BEYOND THE CURRENT PREDICTION: TRACKING DECODING TRAJECTORIES FOR OBJECT HALLUCINATION DETECTION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) can confidently generate objects that are plausible under the generated context but absent from the image. Existing training-free object hallucination detectors mainly score each generated object using signals available at its current generation step. However, autoregressive generation can obscure uncertainty revealed earlier in the decoding process. At each step, the model assigns a probability to the generated token, but once the token is selected, this probability is no longer retained in subsequent decoding. Later predictions instead condition on the selected token itself. Consequently, a token generated with low confidence can shape the subsequent context and make a later hallucinated object appear locally confident despite uncertainty observed earlier in the generation trajectory. Motivated by this observation, we ask whether retaining uncertainty signals from preceding decoding steps can improve object hallucination detection. We introduce History-Aware Risk Tracking (HART), a simple training-free method that estimates hallucination risk using both the current object prediction and uncertainty observed throughout the preceding generation process. Specifically, HART considers both the current object prediction and uncertainty accumulated over preceding decoding steps, with greater weight assigned to more recent steps. We use negative log-likelihood as a simple token-level measure of prediction uncertainty. This formulation retains information that would otherwise no longer be directly available as generation proceeds, without requiring additional training or auxiliary generation. Experiments across multiple LVLMs on MSCOCO and Pascal VOC show that incorporating decoding history improves over current-step scoring and achieves strong object hallucination detection performance compared with existing training-free methods. Our results show that uncertainty from preceding decoding steps provides a complementary signal for object hallucination detection beyond the current object prediction alone.

1 INTRODUCTION

Large vision–language models (LVLMs) have achieved strong performance in visual understanding and language generation, supporting tasks such as image description, visual question answering, and multimodal reasoning (Liu et al., 2023; Zhu et al., 2024). Despite this progress, LVLMs remain prone to *object hallucination*, where generated responses mention objects that are absent from the input image (Rohrbach et al., 2018; Li et al., 2023). Such errors are difficult to detect because hallucinated objects are often fluent and semantically consistent with the surrounding text. An object may become a natural continuation of the generated description and therefore appear highly confident, even when it is not supported by the image. Reliable object-level hallucination detection is therefore important for understanding model failures and enabling online detection during generation.

Recent work has explored training-free hallucination detection using signals available during LVLM inference. Probability-based methods assess the reliability of a generated object using its output likelihood or predictive uncertainty (Malinin & Gales, 2021; Zhou et al., 2024). Other approaches inspect internal model signals, such as visual-attention patterns associated with the generated object (Jiang et al., 2025b), or intermediate representations and cross-modal similarity that reflect how well the object is supported by the visual input (Jiang et al., 2025a; Park & Li, 2026).

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.

More recent work also considers the generated context by measuring how strongly the current object prediction depends on preceding output tokens (Hoang-Xuan et al., 2026). Despite using different signals, these approaches primarily characterize an object using information available when that object is generated. Even when preceding generated tokens are considered, they areused as context for interpreting the current prediction rather than preserving the prediction information that was available when those tokens were originally generated. This can make a particular class of hallucinations difficult to detect. A typical example can arise when the model generates *racket* at an earlier step with relatively high prediction uncertainty. Once *racket* is selected, subsequent decoding conditions on the token itself rather than on the uncertainty under which it was generated. The resulting context may then make *ball* a natural and high-confidence continuation, even when no ball is present in the image. A detector that primarily characterizes *ball* from its current prediction may therefore miss the elevated uncertainty that appeared earlier in the decoding trajectory.

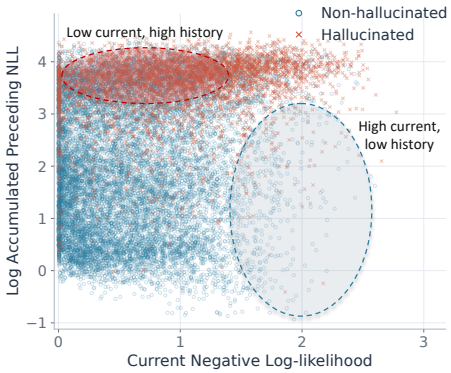


Figure 1: Current-token NLL versus log cumulative preceding-token NLL for LLaVA-1.5-7B on MSCOCO. The highlighted regions illustrate cases in which hallucinated objects have low current-token NLL but high preceding cumulative NLL, while non-hallucinated objects exhibit the opposite pattern.

To examine this effect more systematically, we analyze prediction uncertainty throughout autoregressive generation. At each decoding step t, the model assigns probability p_t to the selected token. We use its negative log-likelihood (NLL), $\ell_t = -\log p_t$, as a simple token-level measure of prediction uncertainty, and compare uncertainty at the current object prediction with that accumulated along the preceding decoding trajectory. Figure 1 reveals a clear mismatch between these two quantities. Some hallucinated objects have low current-token NLL despite substantially higher uncertainty in the preceding trajectory, whereas some non-hallucinated objects exhibit the opposite pattern. This suggests that the current object prediction and the preceding decoding trajectory capture complementary information, and that preserving uncertainty from earlier decoding decisions can provide useful evidence for object hallucination detection.

Motivated by this observations, we introduce History-Aware Risk Tracking (HART), a simple training-free method that retains prediction uncertainty throughout autoregressive generation. HART evaluates each generated object using both the uncertainty of the current prediction and uncertainty observed along the preceding decoding trajectory. To place greater emphasis on decoding steps that are more relevant to the current prediction, HART applies an exponential decay to the preceding uncertainty signals, assigning larger weights to more recent steps. The resulting historical signal is then combined with the current prediction to obtain an object-level hallucination score. HART operates directly on signals produced during the original image-conditioned decoding process and requires neither additional training nor auxiliary generation.

Experiments across multiple LVLMs on MSCOCO and Pascal VOC show that HART consistently improves object hallucination detection over current-step scoring and performs strongly against existing training-free baselines. Further analysis shows that current and preceding uncertainty provide complementary information for hallucination detection. Since HART updates its decoding history incrementally, it can also be applied online as generation proceeds. Our main contributions are summarized as follows:

- We identify an overlooked limitation of current-step hallucination detection in autoregressive LVLMs. Once a token is generated, the prediction risk associated with that decoding decision is no longer preserved, even though the token can continue to influence later predictions.
- We propose HART, a simple training-free detector that records token-level prediction risk during decoding and combines current risk with a decayed accumulation of preceding risks, preserving information about the generation trajectory with minimal additional overhead.

- We conduct extensive experiments across multiple LVLMs on MSCOCO and Pascal VOC, showing that history-aware risk tracking consistently improves object-level hallucination detection and performs strongly against existing training-free baselines without additional training or auxiliary generation.

2 Related Work

2.1 Object Hallucination in Vision-Language Models

Object hallucination refers to the generation of objects that are absent from or unsupported by the visual input. The problem was studied early in image captioning, where CHAIR was introduced to measure whether generated object mentions are grounded in the image (Rohrbach et al., 2018). With the emergence of LVLMs, Li et al. (2023) showed that object hallucination remains prevalent and introduced POPE for systematic evaluation. Subsequent work has further developed evaluation protocols for identifying unsupported object mentions in open-ended generations (Petryk et al., 2024). Prior studies have also connected object hallucination to language priors, object co-occurrence, and insufficient use of visual evidence (Zhou et al., 2024; Leng et al., 2024; Huang et al., 2024). These findings have motivated both methods for reducing hallucination during generation and methods for detecting unsupported content from model outputs and internal signals.

2.2 Inference-Time Hallucination Detection

Recent work has explored hallucination detection using signals available from pretrained LVLMs during inference. Probability-based approaches use output likelihood or predictive uncertainty as indicators of unreliable generations. Zhou et al. (2024), for example, use the negative log-probability of generated object tokens as a confidence-based signal for object hallucination. Other approaches exploit internal model signals. Internal Confidence projects intermediate image representations through the language head to estimate their support for generated objects (Jiang et al., 2025a), while Contextual Lens uses contextualized intermediate representations for hallucination detection and grounding (Phukan et al., 2025). Visual attention provides another source of information. SVAR characterizes visual-attention patterns associated with hallucinated object tokens (Jiang et al., 2025b), while GLSim uses global and local cross-modal similarities to assess visual support for generated objects (Park & Li, 2026). Recent work has also examined the influence of generated context. PAS measures how strongly the current object prediction depends on preceding output tokens, showing that stronger dependence on the generated prefix can be associated with hallucination (Hoang-Xuan et al., 2026). These approaches demonstrate that probability, visual, representation, and contextual signals provide useful evidence for object hallucination detection. Our work considers a complementary source of information by retaining prediction uncertainty observed at preceding decoding steps and using it together with the current object prediction.

3 Preliminaries

3.1 Autoregressive LVLM Generation

Given an image I and a textual prompt x, a large vision–language model generates a response autoregressively as a sequence of tokens

$$Y = (y_1, y_2, \ldots, y_T).$$

At each decoding step t, the model predicts the next token conditioned on the image, the prompt, and all previously generated tokens. We denote the probability assigned to the generated token y_t by

$$p_t = p(y_t \mid I, x, y_{<t}).$$

This probability reflects the model's confidence in the current prediction. As generation proceeds, subsequent predictions condition on the selected tokens, while the uncertainty associated with earlier predictions is not explicitly retained in the generated prefix

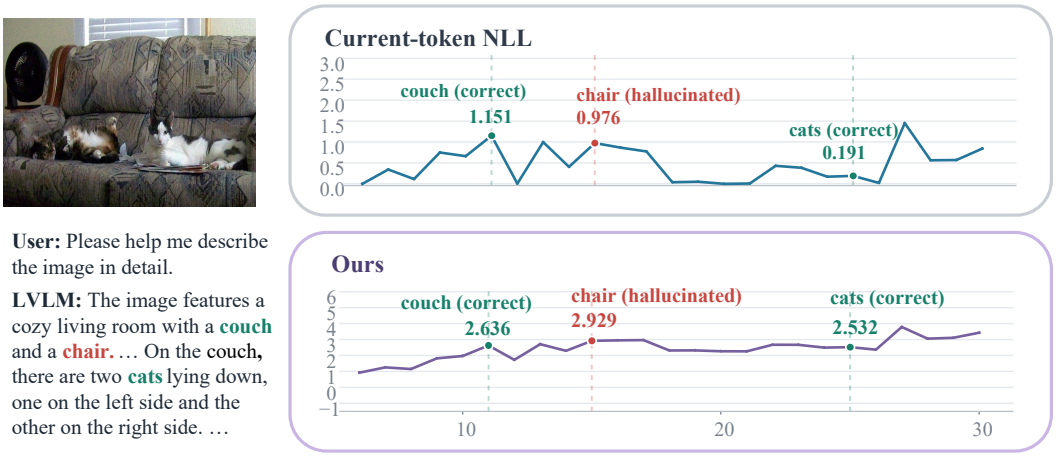


Figure 2: Illustration comparing current-token NLL with HART scores during autoregressive decoding. The hallucinated object "chair" has a lower current-token NLL than the preceding grounded object, whereas HART assigns it a higher score by incorporating preceding generation uncertainty.

3.2 Object-Level Hallucination Detection

We consider object mentions identified in the generated response according to the benchmark-specific evaluation protocol. Let

$$\mathcal{O} = \{o_1, o_2, \ldots, o_K\}$$

denote the sequence of evaluated object categories, ordered by their first occurrence in the generated response. If the same object category is mentioned multiple times, we retain only its first occurrence for evaluation.

For each object o_k, we define its decoding position $t_k \in \{1, \ldots, T\}$ as the position of the first token of its first occurrence. For each image, the evaluation protocol provides a set of object categories that are visually present. An evaluated object o_k is labeled as hallucinated if its category is absent from this ground-truth object set.

The goal of object-level hallucination detection is to assign each evaluated object o_k a hallucination detection score $S(o_k)$, such that hallucinated objects receive higher scores than visually supported objects. We focus on a training-free setting in which the pretrained LVLM remains frozen and the detector uses only signals available during the original image-conditioned generation process. Since generation proceeds sequentially, these signals may come from both the current object prediction and preceding decoding steps. No additional model training or auxiliary generation is required.

4 History-Aware Risk Tracking

4.1 Overview

HART assesses the hallucination risk of each generated object using both the current object prediction and the generation uncertainty observed along the preceding decoding history. At each decoding step, we use the NLL of the generated token as a simple probability-based measure of uncertainty, following prior work on object hallucination in LVLMs (Zhou et al., 2024). Using this basic signal also allows us to directly examine the additional information gained by retaining generation uncertainty over time. We retain this token-level uncertainty throughout the decoding trajectory because uncertainty observed at earlier steps may influence subsequent generation while no longer being reflected in the current object probability. HART applies exponential decay to the preceding uncertainty, assigning greater weight to more recent decoding steps because they are more directly related to the context leading to the current object prediction. The resulting historical uncertainty is then combined with the current-token NLL to form the hallucination detection score for each generated object.

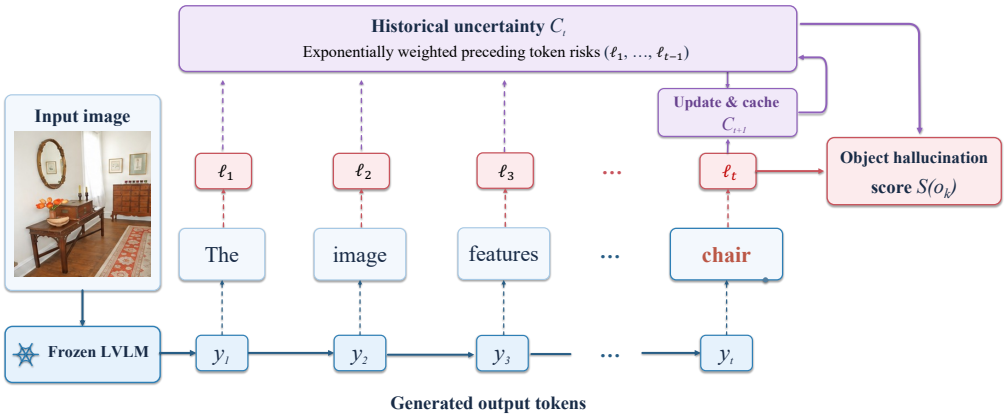


Figure 3: Overview of History-Aware Risk Tracking (HART). At each decoding step, HART measures prediction risk from the generated-token probability and retains risks from preceding decoding steps using mass-preserving exponential decay. When an object is generated, the resulting historical risk is log-transformed and combined with the current prediction risk to obtain the object-level hallucination score.

4.2 Current Prediction Uncertainty

Using the generated-token probability p_t defined in Section 3, we measure prediction uncertainty at decoding step t using negative log-likelihood (NLL), $\ell_t = -\log p_t$. A lower-probability token has a larger NLL and therefore indicates greater uncertainty at that decoding step. Prior work has used NLL as a simple probability-based signal for object hallucination detection (Zhou et al., 2024).

For an evaluated object o_k beginning at decoding position t_k, we use the NLL of its first token, ℓ_{t_k}, as its current prediction uncertainty. This quantity reflects the model's uncertainty when the object first appears. However, it only characterizes the current object prediction and does not account for uncertainty observed earlier in the generation process.

4.3 Mass-Preserving Historical Uncertainty

The uncertainty observed before an object is generated can contain information that is not captured by the current object prediction alone. In autoregressive generation, once a token y_t is selected, subsequent predictions condition on the generated token itself rather than on the probability p_t with which it was generated. Consequently, an earlier decoding step can have high uncertainty even when a later object is assigned high probability under the resulting prefix. To retain this information, HART keeps the token-level NLL values from preceding decoding steps and summarizes the uncertainty along the generation trajectory leading to the current object.

Not all preceding steps should contribute equally to the current prediction. Uncertainty observed more recently is closer to the context immediately leading to the current object, whereas uncertainty from more distant steps is less directly related. We therefore weight preceding token-level NLLs using exponential decay, assigning greater weight to recent steps while gradually reducing the contribution of earlier ones.

Specifically, we use exponential decay with half-life h. We set $\gamma = 2^{-1/h}$ and assign a preceding token at position $s < t$ the relative weight $w_{s,t} = \gamma^{\,t-1-s} = 2^{-(t-1-s)/h}$. Thus, the relative weight is halved whenever the distance from the current prediction increases by h tokens.

Applying the decay weights directly would reduce the total weight assigned to the decoding history. This would change not only the relative importance of different historical positions, but also the overall scale of the historical component. We want the decay to control how the historical contribution is distributed across preceding positions without changing its total weight. We therefore normalize the decay weights so that their total remains $t - 1$:

$$\alpha_{s,t} = (t-1)\frac{w_{s,t}}{\sum_{j=1}^{t-1} w_{j,t}}, \qquad \sum_{s=1}^{t-1} \alpha_{s,t} = t-1. \tag{1}$$

We then define the historical uncertainty before step t as

$$C_t = \sum_{s=1}^{t-1} \alpha_{s,t}\ell_s = (t-1)\frac{\sum_{s=1}^{t-1} w_{s,t}\ell_s}{\sum_{s=1}^{t-1} w_{s,t}}. \tag{2}$$

This mass-preserving formulation keeps the total weight assigned to the decoding history unchanged while placing more weight on recent decoding steps. We compute C_t using only response tokens generated before step t. The textual prompt and the current token are not included.

The historical quantity can also be maintained online without storing or repeatedly traversing the full decoding history. We maintain a decayed NLL sum A_t and a decayed weight sum Z_t,

$$A_t = \sum_{s=1}^{t-1} \gamma^{\,t-1-s}\ell_s, \qquad Z_t = \sum_{s=1}^{t-1} \gamma^{\,t-1-s}. \tag{3}$$

Starting from $A_1 = Z_1 = 0$, they can be updated after each generated token as

$$A_{t+1} = \gamma A_t + \ell_t, \qquad Z_{t+1} = \gamma Z_t + 1. \tag{4}$$

The mass-preserving historical uncertainty is then obtained as $C_t = (t-1)A_t/Z_t$. HART therefore requires only constant additional state and computation at each decoding step.

4.4 Object-Level Hallucination Score

For an object o_k beginning at position t_k, HART combines the current prediction uncertainty with the historical uncertainty accumulated before the object is generated. Because C_{t_k} aggregates uncertainty over multiple preceding decoding steps, it can have a much larger scale than the current-token NLL. We apply a logarithm to reduce this scale difference before combining the two terms without changing the ordering of the historical signal.

The final object-level hallucination score is

$$\boxed{S(o_k) = \log C_{t_k} + \ell_{t_k}}. \tag{5}$$

The two terms capture complementary information from the decoding process. The current-token NLL reflects the model's uncertainty when the object first appears, while the historical term summarizes uncertainty observed along the preceding generation trajectory, with greater weight assigned to more recent decoding steps.

This combination allows an object with low current uncertainty to remain distinguishable when the preceding generation trajectory contains elevated uncertainty. At the same time, the current-token NLL retains information specific to the object being generated, so the historical trajectory does not alone determine the object-level score.

Figure 2 illustrates this complementarity. The hallucinated object has a lower current-token NLL than the preceding grounded object, making it difficult to distinguish from the current prediction alone. Its preceding decoding trajectory, however, contains elevated uncertainty. By retaining this historical information, HART assigns the hallucinated object a higher score than the grounded objects, making it easier to identify at the object level.

Because the historical statistics are updated recursively according to Eq. (4), HART requires only constant additional computation and storage at each decoding step. All required probabilities are obtained from the original decoding process, preserving the training-free and single-generation setting.

Table 1: Object hallucination detection performance across representative LVLMs on MSCOCO and Pascal VOC. We report AUROC (%) across four models covering different model families and scales. The average is computed over all model and dataset combinations.

Method	Shikra-7B		InternVL3.5-4B-HF		InternVL3.5-8B-HF		Qwen2.5-VL-7B-Instruct		Average
	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	
NLL Zhou et al. (2024)	54.57	61.70	58.41	68.23	61.75	72.74	60.09	67.87	63.17
Entropy Malinin & Gales (2021)	72.04	66.32	62.96	58.31	57.14	51.82	64.33	61.90	61.85
IC Jiang et al. (2025a)	71.65	69.51	55.91	52.52	56.08	52.12	59.08	59.59	59.56
GLSim Park & Li (2026)	75.52	74.27	66.63	66.50	68.02	61.32	64.65	67.71	68.08
SVAR Jiang et al. (2025b)	72.35	73.84	73.38	78.09	74.29	78.56	**75.70**	78.36	75.57
PAS Hoang-Xuan et al. (2026)	84.40	85.76	73.11	77.22	71.61	76.09	73.82	77.58	77.45
Ours	**86.41**	**86.15**	**76.66**	**80.33**	**75.26**	**79.38**	75.13	**78.57**	**79.74**

Table 2: Object hallucination detection performance (AUROC, %) across different generations and scales of the LLaVA family on MSCOCO and Pascal VOC. The average is computed over all model and dataset combinations.

Method	LLaVA-1.5-7B		LLaVA-1.5-13B		LLaVA-v1.6-Vicuna-13B		LLaVA-OneVision-Qwen2-7B		Average
	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	
NLL Zhou et al. (2024)	56.34	61.93	56.70	62.52	64.96	76.43	58.65	71.44	63.62
Entropy Malinin & Gales (2021)	71.26	67.20	70.60	64.67	66.09	60.83	65.27	59.81	65.72
IC Jiang et al. (2025a)	70.18	67.35	66.63	66.17	70.78	66.20	72.01	75.14	69.31
GLSim Park & Li (2026)	72.50	73.90	77.25	78.11	69.16	70.47	63.60	55.37	70.05
SVAR Jiang et al. (2025b)	81.52	82.87	81.69	83.45	**75.68**	76.44	77.68	79.79	79.89
PAS Hoang-Xuan et al. (2026)	84.14	85.68	83.84	86.18	71.83	77.14	73.46	79.01	80.16
Ours	**86.13**	**85.98**	**86.01**	**86.18**	74.73	**77.58**	**77.70**	**80.12**	**81.80**

5 Experiments

5.1 Experimental Setup

Models. We evaluate HART across a diverse set of LVLMs. For cross-architecture evaluation, we consider Shikra-7B (Chen et al., 2023), InternVL3.5-4B-HF and InternVL3.5-8B-HF (Wang et al., 2025), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). We additionally evaluate multiple generations and model scales within the LLaVA family, including LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024a), LLaVA-v1.6-Vicuna-13B (Liu et al., 2024b), and LLaVA-OneVision-Qwen2-7B (Li et al., 2024).

Benchmarks. Following prior object hallucination detection protocols (Hoang-Xuan et al., 2026), we evaluate on MSCOCO (Lin et al., 2014) and Pascal VOC (Everingham et al., 2010). We use 5,000 images from the MSCOCO `val2014` split and the full Pascal VOC `val2012` split containing 5,823 images. All models are prompted with "*Please help me describe the image in detail.*" Generated object mentions are identified and labeled as grounded or hallucinated following the CHAIR object-level evaluation protocol (Rohrbach et al., 2018).

Baselines. We compare HART with commonly used training-free object hallucination detection baselines, including Negative Log-Likelihood (NLL) (Zhou et al., 2024), Entropy (Malinin & Gales, 2021), Internal Confidence (IC) (Jiang et al., 2025a), Global-Local Similarity (GLSim) (Park & Li, 2026), Summed Visual Attention Ratio (SVAR) (Jiang et al., 2025b), and Prelim Attention Score (PAS) (Hoang-Xuan et al., 2026). All methods are evaluated on the same generated responses and object-level annotations.

Implementation Details. We use a half-life of $h = 32$ for HART across all models and datasets. We report AUROC (%) as the primary evaluation metric. Additional sensitivity analysis of the half-life is provided in Appendix A.3.

5.2 Overall Detection Performance

Table 1 compares HART with existing training-free object hallucination detectors across four representative LVLMs from different model families and scales. HART achieves consistently strong performance across these architectures, with an average AUROC of 79.74, outperforming the strongest

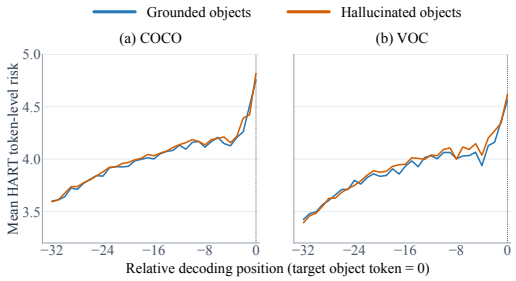


Figure 4: Matched token-risk trajectories before the target object on MSCOCO and Pascal VOC using LLaVA-1.5-7B. Despite having similar risk at the target token, hallucinated objects tend to exhibit higher risk earlier in generation.

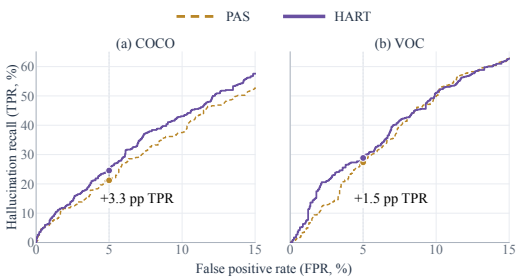


Figure 5: Low-FPR hallucination detection on the more confident 50% of object predictions on MSCOCO and Pascal VOC using LLaVA-1.5-7B. At 5% FPR, HART improves hallucination recall over PAS by 3.3 percentage points on MSCOCO and 1.5 points on Pascal VOC.

Table 3: Component ablation of HART on MSCOCO and Pascal VOC. We report AUROC (%) using LLaVA-1.5-7B.

Variant	Current NLL	Preceding Uncertainty	MSCOCO	Pascal VOC
Current only	✓	×	70.40	64.45
Historical only	×	✓	84.34	85.41
Full HART	✓	✓	**86.13**	**85.98**

competing baseline, PAS, by 2.29 points. HART performs particularly well on Shikra-7B, improving over PAS from 84.40 to 86.41 on MSCOCO and from 85.76 to 86.15 on Pascal VOC. It also improves over the strongest baselines on both datasets for InternVL3.5-4B-HF and InternVL3.5-8B-HF. These results show that HART generalizes well across different LVLM architectures.

Table 2 further evaluates HART across different generations and scales of the LLaVA family. HART maintains strong performance across these models and obtains the highest average AUROC of 81.80, improving over PAS by 1.64 points. On LLaVA-1.5-7B and LLaVA-1.5-13B, HART consistently matches or exceeds the strongest baseline on both datasets. The improvements also extend to newer LLaVA variants. HART achieves the highest Pascal VOC AUROC on LLaVA-v1.6-Vicuna-13B and the highest AUROC on both datasets for LLaVA-OneVision-Qwen2-7B. Overall, the results indicate that the benefit of history-aware risk tracking is maintained across different model generations, backbone architectures, and model scales.

5.3 Analysis of Preceding Generation Uncertainty

We first examine the contributions of the current prediction and the preceding generation uncertainty. As shown in Table 3, using only the current-token NLL leads to a substantial performance drop on both datasets, while the accumulated uncertainty from preceding decoding steps already provides strong detection performance. Combining the two yields the best results, showing that preceding generation uncertainty provides the main performance gain while the current prediction contributes complementary information.

We then examine whether the benefit of preceding generation uncertainty can be explained by decoding position or prefix length alone. As shown in Table 4, decoding position provides a strong signal, but remains less effective than using the uncertainty observed along the actual generation trajectory. Normalizing the accumulated uncertainty by prefix length leads to a clear performance drop. Replacing the actual preceding uncertainty with position-matched uncertainty from other samples further reduces performance. These results indicate that the improvement cannot be explained by decoding position alone and that the uncertainty observed during the sample's own preceding generation provides additional information for hallucination detection.

Table 4: Analysis of historical uncertainty in HART on MSCOCO and Pascal VOC. We report AUROC (%) using LLaVA-1.5-7B.

Variant	Score	MSCOCO	Pascal VOC
Current only	ℓ_{t_k}	70.40	64.45
Position only	$\log(t_k - 1)$	83.73	85.40
Mean history	$\ell_{t_k} + \log\left(\frac{C_{t_k}}{t_k - 1}\right)$	76.46	74.76
Other-sample history	$\ell_{t_k} + \log C^{\text{other}}_{t_k}$	58.37 ± 0.29	55.74 ± 0.61
Actual history	$\ell_{t_k} + \log C_{t_k}$	**86.13**	**85.98**

We next examine whether the preceding decoding trajectory remains informative after controlling for differences at the target object. We use stratified matching to balance hallucinated and grounded objects with respect to their onset position and HART risk at the target step, thereby controlling for differences in both decoding position and target-step risk. We then compare their preceding risk trajectories under these matched conditions. As shown in Figure 4, hallucinated objects exhibit higher risk over the preceding decoding trajectory on both datasets. This indicates that generation history contains information associated with hallucination beyond what is captured by the target-step risk alone. Details of the matching procedure are provided in Appendix A.4.

Finally, we examine objects whose current predictions are relatively confident. As shown in Figure 5, HART maintains strong hallucination detection at low false-positive rates within this subset. This further demonstrates the complementary role of preceding generation uncertainty. Even when the current prediction uncertainty is low, uncertainty observed earlier in the decoding trajectory can still help distinguish hallucinated from grounded objects.

5.4 Limitations

HART has several limitations. First, it uses a fixed exponential decay over preceding decoding steps, although the relevance of earlier uncertainty may vary across objects and contexts. More adaptive weighting could better capture which parts of the decoding trajectory are informative. Second, HART requires access to token-level probabilities, limiting its applicability to closed-source models that expose only generated text. Finally, our evaluation focuses on object hallucination in image-description generation on MSCOCO and Pascal VOC. Whether the same trajectory patterns generalize to other hallucination types, such as unsupported attributes or relations, and to tasks such as visual question answering and multimodal reasoning remains to be studied. Broader evaluation would help determine how generally historical uncertainty transfers across multimodal generation settings.

6 Conclusion

In this work, we introduced HART, a training-free approach for object hallucination detection in LVLMs that incorporates uncertainty observed throughout the preceding generation process when evaluating each generated object. HART combines the current object prediction with a decayed accumulation of preceding token-level uncertainty, allowing information revealed earlier in decoding to remain available for later object predictions. Experiments across diverse LVLM architectures and different generations of the LLaVA family demonstrate strong and consistent detection performance. Our analyses further show that the improvement cannot be explained by decoding position alone. The uncertainty observed along each sample's own preceding generation trajectory provides additional information beyond the current prediction, and hallucinated objects tend to exhibit greater preceding uncertainty even when their current predictions appear similarly confident. These results highlight the value of retaining generation uncertainty over time for object hallucination detection and suggest that the decoding trajectory contains useful information for identifying unreliable model outputs.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design and to assist with the interpretation of experimental results. We also used generative AI tools for brainstorming, editing software code, improving the readability and structure of the manuscript, formatting references, and searching for relevant information and literature.

REPRODUCIBILITY STATEMENT

We provide the methodological details and formulations of HART in the main paper, including the construction of the historical risk and the object-level hallucination score. The experimental setup describes the evaluated models, datasets, baselines, and evaluation protocol. Additional implementation details, hyperparameter analyses, and extended experimental results are provided in the appendix.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2502.13923.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Nhat Hoang-Xuan, Minh Vu, My T. Thai, and Manish Bhattarai. Pas: Prelim attention score for detecting object hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025b.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

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*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 1(8), 2024b.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Seongheon Park and Sharon Li. Glsim: Detecting object hallucinations in lvlms via global-local similarity. *Advances in Neural Information Processing Systems*, 38:37747–37776, 2026.

Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357, 2024.

Anirudh Phukan, Harshit Kumar Morj, Apoorv Saxena, Koustava Goswami, et al. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9661–9675, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

A Appendix

A.1 Token-Level NLL Computation

For HART and the *current-only* baseline, we compute the NLL of the token actually generated by the model at the corresponding object position. This can differ from scoring the canonical object-category token used by the evaluation protocol. For example, if the generated caption contains *woman* while the corresponding evaluation category is *person*, we use the probability assigned to *woman* at that decoding step rather than the probability of *person*. The hallucination label is determined separately by the object-level evaluation protocol and is therefore unaffected by this scoring choice.

A.2 History Window Analysis

We further examine how much of the preceding generation history is needed for effective hallucination detection. Rather than tuning a history window, this analysis asks whether the useful historical signal is concentrated only in the immediately preceding tokens or extends over a broader portion of the decoding trajectory.

Directly truncating the history also reduces the total accumulated historical risk, making different window sizes difficult to compare because of the resulting scale change. We therefore use *mass-normalized* window variants, where the retained historical weights are rescaled to preserve the same total weight as the full history. This isolates the effect of which historical tokens are retained from the effect of simply reducing the total historical contribution.

As shown in Table 5, removing the generation history causes a substantial drop in detection performance, confirming that the preceding decoding trajectory provides information beyond the current prediction alone. Using only the most recent 16 tokens already recovers much of the full-history performance, while extending the window to the most recent 32 tokens essentially matches the full-history result on both datasets.

The location of the retained history also matters. Keeping only the earliest 16 tokens performs consistently worse than keeping the most recent 16 tokens, suggesting that historical signals closer to the current object prediction are more informative. At the same time, the improvement from 16 to 32 recent tokens indicates that the useful historical signal is not confined to only the immediate local context.

Table 5: Effect of the historical range on object hallucination detection using LLaVA-1.5-7B. Windowed variants use mass normalization to preserve the total historical weight of the full-history variant. All values are percentages.

	MSCOCO		Pascal VOC	
History	AUROC	AUPRC	AUROC	AUPRC
No history	70.40	45.56	64.45	48.08
Full history	86.13	63.55	85.98	71.89
Most recent 16	85.89	62.89	85.86	71.61
Most recent 32	86.11	63.49	85.97	71.93
Earliest 16	85.55	62.26	85.69	70.83

Overall, these results show that recent decoding steps contain the strongest historical signal, while a moderately broader history can provide additional information. This pattern is consistent with HART's use of the complete decoding history with greater weight assigned to more recent generation steps, rather than imposing a fixed hard cutoff.

A.3 Sensitivity to the Decay Half-Life

We examine the sensitivity of HART to the half-life h used for weighting preceding decoding steps. All experiments use LLaVA-1.5-7B and the full MSCOCO and Pascal VOC evaluation sets. We

Table 6: Sensitivity of HART to the exponential-decay half-life h. We report AUROC (%) on MSCOCO and Pascal VOC using LLaVA-1.5-7B. Decay weights are normalized to preserve the total weight assigned to the preceding generation history.

Half-life h	MSCOCO	Pascal VOC
4	85.790	85.714
8	86.038	85.906
16	86.125	85.970
32	**86.126**	**85.980**
64	86.111	85.974

use the mass-preserving decay formulation described in Section 4, where the decay weights are normalized so that the total weight assigned to the preceding generation history remains unchanged.

Table 6 reports the AUROC for half-life values ranging from 4 to 64 tokens. Performance becomes stable once the half-life is moderately large. In particular, $h = 16$, $h = 32$, and $h = 64$ yield very similar results on both datasets. The setting $h = 32$, used throughout our main experiments, achieves 86.126 AUROC on MSCOCO and 85.980 on Pascal VOC. These results indicate that HART is not sensitive to the precise choice of the decay half-life within this range.

A.4 Stratified Trajectory Comparison

We use stratification and reweighting to compare the preceding decoding trajectories of hallucinated and grounded objects under comparable target-step conditions. Each object mention is aligned by its onset, defined as the first subtoken of the object, which is assigned relative position 0. We retain objects whose onset occurs at decoding position at least 32, so that every example has a complete trajectory over relative positions $[-32, 0]$.

The matching procedure controls for two variables: the absolute decoding position of the object onset and the HART risk at the onset step. For each dataset, we compute the 20th, 40th, 60th, and 80th weighted percentiles of both variables over the objects included in this analysis. These boundaries divide each variable into five bins, producing a 5×5 stratification.

We restrict the analysis to strata containing both hallucinated and grounded objects, ensuring common support between the two groups. Within each retained stratum c, we compute the weighted mean trajectory separately for hallucinated and grounded objects at each relative position k:

$$\bar{H}^{\mathrm{hal}}_{c,k} \quad \text{and} \quad \bar{H}^{\mathrm{grd}}_{c,k},$$

where the weights account for the number of draws associated with each object instance.

Finally, we reweight both groups using the distribution of hallucinated objects across the retained strata. Let π^{hal}_c denote the normalized hallucinated-object weight in stratum c. The reported trajectories are

$$\mu^{\mathrm{hal}}_k = \sum_c \pi^{\mathrm{hal}}_c \bar{H}^{\mathrm{hal}}_{c,k}, \qquad \mu^{\mathrm{grd}}_k = \sum_c \pi^{\mathrm{hal}}_c \bar{H}^{\mathrm{grd}}_{c,k}.$$

Thus, the grounded trajectory is standardized to the same joint distribution of onset position and target-step HART risk as the hallucinated group. This allows the comparison to focus on differences in the preceding decoding trajectory rather than differences in where the object appears or how risky the target step is. In our analysis, 24 of 25 strata satisfy common support on MSCOCO, while all 25 strata are retained on Pascal VOC.

A.5 Qualitative Analysis

We provide qualitative examples to illustrate the complementary roles of current risk and preceding generation uncertainty in HART.

Hallucinations with low current risk. Figure 6 shows examples from Pascal VOC where the target hallucinated object has relatively low current risk. Based on the current prediction alone, these

objects are therefore difficult to distinguish from grounded predictions. However, the preceding decoding trajectory contains elevated uncertainty, which increases the HART score at the target object. These examples show that uncertainty associated with a hallucinated object can emerge before the object itself is generated and may not be fully reflected in its current risk.

Grounded objects following hallucinations. Figure 7 shows examples from MSCOCO where a hallucinated object is followed shortly by a grounded object. Because the uncertainty accumulated over preceding decoding steps changes gradually, its contribution can remain high for the subsequent object. The current risk provides complementary information in these cases. When the later object is generated with lower current risk, the full HART score decreases relative to the preceding hallucinated object. This helps distinguish grounded objects even when the preceding decoding trajectory contains elevated uncertainty.



Figure 6: Qualitative examples on Pascal VOC where hallucinated objects have relatively low current risk but elevated uncertainty in the preceding decoding trajectory. The red word marks the target hallucinated object. Although the current prediction appears relatively confident, incorporating the preceding generation uncertainty produces a higher HART score at the target object.



Figure 7: Qualitative examples on MSCOCO showing grounded objects generated shortly after hallucinated ones. H denotes a hallucinated object and G denotes the subsequent grounded object. Preceding generation uncertainty remains elevated after the hallucination, while the lower current risk of the grounded object reduces its full HART score.