# A Dual-Perspective Decoding for Hallucination Mitigation in Large Vision-Language Models

**Anonymous ACL submission**

## Abstract

Large Vision-Language Models (LVLMs) face the challenge of object hallucination, where the model generates descriptions of nonexistent objects. This issue primarily arises from the failure of the visual encoder to attend to detailed regions and the tendency of the language model to favor contextual plausibility over visual evidence during generation. In this work, we propose a dual-perspective decoding framework that jointly optimizes text generation from both visual and textual views to address hallucinations caused by image-text misalignment. Our framework aligns generated text with visual content at both the sentence and word levels from the textual perspective, while simultaneously ensuring that visual objects are aligned with their corresponding textual semantics from the visual perspective. Extensive experiments demonstrate that our method significantly reduces object hallucination and achieves superior image-text alignment compared to existing state-of-the-art methods. Notably, our method achieves significant improvements of 7.5% to 19.2% over previous approaches under the CHAIR evaluation metrics, highlighting its effectiveness in enhancing the visual faithfulness of generation.

## 1 Introduction

Large Vision-Language Models (LVLMs) (Dai et al., 2023; Bai et al., 2025; Liu et al., 2023; Zhu et al., 2023; Team, 2024; DeepSeek-AI et al., 2025) have achieved remarkable progress in visual understanding and reasoning tasks (Yang et al., 2025; Wang et al., 2023; Li et al., 2023a; Zhang et al., 2024; Meta, 2025) by harnessing the capabilities of large language models (LLMs) such as Qwen (Bai et al., 2023), LLaMA (Touvron et al., 2023), and GPT (Brown et al., 2020). Despite these advancements, LVLMs are susceptible to a prevalent challenge known as object hallucination, where the model generates descriptions of non-existent objects or fabricates visual details (Zhai et al., 2024; Hu et al., 2023; Li et al., 2023b; Stiennon et al., 2022). This limitation undermines their reliability and applicability, particularly in tasks requiring precise visual-textual alignment.

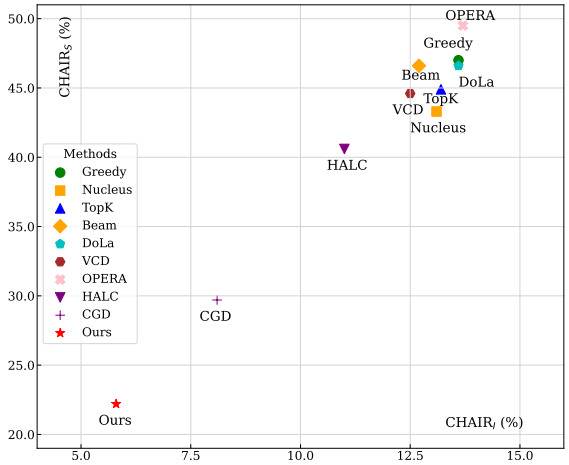



Figure 1: Visualization of the effectiveness of our method on the $CHAIR_S$ and $CHAIR_I$ metrics on the benchmark (Lin et al., 2015). Lower values of $CHAIR_I$ and $CHAIR_S$ indicate better performance. The results demonstrate the significant advantage of our approach, highlighting its superior performance in hallucination mitigation for LLAVA-1.5 compared to other methods.

To address object hallucination, recent studies have proposed two main categories of mitigation strategies: external knowledge-based methods and decoding-based approaches (Huang et al., 2024a; Liu et al., 2024b; Zhou et al., 2024). The former, such as LURE and Woodpecker (Zhou et al., 2024; Yin et al., 2024), leverage annotated data for supervised learning, while the latte including VCD, SID, and HALC (Leng et al., 2023; Chen et al., 2024; Huo et al., 2024) introduces lightweight, training-free mechanisms that enhance image-text alignment during inference. Decoding-based methods have gained increasing attention due to their plug-and-play flexibility, often employing contrastive or

alignment techniques to mitigate hallucination.

Despite significant progress, Large Vision-Language Models (LVLMs) still suffer from object hallucination, which refers to the generation of descriptions that mention objects not actually present in the image. One major cause lies in the limitations of visual encoding. LVLMs often struggle to capture fine-grained visual details due to the way attention is allocated across visual inputs. Although visual encoders can extract semantically meaningful representations, they frequently overlook detailed image regions. This problem is further exacerbated by resolution bottlenecks and the design of projection modules, which may inadvertently discard important fine-grained features during the conversion of images into token representations (Li et al., 2023b; Chen et al., 2024). Another contributing factor is the decoding process based on large language models (LLMs). The attention mechanism in LLMs tends to prioritize linguistic coherence over visual accuracy. As a result, the generated descriptions may include content that is contextually plausible but not supported by the visual input (Zhou et al., 2024; Wang et al., 2024). These misalignment between visual encoding and language decoding weakens the model's ability to produce text grounded in visual content, leading to hallucinated descriptions that deviate from the actual image.

To address the misalignment arising from both visual encoding and language decoding, we propose a dual-perspective decoding framework that enhances image-text alignment and reduces hallucinations during text generation. From the visual perspective, a DINO-based module performs object-level matching, identifying and suppressing hallucinated descriptions by enforcing fine-grained correspondence between detected objects and their textual mentions. From the textual perspective, a CLIP-based evaluator measures global semantic consistency between the generated caption and the image, ensuring that the overall description remains faithful to the visual content. A fusion module then combines the DINO alignment score, the CLIP semantic score, and the LVLM's native generation confidence into a unified re-ranking criterion. By applying this alignment-guided re-ranking to multiple sampled outputs, our method retains the diversity and fluency of decoding-based approaches while substantially reducing reliance on external large models and avoiding additional training.

Overall, our contributions are summarized as follows:

1. We re-examine the causes of object hallucination and introduce a dual-perspective decoding framework that provides fine-grained object alignment from the visual perspective and assesses overall semantic consistency between the generated text and the image.

2. We design a lightweight fusion module that appropriately weights and combines visual-perspective alignment scores from the DINO module, textual-perspective consistency scores from the CLIP evaluator, and the LVLM's native generation likelihood, while incorporating likelihood entropy as an auxiliary metric to promote balanced, high-quality outputs.

3. Experimental results validate the effectiveness of our method, demonstrating significant performance improvements across multiple benchmarks while also demonstrating superior generation quality.

## 2 Related Work

### 2.1 Object Hallucination in LVLMs and Current Mitigation Approaches

Inspired by the success of LLMs, large vision-language models (LVLMs) derived from LLMs demonstrate remarkable performance on a wide range of visual tasks. Mainstream LVLMs, such as LLAVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023), typically comprise three core components: a visual encoder to extract visual features, a projection module to align visual representations with the language model, and an LLM to generate textual outputs.

Object hallucination refers to the generation text by Vision-Language Models (VLMs) that includes objects not faithful to the image content (Rohrbach et al., 2019). In Large Vision-Language Models (LVLMs), this issue primarily manifests as a mismatch between the generated content and the image. Mainstream methods for mitigating hallucination can be broadly categorized into two types: external knowledge leveraging and decoding methods. External knowledge leveraging methods rely on human annotations, hyper models to label training data, or provide feedback. LRV (Liu et al., 2024a) addresses data bias-induced response biases in LLMs by supplementing instruction data. LURE (Zhou et al., 2024) utilize the GPT-4V model to label additional data, which is then used to train reviser for mitigating hallucination. RLAIF-

V (Yu et al., 2024b) makes a significant contribution by employing LVLMs of peers exhibiting comparable or equal capabilities to provide feedback, which is subsequently used to reinforce learning. On the other hand, recently proposed decoding methods leverage the intermediate state distributions of LLMs during the decoding process, the distributions after input distortion, or assistance from other models to optimize the final token distribution. These decoding approaches, which avoid the complex training process required for supplementing knowledge, have become a significant research focus in the field of hallucination mitigation.

### 2.2 Decoding Strategies for Mitigating Object Hallucinations

The decoding method of LVLM is a crucial approach for optimizing the inference stage and serves as an important strategy for mitigating hallucination in generated text. Common basic decoding methods include greedy decoding, nucleus decoding (Holtzman et al., 2020), top-$k$ sampling (Fan et al., 2018), and beam search (Lemons et al., 2022).

Recent works have proposed further exploration from basic decoding methods to enhance performance. Internal methods like DOLA (Chuang et al., 2024) leverage internal state signals to refine decoding results, while contrast decoding methods such as VCD (Leng et al., 2023) and ICD (Kim et al., 2024) contrast output distributions using different prompts or layers. Furthermore, there are several independently proposed methods. HALC (Chen et al., 2024) introduces the visual-alignment (Liu et al., 2024c) module into beam search, aiming to identify the optimal visual context for LVLM inputs. Meanwhile, the CGD (Deng et al., 2024) method employs the text-alignment (Radford et al., 2021) module to assist in selecting among different sampling results to improve vision-language alignment. Our approach adopts a dual-perspective strategy to address vision-language misalignment, leveraging readily available external tools: DINO for visual grounding and CLIP for semantic consistency. By aligning the generated text from both visual and textual perspectives, the model effectively reduces hallucinated content that deviates from the visual input.

## 3 Method

In this section, we introduce the detailed approach of our method, which is designed to mitigate hallucinations in text generation from large vision-language models (LVLMs), the overview of our method is summarized in Figure 2. Our method combines both textual and visual perspectives, with a final score fusion module that iteratively refines the generated captions. We now describe the key components of our approach in detail.

**Textual Perspective:** The first critical component of our approach is aligning the generated caption semantically with the image information. For this, we utilize the CLIP module, which provides a shared embedding space for both text and image, allowing us to measure the consistency between them.

At time $i$, given an image $x_{\text{img}}$ and a sequence of sentences $S = \{s_0, s_1, s_2, \ldots, s_{i-1}\}$ generated by the LVLM, we compute the CLIP score for each sampled sentence. We sample the generated sentences as $\{s_{i1}, s_{i2}, \ldots, s_{ik}\}$ , where k measures the number of sampled time of current generation state. CLIP utilize the image and sampled sentence as input to compute CLIP score, which measures the similarity between the image and the entire sentence in the shared embedding space and defined as:

$$g_{\text{CLIP}}^{sentence}(x_{\text{img}}, s_{ij}) = \cos\left(f_{\text{CLIP}}(x_{\text{img}}), f_{\text{CLIP}}(s_{ij})\right) \quad (1)$$

where: $f_{\text{CLIP}}(x_{\text{img}})$ is the CLIP feature embedding of the image $x_{\text{img}}$, $f_{\text{CLIP}}(s_{ij})$ is the CLIP feature embedding of the generated sentence $s_{ij}$. j represents the $j$-th sampled sentence of current state. A higher cosine similarity indicates a stronger alignment between the image and the sentence. The CLIP score serves as a measure of how well the generated text reflects the image content, helping to identify sentences that may contain hallucinated information that doesn't correspond to any visual element in the image.

Furthermore, to ensure that each object mentioned in the sentence is appropriately grounded in the visual content, we compute the alignment score for each individual word. For each sentence $s_{ij}$, we extract the object terms using a named entity recognition (NER) model (Neumann et al., 2019). For instance, a sentence like "A dog is playing with a ball"would result in the object terms **dog** and **ball**. For each recognized object term $w$ so called

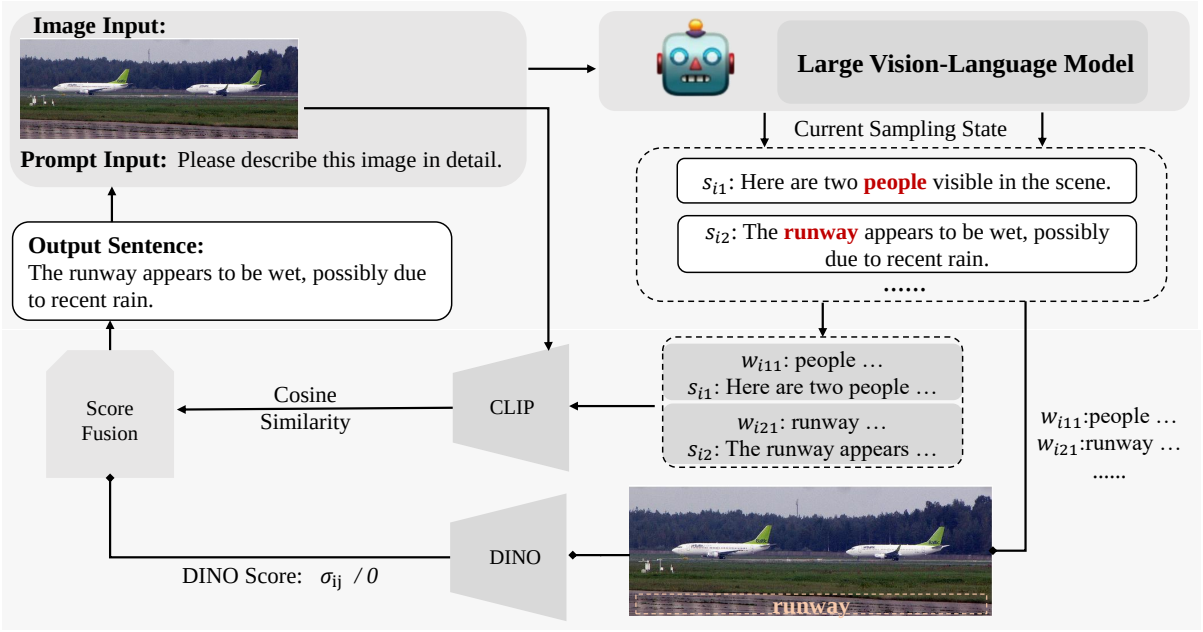



Figure 2: Overview of our method: For the sentences generated by LVLMs, our approach leverages the CLIP module from the textual perspective to evaluate image-text consistency and the DINO module from the visual perspective to ensure fine-grained object alignment. Scores from both modules are then combined with the likelihood probability scores to derive the final scores for comparing the sampled results. Once the optimal sampling candidates are determined, the generation process is iteratively repeated to ensure consistency and quality.

valid word in the sentence $s$, we calculate the cosine similarity between the object embedding in the text and the corresponding object embedding in the image:

$$g_{\text{CLIP}}(x_{\text{img}}, w_{ijl}) = \cos\left(f_{\text{CLIP}}(x_{\text{img}}), f_{\text{CLIP}}(w_{ijl})\right) \quad (2)$$

where $w_{ijl}$ is the $l$-th valid word in the $j$-th sampled sentence. The overall CLIP score for the word is then calculated as the minimum of word-level alignment scores:

$$g^{word}_{\text{CLIP}}(x_{\text{img}}, s_{ij}) = \min_{w_{ijl}} g_{\text{CLIP}}(x_{\text{img}}, w_{ijl}) \quad (3)$$

where $\min_{w_{ijl}}$ denotes the minimum value of the CLIP similarity score across all $w_{ijl}$ in $s_{ij}$. The CLIP score on the sentence, combined with the scores for valid words, ensures that the generated caption is not only consistent with the image as a whole but also correctly grounded for each individual object mentioned in the text. The final CLIP score equation is summuriazed as follows:

$$\begin{aligned} g_{\text{CLIP}}(x_{\text{img}}, s_{ij}) &= \gamma \cdot g^{sentence}_{\text{CLIP}}(x_{\text{img}}, s_{ij}) \\ &+ (1-\gamma) \cdot \text{g}^{word}_{\text{CLIP}}(x_{\text{img}}, s_{ij}) \end{aligned} \quad (4)$$

**Visual Perspective:** While the textual alignment module ensures that the generated captions are semantically aligned with the image, it is equally important to ensure that the specific objects described in the caption correspond to actual objects in the image. The DINO module performs fine-grained object detection and aligns image information with corresponding object terms in the caption, providing an additional layer of visual grounding.

We use the same named entity recognition method as CLIP module to extract objects in generated text. The DINO module then computes an alignment score between the object mentioned in the sentence and the objects present in the image. It computes score as follows:

$$\sigma_{ijl} = \max_{w_{ijl}}(p_{ij,w_{ijl}}), \quad (5)$$

where $w_{ijl}$ represents the $l$-th valid word in the $j$-th sampled sentence, and $p_{ij,w_{ij}}$ denotes the confidence score computed by transformers attention head generated for the word $w_{ij}$ (Liu et al., 2024c). The DINO confidence score represents the visual information detected by DINO from the original image, we define the DINO score as follows:

$$g_{\text{DINO}}(x_{\text{img}}, w_{ijl}) = \sigma_{ijl} \quad (6)$$

In this work, if the DINO confidence score $\sigma_{ijl}$ falls below the threshold, the DINO score is computed as 0; otherwise, it is replaced by a constant $C$. We find this approach to be both simple and

effective in our experiments.

$$g_{\text{DINO}}(x_{\text{img}}, w_{ijl}) = \begin{cases} C, & \text{if } (\sigma_{ijl} > \delta) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $\delta$ is the threshold from the default DINO settings, and the hyperparameter $C$ denotes the weight of the DINO score. This score indicates how well the object is represented in the image.

For each sentence $s_{ij}$, we aggregate the object-level DINO scores to obtain an overall alignment score:

$$g_{\text{DINO}}(x_{\text{img}}, s_{ij}) = \min_{w_{ijl}} g_{\text{DINO}}(x_{\text{img}}, w_{ijl}) \quad (8)$$

where $\min_{w_{ijl}}$ denotes the minimum value of the DINO score across all $w_{ijl}$ in $s_{ij}$. The aggregated score gives a measure of how well the sentence reflects the visual content of the image, particularly with regard to the objects mentioned.

**Score Fusion:** To preserve the generative quality of LVLMs, we introduce the concept of LLM likelihood and utilize this probability as a scoring metric in the subsequent evaluation. Given a premise question text $s = \{t_1, t_2, \ldots, t_m\}$, where $t_i$ denotes the token generated at the $i$-th timestep, we utilize Predictive Entropy (PE) for uncertainty estimation(Kadavath et al., 2022; Duan et al., 2024; Kuhn et al., 2023), which is defined as the entropy of the entire sentence. To mitigate the impact of generation length on predictive entropy and ensure the proper functioning of LVLM, we adopt a variant known as length-normalized predictive entropy as Equation 4. This variant divides the joint log-probability of each sequence by the length of the sequence, as proposed by Malinin and Gales(Malinin and Gales, 2021) in the context of natural language generation (NLG) uncertainty, and has been empirically shown to be advantageous in the work by Kuhn(Kuhn et al., 2023).

$$f_\theta(s_i) = \frac{1}{m} \sum_{i=1}^{m} -\log p_\theta(x_i | s_{<i}) \quad (9)$$

where $\theta$ represents the LVLM parameters and $m$ is the length of the generated sentence.

To combine the strengths of both the CLIP and DINO modules, we propose a score fusion strategy. This module integrates the textual consistency score from CLIP, the visual alignment score from DINO, and the internal likelihood score from the LVLM to compute a final fusion score. The final score $F(x_{\text{img}}, s_{ij})$ for each generated caption $s_{ij}$ is given by:

$$F(x_{\text{img}}, s_{ij}) = (1-\alpha)\big(g_{\text{DINO}}(x_{\text{img}}, s_{ij}) + g_{\text{CLIP}}(x_{\text{img}}, s_{ij})\big) + \alpha \cdot f_\theta(s_{ij}) \quad (10)$$

where $g_{\text{DINO}}(x_{\text{img}}, s_{ij})$ is the visual alignment score computed by DINO, $g_{\text{CLIP}}(x_{\text{img}}, s_{ij})$ is the semantic consistency score computed by CLIP, $f_\theta(s_{ij})$ is the internal likelihood score from the LVLM for caption $s_{ij}$. $\alpha$ is hyperparameter representing the weight of the auxiliary score's influence on the decoding distribution. When $\alpha = 1$, the scorer reduces to greedy decoding.

The fusion score $F(x_{\text{img}}, s_{ij})$ provides a comprehensive measure of caption quality, incorporating both visual alignment and textual consistency. This score is then used to rank the generated captions, and the top candidates are selected for further refinement.

We employ an iterative decoding strategy to refine the generated captions further. At each iteration, the LVLM generates a set of candidate captions for the given image. These candidates are evaluated using the fusion score $F(x_{\text{img}}, s_i)$, and the top $N$ candidates are selected for further refinement. This iterative process continues for $k$ iterations, ensuring that the generated captions gradually improve in terms of both semantic consistency and visual accuracy. Our method's iteration loop is summarized as Algorithm 1 in Appendix F.

## 4 Experiments

In this section, we evaluate the performance of our method on long descriptions, focusing on its effectiveness in mitigating object hallucination while maintaining caption quality. Our experiments include CHAIR, OPOPE, and GPT-4V-assisted evaluations. Additional experimental results and analyses are provided in Appendix B.

### 4.1 Experiment Setups

**Baselines:** To effectively evaluate our method, we include regular greedy decoding, nucleus sampling (Holtzman et al., 2020), top-k sampling (Fan et al., 2018), and beam search (Lemons et al., 2022) as baselines. Additionally, we incorporate state-of-the-art methods specifically designed to mitigate object hallucination (OH), including DoLa (Chuang et al., 2024), OPERA (Huang et al., 2024b), VCD (Leng et al., 2023), CGD (Deng et al.,

| Method | InstructBLIP | | | mPLUG-Owl2 | | | LLAVA-1.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $BLEU \uparrow$ | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $BLEU \uparrow$ | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $BLEU \uparrow$ |
| Greedy | 57.9 | 17.1 | 15.9 | 52.7 | 16.0 | 18.1 | 47.0 | 13.6 | 18.9 |
| Nucleus | 56.1 | 17.0 | 16.4 | 51.9 | 15.6 | 18.1 | 43.3 | 13.1 | 16.4 |
| TopK | 55.8 | 16.9 | 16.5 | 53.1 | 15.9 | 18.1 | 44.9 | 13.2 | 16.3 |
| Beam | 53.2 | 14.8 | 18.7 | 55.8 | 16.1 | 17.1 | 46.6 | 12.7 | 18.3 |
| DoLa | 55.6 | 17.0 | 16.5 | 52.6 | 15.2 | 18.1 | 46.6 | 13.6 | 19.2 |
| VCD | 63.2 | 19.5 | 17.7 | 51.4 | 16.0 | 17.5 | 44.6 | 12.5 | 17.8 |
| OPERA | 51.5 | 15.6 | 18.3 | 48.5 | 16.1 | 17.9 | 49.5 | 13.7 | 18.4 |
| HALC | 61.6 | 18.9 | 18.1 | 51.7 | 15.5 | 17.4 | 40.6 | 11.0 | 19.0 |
| CGD | 42.7 | 10.9 | 16.4 | 35.7 | 8.6 | **19.1** | 29.7 | 8.1 | 18.4 |
| **Ours** | **23.5** | **6.3** | **19.4** | **24.6** | **6.8** | 18.0 | **22.2** | **5.8** | **19.4** |

Table 1: Experimental results of different decoding methods on various LVLMs using the **MSCOCO-CHAIR** (Lin et al., 2015) dataset. The results are reproduced based on the original papers or official code. $C_S$ refers to $CHAIR_S$, $C_I$ refers to $CHAIR_I$ and $B$ refers to BLEU-1 Score. Higher BLEU-1 scores indicate better text generation quality, while lower $CHAIR_S$ and $CHAIR_I$ scores reflect stronger hallucination mitigation. **Bold** values indicate the best performance across other methods.

2024), and HALC (Chen et al., 2024) in our analysis.

**LVLM Backbones:** We conduct our experiments on different LVLMs—InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2023), and mPLUG-Owl2 (Li et al., 2022)—to evaluate our method and all the previously mentioned baselines.

### 4.2 Metrics

**Datasets:** We conduct our experiments mainly on three benchmark:We conducted our experiments primarily on three benchmarks: CHAIR, POPE, and a GPT-4V assisted evaluation. Detailed descriptions of these datasets are provided in Appendix A. The results across all three benchmarks consistently demonstrate the effectiveness of our approach in mitigating object hallucination, while maintaining high-quality text generation.

**CHAIR:** To evaluate the effectiveness of our method in mitigating object hallucination, we follow the standard CHAIR evaluation setting (Rohrbach et al., 2019). For all backbones, we use the prompt "Please describe this image in detail" . The generated parameters for our method are provided in Appendix B, with hyperparameters $\alpha = 0.01$ and $\gamma = 0.5$. The CHAIR results are shown in Table 1. Throughout the experiments, our method achieves state-of-the-art (SOTA) performance in reducing hallucinations across other methods while maintaining caption generation quality. Specifically, our method achieves a **7.5%** to **19.2%** improvement over the previous SOTA under the CHAIR metrics. We observe that our method performs better with backbones exhibiting high levels of hallucination, which can be attributed to the alignment module's effectiveness in mitigating hallucinations. The generated sentences contain an average of 80 to 90 words, with the max new tokens parameter set to 512. Notably, we conduct experiments on different lengths of generated captions in Appendix and evaluate our method on other LVLMs under the CHAIR benchmark in Appendix. These results demonstrate that the superior performance of our method remains consistent across both long and short description generation tasks.

**POPE:** Following the HALC's OPOPE setup (Chen et al., 2024), We conduct the POPE experiment, and the results are presented in Table 2. Throughout the evaluation, our method achieves better results compared to the greedy baseline and the HALC method, despite yielding a lower accuracy score. As noted by HALC, false positives become less reliable in offline POPE testing, and the diversity of described content may introduce biases in true positive samples. Consequently, this can result in deviations in the accuracy metric. Therefore, we primarily utilize precision and $F_{0.2}$ score as reference metrics. According to our experimental results, our approach achieves state-of-the-art performance within HALC's OPOPE framework.

**GPT-4V assisted evaluation:** Following the OPERA(Huang et al., 2024b) protocol, we conduct a GPT-4V assisted evaluation to assess the effectiveness of our method in mitigating hallucinations in generated captions. Notably, we observe that GPT-4V tends to assign higher scores to captions presented second in sequence. To mitigate this bias, we conduct a second round of evaluation where the order of captions in each pair was swapped. The evaluation results, adjusted for order bias, are presented in the Table 3. And the

| Setting | Deooding | InstructBLIP | | | mPLUG-Owl2 | | | LLAVA-1.5 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A\uparrow$ | $P\uparrow$ | $F_{0.2}\uparrow$ | $A\uparrow$ | $P\uparrow$ | $F_{0.2}\uparrow$ | $A\uparrow$ | $P\uparrow$ | $F_{0.2}\uparrow$ |
| Random | Greedy | **76.8** | 94.2 | **91.8** | **75.1** | 92.3 | 90 | **78.4** | 94.8 | 92.6 |
| | HALC | 76.7 | 93.8 | 91.5 | 73.7 | 92.2 | 89.5 | 73.8 | 95.8 | 92.4 |
| | Ours | 71.9 | **94.4** | 90.8 | 72.3 | **95.0** | **91.4** | 73.4 | **96.1** | **92.6** |
| Popular | Greedy | 73.1 | 83.9 | 82.4 | **71.5** | 82.6 | 81 | **74.9** | 85.7 | 84.3 |
| | HALC | **73.3** | 84.2 | 82.7 | 70.2 | 82.1 | 80.3 | 71.4 | 87.7 | 84.9 |
| | Ours | 70.1 | **87.9** | **84.9** | 70.5 | **88.8** | **85.8** | 72.7 | **90.3** | **87.6** |
| Adversarial | Greedy | **72.5** | 82.6 | 81.2 | 68.9 | 76.5 | 75.3 | **73.1** | 81.5 | 80.4 |
| | HALC | 71.2 | 79.4 | 78.2 | 68.4 | 77.5 | 76.1 | 70.3 | 84.6 | 82.2 |
| | Ours | 69.2 | **85.0** | **82.4** | **68.9** | **83.5** | **81.1** | 70.7 | **86.9** | **84.3** |

Table 2: Experimental results of different decoding methods on various LVLMs in the **MSCOCO-OPOPE** (Chen et al., 2024). The results are reproduced using the original papers or official code. A refers to Accuracy, P refers to Precision and $F_{0.2}$ refers to $F_{0.2}$ Score. Higher Accuracy, Precision and $F_{0.2}$ Score scores indicate better quality, whereas lower $CHAIR_S$ and $CHAIR_I$ scores reflect stronger hallucination mitigation. **Bold** values represent the best results among all methods.

| Method | InstructBLIP | | mPLUG-Owl2 | | LLAVA-1.5 | |
|---|---|---|---|---|---|---|
| | C | D | C | D | C | D |
| Greedy | 4.47 | 5.11 | 5.10 | 5.71 | 5.95 | 6.11 |
| **Ours** | **5.08** | **5.93** | **5.83** | **5.74** | **6.27** | **6.34** |
| OPERA | 5.44 | 5.75 | 5.35 | 5.70 | 5.98 | 6.24 |
| **Ours** | **5.90** | **6.04** | **5.78** | **5.71** | **6.11** | **6.29** |
| HALC | 5.95 | **6.34** | 5.51 | **6.29** | 5.10 | 4.91 |
| **Ours** | **6.27** | 6.11 | **6.24** | 6.16 | **6.28** | **6.40** |

Table 3: Experimental results of different decoding methods on GPT4V assist evaluation in OPERA (Chen et al., 2024). The results are reproduced using the OPERA official code.

comprehensive results of GPT-4V assisted evaluation are shown in Appendix. Experimental results demonstrate that our method outperforms existing approaches in both hallucination mitigation and generation quality.

### 4.3 Ablation Study and Analysis

All ablation experiments are conducted using LLaVA-1.5 as the backbone model, with hyperparameters consistent with those described in Section 4.1 for LLaVA-1.5.

**Effectiveness of Modules:** To demonstrate the effectiveness of individual modules and the improvement in hallucination mitigation achieved by combining them, we conducted ablation experiments under four conditions, as shown in Table 4. The results demonstrate that both the DINO and CLIP modules, when used individually, significantly reduce hallucinations, with the effect being more pronounced when using DINO alone. This validates the effectiveness of using DINO and CLIP as alignment mechanisms in our approach. Moreover, when both DINO and CLIP are used together, the performance surpasses that of either module alone, confirming the enhanced effect of their combined supervision.

| Greedy | DINO | CLIP | $\text{CHAIR}_S$ | $\text{CHAIR}_I$ |
|---|---|---|---|---|
| ✓ | | | 47.0 | 13.6 |
| | ✓ | | 24.8 | 6.8 |
| | | ✓ | 38.0 | 11.4 |
| | ✓ | ✓ | **22.2** | **5.8** |

Table 4: Comparison of performance for different modules.

**Granularity of Inputs:** We conduct ablation experiments by using *object*, *attribute*, and *relation* as separate inputs for DINO and CLIP. Additionally, we test various input combinations for CLIP to evaluate the effectiveness of both object and sentence inputs. The greedy setting refers to the greedy decoding baseline used for comparison. The DINO experimental results are shown in Table 6 in Appendix.

Although *attribute* and *relation* are semantically important categories, and previous studies, such as HALC and HalluciDoctor (Yu et al., 2024a), have used *existence*, *attribute*, and *relation* as keywords for hallucination mitigation, our results indicate that the best performance in hallucination mitigation is achieved when only *object* is used as the input for CLIP, as shown in Table 5.

Our analysis reveals that DINO struggles to effectively localize *attribute* and *relation*, resulting in excessive meaningless grounding. This issue is also discussed in R-Bench (Wu et al., 2024). Considering the time efficiency of the DINO module, we ultimately choose to use only *object* as the input for DINO and CLIP.

While CGD has made significant progress in sentence-level CLIP decoding, our findings suggest that both word-level and sentence-level inputs contribute uniquely to hallucination mitigation, as demonstrated in Table 6. The combination of object

| Greedy | Rel | Attr | Obj | $CHAIR_S$ | $CHAIR_I$ |
|---|---|---|---|---|---|
| ✓ | | | | 47.0 | 13.6 |
| | ✓ | | | 45.5 | 14.4 |
| | | ✓ | | 42.4 | 13.1 |
| | | | ✓ | **38.9** | **11.7** |
| | ✓ | ✓ | | 40.3 | 12.6 |
| | ✓ | | ✓ | 40.5 | 12.4 |
| | | ✓ | ✓ | 42.4 | 12.1 |
| | ✓ | ✓ | ✓ | 45.4 | 12.8 |

Table 5: Comparison of performance for Word Categories in CLIP input.

| Greedy | Object | Sentence | $CHAIR_S$ | $CHAIR_I$ |
|---|---|---|---|---|
| ✓ | | | 47.0 | 13.6 |
| | ✓ | | 46.2 | 12.5 |
| | | ✓ | 44.2 | 12.4 |
| | ✓ | ✓ | **38.0** | **11.4** |

Table 6: Comparison of performance under different input for CLIP.

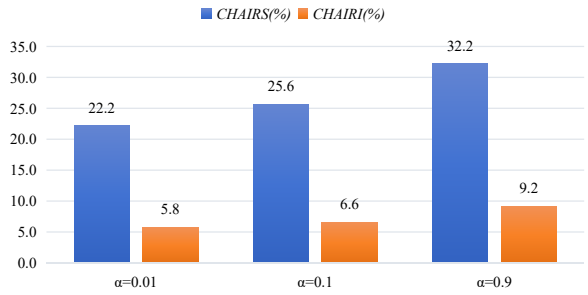



Figure 3: Hallucination ratio under different settings of the hyperparameter $\alpha$

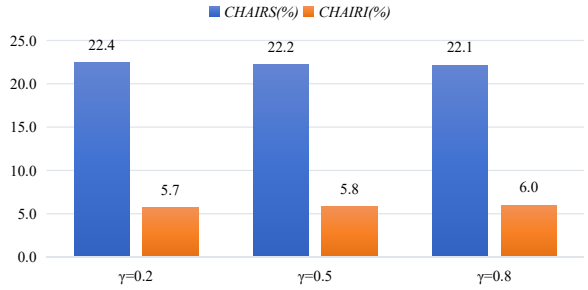



Figure 4: Hallucination ratio under different settings of the hyperparameter $\gamma$

and sentence inputs produces the best performance.

**Hyper Parameters:** Due to the use of multiple modules in our method, we conduct detailed ablation experiments on various hyperparameters of the model.

We first focus on the ratio of the auxiliary score to the likelihood score, which involves the hyperparameter $\alpha$, as shown in the formula. We set $\alpha$ to 0.01, 0.1, and 0.9. The $CHAIR_S$ and $CHAIR_I$ metrics under different $\alpha$ settings are presented in Figure 3. Next, we conduct experiments on the ratio of word-level to sentence-level CLIP scores. Based on Equation 4, we experiment with different values of $\gamma$, set to 0.2, 0.5, and 0.8. The $CHAIR_S$ and $CHAIR_I$ metrics under different $\gamma$ settings are depicted in Figure 4.

Based on the above analysis and considering the results from other experiments, we set $\alpha = 0.01$ and $\gamma = 0.5$ in our experiments.

## 5 Limitations.

While our method demonstrates strong effectiveness in mitigating hallucinations, there are two primary limitations.

(1) While the method performs well across general benchmarks, its effectiveness in specialized domains, such as medical imaging, low-resource languages, or scenes with densely packed objects remains underexplored. Nonetheless, preliminary experiments in the safety domain have yielded promising results. In future work, we plan to further validate its effectiveness across a broader range of application scenarios.

(2) Although our method remains reasonably efficient in practice, there remains room for improvement in decoding speed. We provide a detailed time complexity analysis and discuss potential acceleration strategies in Appendix C.

## 6 Conclusion

Motivated by the misalignment between vision and language during generation, we propose a dual-perspective decoding framework to mitigate hallucinations in large vision-language models (LVLMs). Extensive experiments across multiple benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches in both reducing hallucinations and preserving the semantic integrity of generated captions, achieving a 7.5%–19.2% improvement in CHAIR metrics. We further observe that the effectiveness of hallucination mitigation strongly depends on the alignment module, particularly for models with a high tendency to hallucinate. Moreover, our method maintains its superior performance across both short and long description generation tasks. Importantly, it achieves these improvements without requiring additional training or external data, making it a practical and readily deployable solution for existing LVLMs.

## References


Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2000. *Longman Grammar of Spoken and Written English*. Longman, Harlow, England.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *Preprint*, arXiv:2403.00425.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. *Preprint*, arXiv:2309.03883.

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. *Preprint*, arXiv:2305.06500.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *Preprint*, arXiv:2402.15300.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *Preprint*, arXiv:2307.01379.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *Preprint*, arXiv:1805.04833.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *Preprint*, arXiv:2309.02301.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024b. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *Preprint*, arXiv:2408.02032.

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*.

Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. 2024. Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions. *Preprint*, arXiv:2311.00233.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Sofia Lemons, Carlos Linares López, Robert C. Holte, and Wheeler Ruml. 2022. Beam search: Faster and monotonic. *Preprint*, arXiv:2204.02929.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022.

mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *Preprint*, arXiv:2205.12005.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. *Preprint*, arXiv:2002.07650.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/the-llama-4-herd/`. Accessed: 2025-05-14.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

LLaVA Team. 2024. Llava-Next: Exploring large vision-language models. Accessed: 2024-05-28.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. 2024. Vigc: Visual instruction generation and correction. *Preprint*, arXiv:2308.12714.

Zhen Wang, Jun Xiao, Yueting Zhuang, Fei Gao, Jian Shao, and Long Chen. 2023. Learning combinatorial prompts for universal controllable image captioning. *Preprint*, arXiv:2303.06338.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. *Preprint*, arXiv:2406.16449.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12).

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *Preprint*, arXiv:2311.13614.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *Preprint*, arXiv:2405.17220.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. 2024. Halle-control: Controlling object hallucination in large multimodal models. *Preprint*, arXiv:2310.01779.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024. Gpt4roi: Instruction tuning large language model on region-of-interest. *Preprint*, arXiv:2307.03601.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. *Preprint*, arXiv:2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

# A Datasets

**CHAIR:** The Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2019) tool is specifically designed to assess hallucinations in image captioning tasks. It quantifies hallucinations by evaluating how many objects mentioned in the caption are absent from the ground truth label set. CHAIR provides two distinct evaluation metrics: $CHAIR_S$, which measures the proportion of hallucinated sentences relative to the total number of sentences, and $CHAIR_I$, which evaluates the proportion of hallucinated objects relative to the total number of generated objects. Lower scores on either metric indicate fewer hallucinations. We also evaluate the methods using BLEU (Papineni et al., 2002), a caption-related metric that measures the similarity between generated and ground truth captions. Higher BLEU scores, specifically BLEU-1, indicate better generation quality.

**OPOPE:** Polling-based Object Probing Evaluation (POPE) is a method specifically designed to assess hallucination issues in LVLM. POPE focuses on evaluating object hallucination by utilizing an essay-style prompt in the format: "Is there a <object> in the image?" to pose visual question answering (VQA) queries to the model. The complete POPE test is divided into three splits: Random Split: Objects are randomly selected from the entire dataset for evaluation. Popular Split: This split assesses the presence of objects that most frequently appear in the dataset. Adversarial Split: This evaluates the model's ability to identify objects that are highly relevant to those present in the image.

We adopt the OPOPE evaluation method proposed by HALC to assess hallucination under descriptive conditions rather than simple "yes" or "no" answers. This approach enables our method to be evaluated in a long-sentence generation environment. In practice, OPOPE employs the prompt "Please describe this image in detail" to generate captions. OPOPE then checks whether the sampled positive and negative objects appear in the generated captions to compute the POPE scores. To ensure consistency, we used the $F_{0.2}$ score, as proposed by HALC, where false negatives (FN) and the resulting recall are given less weight due to their limited trustworthiness in offline checks. Additionally, we used the same parameters and generated captions of the same average length as CHAIR.

**GPT-4V assisted evaluation:** We adopt the GPT-4V assisted evaluation method proposed by OPERA to assess the generation quality and hallucination phenomena of our approach compared to other decoding methods. Specifically, we randomly sample 500 images from the MSCOCO validation set and use decoding methods to generate descriptions for these images. The caption generation parameters and prompt we use are the same as CHAIR experiment. The evaluation involves presenting GPT-4V with the image and the corresponding descriptions generated using two decoding methods. GPT-4V is subsequently prompted to assign a score ranging from 0 to 10 for each description, evaluating two key aspects: correctness (C) and detailedness (D).

# B Experimentation Details

## B.1 Experiment Setups

The main generation parameters are configured as follows: the maximum number of new tokens is set to 512, top-$k$ to 5, top-$p$ to 1, and the temperature to 1. Our method targets hallucination mitigation in captions comprising multiple sentences; therefore, the maximum new tokens parameter is set to 512 to evaluate its effectiveness in long-caption scenarios. This generation length is aligned with the standard configuration in mainstream methods. The remaining parameters follow the default settings of the sampling method implemented in the HuggingFace Transformers library[1].

## B.2 Generation Length Comparison

In our experiments, similar to mainstream methods, we use 512 tokens for caption generation. Additionally, to ensure a fair comparison with other decoding methods, such as HALC and OPERA, we conduct experiments on the CHAIR benchmark with a max new tokens setting of 64, as shown in Table 7. Experimental results demonstrate that our method attains optimal performance at this generation length.

# C Time Analysis.

Figure 5 demonstrates that the best results are achieved with a sampling time of 3. To optimize generation efficiency, we set the sampling time to 3 for all experiments. Table 9. The experimental

[1] https://huggingface.co/docs/transformers

| Method | Instuctblip | | mPLUG-Owl2 | | LLAVA-1.5 | |
|---|---|---|---|---|---|---|
| | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ |
| Greedy | 30.9 | 12.3 | 23.2 | 8.3 | 20.8 | 6.8 |
| VCD | 30.3 | 12.6 | 27.3 | 9.7 | 23.3 | 7.90 |
| OPERA | 30.0 | 11.7 | 22.1 | 7.6 | 21.1 | 6.7 |
| HALC | 30.0 | 11.4 | 17.3 | 7.4 | 13.8 | 5.5 |
| **Ours** | **21.8** | **8.1** | **16.4** | **5.9** | **11.5** | **4.2** |

Table 7: Experimental results of various methods with a 64 max new tokens setting on different LVLMs in the **MSCOCO-CHAIR** dataset. Results are reproduced using the original papers and official code.

| Method | MiniGPT-4 | | | LLAVA-Next | | |
|---|---|---|---|---|---|---|
| | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $BLEU \uparrow$ | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | $BLEU \uparrow$ |
| Greedy | 40.6 | 14.1 | 16.7 | 19.8 | 6.2 | 16.6 |
| Nucleus | 34.0 | 12.5 | 17.3 | 23.0 | 7.9 | 16.3 |
| TopK | 35.0 | 12.5 | 17.1 | 21.2 | 7.1 | 16.4 |
| Beam | 32.2 | 11.9 | 17.1 | 15.5 | 5.5 | 16.8 |
| DoLa | 31.8 | 11.6 | 17.0 | 17.8 | 6.1 | **16.8** |
| VCD | 35.7 | 13.8 | **18.1** | 21.4 | 7.3 | 16.4 |
| OPERA | 36.4 | 12.7 | 17.0 | 17.8 | 6.1 | 16.8 |
| HALC | 34.3 | 11.8 | 16.8 | 16.6 | 6.3 | 16.7 |
| **Ours** | **21.0** | **8.2** | 16.2 | **14.1** | **4.7** | 16.2 |

Table 8: Experimental results of different methods on MiniGPT-4 and LLAVA-Next in the **MSCOCO-CHAIR**

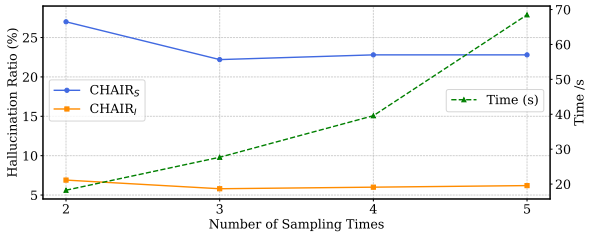



Figure 5: Performance of our method sampling times in range of 2 to 5

| Method | Decoding Time |
|---|---|
| Greedy | **3.90** |
| HALC | 89.88 |
| CGD | 19.13 |
| Ours | 27.68 |

Table 9: Comparison of time cost of different decoding method. The parameters are configured to the official settings.

parameters for each method are selected based on their best performance. The results indicate that our method achieves state-of-the-art hallucination mitigation while maintaining competitive generation efficiency.

Based on the (Biber et al., 2000), nouns comprise approximately 25% of generated words. Since sentence-level decoding is independent, these steps can be **parallelized**, enabling a tractable estimation of time cost. Assuming an average sentence length of $m$ words, and that each sentence triggers one additional CLIP evaluation, the average per-token time cost can be approximated as:

$$T_{\text{LVLM}} + 0.25 \times (T_{\text{DINO}} + T_{\text{CLIP}}) + \frac{1}{m} T_{\text{CLIP}}$$

Here, $T_{\text{LVLM}}$ denotes the time required for the base vision-language model to decode one token. The term $0.25 \times (T_{\text{DINO}} + T_{\text{CLIP}})$ reflects the fact that roughly 25% of tokens (nouns) are grounded using DINO and undergo additional CLIP validation, while the $\frac{1}{m} T_{\text{CLIP}}$ accounts for sentence-level scoring applied once per sentence.

In practice, since $T_{\text{DINO}}$ and $T_{\text{CLIP}}$ are significantly smaller than $T_{\text{LVLM}}$, the overall time cost is close to standard greedy decoding. Therefore, despite the integration of two alignment modules, the expected runtime overhead remains minimal due to both their low per-call latency and the parallelizable nature of the added operations.

## D CHAIR Results

We conduct CHAIR experiments on other mainstream LVLMs, including Minigpt4 (Zhu et al., 2023) and LLAVA-Next (Team, 2024), which are less commonly used with CHAIR compared to models such as LLAVA-1.5, Instructblip, and mPLUG-Owl2. For Minigpt4, we use Llama2 as its large language model, and for LLAVA-Next, we use the Vicuna-7B version[2]. The experimental results are shown in Table B.1. These experiments demonstrate the generalizability of our method,

[2] https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b

| Greedy | Rel | Attr | Obj | $CHAIR_S$ | $CHAIR_I$ |
|---|---|---|---|---|---|
| ✓ | | | | 47.0 | 13.6 |
| | ✓ | | | 47.4 | 13.6 |
| | | ✓ | | 46.6 | 13.9 |
| | | | ✓ | **22.2** | **5.8** |

Table 10: Comparison of $CHAIR_S$ and $CHAIR_I$ for different DINO inputs, **Bold** values represent the best results.

highlighting its ability to mitigate hallucinations even when applied to more advanced models.

## E Ablation Study and Analysis

We also conduct experiment on different granularity inputs for DINO, which contains *object*, *attribute* and *relation*. The experimental results are presented in Table 10. Our analysis reveals that DINO struggles to effectively localize *attribute* and *relation*, resulting in excessive meaningless grounding.

## F Algorithm

We summarize the process of our method in Algorithm 1.

**Algorithm 1** Our method's Algorithm

**Input:** LVLM parameterized by $\theta$, sampling times $k$, candidates number $N$, weight hyperparameter $\alpha$, image input $x_{img}$ and text prompt $s_0$
**Parameter:** $\theta$, $k$, $N$, $\alpha$
**Output:** $y$

Let $t = 0$
$\text{SET}_0 \leftarrow \{\langle \text{Input}(\text{x}_{img}, s_0) \rangle\}$
**while** $\text{SET}_t$ is not empty **do**
  $\text{SET}_{t+1} \leftarrow \emptyset$
  **for all** candidate in $\text{SET}_i$ **do**
    **repeat**
      Sample $s \sim \text{LVLM}_\theta(s_{t+1}|x, s_0, s_1, \ldots, s_{ij})$
      $F(x_{\text{img}}, s_{ij}) = (1-\alpha)\big(g_{\text{DINO}}(x_{\text{img}}, s_{ij})) + g_{\text{CLIP}}(x_{\text{img}}, s_{ij})\big) + \alpha \cdot f_\theta(s_{ij})$
      $\text{SET}_{t+1} \leftarrow \text{SET}_{t+1} \cup \{x, s_0, s_1, \ldots, s_{ij}\}$
    **until** $k$ times
  **end for**
  Rank $\text{SET}_{t+1}$ by F(s)
  $\text{SET}_{t+1} \leftarrow$ Top N candidates in $\text{SET}_{t+1}$
  $t \leftarrow t + 1$
**end while**
$y = \arg\max(\text{SET})$
**return** $y$

## G Comprehensive GPT-4V Assisted Evaluation

Following the GPT-4V assisted evaluation proposed by OPERA, we conduct experiments on mainstream LVLMs such as LLAVA-1.5, InstructBLIP, and Mplug-Owl2. Two aspects are evaluated: correctness (C) and detailedness (D), both scored by GPT-4V. Since we observe that GPT-4V tends to assign higher scores to captions presented second in sequence, we construct prompts in both orders: the original prompt order, as used in OPERA's official code, where baseline captions appear first followed by our method's captions, and the reverse prompt order, where our method's captions appear first followed by baseline captions. Experimental results from Table 11 to Table 13 demonstrate that our method outperforms existing approaches in both hallucination mitigation and generation quality.

| Order | Method | InstructBLIP | | mPLUG-Owl2 | | LLAVA-1.5 | |
|---|---|---|---|---|---|---|---|
| | | C | D | C | D | C | D |
| Original Order | Greedy | 4.63 | 5.12 | 5.25 | **5.62** | 6.05 | 6.07 |
| | **Ours** | **5.73** | **5.89** | **5.63** | 5.61 | **6.14** | **6.28** |
| | Difference | +1.10 | +0.77 | +0.38 | -0.01 | +0.09 | +0.17 |
| Reverse Order | **Ours** | **6.42** | **5.96** | **6.02** | **5.87** | **6.39** | **6.4** |
| | Greedy | 4.3 | 5.09 | 4.95 | 5.8 | 5.85 | 6.14 |
| | Difference | +2.17 | +0.87 | +1.07 | +0.07 | +0.54 | +0.26 |

Table 11: Experimental results of comparing between our decoding method and greedy decoding methods on GPT4V-assist benchmark in OPERA paper. The "original order" refers to the prompt where greedy captions appear first, followed by our method's captions. In contrast, the "reverse order" refers to the prompt where our method's captions appear first, followed by greedy captions.

| Order | Method | InstructBLIP | | mPLUG-Owl2 | | LLAVA-1.5 | |
|---|---|---|---|---|---|---|---|
| | | C | D | C | D | C | D |
| Original Order | OPERA | 5.25 | 5.79 | 5.55 | **5.82** | 5.97 | 6.09 |
| | **Ours** | **6.02** | **6.05** | **5.56** | 5.81 | **6.03** | **6.18** |
| | Difference | +0.77 | +0.26 | +0.01 | -0.01 | +0.06 | +0.09 |
| Reverse Order | **Ours** | **5.77** | **6.02** | **6.00** | **6.5** | **6.19** | **6.4** |
| | OPERA | 5.63 | 5.70 | 5.14 | 5.58 | 5.99 | 6.39 |
| | Difference | +0.14 | +0.32 | +0.86 | +0.92 | +0.20 | +0.01 |

Table 12: Experimental results of comparing between our decoding method and OPERA decoding method on GPT4V-assist benchmark. The "original order" and "reverse order" correspond to the same content as described in Table 11.

## H Case Study

We present case studies on hallucination mitigation using the COCO 2014 validation dataset, as shown in Figure 6. The table compares captions gener-

| Order | Method | InstructBLIP | | mPLUG-Owl2 | | LLAVA-1.5 | |
|---|---|---|---|---|---|---|---|
| | | C | D | C | D | C | D |
| Original Order | HALC | 6.05 | **6.28** | **6.03** | **6.18** | 5.21 | 4.89 |
| | **Ours** | **6.14** | 6.07 | 5.97 | 6.13 | **6.13** | **6.41** |
| | Difference | +0.09 | -0.21 | -0.06 | -0.05 | +0.92 | +1.52 |
| Reverse Order | **Ours** | **6.39** | 6.14 | **6.5** | 6.19 | **6.42** | **6.38** |
| | HALC | 5.85 | **6.40** | 5.99 | **6.39** | 4.98 | 4.93 |
| | Difference | +0.54 | -0.26 | +0.51 | -0.20 | +1.44 | +1.45 |

Table 13: Experimental results of comparing between our decoding method and HALC decoding method on GPT4V-assist benchmark. The The "original order" and "reverse order" correspond to the same content as described in Table 11.

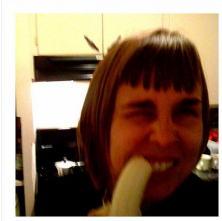
| Image | Greedy | HALC | CGD | Ours |
|---|---|---|---|---|
|  | The image features a delicious breakfast plate filled with various fruits and waffles. The plate is placed on a dining table, and it contains a generous portion of blueberries, apples, and bananas. The fruits are spread across the plate, with some blueberries on the left. | The image features a delicious breakfast plate filled with various fruits and waffles. The plate is placed on a dining tablecloth, and it is adorned with a blue floral pattern. The breakfast consists of two waffles, each topped with blueberries and apples. | The image shows a plate filled with delicious food consisting of pancakes topped with sliced apples, blueberries, and syrup. The blueberries are spread out across the plate, with some closer to the center and others closer to the edges. This combination of flavors is not only visually appealing but also looks like a tasty and healthy breakfast option. | The image displays a plate of delicious food, containing waffles, fruit, and syrup on a dining table. The waffles are garnished with fresh blueberries, adding a touch of color and flavor to the dish. There are multiple blueberries on the plate, distributed evenly throughout the waffles. The presentation of this scrumptious breakfast is quite appetizing and inviting. |
|  | The image features a woman with short hair, wearing a black shirt, and holding a banana in her mouth. She appears to be making a funny face while eating the banana. The woman is standing in a kitchen, with a refrigerator visible in the background. | The image features a woman with short hair, wearing a black shirt, and holding a banana in her mouth. She appears to be making a funny face while eating the banana. The woman is standing in a kitchen, with a refrigerator visible in the background. | In the image, a woman is sitting in a kitchen, chewing on a banana. She is leaning back with the banana in her mouth, and her eyes are closed. In the background, there is an oven and a microwave visible, which are likely part of the kitchen appliances. Additionally, a knife can be seen placed nearby on the counter. | She is wearing a black shirt and is holding a half-eaten banana in her mouth. The woman is looking at the camera, showcasing her unique style and humor. In the background, there are a few kitchen items visible. An oven is placed near the top left corn is placed near the top left corner of the scene, and a cup is located on the right side. |
|  | The image depicts a small, cluttered kitchen with a wooden dining table surrounded by chairs. The table is covered with papers, and there are several chairs placed around it. A refrigerator is located on the left side of the room, and a microwave can be seen above. | The image depicts a small, cluttered kitchen with a wooden dining table surrounded by chairs. The table is covered with papers, and there are several chairs placed around it. A person is standing near the table, possibly working on the papers. | The image depicts a large residential garage turned into a multipurpose room, featuring various appliances and furniture. There is a refrigerator on the left side of the room, surrounded by numerous chairs and dining tables. Some of the chairs are closer to the refrigerator, while others are situated around the dining tables. | The image depicts a small home office space, likely an unfinished garage, set up for various activities. It is equipped with a refrigerator situated on the left side of the room and a white table in the center. The table is surrounded by four wooden chairs, with two chairs positioned on the left side, one on the right side, and one chair at the head of the table. A variety of objects and supplies can be found throughout the space. |
|  | The image features a woman standing in front of a video game display, possibly at a convention or a store. She is wearing a white shirt and appears to be observing the game. The display consists of a large TV screen and a Wii console, with a remote control placed nearby. A chair is positioned in the background, and a handbag can be seen placed on the floor. | The image features a woman standing in front of a video game display, possibly at a convention or a store. She is wearing a white shirt and appears to be observing the game. The display consists of a large TV screen and a Wii console, with a man's image on the screen. There are also a few chairs in the area, with one located near the center of the scene and another towards the right side. A chair is also present in the background. | The scene features a person standing in front of a video game display, which includes a Nintendo Wii gaming console with a TV screen attached. The display is set up in a booth-like area to attract visitors, and there is a person positioned in the background of the display, potentially working behind the counter. A chair is positioned in the background, and a handbag can be seen placed on the floor. | The image displays a busy event featuring a large screen in the center, which appears to be a Nintendo Wii game. Numerous individuals can be seen playing games at the event, with some standing around and enjoying the experience. The main display features a black and white image of a man playing with a Nintendo Wii, likely on a television screen or a large monitor. A row of figures, representing the Wii players, are also present, likely set up on the front of the screen for an interactive element at the event. |

Figure 6: A comparison of text generated by Greedy Search, HALC, CGD, and our proposed method, using examples from the COCO 2014 validation dataset with LLaVA-1.5. The hallucinated parts are highlighted in red.

ated by Greedy, HALC, CGD, and our proposed approach for the images in the leftmost column. Notably, our method generates longer and more detailed captions. Hallucinated content in the descriptions is highlighted in red.