LOOKBACK: Where and How to Score LVLM Responses via Visual Reference Usage

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) integrate visual perception with language generation, enabling responses that span image understanding and complex reasoning. However, LVLMs do not just inherit the text-level hallucinations; they also hallucinate against the image, producing fluent responses ungrounded in what they see. This makes LVLM response scoring inherently harder, and our diagnostics show that existing confidence-based metrics adopted from LLMs are insufficient for LVLMs. Specifically, removing the input image barely changes confidence-based selection, suggesting that output-space confidence primarily captures textual plausibility rather than agreement with the image. To address this gap, we propose LOOKBACK, a training-free LVLM response scoring method that augments token likelihood with visual lookback score, a lightweight measure of how strongly each response token refers to image tokens. Across four benchmarks and three models, LOOKBACK consistently improves Best-of-N selection over existing baselines with negligible additional overhead.

1 Introduction

With the recent success of Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023), Large Vision-Language Models (LVLMs) have integrated visual understanding with text generation, demonstrating strong capability from image perception to complex reasoning (Dai et al., 2023; Liu et al., 2024; Zhu et al., 2024; Bai et al., 2025; Zhu et al., 2025). Despite these advances, LVLMs inherit a fundamental challenge of autoregressive generation: they can produce responses that are fluent and plausible yet factually incorrect. In multimodal settings, this challenge is further complicated by visual hallucination, where a response asserts objects, attributes, or relations that are not supported by the image (Rohrbach et al., 2018; Li et al., 2023; Leng et al., 2024; Guan et al., 2024). Consequently, reliably identifying which responses are faithful to the image becomes an important problem for LVLMs.

A common setting for improving reliability is *Best-of-N* selection, originating from the LLM literature: multiple candidate responses are sampled, and the best one is selected (Wang et al., 2023b; Gui et al., 2024; Snell et al., 2025). The core of Best-of-N is the selection criterion that scores each candidate, and a prominent way is to train reward models or verifiers to assess response quality (Cobbe et al., 2021; Lightman et al., 2024). This idea has recently been extended to LVLMs through multimodal reward models that evaluate the visual correctness of candidates (Wang et al., 2025; Chen et al., 2025). While effective, these methods require additional models, task-specific supervision, or preference annotations, which limits their applicability.

A natural auxiliary-free direction is the model's own output-space confidence, *i.e.*, how plausible a response is under its output distribution. In the LLM literature, such confidence signals have proven a useful selection criterion that requires no external verifier (Wang and Zhou, 2024; Kang et al., 2025). This makes the model's own output distribution an appealing candidate for scoring LVLM responses, as it requires no auxiliary model. However, LVLM response scoring differs from its text-only counterpart in one crucial respect: response quality depends on an additional source of evidence, *the image*. Through a diagnostic analysis, we find that confidence-based selection remains nearly as effective even when the confidence is computed without conditioning on the image. This reveals a gap between image-conditioning and image-sensitivity: the confidence score reflects a response's textual plausibility rather than its agreement with the image. Since visual hallucination arise from a mismatch between the response and the image, such an image-insensitive signal cannot reliably detect them.

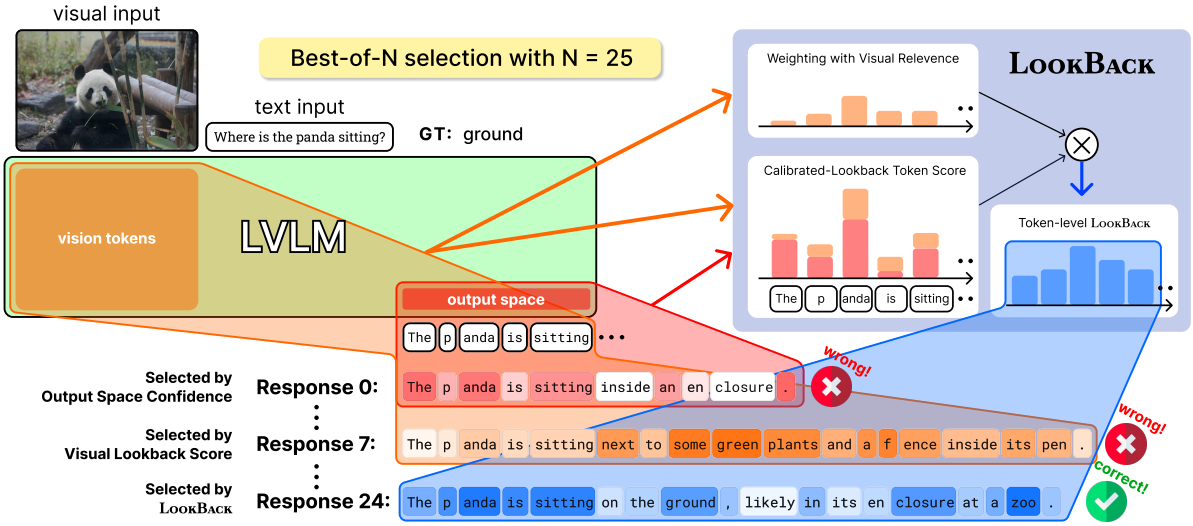

Figure 1: **Overview of LookBack: combining output-space confidence with visual reference usage.** Given an image and a question, an LVLM generates multiple candidate responses and our goal is to select the proper one. Output-space confidence can assign high scores to fluent but visually unsupported tokens, while visual reference usage highlights tokens that directly look back to the image but does not by itself ensure response correctness. LookBack (Ours) combines these two signals at the token level by calibrating token confidence with visual reference usage, and aggregates token scores at the response level with visual relevance weights.

To close this gap, we incorporate an attention-based measure of *visual lookback score* that captures how strongly each generation step attends back to vision tokens. This score serves as a lightweight proxy for *visual reference usage*. Specifically, we propose LookBack, a training-free LVLM response scoring method that calibrates output-space confidence using visual lookback scores. At the token level, LookBack combines each token's likelihood with its visual lookback score to form a *lookback calibrated token score*. At the response level, it aggregates these token scores under a visual relevance distribution, giving greater weight to tokens with stronger visual reference usage. This distinction is crucial because LVLM responses contain both visually diagnostic content words and generic fluency words: lookback calibrated token scoring captures whether each word is visually supported, while response-level aggregation determines how much visually relevant words contribute to the LookBack.

As a result, LookBack favors responses whose high-confidence tokens are strongly tied to the visual input, without requiring external verifiers, additional training, or extra inference passes. We evaluate LookBack in Best-of-N selection across four benchmarks that require visual understanding and three representative LVLMs: LLaVA-1.5-7B (Liu et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-8B (Zhu et al., 2025). Across these settings, LookBack achieves the highest model-wise average performance for all three LVLMs and generally improves over both linguistic- and vision-side baselines.

2 Related Works

Output-space response scoring in LLMs. Stochastic decoding in LLMs yields multiple candidate responses that may follow different inference paths (Kadavath et al., 2022; Qiu and Miikkulainen, 2024), making *Best-of-N* selection a natural strategy to improve the performance and reliability (Gui et al., 2024; Snell et al., 2025). Existing methods typically score candidates from the generated output alone, using answer consistency (Wang et al., 2023b), external reward models or verifiers (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024), or training-free model-intrinsic signals (Wang and Zhou, 2024; Kang et al., 2025; Lee et al., 2025; Gwak et al., 2025). However, the output-space score is less reliable in LVLMs, since fluent and confident responses can still contain visual claims unsupported by the image. LookBack retains the lightweight, model-intrinsic nature of such scores while incorporating visual reference usage.

Response scoring in LVLMs. Unlike LLM response scoring, LVLM response scoring must evaluate both linguistic plausibility and visual faithfulness simultaneously. Prior work has addressed this

challenge using external multimodal evaluators or reward models, such as VLM-as-a-judge models, multimodal critics, and process reward models (Lee et al., 2024; Xiong et al., 2025; Wang et al., 2025; Chen et al., 2025). Other approaches estimate image–response alignment through cross-modal matching or calibrated visual constraints (Hessel et al., 2021; Zhou et al., 2024), or analyze LVLM reliability through uncertainty, hallucination detection, and visual-grounding signals (Leng et al., 2024; Park et al., 2026). While these works emphasize the need for visual evidence, they often introduce external models, additional training, or extra scoring passes. In contrast, LOOKBACK scores responses using token likelihood and visual lookbackm which can be freely obtained during generation, providing a training-free and efficient alternative for LVLM response selection.

3 Motivation

To identify a reliable scoring function for LVLM response, we begin with output-space confidence, as it has shown consistent effectiveness in LLM scoring across inference-time selection (Kang et al., 2025) and training-time reward (Zhao et al., 2026). Our investigation *asks whether this signal remains reliable when response quality depends not only on linguistic plausibility, but also on visual input.*

We instantiate this investigation in a Best-of-N (BoN) response selection setting, where multiple candidate responses are generated for the same input and then ranked by a scoring function. Our analysis covers both discriminative and generative LVLM responses on MS-COCO (Lin et al., 2014) dataset: VQAv2 (Goyal et al., 2017) for visual question answering and CHAIR (Rohrbach et al., 2018) for object hallucination evaluation. For each benchmark, we randomly sample 1,000 instances and generate $N = 25$ candidate responses with LLaVA-1.5-7B (Liu et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025). As a representative output-space confidence scorer, we adopt Self-Certainty (SC) (Kang et al., 2025), a state-of-the-art BoN scoring method based on the token-averaged KL divergence between the model's predicted distribution and a uniform distribution.

3.1 Is Output-Space Confidence Sensitive to Visual Input?

We first examine whether SC is affected by the visual input by comparing its score distributions under image-present and image-absent conditions. For each generated response, we compute SC under two conditions: with the input image (SC w/ image) and without the input image (SC w/o image). Figure 2 shows that across all model–benchmark combinations, the SC distributions under the two conditions are nearly identical in shape and mean, suggesting that SC scores are largely unaffected by the removal of the image at the population level.

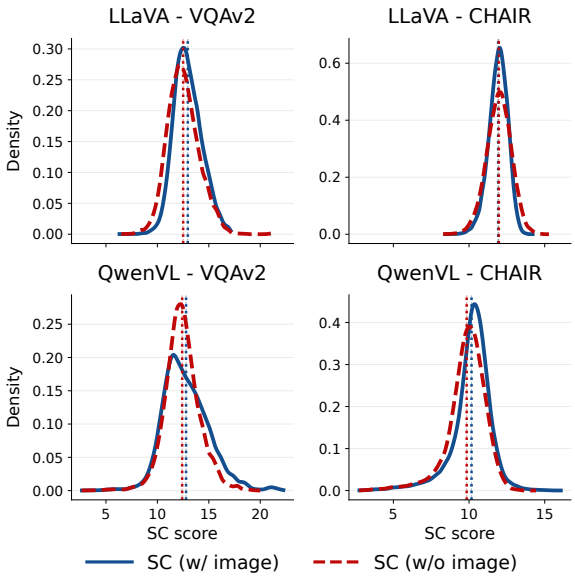

Figure 2: **SC score distributions remain similar with and without the image.** SC score distributions under image-present and image-removed conditions highly overlap, with closely aligned means across setups.

To further assess this weak image sensitivity at the selection level, we measure the *top-1 agreement ratio*: the fraction of instances for which SC w/ image and SC w/o image select the same top-ranked candidate from $N = 25$, directly reflecting whether the input image influences Best-of-N selection under SC. The solid bars in Figure 3 report this agreement. If the image were a critical determinant of response quality, we would expect the two scorers to select different top-1 candidates. However, we observe agreement ratios of 0.36–0.64, far above the random baseline of $1/N = 0.04$, indicating that removing the image barely changes which response SC selects.

These results reveal a gap between *image-conditioning* and *image-sensitivity*: although LVLMs are conditioned on the image during generation, their output-space confidence can remain largely unchanged even without the image. Thus, output-space confidence alone cannot reliably indicate whether a response is grounded in the image.

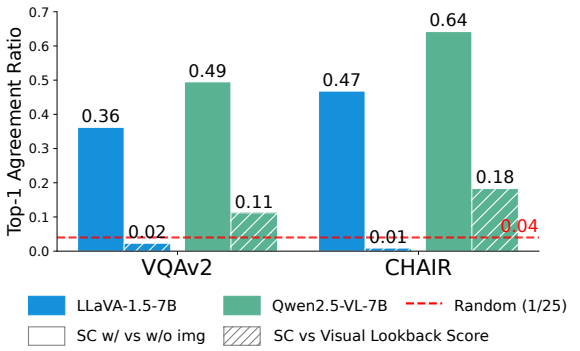

Figure 3: **Top-1 agreement with image-conditioned SC: image-removal consistency and visual-lookback complementarity.** We report the fraction of instances in which two scorers select the same top-ranked response from $N = 25$ candidates.

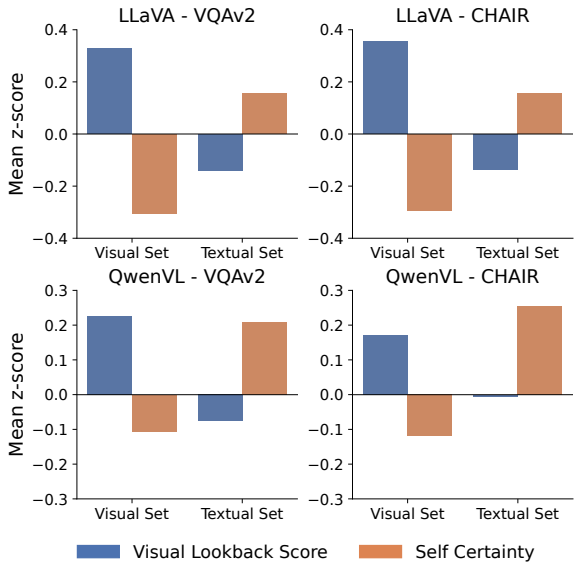

Figure 4: **Token-level POS analysis of visual lookback score and Self-Certainty.** We group generated words into a visual set and a textual set, and report the mean z-score of each signal within each group. Visual lookback scores are higher for the visual set, whereas Self-Certainty is higher for the textual set.

3.2 Visual Lookback Complements Token-Level Confidence

The weak image-sensitivity of output-space confidence raises a natural question: *what model-internal signal can reflect whether each generation step consults the visual input?*

Unlike text-only LLM, LVLMs receive the image as explicit vision tokens in the input context. Although visual information may be propagated implicitly through image-conditioned text representations and previous output states, the original vision tokens remain available in the causal context throughout decoding. Namely, each response-token prediction can therefore directly *look back* to the visual input. We quantify this behavior by the fraction of attention directed from each output token to the vision tokens. We call this token-level quantity the *visual lookback score* A_t, which serves as a lightweight proxy for *visual reference usage*; its formal definition is given in Section 4.1.

To examine whether this signal indeed captures a token-level visual reference, we analyze how it varies across different types of generated words. If visual lookback score captures how much a generation step refers back to visual evidence, it should be higher for tokens that express visually referential content than for tokens that primarily serve grammatical or discourse functions.

To this end, we conduct a token-level Part-of-Speech (POS) analysis on the same generated responses from Sec. 3.1. Generated words are tagged with spaCy and grouped into two sets following prior works (Chen et al., 2024; Dong et al., 2025; Min et al., 2025): a *visual set*, containing words more likely to express visually referential content—nouns, proper nouns, adjectives, and numerals—and a *textual set*, containing words that primarily serve grammatical or discourse roles, such as auxiliaries, determiners, pronouns, and conjunctions. Over the same word groups, we also analyze SC to test whether output-space confidence favors visually referential words despite its weak image sensitivity at the response-selection level. For words split into multiple subtokens, we use the mean subtoken score as the word-level score, and independently z-score normalize each score dimension over all analyzed output tokens before word-level aggregation.

As shown in Figure 4, across all model-benchmark combinations, words in the visual set exhibit above-average visual lookback score but below-average SC, whereas words in the textual set show the opposite pattern. This opposing pattern suggests that visual lookback score captures a token-level tendency complementary to SC: SC is higher for linguistically predictable words, whereas visual lookback score is higher for words that are more visually referential. A fine-grained per-POS breakdown in Figure 8 further shows that this trend is not driven by a single POS category, but broadly appears across content-oriented categories.

This token-level complementarity also appears at the response level. The hatched bars in Figure 3 show that SC and visual lookback score rarely select the same top-ranked response, with the lowest

agreement ratio reaching 0.01, below the random baseline of 0.04. This indicates that the two signals induce substantially different response rankings.

Figure 1 illustrates this behavior qualitatively. SC assigns high scores to fluent, linguistically predictable tokens including function words and assertion markers, whereas visual lookback score highlights object names, attributes, actions, and spatial relations that require direct consultation of the image. This divergence is particularly pronounced when the model generates confident but visually unsupported assertions: confidence rewards linguistic fluency regardless of visual grounding, whereas visual lookback score reflects whether the prediction step directly attended to the visual input.

Overall, these observations suggest that output-space confidence and visual lookback score capture complementary aspects of LVLM generation. Output-space confidence is useful for identifying linguistically likely responses, but it may remain high even when a response is driven primarily by language priors rather than visual evidence. Visual lookback score captures whether generation steps attend back to the explicit vision tokens, but high visual attention alone does not guarantee correctness. Therefore, a visually grounded response score should combine token-level confidence with token-level visual lookback score, while placing greater emphasis on response positions where visual evidence is most relevant.

4 Method

Based on the observations in Section 3, our goal is to define a response score that converts output-space confidence into vision-aware confidence using only model-internal signals, without extra inference passes. Specifically, we design the score according to following objectives:

O1. **Preserve language confidence.** The score should retain token-level model confidence, so that implausible or low-probability responses are not favored.

O2. **Incorporate visual reference usage.** Tokens that are likely under the output distribution should be further assessed by how strongly they interact with the visual reference.

O3. **Aggregate according to visual relevance.** Since not all output tokens are equally informative for visual grounding, visually relevant token positions should contribute more.

Preliminary. Let p_θ denote an LVLM parameterized by θ. Given a text query $\mathbf{x}$ and vision tokens $\mathbf{v}$ encoded from an input image $\mathcal{I}$, the model generates a response $\mathbf{y} = (y_1, \ldots, y_T)$ autoregressively:

$$p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{v}) = \prod_{t=1}^{T} p_\theta(y_t \mid \mathbf{x}, \mathbf{v}, y_{<t}), \quad (1)$$

and we denote the per-token probability as $p_t = p_\theta(y_t \mid \mathbf{x}, \mathbf{v}, y_{<t})$. Under stochastic decoding, we sample N candidate responses $\{\mathbf{y}^{(i)}\}_{i=1}^{N}$ for the same input $(\mathbf{x}, \mathbf{v})$, and select the best candidate under a training-free score S:

$$\mathbf{y}^* = \arg\max_i \; S(\mathbf{y}^{(i)} \mid \mathbf{x}, \mathbf{v}). \quad (2)$$

4.1 Visual Lookback Score

For each generated token y_t, we estimate its *visual lookback score*: how strongly the prediction step attends back to the original vision tokens. Let C_t denote the full causal context at the prediction step for y_t, comprising text query tokens, vision tokens, and previous output tokens. Let $\mathcal{P}_v \subset C_t$ denote the positions corresponding to vision tokens $\mathbf{v}$. Let $a_{t,k}^{(\ell,h)}$ be the attention weight from the query position predicting y_t to context position k at layer ℓ and head h. We define the token-level *visual lookback score* A_t as:

$$A_t = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \frac{\sum_{k \in \mathcal{P}_v} a_{t,k}^{(\ell,h)}}{\sum_{k \in C_t} a_{t,k}^{(\ell,h)}}, \quad (3)$$

where L and H denote the number of layers and attention heads, respectively. Since C_t is the full causal context, the denominator equals one under softmax normalization; we retain the ratio form to make explicit that A_t measures the fraction of attention directed to vision tokens. As A_t is obtained from the attention weights produced during the generation forward pass, it does not require additional inference or external model. A larger A_t indicates greater *visual reference usage*, meaning that the prediction step allocates a larger fraction of its attention to the explicit vision tokens.

4.2 Lookback-Calibrated Token Score

Since p_t and A_t capture complementary aspects of token generation, we define a token-level score that preserves output-space confidence while accounting for visual lookback score. We define the lookback-calibrated token score u_t as

$$u_t = \log(p_t) + \alpha \log(A_t), \quad (4)$$

equivalently written as $u_t = \log(p_t \cdot A_t^{\alpha})$. The hyperparameter α controls the strength of this lookback calibration. When $\alpha = 0$, u_t reduces to pure output-space confidence $\log p_t$ (O1). When $\alpha > 0$, u_t is larger for tokens that are both highly probable and strongly attend to vision tokens, and smaller for tokens whose confidence is weakly supported by the visual reference (O1, O2).

4.3 Weighting Tokens with Visual Relevance

To obtain a response-level score, we aggregate token-level scores $\{u_t\}_{t=1}^{T}$ under a weight distribution q over output token positions:

$$S(\mathbf{y} \mid \mathbf{x}, \mathbf{v}) := \mathbb{E}_{t\sim q}\big[u_t\big] = \sum_{t=1}^{T} q(t)\, u_t. \quad (5)$$

Uniform averaging corresponds to $q(t) = \frac{1}{T}$, but it assumes every token is equally informative, which is often not the case. In visually grounded responses, tokens such as object names, attributes, counts, and relations are generally more diagnostic than function words or generic phrases. The aggregation distribution should therefore concentrate on positions that interact more strongly with the visual reference, motivating q derived from A_t.

We define q_λ as a visual relevance distribution over output tokens, derived as the solution to an entropy-regularized relevance maximization:

$$q_\lambda = \arg\max_{q\in\Delta_T} \Big[\lambda\, \mathbb{E}_{t\sim q}\big[\log(A_t)\big] + H(q)\Big], \quad (6)$$

where Δ_T is the probability simplex over T output positions, and $H(q) = -\sum_t q(t)\log\big(q(t)\big)$ is the entropy of q. The first term encourages q to concentrate on positions with high visual lookback score, while the entropy term keeps the distribution smooth and prevents hard selection. The hyperparameter λ controls how sharply q concentrates on high-A_t positions.

Proposition 1. *The solution to the entropy-regularized relevance maximization is*

$$q_\lambda(t) = \frac{A_t^{\lambda}}{\sum_{j=1}^{T} A_j^{\lambda}}. \quad (7)$$

Proof. See Appendix A.2. □

When $\lambda = 0$, q_λ is uniform; as λ increases, q_λ assigns more mass to positions with larger visual lookback score (O3).

Final score. Instantiating S with q_λ gives the final response score LOOKBACK:

$$\begin{aligned} S(\mathbf{y} \mid \mathbf{x}, \mathbf{v}) &= \mathbb{E}_{t\sim q_\lambda}\big[\log\big(p_t\big) + \alpha \log\big(A_t\big)\big] \\ &= \sum_{t=1}^{T} \frac{A_t^{\lambda}\big(\log\big(p_t\big) + \alpha\log\big(A_t\big)\big)}{\sum_{j=1}^{T} A_j^{\lambda}}. \end{aligned} \quad (8)$$

LOOKBACK measures the expected lookback calibrated token score under a visual relevance distribution. Intuitively, it evaluates model confidence in a response while emphasizing tokens that interact more strongly with the explicit visual reference.

5 Experiments

5.1 Setups

Benchmarks and metrics. We evaluate on four image-grounded benchmarks suited to Best-of-N response selection: VQAv2 (Goyal et al., 2017), CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023a), and HallusionBench (Guan et al., 2024). We chose this suite so that it (1) requires visual evidence beyond language plausibility, (2) provides sufficient candidate diversity and selection headroom, and (3) covers complementary response formats, including short-answer QA, open-ended generation, and discriminative visual-grounding evaluation. For a consistent main comparison, we report a single higher-is-better primary metric per benchmark: accuracy for VQAv2, F1 for CHAIR and AMBER, and GPT-evaluated correctness for HallusionBench. See Appendix B.1 for details.

Models and baselines. For the experiments, we consider three representative LVLMs: LLaVA-1.5-7B (Liu et al., 2023), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-8B (Zhu et al., 2025). Also, to evaluate the effectiveness of LOOKBACK, we consider two categories of baselines: language side and vision side. (1) *Language side*: Self-Certainty (SC) (Kang et al., 2025) measures model's token distributional confidence leveraging KL divergence with uniform distribution. Universal Self-Consistency (USC) (Chen et al., 2023) select most consistent response by prompting all candidates responses to model. (2) *Vision side*: CLIP-Score (Hessel et al., 2021) leverages a pretrained vision-language encoder to measure the cosine similarity between the text embedding of a generated response and the corresponding image embedding. VAUQ (Park et al., 2026) estimates model uncertainty by masking a fixed proportion of visual attention weights and computing the entropy of the

Table 1: **Main results.** Best-of-N selection performance for $N = 5$ and $N = 25$ across four benchmarks and three LVLMs. Average denotes the average score over all benchmark–N combinations within each model. Higher scores (↑) indicate better performance. The best results in each setup are **bolded**, and the second-best are underlined.

Model	Method	VQAv2		CHAIR		HallusionBench		AMBER		Average
		$N = 5$	$N = 25$	$N = 5$	$N = 25$	$N = 5$	$N = 25$	$N = 5$	$N = 25$	
LLaVA-1.5-7B	Random	62.57	60.40	70.20	71.52	44.84	43.82	63.90	64.08	60.16
	SC	65.93	65.70	72.76	73.53	47.21	46.19	64.10	63.45	62.36
	USC	63.78	61.62	71.07	70.87	43.09	43.20	64.61	63.96	60.28
	CLIPScore	64.07	62.27	71.89	71.93	43.15	43.65	**66.08**	65.25	61.03
	VAUQ	66.03	66.53	72.48	73.15	45.01	45.35	64.28	64.22	62.13
	Ours	**66.63**	**67.60**	**74.03**	**74.43**	**47.38**	**48.22**	65.33	**65.74**	**63.67**
Qwen2.5-VL-7B	Random	63.40	62.23	74.18	73.22	57.19	57.36	74.19	73.84	66.95
	SC	66.33	65.47	74.41	71.97	57.36	58.04	74.43	74.30	67.78
	USC	66.88	66.73	74.63	73.96	56.80	56.91	74.33	74.64	68.11
	CLIPScore	64.07	62.17	74.47	74.80	57.36	58.21	74.17	74.35	67.45
	VAUQ	66.67	66.83	75.15	73.24	57.02	57.19	74.77	74.50	68.17
	Ours	**68.23**	**67.47**	**75.42**	**75.28**	**60.07**	**61.93**	**74.98**	**75.92**	**69.91**
InternVL3-8B	Random	62.17	61.33	77.88	77.78	56.35	57.19	79.82	79.45	69.00
	SC	66.70	66.67	78.35	78.24	56.35	56.85	81.66	81.67	70.81
	USC	**70.33**	69.39	78.34	**78.28**	58.15	58.38	80.04	79.72	71.58
	CLIPScore	63.47	63.47	78.32	78.27	**59.39**	**59.39**	79.68	79.98	70.25
	VAUQ	68.23	69.77	78.49	**78.28**	57.19	57.19	81.47	81.65	71.53
	Ours	**70.33**	**72.57**	**79.37**	78.22	57.02	56.85	**81.77**	**82.07**	**72.27**

resulting output distribution. To evaluate the gain of each method, we additionally consider the simplest baseline, random selection (Random).

Implementation details. For all models, to ensure sufficient diversity among candidates, we generate $N = 25$ candidate responses per input using nucleus sampling (Holtzman et al., 2019) with temperature=1.2 and top-p=0.9. LOOKBACK has two hyperparameters, α and λ, which control the strength of token-level visual lookback calibration and response-level visual relevance weighting, respectively. We set these hyperparameters on a per-model basis. Specifically, we use $(\alpha, \lambda) = (7.0, 1.5)$ for LLaVA-1.5-7B, $(0.25, 1.25)$ for Qwen2.5-VL-7B, and $(0.25, 1.25)$ for InternVL3-8B. See Appendix B.2 for details.

5.2 Main Results

Table 1 reports Best-of-N selection performance for $N = 5$ and $N = 25$ across four benchmarks and three LVLMs. Across all setups, LOOKBACK improves the average score from 65.37% to 68.62%, achieving a 4.97% relative gain over random selection.

While SC achieves competitive performance as an output-space baseline, LOOKBACK improves over SC in most settings and achieves higher average scores. This suggests that visually grounded response selection benefits from measuring not only how confident a response is, but whether that confidence is supported by visual reference usage.

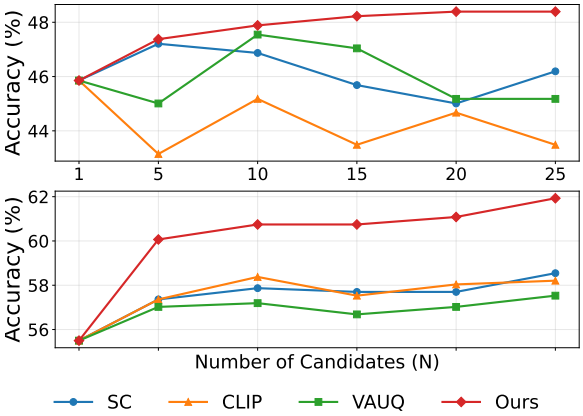

Figure 5: **Best-of-N scaling.** We report Best-of-N selection performance as the candidate pool size increases from N=1 to N=25 on HallusionBench with LLaVA-1.5-7B (top) and Qwen2.5-VL-7B (bottom).

USC is also competitive in InternVL3-8B, but its gains are less stable across models and candidate budgets. Since USC relies on the LVLM itself to select candidate responses, its effectiveness may depend on the model's own selection ability, not just on the quality of the sampled candidates.

Compared with VAUQ which estimates reliability from a perturbed visual condition, LOOKBACK uses visual evidence in a more candidate-specific way; LOOKBACK directly measures whether high-confidence tokens look back to the visual input.. This makes LOOKBACK well suited to Best-of-N selection, where the score must distinguish candidates whose confident claims are visually supported from those that are merely plausible.

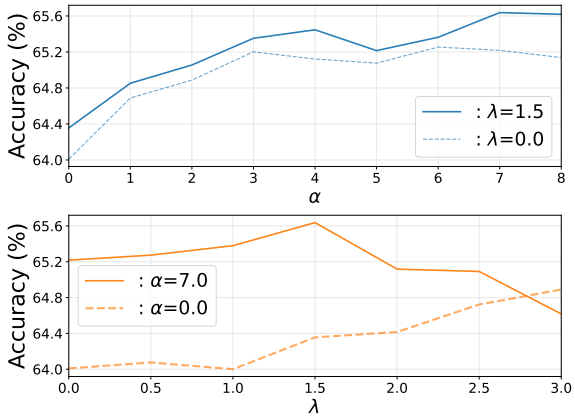

Figure 6: **Ablation study.** We ablate token-level calibration strength α and the response-level visual weighting strength λ on AMBER with LLaVA-1.5-7B. Top: performance when varying α with fixed λ. Bottom: performance when varying λ with fixed α. Dashed lines show the corresponding variants without the other component.

5.3 Scaling Results

We further investigate how each scoring method benefits from varied candidate budget. Figure 5 shows Best-of-N accuracy as N increases from 1 to 25 on HallusionBench for LLaVA-1.5-7B and Qwen2.5-VL-7B. LOOKBACK maintains a consistent advantage over all baselines across the full range of N. While SC improves gradually with N, LOOKBACK maintains a clear margin as the candidate pool grows, especially on Qwen2.5-VL-7B. CLIPScore and VAUQ do not consistently benefit from larger candidate pools. This suggests that scaling Best-of-N requires a reliable grounding-aware score, rather than simply increasing the number of candidates. Overall, LOOKBACK shows stable performance gains as N increases.

5.4 Additional Analyses

Ablation study. Figure 6 ablates the two hyperparameters of LOOKBACK on AMBER with LLaVA-1.5-7B: α for token-level visual lookback calibration and λ for response-level visual relevance weighting. In the top panel, increasing α yields modest gains when $\lambda = 0$ because aggregation remains uniform. With $\lambda = 1.5$, however, the same increase leads to stronger performance, showing that token-level calibration is more effective when visually relevant tokens are weighted more heavily. In the bottom panel, increasing λ improves performance when $\alpha = 0$, but with $\alpha = 7$, performance peaks at a moderate λ and declines as the aggregation becomes too sharp. These results show that token-level lookback calibration and visual relevance-based aggregation are complementary.

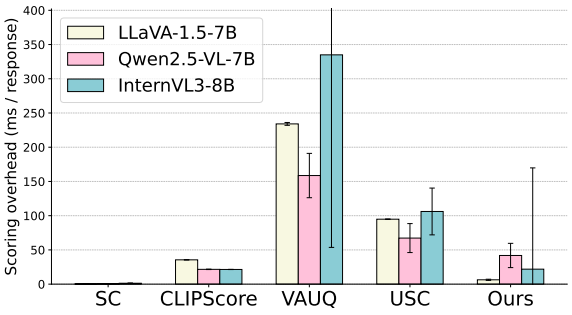

Figure 7: **Scoring overhead.** We compare the average scoring time per response across three LVLMs on CHAIR. Error bars indicate standard deviation.

Scoring overhead. To evaluate the computational efficiency of LOOKBACK, we measure post-generation scoring overhead in milliseconds per response on CHAIR across three LVLMs. As shown in Figure 7, LOOKBACK adds only modest overhead and remains substantially cheaper than VAUQ and USC, which require perturbation-based uncertainty estimation or additional candidate-comparison passes. This shows that LOOKBACK provides a favorable efficiency–grounding trade-off: unlike SC, it incorporates an explicit visual-grounding signal, yet unlike heavier visual scorers, it obtains this signal directly from the LVLM's internal likelihood and attention statistics without external models or expensive post-hoc verification. See Appendix B.3 for measurement setup and hardware configuration.

6 Conclusion

We presented LOOKBACK, a training-free response scoring method for LVLMs that calibrates output-space confidence with visual lookback. Our motivation analysis shows that confidence-based scoring can remain largely insensitive to the image, even when computed under image-conditioned generation, and therefore tends to capture textual plausibility rather than visual grounding. To address this gap, LOOKBACK combines token likelihood with a token-level visual lookback score and aggregates these calibrated token scores according to a visual relevance distribution. Across multiple LVLMs and visual understanding benchmarks, LOOKBACK consistently improves Best-of-N selection over both linguistic and vision side baselines, while requiring no auxiliary verifier, training, or extra inference passes.

Limitations and Future Work

LOOKBACK requires access to internal attention weights, which limits its applicability to black-box LVLMs. In addition, visual lookback is only a proxy for visual reference usage, not a guarantee of factual correctness: a model can strongly attend to image tokens while still generating an incorrect or unsupported response. The reliability of this proxy may also depend on model-specific attention behavior, as attention distributions and their calibration can vary across LVLM architectures.

In addition, our current evaluation focuses on image-grounded Best-of-N selection with relatively concise responses. Whether LOOKBACK generalizes to long-form multimodal reasoning, multi-image or video inputs, and more complex grounding scenarios remains an important direction for future work. Interestingly, our formulation suggests a broader perspective on source-aware scoring for generative models, beyond LVLM response selection. The same principle may apply whenever a response should be grounded in a designated part of the input context, such as retrieved documents in RAG, tool outputs, or instruction tokens for prompt following. In such settings, a reliable scorer should assess not only whether a response is likely under the model's output distribution, but also whether high-confidence generation steps refer back to the intended source. We believe this perspective opens a path toward lightweight, model-internal scoring methods for source-grounded generation beyond the visual domain.

Broader Impact and Ethical Implications

LOOKBACK is intended to improve the reliability of LVLM response selection by favoring candidates whose high-confidence tokens are more strongly tied to the visual input. This direction may be beneficial in applications where visually unsupported responses can mislead users, such as image-based assistance, educational tools, accessibility interfaces, and content analysis systems. Because LOOKBACK requires neither additional training nor an auxiliary evaluator, it may also provide a lightweight mechanism for improving response selection in resource-constrained settings.

At the same time, LOOKBACK should not be interpreted as a guarantee of correctness or safety. Visual lookback is only a proxy for reference usage: a response can strongly attend to image tokens while still being incorrect, biased, or harmful. These risks are especially consequential in high-stakes applications, including medical, legal, surveillance, and accessibility-critical settings. We therefore recommend using LOOKBACK only as a response-selection aid, together with task-specific validation and human oversight.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xinquan Chen, Bangwei Liu, Xuhong Wang, Yingchun Wang, and Chaochao Lu. 2025. Vrprm: Process reward modeling via visual reasoning. *arXiv preprint arXiv:2508.03556*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *International Conference on Computer Vision (ICCV)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin Gui, Cristina Gârbacea, and Victor Veitch. 2024. Bonbon alignment for large language models and the sweetness of best-of-n sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Minju Gwak, Guijin Son, and Jaehyung Kim. 2025. Revisiting the uniform information density hypothesis in llm reasoning traces. *arXiv preprint arXiv:2510.06953*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dongseok Lee, Jimyung Hong, Dongyoung Kim, and Jaehyung Kim. 2025. Training-free llm verification via recycling few-shot examples. *arXiv preprint arXiv:2506.17251*.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Seongheon Park, Changdae Oh, Hyeong Kyu Choi, Xuefeng Du, and Sharon Li. 2026. Vauq: Vision-aware uncertainty quantification for lvlm self-evaluation. *arXiv preprint arXiv:2602.21054*.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2026. Learning to reason without external rewards. In *International Conference on Learning Representations (ICLR)*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Additional Theoretical Analysis

A.1 Log-Linear Product-of-Experts Interpretation

The grounded token score can be written as

$$u_t = \log(p_t) + \alpha \log(A_t) = \log\left(p_t \cdot A_t^{\alpha}\right). \quad (9)$$

This resembles a product-of-experts-style score (Hinton, 2002), where output-space confidence p_t and visual lookback score A_t serve as complementary factors that are combined multiplicatively into an unnormalized token-level score. The hyperparameter α controls the relative strength of the visual factor. Thus, tokens with low visual lookback score receive a lower combined score even when their output probability is high.

A.2 Proof of Proposition 1

We prove that q_λ defined by

$$q_\lambda = \arg\max_{q \in \Delta_T} \Big[\lambda\, \mathbb{E}_{\tau \sim q}\big[\log(A_\tau)\big] + H(q)\Big] \quad (10)$$

has the closed-form solution

$$q_\lambda(t) = \frac{A_t^{\lambda}}{\sum_{j=1}^{T} A_j^{\lambda}}. \quad (11)$$

Here,

$$\Delta_T = \left\{ q \in \mathbb{R}_{\geq 0}^{T} : \sum_{t=1}^{T} q(t) = 1 \right\} \quad (12)$$

is the set of distributions over output positions.

Proof. Expanding the objective using the definition of $H(q)$, the problem becomes

$$\max_{q \in \Delta_T} \sum_{t=1}^{T} q(t)\Big[\lambda \log(A_t) - \log(q(t))\Big]. \quad (13)$$

Assuming $A_t > 0$ for all t, this is a strictly concave maximization over the probability simplex Δ_T, so a unique solution exists. Introducing a Lagrange multiplier η for the constraint $\sum_t q(t) = 1$, the Lagrangian is

$$\mathcal{L} = \sum_{t=1}^{T} q(t)\lambda \log(A_t) - \sum_{t=1}^{T} q(t)\log(q(t)) + \eta\left(\sum_{t=1}^{T} q(t) - 1\right). \quad (14)$$

Setting $\partial\mathcal{L}/\partial q(t) = 0$ gives

$$\lambda \log(A_t) - \log(q(t)) - 1 + \eta = 0, \quad (15)$$

which implies

$$\log(q(t)) = \lambda \log(A_t) + (\eta - 1). \quad (16)$$

Therefore,

$$q(t) \propto A_t^{\lambda}. \quad (17)$$

Applying the normalization constraint $\sum_t q(t) = 1$ yields

$$q_\lambda(t) = \frac{A_t^{\lambda}}{\sum_{j=1}^{T} A_j^{\lambda}}. \quad (18)$$

□

Interpretation. The exponent λ controls the sharpness of q_λ. When $\lambda = 0$, q_λ is the uniform distribution over output positions, recovering uniform averaging. As $\lambda \to \infty$, q_λ concentrates on the position with the largest A_t, approaching hard selection of the most visually attended token. For finite $\lambda > 0$, q_λ acts as a soft selector that prefers visually attended positions while maintaining nonzero weight on all positions.

A.3 Geometric Mean Interpretation

We show that the proposed score $S(\mathbf{y} \mid \mathbf{x}, \mathbf{v})$ can be interpreted as the log of a visual-relevance-weighted geometric mean. This provides a justification for using a normalized weighted average rather than an unnormalized sum over output tokens.

Corollary A.0.1. *Let q_λ be as defined in Proposition 1. Then*

$$\exp\Big(S(\mathbf{y} \mid \mathbf{x}, \mathbf{v})\Big) = \prod_{t=1}^{T} \left(p_t \cdot A_t^{\alpha}\right)^{q_\lambda(t)}. \quad (19)$$

That is, $\exp(S)$ is the visual-relevance-weighted geometric mean of the token-level product scores $p_t \cdot A_t^{\alpha}$.

Proof. By definition,

$$S(\mathbf{y} \mid \mathbf{x}, \mathbf{v}) = \sum_{t=1}^{T} q_\lambda(t)\Big[\log(p_t) + \alpha \log(A_t)\Big]. \quad (20)$$

Since

$$\log(p_t) + \alpha \log(A_t) = \log\left(p_t \cdot A_t^{\alpha}\right), \quad (21)$$

Table 2: **Oracle performance across benchmarks and LVLMs.** We report Oracle Best-of-N selection performance for $N = 5$ and $N = 25$. Average denotes the average score across the four benchmarks for each candidate budget. Higher scores (↑) indicate better performance.

Benchmark	LLaVA-1.5-7B		Qwen2.5-VL-7B		InternVL3-8B	
	$N = 5$	$N = 25$	$N = 5$	$N = 25$	$N = 5$	$N = 25$
VQAv2	84.07	92.50	82.83	91.17	88.60	94.80
CHAIR	84.17	90.55	86.11	91.47	86.73	91.72
HallusionBench	80.71	94.42	83.42	92.89	80.54	91.71
AMBER	76.96	83.83	84.97	89.91	88.81	92.76
Average	81.48	90.33	84.33	91.36	86.17	92.75

we have

$$S(\mathbf{y} \mid \mathbf{x}, \mathbf{v}) = \sum_{t=1}^{T} q_\lambda(t) \log \left(p_t \cdot A_t^\alpha\right). \tag{22}$$

Exponentiating both sides gives

$$\exp\Big(S(\mathbf{y} \mid \mathbf{x}, \mathbf{v})\Big) = \prod_{t=1}^{T} \Big(p_t \cdot A_t^\alpha\Big)^{q_\lambda(t)}. \tag{23}$$

□

Implication. A standard log-likelihood sum, $\sum_t \log(p_t)$, scales with response length T, making scores difficult to compare across candidates of different lengths. In contrast, the proposed score is equivalent to maximizing the visual-relevance-weighted geometric mean

$$\prod_{t=1}^{T} \left(p_t \cdot A_t^\alpha\right)^{q_\lambda(t)}, \tag{24}$$

where the exponents satisfy $\sum_t q_\lambda(t) = 1$. Thus, the score compares candidates using a normalized weighted average of token-level product scores rather than an unnormalized sum. Unlike uniform length normalization with $q(t) = 1/T$, the weighting q_λ adapts the normalization to visual relevance.

B Additional Experimental Details

B.1 Benchmark Details

We select benchmarks that are informative for Best-of-N response selection. In particular, the candidate pool should contain meaningful variation in response quality, so that a response scorer has non-trivial room to improve over random selection. We therefore consider both candidate diversity and oracle–average gaps when constructing the evaluation suite. Table 2 reports oracle Best-of-N performance, where the oracle selects the highest-scoring response among the sampled candidates for each input. The large gap between oracle and average performance indicates that the sampled candidate pools contain substantial selection headroom.

For the main comparison, we report a single higher-is-better metric for each benchmark. This avoids mixing heterogeneous native metrics in the main table while still allowing a consistent comparison of response scorers. We provide benchmark-specific details below.

VQAv2. VQAv2 (Goyal et al., 2017) is a short-answer visual question answering benchmark where each question is paired with multiple human annotations. We use VQAv2 to evaluate whether a scorer can select answers that are grounded in the visual content of the image. For each generated response, we parse the predicted answer and evaluate it against the ground-truth annotations. The primary metric is accuracy:

$$\mathrm{Acc}(ans) = \min\left\{\frac{\#\text{ human annotators that provided } ans}{3}, 1\right\}.$$

Dataset-level accuracy is computed by averaging the per-question scores.

CHAIR. CHAIR (Rohrbach et al., 2018) evaluates object hallucination in open-ended image descriptions. Following the CHAIR setup, generated captions are parsed into object mentions and matched to MS-COCO object categories using the benchmark's synonym mapping. For each image, let G_i denote the set of ground-truth objects and P_i denote the set of predicted object mentions extracted from the generated response. A predicted object is counted as hallucinated if it does not appear in G_i. Although CHAIR conventionally reports CHAIR_S and CHAIR_I, we report F1 as the primary metric to obtain a single higher-is-better score for Best-of-N selection:

$$\mathrm{Prec}_i = \frac{|P_i \cap G_i|}{|P_i|}, \qquad \mathrm{Rec}_i = \frac{|P_i \cap G_i|}{|G_i|},$$

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{Prec}_i \cdot \mathrm{Rec}_i}{\mathrm{Prec}_i + \mathrm{Rec}_i}, \qquad \mathrm{F1} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathrm{F1}_i.$$

When the denominator is zero, the corresponding precision, recall, or F1 term is set to zero.

AMBER. AMBER (Wang et al., 2023a) is a multidimensional hallucination benchmark for LVLMs. In the generative setting, AMBER evaluates whether generated responses contain visual

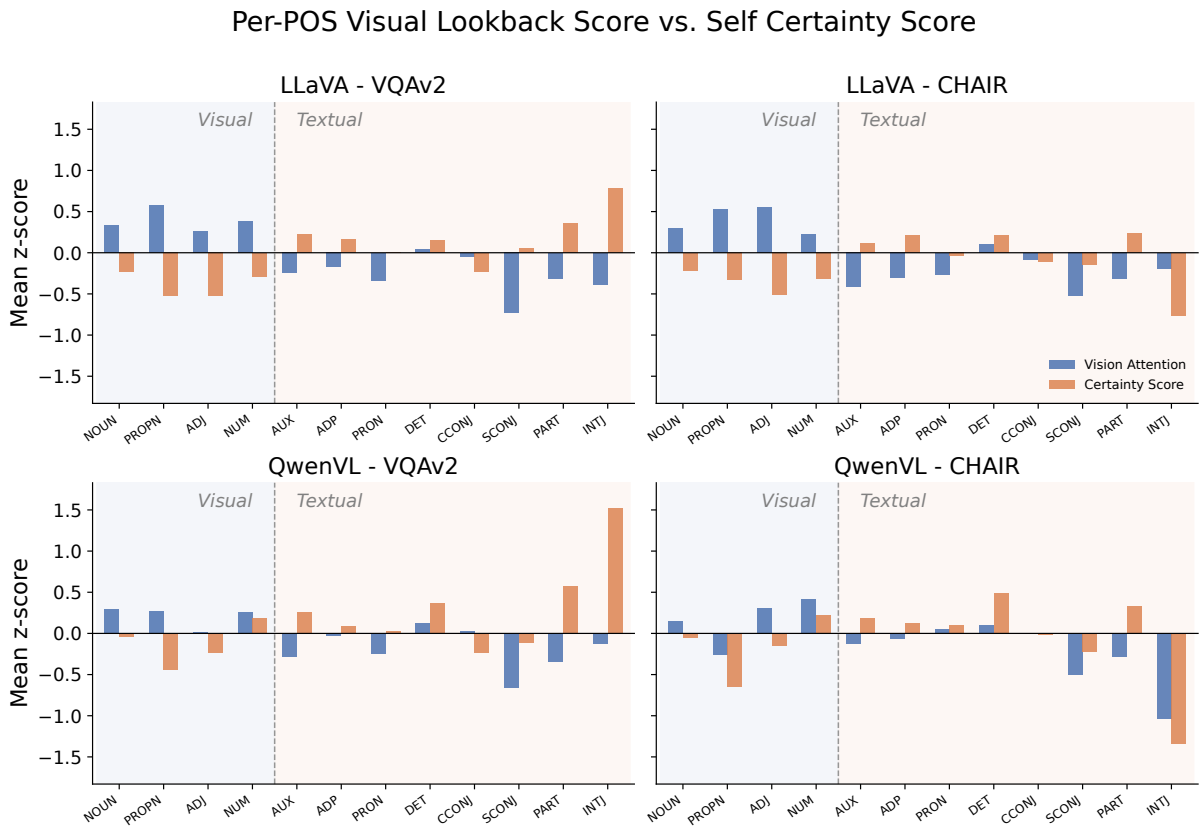

Figure 8: **Visual lookback and SC emphasize different POS groups.** We report mean z-scored Visual lookback and Self-Certainty (SC) scores for each POS category. Across models and benchmarks, visual POS categories tend to receive higher Visual lookback scores, whereas textual categories tend to receive higher SC scores.

instances that are supported by the image. Unlike CHAIR, which focuses on object hallucination, AMBER covers a broader set of visual error types, including object existence, attributes, and relations. For each sample, we compare generated visual instances against the ground-truth visual instance set. Instances absent from the ground-truth set are treated as hallucinated. For consistency with CHAIR and the main table, we report F1 as the primary higher-is-better metric:

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{Prec}_i \cdot \mathrm{Rec}_i}{\mathrm{Prec}_i + \mathrm{Rec}_i}, \qquad \mathrm{F1} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathrm{F1}_i.$$

This provides a unified measure that rewards both avoiding hallucinated visual claims and covering ground-truth visual content.

HallusionBench. HallusionBench (Guan et al., 2024) evaluates visually grounded reasoning under cases where language priors and visual evidence can conflict. We use it as a discriminative visual-grounding benchmark. For each candidate response, we use a GPT-based evaluator to judge whether the response is visually grounded and consistent with the reference answer. The evaluator returns a correctness label for each response, and we treat uncertain or invalid judgments as incorrect. The primary metric is GPT-evaluated correctness:

$$\text{Correctness} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbf{1}\{\text{Judge}(I_i, q_i, y_i) = \texttt{Correct}\}.$$

This single accuracy-style metric allows HallusionBench to be compared with the other benchmarks under the same Best-of-N selection framework.

B.2 Baseline Implementation Details

For USC, when the method failed to produce a valid selection, we fell back to the first sampled response. To reduce the order bias in USC methods, we report the average accuracy over three independent runs ($m = 3$).

For VAUQ, we follow the layer choices specified in the original paper: 10–25 for LLaVA-1.5-7B, 12–26 for Qwen2.5-VL-7B, and 10–25 for InternVL3-8B. We use fixed masking ratios of $K = 0.6$, 0.5, and 0.4 for the three models, respectively, and set $\alpha = 1$ throughout all experiments. We implement VAUQ following its attention-knockout-based visual masking procedure as mentioned in paper.

Table 3: **Hardware and software configuration for scoring overhead measurement.**

	LLaVA-1.5-7B	Qwen2.5-VL-7B / InternVL3-8B
GPU	NVIDIA RTX A6000	NVIDIA H200
CPU	AMD EPYC 9354	Intel Xeon Platinum 8568Y+
CUDA	11.7	12.4
cuDNN	8.5.0	9.1.0
Python	3.10.15	3.9.21
PyTorch	2.0.1+cu117	2.6.0+cu124
Transformers	4.31.0	4.57.6
Random seed	42	42

B.3 Scoring Overhead Measurement

We measure post-generation scoring overhead in milliseconds per response. Candidate generation is excluded from timing: all scorers are given the same generated candidate responses, and we measure only the additional wall-clock time required to assign scores after generation. Measurements are conducted on CHAIR for LLaVA-1.5-7B, Qwen2.5-VL-7B, and InternVL3-8B.

For SC and LOOKBACK, scoring uses model-internal quantities from the LVLM scoring pass, including token probabilities and attention weights. For CLIPScore, we include the external CLIP encoder forward pass. For VAUQ, we include visual-attention perturbation and uncertainty computation. For USC, which scores all candidates from the same input in one teacher-forcing forward pass, we divide the total candidate-set scoring time by the number of candidates. We report mean milliseconds per response with standard deviation across evaluated samples.

Due to resource availability, LLaVA-1.5-7B is measured on an NVIDIA RTX A6000, while Qwen2.5-VL-7B and InternVL3-8B are measured on an NVIDIA H200. Therefore, absolute latency values should be interpreted as scorer-overhead comparisons within each model, rather than as direct speed comparisons across LVLMs.

C Detailed Quantitative Results

C.1 PoS

In Section 3.2, we group generated words into a visual-reference-prone content set and a textual-function set to compare visual lookback score with output-space confidence. Here, we provide a fine-grained breakdown by individual POS tags.

We use the same generated responses and scoring setup as in Section 3.2. Each word is assigned a POS tag using spaCy. For words split into multiple subtokens, we average the corresponding subtoken scores to obtain a word-level score. Visual lookback score and Self-Certainty are independently standardized over all analyzed output words within each model–benchmark setup, and we report the mean z-score for each POS category.

Figure 8 shows that content-oriented POS categories, such as nouns, proper nouns, adjectives, and numerals, generally exhibit higher visual lookback score than textual-function categories. In contrast, textual-function categories often show relatively higher certainty, reflecting their linguistic predictability. This fine-grained analysis supports the coarse POS grouping used in the main text. At the same time, some rare or context-dependent categories exhibit noisier behavior, which motivates our use of grouped POS sets rather than drawing conclusions from individual POS tags.

D Additional Qualitative Results

In Figures 9 and 10, we additionally present the qualitative examples to compare SC, Visual look-back score, and our final score.

Q: How many cars are parked? | **GT:** 1

(1) Selected by SC — Incorrect

There are two cars park ed next to each other .

(2) Selected by Visual Lookback Score — Incorrect

There are a couple of cars park ed near the house or tr uck .

(3) Selected by Ours — Correct

There is only one car park ed in the image .

Figure 9: **Token-level comparison of response scorers.** SC favors a fluent but incorrect response, and Visual Lookback Score highlights visually referential tokens but remains insufficient on its own. LOOKBACK combines confidence and visual lookback to select the visually grounded response. Darker highlights indicate higher token-level scores.

Q: Do the two seem comfortable with each other? | **GT:** yes

(1) Selected by SC — Incorrect

Based on the image , it appears that the man and the little boy are comfortable with each other . They are playing with the blue to oth br ush together , suggesting that they feel secure and at ease in each other ' s presence . This interaction might indicate a close relationship , such as a father and his child . Although

(2) Selected by Visual Lookback Score — Incorrect

In the image , a young child can be seen sh irt less with a to oth br ush in its mouth near a person . The parent or care g iver , a man , is standing in close proxim ity to the child and his raz or st rap who ' s br ushing his teeth . It is difficult to gauge

(3) Selected by Ours — Correct

Yes , the man and the baby seem comfortable with each other , as they appear to be enjo ying a pleasant and relax ed moment . The baby seems at ease , holding a to oth br ush near the man ' s ear , indicating trust and familiar ity between the two .

Figure 10: **Token-level behavior on a visually complex example.** Visual lookback Score broadly highlights image-referential tokens, whereas SC places large scores on a few confident tokens. LOOKBACK instead emphasizes question-relevant visual evidence, selecting the correct response. Darker highlights indicate higher token-level scores.

E Usage of AI assistants

In preparing this work, we utilized AI-based writing assistants in a limited way to suggest alternative phrasings, correct grammatical errors, and improve the readability. All research ideas, experimental designs, implementation details, and reported results were determined, implemented, and verified by the authors, and all model outputs and quantitative results in the paper were generated and checked through our own code and experiments. AI assistants did not generate any factual content reported in the paper (e.g., experimental results, dataset statistics, or citations), ensuring the originality, scientific contributions, technical content, methodology, and experimental findings are entirely attributable to the authors.