswami, K., et al. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *NACCL*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models. In *ICLR*, 2023.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *EMNLP*, 2018.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models. In *ICML*, 2025.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. In *ACL Findings*, 2024.

Tong, P., Brown, E., Wu, P., Woo, S., IYER, A. J. V., Akula, S. C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., XiXuan, S., et al. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024.

Yang, B., He, L., Liu, K., and Yan, Z. Viassist: Adapting multi-modal large language models for users with visual impairments. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 2024.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., and Huang, F. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024.

Zhai, B., Yang, S., Zhao, X., Xu, C., Shen, S., Zhao, D., Keutzer, K., Li, M., Yan, T., and Fan, X. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption. *https://openreview.net/forum?id=9Ebi1euQZQ*, 2023.

Zhang, R., Zhang, H., and Zheng, Z. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

Contents

A. Implementation Details

We implement our method using greedy decoding with a maximum generated token length of 512. The layer indices (l, l'), the number of selected patches K, and the weighting parameter w used for computing the final score are selected based on a separate validation set, as detailed in Table 6. For multi-token objects, we use the first token to compute the scores and consider the first occurrence of each object for hallucination detection. The total number of generated objects is shown in Table 7. For all experiments, we report the average over three different random seeds. All experiments are conducted using Python 3.11.11 and PyTorch 2.6.0 (Paszke, 2019), on a single NVIDIA A6000 GPU with 48GB of memory.

Table 6. Hyperparameters.

Model	Hyperparameters		
	Layer indices	K	w
LLaVA-1.5-7b	(32, 31)	32	0.6
LLaVA-1.5-13b	(40, 38)	32	0.6
MiniGPT-4	(32, 30)	4	0.5
Shikra	(30, 27)	16	0.6

Table 7. Number of generated objects.

Model	MSCOCO		Objects365	
	Real	Hallu.	Real	Hallu.
LLaVA-1.5-7b	14,910	4,121	14,357	9,850
LLaVA-1.5-13b	15,372	3,687	15,086	9,672
MiniGPT-4	11,642	2,282	11,603	7,222
Shikra	15,724	4,727	15,063	10,350

B. Additional Qualitative Results

B.1. Qualitative Results

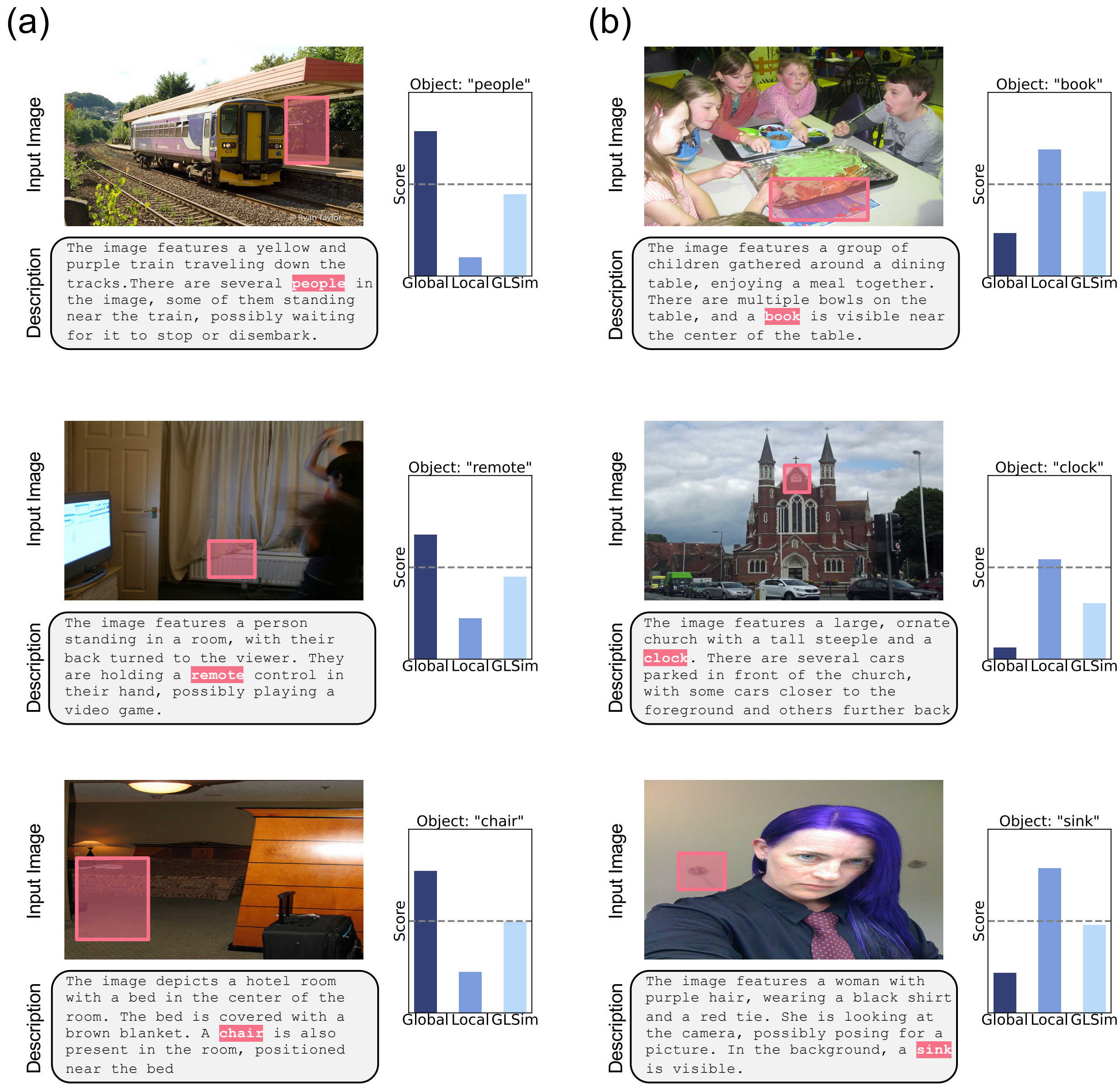


Figure 6. **Additional qualitative evidence.** In the generated descriptions, hallucinated objects are highlighted in red. The localized image regions are shaded with the same color as their corresponding objects. The gray line shows a threshold value τ. If an object's score is lower than the threshold τ, we consider it a hallucination. In (a), the local score successfully compensates for the failure of the global score, while in (b), the global score offsets the limitations of the local score.

B.2. Object Grounding

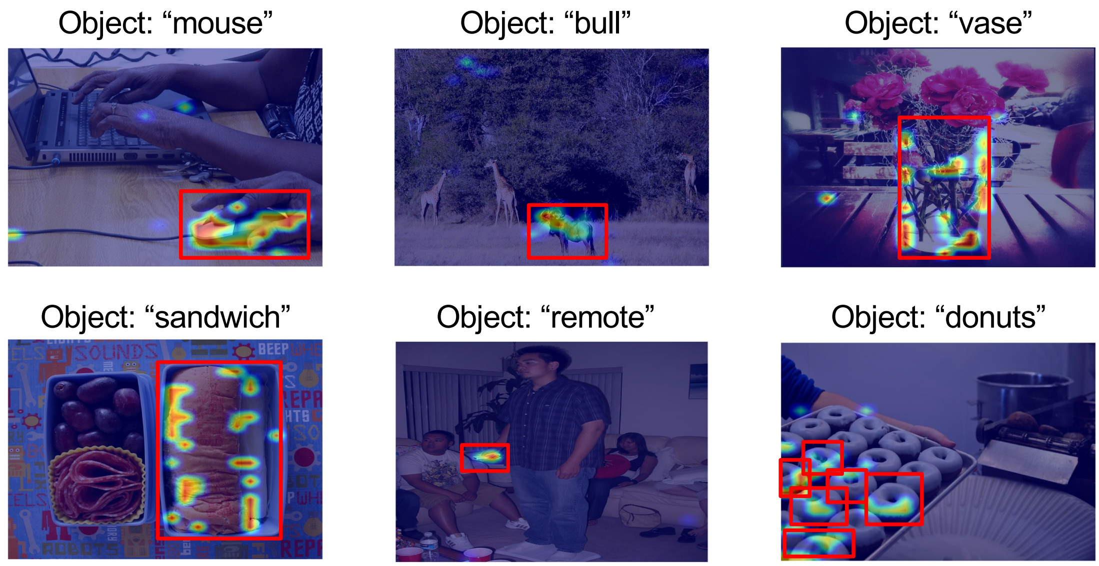


Figure 7. **Object grounding results for real objects.** We visualize the Top-K Logit Lens probabilities at the 32nd layer of LLaVA-1.5-7B. Ground-truth bounding boxes are shown in red.

Figure 8. **Object grounding results for hallucinated objects.** We visualize the Top-K Logit Lens probabilities at the 32nd layer of LLaVA-1.5-7B.

C. Related Works

C.1. Baselines

Negative Log-likelihood. Zhou et al. (2024) represents the probability of autoregressive decoding for each object token as $p(o \mid \mathbf{y}_{<j}, \mathbf{v})$, where j denotes the positional index of object o. For each object o, the corresponding hallucination score is defined as:

$$s_{\text{nll}} = -\log p(o \mid \mathbf{y}_{<j}, \mathbf{v}). \tag{6}$$

To align with the definition in Equation (1), we use $s'_{\text{nll}} = -s_{\text{nll}}$.

Entropy. We further investigate the object-level hallucination score by estimating the entropy (Malinin & Gales, 2021) of the token probability distribution at position j:

$$s_{\text{entropy}} = -\sum_{y \in \mathcal{V}} p(y \mid \mathbf{y}_{<j}, \mathbf{v}) \log p(y \mid \mathbf{y}_{<j}, \mathbf{v}). \tag{7}$$

To align with the definition in Equation (1), we use $s'_{\text{entropy}} = -s_{\text{entropy}}$.

Internal Confidence. Jiang et al. (2025a) apply the logit lens to image representations, enabling the analysis of how visual features are transformed into textual predictions. To quantify the model's confidence for object hallucination detection, the internal confidence score is computed as the maximum softmax probability of the object word o across all image representations and layers. Following the notations introduced in Section 4.2, the hallucination score is defined as:

$$s_{\text{IC}} = \max_{l \in [L]} \max_{i \in [N]} \text{VLL}_l(v_i)[o], \tag{8}$$

where L denotes the total number of layers and N denotes the total number of image patches.

Summed Visual Attention Ratio (SVAR). The Visual Attention Ratio (VAR) quantifies the interaction of a generated token o with visual information by summing its attention weights assigned to image tokens in a specific attention head h and layer ℓ:

$$\text{VAR}^{(\ell,h)}(o) \triangleq \sum_{i=1}^{N} A^{(\ell,h)}(o, v_i), \tag{9}$$

where $A^{(\ell,h)}(o, v_i)$ represents the attention weight from object token o to image token v_i at h-th head in l-th layer. Building on this, Jiang et al. (2025b) define the Summed Visual Attention Ratio (SVAR), which measures the overall visual attention by averaging VAR scores across all heads and summing over a range of layers. Specifically, for an object token o within layers ℓ_5 to ℓ_{18}, SVAR is computed as:

$$s_{\text{SVAR}} = \frac{1}{H} \sum_{\ell=5}^{18} \sum_{h=1}^{H} \text{VAR}^{(\ell,h)}(o), \tag{10}$$

where H denotes the total number of attention heads.

Contextual Lens. To detect sentence-level hallucination, Phukan et al. (2025) compute the maximum cosine similarity between the average embedding of the generated description at a specific layer l_T and each image embedding at layer l_I.

$$\text{Sentence-level Score} = \max_{i \in [N]} \text{sim}\Big(\frac{1}{M} \sum_{j=1}^{m} h_{l_T}(y_j), h_{l_I}(v_i)\Big). \tag{11}$$

To compute the object-level hallucination score, we modify the original score with:

$$s_{\text{CL}} = \max_{i \in [N]} \text{sim}(h_{l_T}(o), h_{l_I}(v_i)). \tag{12}$$

C.2. Extended Literature Review

Sentence-level hallucination detection in LLMs and LVLMs aims to classify an entire generation as either hallucinated or correct, providing a coarse-grained assessment of factuality (Huang et al., 2025). A plethora of work addresses sentence-level hallucination detection in large language models (LLMs) by designing uncertainty scoring functions, such as utilizing token generation probabilities (Ren et al., 2023), prompting LLMs to quantify their confidence (Lin et al., 2022), and evaluating consistency across multiple responses (Manakul et al., 2023). Specifically, internal state-based methods leverage latent model embeddings (Azaria & Mitchell, 2023), employing techniques such as contrast-consistent search (Burns et al., 2023), identifying hallucination-related subspaces (Du et al., 2024), or reshaping the latent space for hallucination detection (Park et al., 2025).

Recently, reference-free sentence-level hallucination detection for large vision-language models (LVLMs) has attracted research attention. Li et al. (2024) first compare uncertainty quantification methods from LLMs for application to LVLMs. Inspired by Kuhn et al. (2023), VL-Uncertainty (Zhang et al., 2024) estimates uncertainty by measuring prediction variance across semantically equivalent but perturbed prompts.

In contrast to these works, we propose the hallucination scoring function for *object-level hallucination detection* in LVLMs, which provides a fine-grained assessment by localizing hallucinations within generations rather than classifying entire outputs. Our method leverages latent embeddings from both visual and textual modalities and explores intrinsic metrics tailored to LVLMs.

D. Further Ablation Studies

D.1. Visualization of Score Distributions

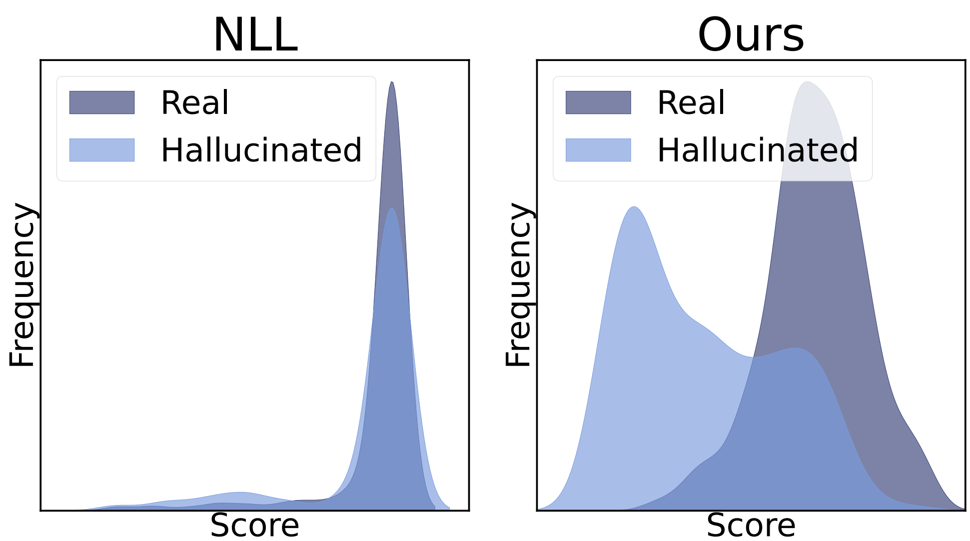


Figure 9. Score distribution for negative log-likelihood (Zhou et al., 2024) *vs.* our method.

We provide score distribution for the negative log-likelihood (NLL) (Zhou et al., 2024) and GLSIM (ours) in Figure 9. Our approach exhibits a more distinct separation between the real and hallucinated data distributions. This enhanced separation highlights the effectiveness of our global and local scoring designs, as well as their combination strategy, contributing to more reliable detection performance compared to existing method.

D.2. Comparison with POPE

To compare with prompting-based methods such as POPE (Li et al., 2023b), we extract objects from the generated captions produced using the prompt "Describe this image in detail." We then convert these objects into a set of Yes-or-No short-form questions. We prompt the LVLM with the following template:

Input prompt for POPE evaluation

Prompt:
Q: Is there a {object} in the image?
A:

Responses containing "Yes" are labeled as 1 (real), and all others as 0 (hallucinated). For evaluation, each question is labeled using the ground-truth object annotations from the MSCOCO dataset, following the same procedure as our main pipeline. We select the decision threshold τ that maximizes the F1 score. We report accuracy (ACC), precision (PR) for real (true) and hallucinated (false) objects, recall, and F1 score with % in Table 8. To compare computational efficiency, we also report the average inference time (Time) required to detect object hallucinations per image. For LLaVA-1.5-7B, our method achieves 2.8% improvement in the precision of real objects and a substantial 1.0% improvement in the precision of hallucinated objects compared to POPE. From a computational perspective, our method requires only a *single* forward pass for generation, whereas prompting-based methods like POPE require $(1 + C)$ forward passes—one for generating the description and C for the number of object-level verification prompts. Notably, our method reduces inference time by 6.5 seconds per generated description compared to POPE, demonstrating substantial efficiency gains. These results demonstrate the superior effectiveness and computational efficiency of our method in detecting object hallucinations, even when compared to prompting-based approach, particularly in accurately filtering hallucinated content while maintaining precision on real objects.

Table 8. Comparison with POPE on MSCOCO.

Model	**Method**	Time (s)	ACC ↑	PR (True) ↑	PR (False) ↑	Recall ↑	F1 ↑
LLaVA-1.5-7B	POPE (Li et al., 2023b)	8.1	79.5	80.3	70.8	96.7	87.8
	Ours	1.6	81.8	83.1	71.8	96.6	89.3
LLaVA-1.5-13B	POPE (Li et al., 2023b)	9.3	80.9	81.2	75.0	98.6	89.1
	Ours	1.8	83.6	86.4	75.2	95.0	90.5

D.3. Multi-token Objects

Table 9. Comparison of token selection strategies for multi-token objects.

Token	**LLaVA-1.5-7B**	**LLaVA-1.5-13B**
First	83.7	84.8
Last	83.3	84.0
Average	83.4	84.2

To compute the visual logit lens probability for object grounding and the embedding similarity for hallucination detection, we default to using the first token of multi-token objects. To evaluate the impact of this design choice, we conduct an ablation study comparing three strategies: (1) using the first token, (2) using the last token, and (3) taking the average across all tokens. Results on the MSCOCO dataset in Table 9 show that the first-token strategy is most effective, since the first token often captures the core semantic meaning of the object.

D.4. Distance Metric

We investigate the impact of the choice of distance metric for computing embedding similarity on overall performance on the MSCOCO dataset, as shown in Table 10. Specifically, we compare cosine similarity and L2 distance (*i.e.*, Euclidean distance) as the underlying metric for our GLSIM score. On the LLaVA-1.5-7B model, L2 distance yields slightly better performance, improving AUROC by 0.3%. In contrast, on the Shikra model, cosine similarity outperforms L2 distance by 1.7%. These results suggest that the effectiveness of a distance metric may depend on the model's training strategy and architecture. Nevertheless, both metrics consistently outperform the baselines in Table 1, demonstrating the robustness of our method across different metric designs.

Table 10. Ablation on the impact of distance metric.

	Metric	LLaVA	Shikra
Global	L2	80.2	77.2
	Cosine	79.3	78.9
Local	L2	79.9	75.6
	Cosine	78.8	76.8
G & L	L2	84.0	81.3
	Cosine	83.7	83.0

E. Layer-wise Performance Matrix

We provide the full performance (AUROC) matrix across all combinations of image and text embedding layers (l, l') on the MSCOCO dataset, illustrating how the composition of layers influences hallucination detection performance.

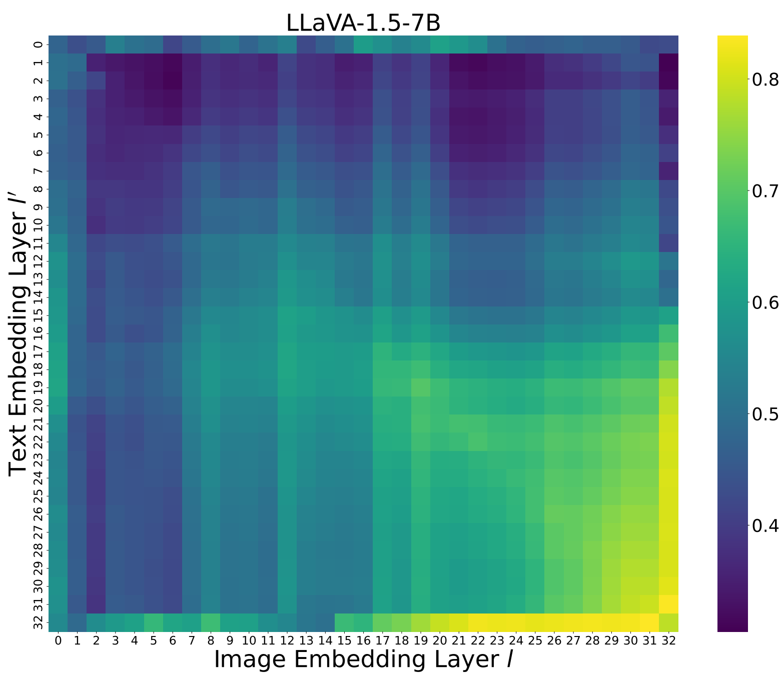


Figure 10. Performance matrix of LLaVA-1.5-7B.

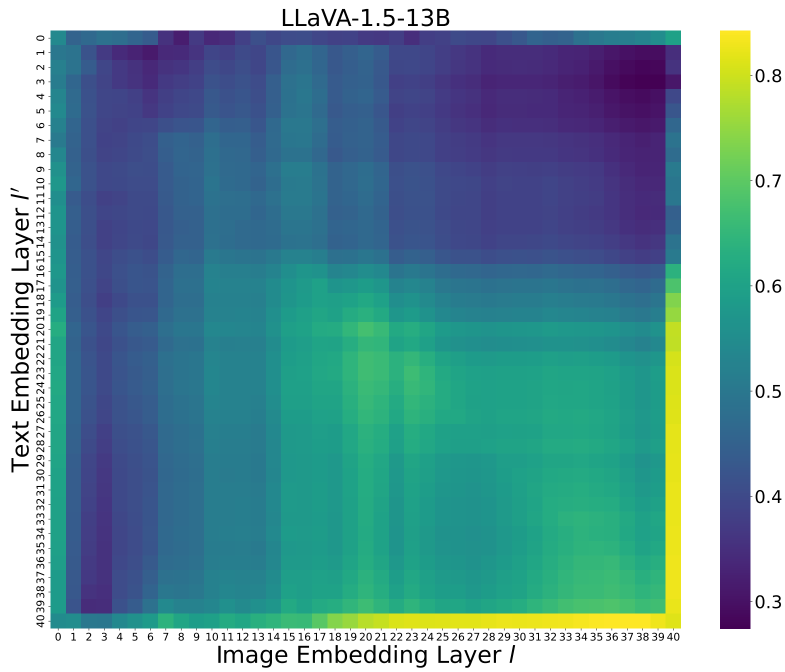


Figure 11. Performance matrix of LLaVA-1.5-13B.

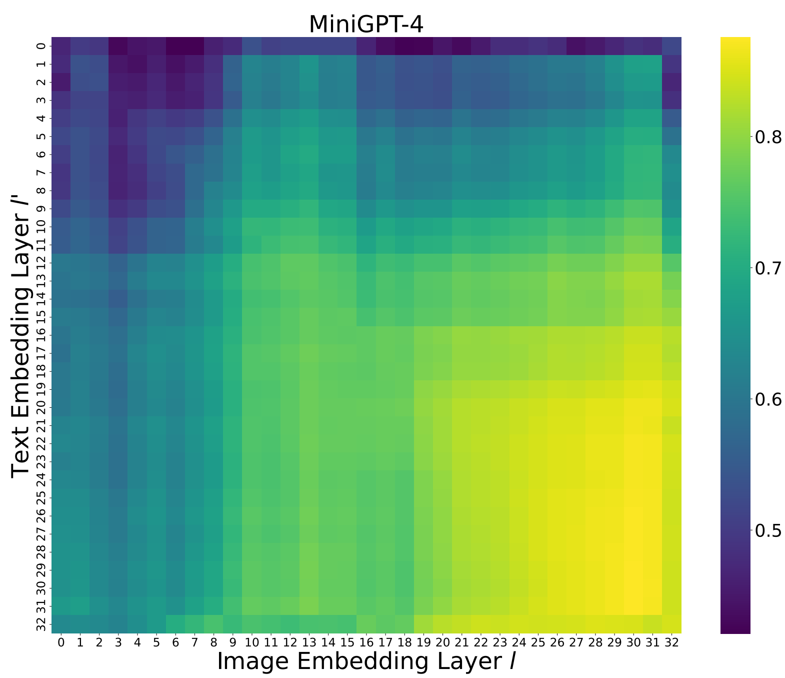


Figure 12. Performance matrix of MiniGPT-4.

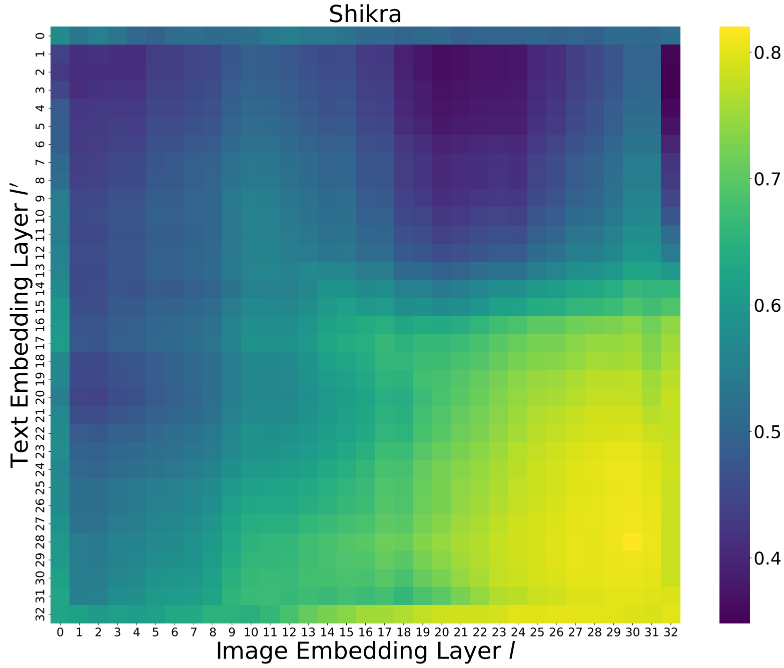


Figure 13. Performance matrix of Shikra.

F. Broader Impacts

Ensuring the reliability of LVLMs is critical as they are increasingly deployed in high-stakes domains such as autonomous navigation, medical diagnosis, and accessibility applications. This work addresses the critical challenge of object hallucination detection, which identifies objects mentioned in generated outputs that are not present in the input image. We propose a practical, training-free method that combines global and local signals from pre-trained LVLMs to enhance hallucination detection. Our research not only advances the technical frontier in this area, but also contributes to the development of trustworthy AI systems, fostering confidence in the deployment of LVLMs in safety-critical applications.