

Why LVLMs Are More Prone to Hallucinations in Longer Responses: The Role of Context

Ge Zheng^{1,2*} Jiaye Qian^{2*} Jiajin Tang² Sibe Yang^{1†}

¹School of Computer Science and Engineering, Sun Yat-sen University ²ShanghaiTech University

Project Page: <https://github.com/SooLab/HalTrapper>

Abstract

Large Vision-Language Models (LVLMs) have made significant progress in recent years but are also prone to hallucination issues. They exhibit more hallucinations in longer, free-form responses, often attributed to accumulated uncertainties. In this paper, we ask: Does increased hallucination result solely from length-induced errors, or is there a deeper underlying mechanism? After a series of preliminary experiments and findings, we suggest that the risk of hallucinations is not caused by length itself but by the increased reliance on context for coherence and completeness in longer responses. Building on these insights, we propose a novel “induce-detect-suppress” framework that actively induces hallucinations through deliberately designed contexts, leverages induced instances for early detection of high-risk cases, and ultimately suppresses potential object-level hallucinations during actual decoding. Our approach achieves consistent, significant improvements across all benchmarks, demonstrating its efficacy. The strong detection and improved hallucination mitigation not only validate our framework but, more importantly, re-validate our hypothesis on context. Rather than solely pursuing performance gains, this study aims to provide new insights and serves as a first step toward a deeper exploration of hallucinations in LVLMs’ longer responses.

1. Introduction

Recently, Large Vision-Language Models (LVLMs) [3, 7, 8, 12, 18, 48, 98] have made significant strides in developing general-purpose foundation models, achieving new, unprecedented capabilities. These models facilitate dynamic, context-driven interactions centered on the image content through open-ended conversations with users, given the input image and user instructions. Their impressive generative capabilities allow them to address various traditional

*Equal contribution.

†Corresponding author is Sibe Yang.

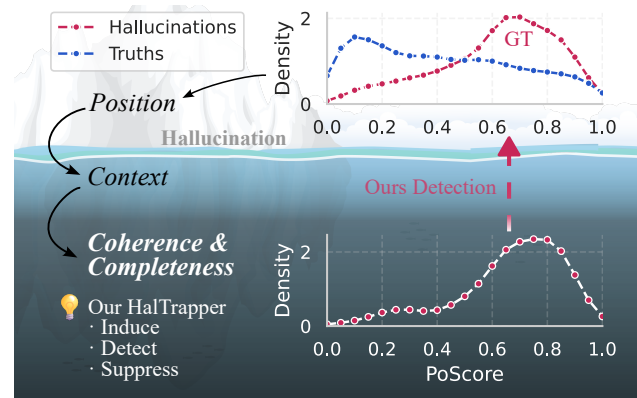


Figure 1. Left: Our three main findings and the three steps of our HalTrapper method. Right: The distribution of hallucination locations detected by our HalTrapper is close to the true distribution of hallucinations, indicating that our method, to some extent, captures the essence of LVLM hallucinations.

vision tasks [5, 20, 34, 35, 38, 49, 56, 58, 63–65, 88, 95, 96, 99, 100] within a unified framework and seamlessly handle more comprehensive tasks [15, 16, 23, 51, 60, 79, 86, 93] that require world knowledge and complex reasoning, such as visual question answering [2, 26, 59, 62], video-based reasoning [6, 9, 37, 41] and mathematical reasoning [54, 74]. However, LVLMs also grapple with the hallucination issue [27, 57, 92, 97], a serious and well-recognized challenge in deploying them in real-world scenarios [21, 36, 45, 52], due to their propensity for erroneous generation.

Hallucination in LVLMs specifically refers to the discrepancy between the generated textual responses and the actual visual content and user instruction received, resulting in the production of irrelevant or non-existent objects, attributes, and other details. Various approaches have been proposed to reduce hallucinations, including filtering more reliable training data [46, 90, 97] or using specialized contrastive training materials [28] to re-fine-tune the model, thereby minimizing factually incorrect outputs. Rather than relying on costly, data-intensive solutions, recent approaches propose training-free strategies, such as

contrastive decoding to contrastive model responses with their error-prone versions [31, 33, 76], rolling back uncertain outputs [25], or enhancing attention to visual content [50]. This has significantly mitigated the hallucination phenomenon, particularly in answering visual questions and identifying specific object hallucinations. However, most of these efforts primarily focus on short responses, while hallucinations in long-form generation remains underexplored.

In this paper, we explore a seemingly straightforward—even widely taken for granted—phenomenon: LVLMs are more prone to hallucinations in longer, free-form textual responses compared to shorter answers. As shown in Fig. 1, the frequency of hallucinated objects correlates with their position in the output token sequence, with a higher likelihood of appearing at later positions. Previous work [97] has also observed similar phenomena, simply attributing the issue to autoregressive text generation, where increasing length leads to accumulated hallucinations and greater uncertainties. However, beneath the intuitive manifestation of length (like an iceberg), deeper factors (beneath the surface) have yet to receive adequate attention: *Is the increased hallucination merely a result of the cumulative errors due to length itself, or does it arise from a deeper underlying mechanism?*

Motivated by this, this paper presents the first and preliminary attempt to explore the underlying factors through a three-step analysis approach:

- Phenomenon discovery to propose hypotheses (Sec. 3).
- Preliminary statistics to analyze hypotheses (Sec. 4).
- Hypothesis application to detect and mitigate hallucinations, thereby re-validating it (Sec. 5).

Phenomenon Discovery: Context may be a potential factor. Since free-form textual responses lack a predefined answer set or clear response forms, LVLMs rely heavily on context, including user instructions, visual input, and especially prior textual outputs. Consequently, we investigate the effect of context (see Sec. 3), specifically by modifying either the image or text context and observing marked shifts in the distribution of the hallucination-length curve, which indicates that hallucinations appear at earlier positions.

Hypothesis Analysis: Contextual coherence and completeness induce hallucinations. Based on this observation, we hypothesize that contextual cues influence hallucinations along two key dimensions:

- **Contextual coherence** drives LVLMs to maintain consistency with prior outputs while avoiding redundancy through distinct generation. The former focuses attention on contextual image content, while the latter shifts it to new information, potentially leading to dispersed attention, confusion, and hallucinations (see Sec. 4.1). Non-hallucinated tokens exhibit clear, focused attention, whereas hallucinated tokens show dispersed patterns. Notably, hallucinated tokens share highly similar attention

distributions (see Fig. 3), suggesting LVLMs may be forced to attend to the same ungrounded, fragmented regions when balancing contextual and distinct content fails.

- **Contextual completeness** requires responses to incorporate comprehensive content while maintaining a logically coherent linguistic structure. However, when available recognized content is insufficient, LVLMs may employ contextual extrapolation as a compensatory strategy, potentially leading to hallucinated outputs (see Sec. 4.2). As contextual completeness increases, hallucinations tend to appear earlier in the response (see Fig. 4). Furthermore, contextual extrapolation seems to follow inherently fixed patterns, with different sets of prompts repeatedly generating overlapping hallucinated tokens.

Application and Re-validation. To further validate the hypotheses, we propose HalTrapper—a novel “**induce-detect-suppress**” framework that directly induces hallucinations by applying the two hypotheses, leverages the induced instances to detect high-risk cases *early to nip them in the bud*, and ultimately suppress potential hallucinations during *the actual decoding stage*.

- **Induction:** (1) Imposing new, coherent outputs on an already complete response induces intra-response hallucinations. (2) Explicitly guiding imagination both based on and beyond recognized objects induces external expansion hallucinations.
- **Detection:** (1) Building on our coherence findings in Fig. 3, we identify hallucinations by analyzing attention similarity with induced intra-response hallucinations. (2) Building on our completeness findings in Fig. 4, we collect potential hallucinations by identifying objects that frequently appear under different imagination prompts. (3) Interestingly, our detection results align with the original hallucination distribution in Fig. 1, suggesting that context-induced and detected hallucinations mirror those seemingly driven by length, re-validating context is one of the potential factors beneath the iceberg of length.
- **Suppression:** Given the detected potential hallucinations, we can directly suppress their likelihood to mitigate hallucinations. Inspired by contrastive decoding [31, 32, 76], we innovatively treat detected hallucinated objects as contrastive context tokens to their probability in the contrastive branch, thereby reducing their likelihood in the original decoding branches.

To sum up, our contributions are as follows:

- We are the first to explore the underlying factors beneath the intuitive length-hallucination correlations, and identify context as the potential factor.
- We introduce a novel hypothesis based on coherence and completeness, and validate it through statistical analysis, hallucination detection, and suppression.
- Our exploration reveals novel insights, including the sim-

ilarity in image attention patterns of hallucinated objects and the repetition of hallucinations across prompts.

- Building on the hypothesis, we propose a novel “induce-detect-suppress” framework, which re-validates our hypothesis while achieving competitive performance on public benchmarks.

2. Related Work

2.1. Large Vision-Language Models

The success of large language models (LLMs) [1, 4, 13, 71] establishes the foundation for the development of large visual-language models (LVLMs) [3, 17, 47, 98]. Recent approaches typically adopt a unified framework, where a pre-trained visual encoder extracts visual features, which are then mapped to the LLM embedding space via either linear layers [12, 47] or Q-Former [3, 17, 98], and subsequently processed with text inputs. While LVLMs demonstrates remarkable capabilities in visual understanding [2, 10, 14, 26, 43, 55, 59, 61, 67–69, 80–85, 94] and reasoning tasks [29, 53, 89] through supervised fine-tuning [22, 24, 42, 47, 91], hallucinations remains a prominent challenge [33, 40, 57, 97]. Existing studies [19, 30, 70, 77] on the internal mechanisms of LVLMs have yet to provide a thorough explanation of the nature of hallucinations, particularly in long-form responses. This work sheds light on hallucinations in long-form generation in LVLMs.

2.2. Hallucinations in LVLMs

Unlike hallucination in LLMs, which refers to the generation of factually incorrect or meaningless content, hallucinations in LVLMs are more concerned with discrepancies between the generated content and the provided visual inputs. Early studies [40, 57] adapt the definition of hallucinations from the captioning task to the context of LVLMs. Subsequent research [25, 32, 46, 97] conduct preliminary analyses of hallucinations, investigating factors such as language priors [32, 46], co-occurrence patterns [32, 97], uncertainty [97], and positional dependencies [97].

Several approaches [25, 28, 31, 32, 46, 50, 76, 87, 90, 97] are proposed to mitigate hallucinations in LVLMs through training. These methods include curating high-quality training datasets [97], integrating specialized contrastive training signals [28], and employing revisor models designed to correct hallucinated outputs [46, 87]. In contrast, other studies [25, 31, 32, 50, 76] explore training-free strategies as alternatives to resource-intensive training approaches. VCD [32] introduces the contrastive decoding (CD) [39] method to suppress hallucinations, gaining significant attention in the field. Subsequent methods [31, 50, 76] further design various contrastive conditions to induce hallucinations from new perspectives. Additionally, OPERA [25] identifies the overreliance on knowledge aggregation posi-

tions within the text attention mechanism as a key cause of hallucinations and suggests a rollback strategy to address this issue. Furthermore, PAI [50] strengthens the impact of image attention on model outputs, effectively reducing hallucinations.

3. Is Context a Deeper Underlying Factor?

In this section, we conduct exploratory experiments to investigate the underlying factor influencing hallucination beyond generation length. We first introduce PoScore to represent hallucination positions and reproduce the widely recognized phenomenon that hallucinations tend to occur in longer responses (Sec. 3.1). Subsequently, we modify either image or text context and analyze their effects on hallucination distribution, thereby identifying context as a potential underlying factor (Sec. 3.2).

Default Experimental Settings. Our default experimental setup (in Sec. 3 and Sec. 4) evaluates the LLaVA v1.5 7B [47], Qwen VL Chat [3], and MiniGPT-4 [98] on a randomly sampled set of 500 COCO [44] images for statistical analysis. Additional experimental details are presented in Appendix A.

3.1. Hallucinations Linked to Length.

When leveraging LVLMs for dialogue or question-answering tasks, a notable phenomenon is that hallucinations tend to occur more frequently in the later positions of the response. To quantitatively analyze this phenomenon, we define the relative position score for each generated object as follows, consistent with previous work [97]:

$$\text{PoScore}_{s,i} = \frac{\text{Index}(o_{s,i})}{N_s} \quad (1)$$

where $o_{s,i}$ denotes the i^{th} object in the response of the s^{th} sample, and N_s represents the length of the s^{th} sample. We visualize the PoScore distributions for hallucinated and non-hallucinated objects for the LLaVA model in Fig. 1, with additional results from other models provided in Fig. 7 in Appendix. The results reveal a marked increase in the frequency of hallucinations as the response lengthens, aligning with findings from previous studies [78, 97].

3.2. Hallucinations Beyond Length.

Moving beyond these prior observations, we delve deeper by posing a critical question: *Is the increased hallucination merely a result of the cumulative errors due to length itself, or does it arise from a deeper underlying mechanism?* In light of the critical role that context plays in free-form responses, we design the following two context modification strategies and analyze the changes in hallucination positions (PoScore) to investigate the effect of context:

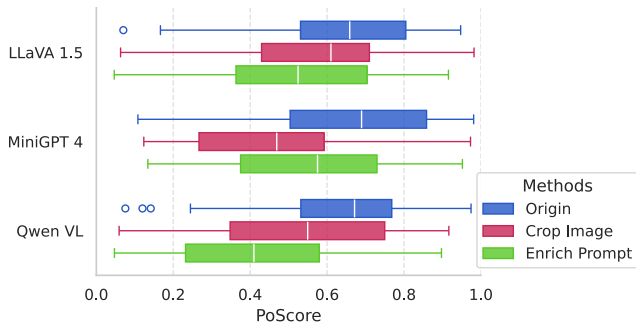


Figure 2. Statistical analysis of hallucination positions under context modifications. Both cropping the image and enriching the prompt lead to earlier hallucination occurrences.

- **Crop the image input** into centered squares, retaining approximately one-third of the original area, and re-annotate accordingly.
- **Enrich the text input** by adding two sentences that describe the image, and then prompt to describe other details.

The results in Fig. 2 show that hallucinations tend to occur earlier in the generation process across both settings, challenging the widely held belief that they are more likely to appear in the later stages. These findings underscore the complexity of hallucinations, revealing that context plays a significant role in their occurrence, rather than attributing them solely to generation length.

4. Coherence and Completeness

This section delve into the mechanisms through which context influences hallucinations by employing a hypothesis-verification framework. Our analysis focuses on two key aspects: contextual coherence (Sec. 4.1) and contextual completeness (Sec. 4.2). Finally, we link back to text and image manipulation experiments in Sec. 3.2, providing explanations with these factors (Sec. 4.3).

4.1. Coherence: Avoidance of Internal Repetition

Contextual coherence drives the model to maintain consistency with previous outputs while avoiding redundant repetition of both the input and prior content. Based on this, we propose and validate a hypothesis on hallucination occurrence.

Hypothesis. The two aspects of contextual coherence in image attention are conflicting: attention is required to focus on relevant regions for consistency with previous outputs, while also shifting to new areas to avoid repetition. This tension leads to dispersed attention and hallucinations.

Experimental settings. To validate our hypothesis, we analyze both individual attention and pairwise attention comparisons. Specifically, we analyze the image attention maps of hallucinated objects \mathcal{H} and non-hallucinated objects \mathcal{N} , with representative results shown in Fig. 3 (right).

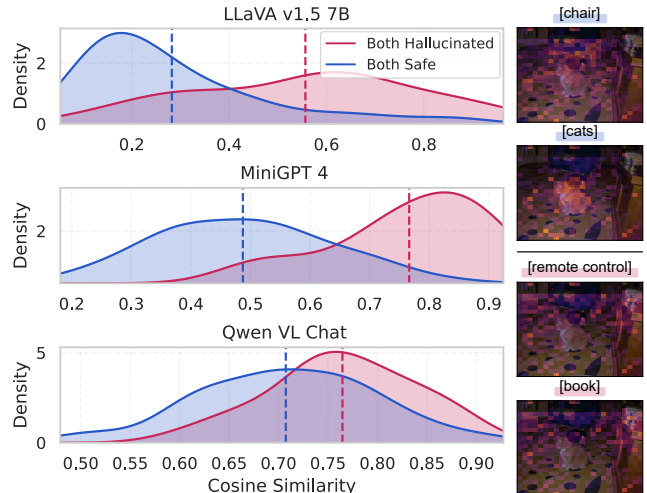


Figure 3. Statistical analysis related to contextual coherence. Within the same caption, hallucinated object pairs exhibit higher attention similarity scores than non-hallucinated pairs.

tionally, we quantify the intra-set attention similarity of objects within the same response, denoted by $S_{\mathcal{H}}$ and $S_{\mathcal{N}}$, as follows:

$$\begin{aligned} S_{\mathcal{H}} &= \{\text{sim}(A_{s,i}, A_{s,j}) \mid o_{s,i}, o_{s,j} \in \mathcal{H}\}, \\ S_{\mathcal{N}} &= \{\text{sim}(A_{s,i}, A_{s,j}) \mid o_{s,i}, o_{s,j} \in \mathcal{N}\} \end{aligned} \quad (2)$$

where $A_{s,i}$ and $A_{s,j}$ represent the image attention maps of the i^{th} and j^{th} objects in the response for the s^{th} image, and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. Fig. 3 (left) illustrates the distributions of $S_{\mathcal{H}}$ and $S_{\mathcal{N}}$.

Results. Qualitative analysis (right panel of Fig. 3) indicates that when the model successfully identifies real objects, it concentrates on the relevant regions. Conversely, if the model fails to recognize a novel object, its attention disperses and distracting information, leading to hallucinations. Quantitative results (left panel of Fig. 3) show a clear difference between the distributions of $S_{\mathcal{H}}$ and $S_{\mathcal{N}}$. Specifically, hallucinated objects exhibit higher attention similarity, while real objects show lower values. This further indicates that hallucinated objects typically manifest diffuse, noisy attention patterns, making attention similarity a robust metric for their detection.

4.2. Completeness: External Extrapolation

Contextual completeness comprises two key dimensions: the informational dimension, which demands a thorough and comprehensive response, and the structural dimension, which ensures the response is logically coherent and grammatically sound. Building on this, we propose the following hypotheses regarding the occurrence mechanism and inherent tendency of hallucination.

Hypothesis. (a) Occurrence: When a response includes correctly identified objects but remains incomplete in informative or structural aspect, the model compensates by ex-

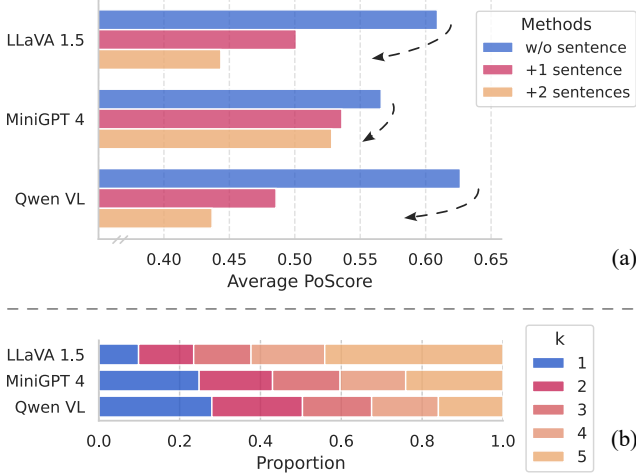


Figure 4. Statistical analysis related to contextual completeness: (a) Hallucination positions shift progressively earlier as more image information is included in the prompts. (b) Similar hallucinations consistently recur across varied prompts for the same image.

panding imagined details, i.e., hallucinations. (b) Tendency: These hallucinations from external extrapolation rely on multimodal context, particularly visual inputs.

Experimental settings. We conduct two separate experiments for validation as follows:

(a) We validate the role of completeness by analyzing its correlation with hallucination positions. Specifically, we extend text manipulation experiment in Sec. 3.2 by incrementally adding image descriptions to the prompt and visualizing the average PoScore in Fig. 4(a).

(b) We further investigate the consistency and image-related properties of hallucinated objects across different prompts. Specifically, we apply five prompts to each image and compute the proportion of repeated hallucinated objects. Formally, let \mathcal{H}_{s_k} represent the set of hallucinated objects generated by the k^{th} prompt for the s^{th} sample, with the complete hallucination set given by $\mathcal{H}_s = \bigcup_{k=1}^5 \mathcal{H}_{s_k}$. We count the occurrence of each hallucinated object $h \in \mathcal{H}_s$ as $c_s(h) = \sum_{k=1}^5 \mathbb{1}(h \in \mathcal{H}_{s_k})$, where $\mathbb{1}$ is the indicator function. Then we calculate $N(k)$, the number of hallucinated objects that appear $k \in [1, 2, 3, 4, 5]$ times over all samples, along with its proportion $R(k)$ shown in Fig. 4(b):

$$N(k) = \sum_s \sum_{h \in \mathcal{H}_s} k \cdot \mathbb{1}(c_s(h) = k), \quad (3)$$

$$R(k) = \frac{N(k)}{\sum_{k=1}^5 N(k)}$$

Results. (a) The results in Fig. 4(a) indicate that as more enriched sentences are incorporated, leading to a more comprehensive context, hallucinations occur at earlier positions. This is because the diminishing content available for generation makes it increasingly challenging for LVLMs to accurately identify details for a complete and coherent response.

(b) The proportion presented in Fig. 4(b) demonstrate that all models exhibit a high degree of repetitiveness in hallucinated objects, with objects appearing in only one response accounting for merely 30% on average. Given the variations in both questions and preceding responses, the repeated hallucinated objects are often closely tied to the image context, aligning with our qualitative analysis in Appendix E.

4.3. Link Back to Phenomenon in Sec. 2.2

Explaining Text Manipulation Experiments. Revisiting the text manipulation experiments, we find that contextual coherence and completeness provides an intuitive explanation for this behavior. When additional descriptions of real objects are incorporated, the model tend to avoid redundancy and maintain coherence, thereby reducing the number of objects to describe. Consequently, the model turns to uncertain or unverified objects more quickly to ensure completeness, leading to earlier hallucinations.

Explaining Image Manipulation Experiments. Contextual completeness offers a compelling explanation for the image manipulation experiments. Similarly, cropping images systematically reduces the number of recognizable objects, forcing the model to hallucinate earlier in order to maintain contextual completeness.

5. Re-Validation via Detection and Suppression

To rigorously validate our hypothesis, we extend the findings from Section 4 to practical application of hallucination detection and suppression. Specifically, we propose HalTrapper, which introduces a novel “induce–detect–suppress” strategy (see Fig. 5). The induce–detect stages leverage Internal Grounding (IG) and External Expansion (EE) techniques for hallucination detection (Sec. 5.1), and can be easily adapted with Contrastive Contextual Decoding (CCD) for suppression (Sec. 5.2).

5.1. Hallucination Induction-Detection

5.1.1. Internal Grounding

In Sec. 4.1, we demonstrate that the attention similarity between paired objects serves as an effective indicator for distinguishing hallucinated pairs from non-hallucinated ones. Building on this insight, we propose the Internal Grounding (IG) method, which adopts an *induce-then-detect* paradigm to detect hallucinated objects in model responses.

Induction. A key component of IG is the selection of reference objects, which serve as anchors for similarity computation. Instead of using naturally generated objects, we induce the model to generate additional objects following the initial response, which are more prone to hallucination. Specifically, given an input image and the model’s initial response, we replace the EOS token in the generated output with an additional cue, “*There is also*”. Since the initial responses inherently covers a considerable number of

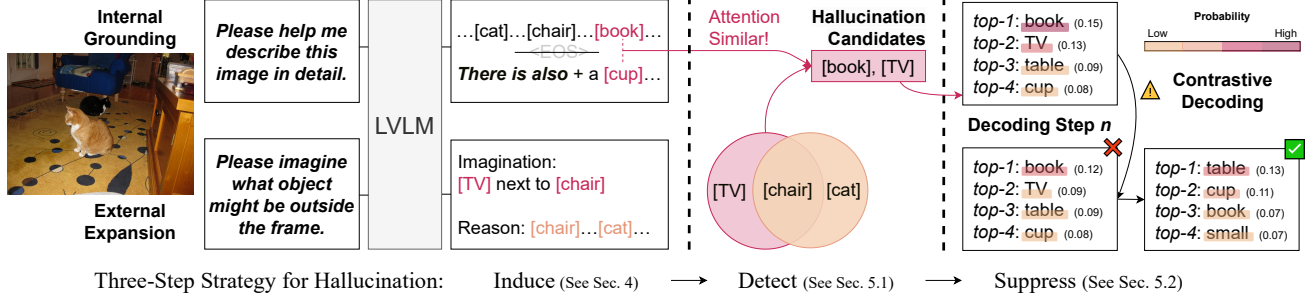


Figure 5. **Overview of HalTrapper:** It consists of two branches leveraging *coherence* and *completeness* insights. One generates captions with an appended “There is also” prompt to induce potential hallucinated objects, detected via high attention similarity between caption and induced tokens. The other prompts the LVLm to imagine surrounded content beyond the image to identify consistent hallucinations. With detected hallucinated objects, HalTrapper further suppresses hallucinations through Contrastive Contextual Decoding.

the identified objects, when completeness is compromised, the model tends to externally extrapolate to compensate, thereby restoring completeness (see Sec. 4.2). The resulting object serves as the reference object and is denoted as o_s^{ref} for the s^{th} sample.

Detection. We then compute the attention similarity scores IGScore between the induced hallucinated objects o_s^{ref} and the preceding objects, filtering out potential hallucination candidates S_{IG} with high similarity:

$$\begin{aligned} \text{IGScore}_{s,i} &= \text{sim}(A_s^{ref}, A_{s,i}) \\ S_{IG} &= \{o_{s,i} \mid \text{IGScore}_{s,i} > \theta_{IG}\} \end{aligned} \quad (4)$$

where θ_{IG} denotes the threshold. Notably, the proposed method remains robust even when the reference object is real, as the similarity scores between non-hallucinated objects are typically low, effectively preventing real objects from being misclassified as hallucinations.

5.1.2. External Expansion

Another observation is that hallucinated objects exhibit consistency across identical visual inputs (Sec. 4.2). Based on this property, we propose the External Expansion (EE) method, explicitly *inducing* the imagination related to the image, treating them as *detected* potential hallucinations.

Induction. Considering that hallucinations from external extrapolation rely on image context, we first prompt with “Please imagine what object might be outside the frame” to induce image-related associations and capture potential hallucinations. However, directly extracting hallucinated objects from the response leads to false positives, as the model might imagine objects present in the image. To address this, we design a reason-then-imagine prompt to filter out such existing objects (see Appendix C.2). It explicitly guides the model in distinguishing between recognized objects and imagined ones. Furthermore, it utilizes reliable intermediate steps to enable context-driven reasoning, thereby improving response fidelity.

Detection. We introduce EEScore, based on the principle that an object’s presence in the imagination set improves the

likelihood of it being perceived as a hallucination, while its presence in the reason set reduces this likelihood. Specifically, we define the imagination set and the reason set at direction $d \in \mathcal{D}$ as $S_{I,d}$ and $S_{R,d}$, respectively. The final set of potential hallucinations is formulated as follows:

$$\begin{aligned} \text{EEScore}_{s,i} &= \sum_{d \in \mathcal{D}} [\mathbb{1}(o_{s,i} \in S_{I,d}) - \mathbb{1}(o_{s,i} \in S_{R,d})] \\ S_{EE} &= \{o_{s,i} \mid \text{EEScore}_{s,i} > \theta_{EE}\} \end{aligned} \quad (5)$$

Finally, we combine the potential hallucinations detected by the IG and EE methods as follows:

$$S_{induction} = S_{IG} \cup S_{EE} \quad (6)$$

5.2. Hallucination Suppression

Preliminaries. Let θ denote the parameters of an LVLm. Given an input image v and a text prompt x , the model autoregressively generates a response y of length L . Formally, the decoding process can be formulated as follows:

$$p_{\theta}(y|v, x) = \prod_{i=1}^L p_{\theta}(y_i|v, x, y_{<i}) \quad (7)$$

where y_i and $y_{<i}$ represent the token at position i and preceding tokens before position i , respectively, and $p_{\theta}(y_i|v, x, y_{<i}) \propto \exp \text{logit}_{\theta}(y_i|v, x, y_{<i})$ denotes the conditional probability distribution of the next token y_i given the preceding tokens $y_{<i}$.

Based on this formulation, we introduce contrastive decoding (CD), originally proposed by [39]. CD utilizes an amateur model as a contrastive reference to optimize the decoding objectives while maintaining plausibility constraint. Recently, [31, 32, 76] apply CD to LVLms, leveraging hallucination-amplifying branches as contrastive signals to mitigate hallucinations. Specifically, the CD process, with the new model θ' as the contrastive branch and all other in-

Model	Metric	AUROC	TPR@5%FPR	F1 _{max}	Acc.
LLaVA v1.5	PoScore	70.7	4.3	38.3	70.7
	Top Logit	64.0	13.0	32.2	61.9
	Logits' Entropy	67.7	16.6	36.6	71.4
	Image Attn. Ratio	44.9	6.0	27.3	32.0
	IG Score	82.3	43.3	54.8	86.3
MiniGPT 4	EE Score	77.5	-	46.1	72.9
	PoScore	70.5	12.2	35.4	66.2
	Top Logit	65.6	22.9	37.0	76.5
	Logits' Entropy	65.5	22.1	35.3	75.9
	Image Attn. Ratio	64.3	7.7	31.9	57.9
Qwen VL	IG Score	76.6	34.0	48.6	80.7
	EE Score	60.5	-	30.0	46.5
	PoScore	71.1	4.8	34.4	65.8
	Top Logit	71.5	19.6	36.1	77.7
	Logits' Entropy	70.7	23.3	36.6	73.9
Qwen VL	Image Attn. Ratio	57.3	6.8	26.9	41.4
	IG Score	76.2	33.3	43.8	84.6
	EE Score	81.3	-	46.3	73.0

Table 1. Quantitative results for hallucination detection. The best performances within each setting are **bolded**.

puts unchanged, is expressed as follows:

$$p_{cd}(y_i|v, x, y_{<i}) = \text{softmax}[(1 + \alpha)\text{logit}_\theta(y_i|v, x, y_{<i}) - \alpha\text{logit}_{\theta'}(y_i|v, x, y_{<i})] \quad (8)$$

where $p_{\theta'}(x_i|v, x, y_{<i}) \propto \exp \text{logit}_{\theta'}(x_i|v, x, y_{<i})$. It also employs a truncation of the probability distribution following [32].

Contrastive Contextual Decoding (CCD). Building on the previously introduced induce-detect stages, a simple CD-based extension CCD enables hallucination suppression. Unlike previous CD methods, CCD explicitly integrates a prior for potential hallucination objects, aiming to reduce their likelihood in response. Specifically, we encode potential hallucinated objects as text tokens, referred to as Contrastive Contextual Tokens (CCT) x_{cct} . We then concatenate CCT with the image input to construct a contrastive branch, with model parameters and other inputs unchanged. The CCD process can be formally expressed as follows:

$$p_{ccd}(y_i|v, x, y_{<i}) = \prod_{i=1}^L p_{ccd}(y_i|v, x_{cct}, x, y_{<i}) \quad (9)$$

We then detail the modifications applied to the CD process as follows:

$$p_{ccd}(y_i|v, x_{cct}, x, y_{<i}) = \text{softmax}[(1 + \alpha)\text{logit}_\theta(y_i|v, x, y_{<i}) - \alpha\text{logit}_\theta(y_i|v, x, x_{cct}, y_{<i})] \quad (10)$$

By treating CCT tokens as complementary to image content, the model naturally increases the likelihood of potential hallucinated objects and their associated terms in the contrastive branch, thereby effectively reducing their occurrence in the final generation.

6. Experiments

Datasets and Benchmarks. To demonstrate the effectiveness of our HalTrapper, we use images from COCO [44]

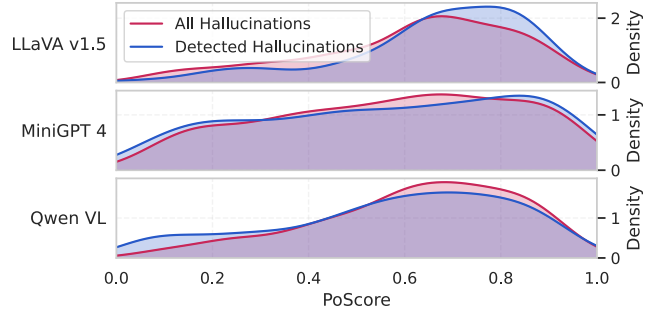


Figure 6. Comparison between the positional distribution of hallucinations detected by our method and the overall hallucination distribution, demonstrating a high degree of alignment.

and AMBER [72] datasets. Detailed descriptions can be found in the Appendix C.1.

Base Models. We select LLaVA v1.5 7B [47], MiniGPT-4 [98], and Qwen VL Chat [3] as our main baselines for our study. We also evaluate more recent models Qwen2 VL 7B [75] and Janus Pro 7B [11] on AMBER, which has higher annotation qualities.

Implementation Details. For all experiments, the maximum number of newly generated tokens is set to 512. Following prior mainstream studies on CD [32, 76], we adapt $\alpha = 1.0$ and $\beta = 0.1$. See Appendix C.3 and C.4 for details on CCT construction and more hyperparameters.

6.1. Detection

Metrics. Inspired by [66], we adapt AUROC (Area Under the ROC Curve) and TPR@5%FPR (the True Positive Rate at 5% False Positive Rate) as our primary metrics for hallucination detection. AUROC quantifies the model’s overall discriminative ability across all classification thresholds, while TPR@5%FPR is suitable for scenarios with strict requirements on the false positive rate. We also report the F1 Score and Accuracy at the threshold that *maximizes* the F1.

Baseline Methods. For each generated object $o_{s,i}$, we first employ PoScore [97] as a basic metric. We also propose two uncertainty-based metrics: Top Logit and Logits’ Entropy. The Top Logit is the maximum value of the logits when generating $o_{s,i}$, while Logits’ Entropy refers to the entropy of the logits at that moment. Additionally, we employ an Attention-based metric called the Image Attention Ratio, defined as the ratio of the model’s attention score on the image to its total attention score when generating $o_{s,i}$.

Results. The quantitative results of hallucination detection are presented in Table 1. As shown, our approach demonstrates significant improvements across all evaluation settings. For IG, in terms of the AUROC metric, our method outperforms the best baseline PoScore by 5%–12%. This indicates that our method enhances performance across the entire classification curve. Considering that our IG method originates from within the model, this indicates that the model indeed exhibits significant similar attention pat-

Decoding	Method	LLaVA v1.5 7B [47]					
		C _S ↓	C _I ↓	Prec.	Recall	F1	Len
Greedy	ICD [76]	51.4	14.7	73.4	81.0	77.0	102.1
	CODE [31]	50.0	13.7	75.8	76.9	76.4	88.3
	Vanilla Ours	52.2 41.6	14.6 11.9	73.7 78.7	80.3 80.1	76.9 79.4	100.8 100.0
Nucleus	VCD [32]	58.2	16.9	70.8	78.8	74.6	103.2
	ICD [76]	55.0	16.5	70.9	77.9	74.2	102.1
	CODE [31]	54.2	16.4	72.3	76.2	74.2	91.6
Beam Search	Vanilla Ours	58.6 48.6	18.8 14.5	68.1 74.6	76.4 77.7	72.0 76.1	105.2 100.9
	OPERA [25]	53.6	15.7	72.4	77.6	74.9	98.8
	Vanilla Ours	55.6 45.2	15.8 12.1	72.8 78.9	81.0 81.2	76.7 80.0	104.2 101.8

Table 2. Results on CHAIR. Lower CHAIR_S, CHAIR_I, and higher precision, recall and F1 indicate fewer hallucinations. The best performances within each setting are **bolded**.

tern in certain hallucination scenarios. Additionally, for TPR@5%FPR, our method improves by at least 10% compared to the baselines. This highlights the substantial potential of the EE metric in inducing hallucinations. Given that MiniGPT 4 is trained only on the image interface, its ability to follow instructions is relatively limited, which may account for the lack of improvement in the EE metric.

We also visualized the qualitative results of hallucination positions distribution detected by our method, with the overall distribution of hallucination positions, as shown in Fig. 6. It demonstrates that our method accurately captures the hallucination distribution, closely aligning with the overall pattern observed in captions. This further indicates that although we claim that our method is designed for long-text scenarios, its effectiveness is not merely dependent on the length of the generated text. Instead, our approach effectively captures an intrinsic mechanism underlying LVLM hallucinations, which is beyond text length. Therefore, our study not only validates the applicability of our method but also provides a new perspective for understanding the formation mechanism of LVLM hallucinations.

6.2. Suppression

Metrics. CHAIR [57] is commonly used to quantify hallucinations in model-generated captions based on COCO. Besides CHAIR, we also report several classic metrics, including Precision, Recall, F1, and the average length of the captions. For AMBER [72], following the approach outlined in their paper, we report CHAIR, Cover, Hal, and Cog. As we primarily focus on long context scenarios, we conduct full evaluations only on its generative subset and reported the results accordingly. We also conduct experiments on POPE and GPT-4o, please refer to the Appendix D.2 and D.4.

Baseline Methods. We compare our HalTrapper with VCD [32], ICD [76], CODE [31], and OPERA [25].

CHAIR Evaluation. As shown in Tables 2 and 3. HalTrapper significantly reduces CHAIR while maintaining Recall

Decoding	Method	MiniGPT 4 [98]			Qwen VL Chat [3]		
		C _S ↓	C _I ↓	Prec.	C _S ↓	C _I ↓	Prec.
Greedy	Vanilla	39.6	14.7	76.6	43.4	13.5	75.8
	ICD [76]	42.6	14.7	76.3	50.4	14.4	73.7
	CODE [31] Ours	32.8 28.6	13.6 10.7	81.2 83.1	40.4 38.6	12.5 10.2	78.9 80.9
Nucleus	Vanilla	37.2	14.6	77.1	44.8	13.6	76.3
	VCD [32]	39.6	14.9	76.6	47.4	14.1	74.3
	ICD [76]	41.4	14.9	76.1	52.6	15.0	73.0
Beam Search	CODE [31] Ours	36.6 29.0	14.0 11.5	79.5 82.1	43.6 42.4	14.5 11.3	75.4 79.3
	Vanilla	38.8	13.8	78.0	41.4	11.6	79.0
	OPERA [25] Ours	43.0 37.6	14.9 13.7	75.8 78.3	42.8 34.2	12.5 9.7	76.9 82.7

Table 3. More results on CHAIR with MiniGPT-4 and Qwen VL.

Model / Method	CHAIR↓	Cover↑	Hal↓	Cog↓
LLaVA v1.5 7B [47]	11.2	50.2	47.9	4.6
+ VCD [32]	8.9	51.2	38.1	4.4
+ ICD [76]	8.6	51.1	37.3	3.9
+ CODE [31]	9.0	51.1	39.5	4.3
+ Ours	8.0 (3.2↓)	51.5 (1.3↑)	36.3 (11.6↓)	3.8 (0.8↓)
Qwen2 VL [75]	6.6	71.8	50.3	4.6
+ VCD [32]	7.3	70.6	53.2	4.6
+ ICD [76]	8.2	74.9	74.9	9.1
+ CODE [31]	7.6	71.6	56.3	5.1
+ Ours	5.6 (1.0↓)	70.9	46.1 (4.2↓)	3.8 (0.8↓)
Janus Pro 7B [11]	6.3	65.6	37.5	2.0
+ VCD [32]	5.5	66.2	32.5	2.1
+ ICD [76]	6.1	67.1	36.3	2.5
+ CODE [31]	6.0	65.3	33.6	1.6
+ Ours	5.4 (0.9↓)	66.5 (0.9↑)	32.7 (4.8↓)	1.8 (0.2↓)

Table 4. Results on AMBER [73] generative task. ↓ indicates lower is better.

with minimal negative impact. Across all experiments on CHAIR_S and CHAIR_I, HalTrapper achieves significant improvements. Notably, in Table 2, our approach consistently improves CHAIR_S by over 10% and CHAIR_I by 2.5%. This demonstrates that the hallucination candidates identified by our IG and EE metrics are of high quality, enabling the inclusion of a large number of hallucinated objects while minimizing the presence of non-hallucinated ones. This, in turn, provides validation of the effectiveness of our IG and EE metrics in detecting hallucinations, further highlighting the universality and practical significance of our findings.

AMBER Evaluation. As shown in the Table 4, HalTrapper continues to demonstrate performance improvements on latest models. **Ablation Study.** See Appendix D.1 for more details on the ablation study.

7. Conclusion

In this paper, we propose a novel method for eliminating hallucinations in Large Vision-Language Models through two mechanisms: external spatial expansion and internal visual grounding. Our HalTrapper introduces a simple, zero-shot hallucination detection and suppression technique that achieves significant improvements across all benchmarks, with no additional training required. Our approach consistently delivers substantial improvements across all benchmarks, validating its effectiveness.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No.62206174).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 3
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3, 7, 8
- [4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [5] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibe Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10297–10318, 2024. 1
- [6] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 1
- [7] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1
- [9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1
- [10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3
- [11] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 7, 8
- [12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3
- [13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3
- [14] Qiyuan Dai and Sibe Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13711–13722, 2024. 3
- [15] Qiyuan Dai and Sibe Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em, 2025. 1
- [16] Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibe Yang. Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement, 2025. 1
- [17] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 3
- [18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1
- [19] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2025. 3
- [20] Weifeng Ge, Sibe Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning, 2018. 1
- [21] Xiang He, Sibe Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8417–8424, 2019. 1
- [22] Zijian He, Yuwei Ning, Yipeng Qin, Guangrun Wang, Sibe Yang, Liang Lin, and Guanbin Li. Vton 360: High-fidelity virtual try-on from any viewing direction, 2025. 3
- [23] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibe Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023. 1
- [24] Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibe Yang. MVTokenflow: High-quality 4d content generation using multiview token flow. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [25] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and

- retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 2, 3, 8
- [26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 3
- [27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 1
- [28] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 1, 3
- [29] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3
- [30] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s” up” with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 3
- [31] Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*, 2024. 2, 3, 6, 8
- [32] 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. *arXiv preprint arXiv:2311.16922*, 2023. 2, 3, 6, 7, 8
- [33] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 2, 3
- [34] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1
- [36] Jinpeng Li, Haiping Wang, Jiabin chen, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibe Yang, Yuan Li, Wenping Wang, Zhen Dong, and Bisheng Yang. Cityanchor: City-scale 3d visual grounding with multi-modality LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [37] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1
- [38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1
- [39] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022. 3, 6
- [40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 3
- [41] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 1
- [42] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibe Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 482–493, 2024. 3
- [43] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibe Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2022. 3
- [44] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 3, 7
- [45] Zhenxiang Lin, Xidong Peng, Peishan Cong, Ge Zheng, Yujing Sun, Yuenan Hou, Xinge Zhu, Sibe Yang, and Yuexin Ma. Wildrefer: 3d object localization in large-scale dynamic scenes with multi-modal visual data and natural language. In *ECCV (46)*, pages 456–473, 2024. 1
- [46] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3, 7, 8
- [48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

- [49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1
- [50] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024. 2, 3
- [51] Xuyang Liu, Bingbing Wen, and Sibeï Yang. Ccq: Cross-class query network for partially labeled organ segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1755–1763, 2023. 1
- [52] Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibeï Yang, Wenping Wang, Jingyi Yu, Xuming He, and Yuexin Ma. Realdex: towards human-like grasping for robotic dexterous hand. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. 1
- [53] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022. 3
- [54] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1
- [55] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [57] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1, 3, 8
- [58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [59] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 1, 3
- [60] Cheng Shi and Sibeï Yang. *Spatial and Visual Perspective-Taking via View Rotation and Relation Reasoning for Embodied Reference Understanding*, page 201–218. Springer-Verlag, Berlin, Heidelberg, 2022. 1
- [61] Cheng Shi and Sibeï Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3
- [62] Cheng Shi and Sibeï Yang. Logoprompt:synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1
- [63] Cheng Shi and Sibeï Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. In *The Twelfth International Conference on Learning Representations*, 2024. 1
- [64] Cheng Shi, Yulin Zhang, Bin Yang, Jiajin Tang, Yuexin Ma, and Sibeï Yang. Part2object: Hierarchical unsupervised 3d instance segmentation. *arXiv preprint arXiv:2407.10084*, 2024.
- [65] Cheng Shi, Yuchen Zhu, and Sibeï Yang. Plain-det: A plain multi-dataset object detector. In *Computer Vision – ECCV 2024*, pages 210–226, Cham, 2025. Springer Nature Switzerland. 1
- [66] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7
- [67] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibeï Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23570–23580, 2023. 3
- [68] Jiajin Tang, Ge Zheng, and Sibeï Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023.
- [69] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibeï Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023. 3
- [70] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 3
- [71] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [72] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 7, 8

- [73] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 8
- [74] Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 1
- [75] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7, 8
- [76] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Bie-mann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. 2, 3, 6, 7, 8
- [77] Yifan Wang, Yifei Liu, Yingdong Shi, Changming Li, Anqi Pang, Sibe Yang, Jingyi Yu, and Kan Ren. Discovering influential neuron path in vision transformers. In *International Conference on Representation Learning*, pages 25244–25272, 2025. 3
- [78] Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, and Debin Zhao. Toward a stable, fair, and comprehensive evaluation of object hallucination in large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3
- [79] Yu Wu, Yana Wei, Haozhe Wang, Yongfei Liu, Sibe Yang, and Xuming He. Grounded image text matching with mismatched relation reasoning, 2023. 1
- [80] Sibe Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4643–4652, 2019. 3
- [81] Sibe Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4140–4149, 2019.
- [82] Sibe Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *Computer Vision – ECCV 2020*, pages 589–605, Cham, 2020. Springer International Publishing.
- [83] Sibe Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9949–9958, 2020.
- [84] Sibe Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2021.
- [85] Sibe Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11261–11270, 2021. 3
- [86] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibe Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. *arXiv preprint arXiv:2412.01550*, 2024. 1
- [87] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 3
- [88] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1
- [89] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 3
- [90] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2407.11422*, 2024. 1, 3
- [91] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibe Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *ACM Transactions on Graphics*, 42:1–16, 2023. 3
- [92] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023. 1
- [93] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*, 2023. 1
- [94] Hong-Yu Zhou, Chixiang Lu, Sibe Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3479–3489, 2021. 3
- [95] Hong-Yu Zhou, Chixiang Lu, Sibe Yang, and Yizhou Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2230–2238, 2021. 1
- [96] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibe Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023. 1
- [97] 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. *arXiv preprint arXiv:2310.00754*, 2023. 1, 2, 3, 7
- [98] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language

understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 3, 7, 8

- [99] Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibe Yang. Rethinking query-based transformer for continual image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4595–4606, 2025. 1
- [100] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 1