

ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models

Yeji Park[†], Deokyeong Lee[†], Junsuk Choe^{*}, Buru Chang^{*}

Sogang University

{yjparkm, plmft, jschoe, buru}@sogang.ac.kr

Abstract

Hallucinations in Multimodal Large Language Models (MLLMs) where generated responses fail to accurately reflect the given image pose a significant challenge to their reliability. To address this, we introduce ConVis, a novel training-free contrastive decoding method. ConVis leverages a text-to-image (T2I) generation model to semantically reconstruct the given image from hallucinated captions. By comparing the contrasting probability distributions produced by the original and reconstructed images, ConVis enables MLLMs to capture visual contrastive signals that penalize hallucination generation. Notably, this method operates purely within the decoding process, eliminating the need for additional data or model updates. Our extensive experiments on five popular benchmarks demonstrate that ConVis effectively reduces hallucinations across various MLLMs, highlighting its potential to enhance model reliability. Source code is available at <https://github.com/yejipark-m/ConVis>

Introduction

Multimodal Large Language Models (MLLMs) (Dai et al. 2023; Liu et al. 2024b) are advanced language models capable of understanding both images and text, such as image captioning and visual question answering (VQA). While MLLMs have achieved significant success that utilize both visual and textual information, the issue of *hallucination*, where the models generate responses that do not align with the given image, has greatly undermined their reliability (Liu et al. 2023a; Sun et al. 2024). This problem poses a significant obstacle to adopting MLLMs in critical fields where reliability is crucial. For instance, in medical applications, it could lead to incorrect diagnoses (Liu et al. 2023b), while in MLLM-based autonomous systems, it might result in erroneous interpretations (Shao et al. 2024).

Recent research has been actively conducted to address this. WoodPecker (Yin et al. 2023) and LURE (Zhou et al. 2024) reduce hallucinations by post-processing the generated responses. Datasets such as LRV-Instruction (Liu et al. 2023a) and RLHF-V (Yu et al. 2024) have been proposed to mitigate hallucinations through instruction tuning of MLLMs. However, these studies often rely on external

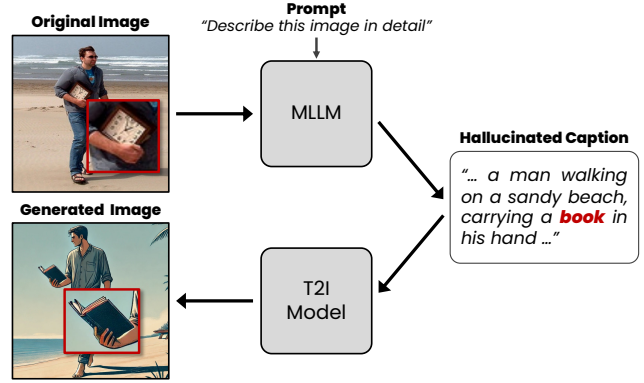


Figure 1: The text-to-image model visualizes hallucinations (e.g., ‘book’) in the semantically reconstructed images based on the hallucinated caption, exhibiting differences (e.g., missing ‘clock’) from the original image.

APIs like GPT-3.5, require costly human feedback collection, and necessitate additional training of MLLMs.

In contrast, this paper focuses on decoding strategies that reduce hallucinations by intervening solely in the decoding process, without the need for additional data or model training. The following studies fall into this category: OPERA (Huang et al. 2024) imposes penalties on token generation that does not reference visual tokens. VCD (Leng et al. 2024) creates contrasting distributions using distorted images to reduce the model’s reliance on statistical biases and priors that lead to hallucinations. HALC (Chen et al. 2024) corrects hallucinations by leveraging cues provided by visual information from various fields of view.

In this study, we propose a contrastive decoding method called **ConVis** (Contrastive Decoding with Hallucination Visualization), which can be applied to any existing MLLM without additional training. Inspired by the previous work (Kim et al. 2024), ConVis leverages text-to-image (T2I) generation models, specifically Hyper-SDXL (Ren et al. 2024), to capture visual contrast signals. The process begins with the MLLM generating a caption for the input image, after which the T2I model reconstructs an image based on this caption. As shown in Figure 1, if the generated caption contains hallucinations (e.g., a *book*), there will be vi-

[†]These authors contributed equally.

^{*}Corresponding authors.

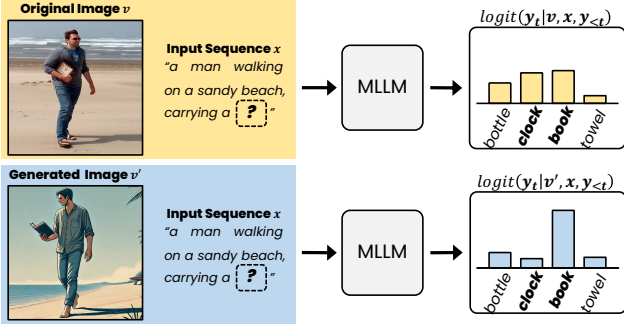


Figure 2: The original and reconstructed image generate the contrastive logit distribution for the hallucinated tokens (e.g., ‘book’). The reconstructed image tends to amplify the logits of tokens corresponding to the visualized hallucination.

sual discrepancies between the original and reconstructed images (e.g., a missing *clock*). ConVis then uses the original and reconstructed images to compare the probability distributions (Figure 2), capturing visual contrast signals that highlight hallucinations. Based on these signals, ConVis penalizes the generation of hallucinations during the decoding process, reducing the hallucinations.

To validate the effectiveness of ConVis, we conducted experiments across five benchmarks: CHAIR (Rohrbach et al. 2018), HallusionBench (Guan et al. 2024), POPE (Li et al. 2023c), MME (Fu et al. 2023) and LLaVA-Bench (Liu et al. 2024b). The results consistently demonstrated that our decoding method reduces hallucinations while maintaining overall response generation performance across various MLLMs, including LLaVA-1.5 (Liu et al. 2024a), MiniGPT-4 (Zhu et al. 2024), and mPLUG-Owl2 (Ye et al. 2024).

Our contributions can be summarized as follows: (1) Propose ConVis, a novel contrastive decoding method that visualizes hallucinations using a T2I model. To the best of our knowledge, this is the first time a T2I model has been employed to mitigate hallucinations through a decoding strategy. (2) Conduct extensive experiments to validate the effectiveness of ConVis in reducing hallucinations. (3) Provide insights into how T2I models can serve as a valuable source of visual contrastive signals in decoding methods aimed at mitigating hallucinations.

Related Work

Multimodal Large Language Models

The emergence of LLMs has revolutionized the paradigm of Natural Language Processing (NLP). The significant success of LLMs in the NLP field has led to research on leveraging LLMs in the visual domain. Consequently, MLLMs that can simultaneously handle visual and textual data have recently been proposed. Specifically, to process visual information, LLaVA (Liu et al. 2024b) uses a CLIP vision encoder (Radford et al. 2021) and a linear layer to project images into the LLM’s input embedding space. MiniGPT-4 (Zhu et al. 2024) employs a Q-Former (Li et al. 2023a) and a linear layer to project images into the LLM’s input

embedding space. Additionally, mPLUG-Owl2 (Lai et al. 2024) introduces a modality-adaptive module that preserves modality-specific features, allowing the model to excel in both multimodal and NLP tasks.

However, despite these efforts, misalignment between modalities can still occur for various reasons, leading to generated responses that do not correspond to the visual information. This phenomenon, known as hallucination, undermines the reliability of MLLMs and poses a significant challenge to their application in real-world scenarios.

Hallucination Mitigation

To address the hallucination problem in MLLMs, several studies have been proposed recently. Lure (Zhou et al. 2024) and Woodpecker (Yin et al. 2023) employ post-processing methods to revise generated responses, either by training a revisor or using GPT-3.5-turbo (Brown et al. 2020). Fine-tuning approaches (Liu et al. 2023a; Yu et al. 2024) mitigate hallucinations through instruction tuning with additional data, but they require significant data collection and training resources. Given the large number of parameters in MLLMs, this is computationally inefficient.

Therefore, methods for improving the decoding process have recently received great attention due to the advantage that they do not require additional training. Specifically, OPERA (Huang et al. 2024) explores aggregation patterns that cause hallucinations. OPERA utilizes this insight to suppress the generation of tokens that exhibit these patterns. VCD (Leng et al. 2024) leverages the characteristic that the model tends to prioritize prior knowledge over visual information when responding to distorted images. As a result, the responses to the distorted image and the original image show significant differences in hallucinated tokens, and VCD contrasts these to mitigate the hallucinations. HALC (Chen et al. 2024) observes that when images with varying fields of view are input into the MLLM, the probability changes for ground truth tokens are much greater than for hallucinated tokens. This observation helps identify visual context candidates that clearly depict objects, and by contrasting these candidates, HALC reduces hallucinations.

Unlike existing techniques, we propose a new decoding method that utilizes a T2I model. Specifically, our approach visualizes hallucinations in the initially generated caption using a T2I model, then contrasts the responses generated from the reconstructed image with those from the original image. Through this process, we contrast distributions of the hallucinated tokens and effectively mitigate hallucinations.

Methodology

Preliminaries

Response Generation. The MLLM generates a response y corresponding to a given input image v and instruction text x . The input image is projected into visual tokens through an image encoder, and these tokens, along with the tokens corresponding to the instruction text, are fed into the LLM. The response is generated through autoregressive decoding according to the following equation:

$$y_t \sim p_\theta(\cdot | v, x, y_{<t}) \propto \exp(f_\theta(\cdot | v, x, y_{<t})), \quad (1)$$

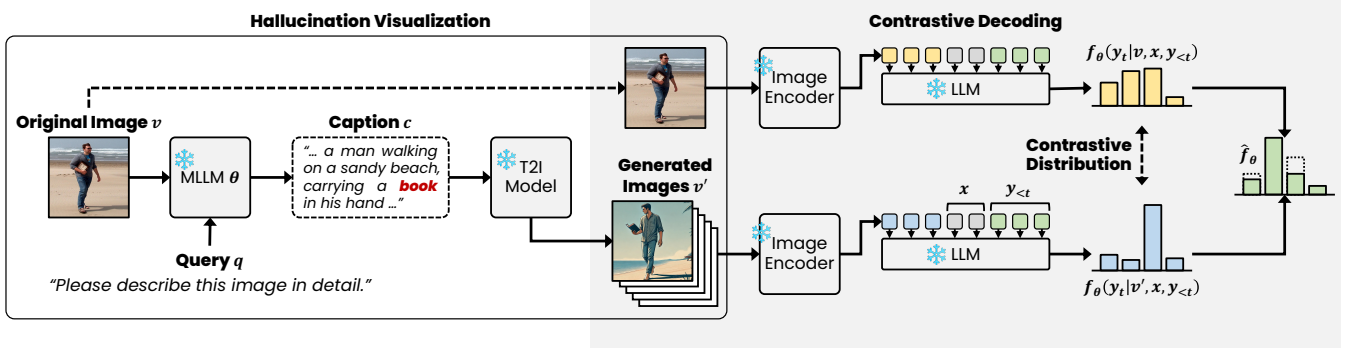


Figure 3: The original and generated image produce the contrastive distribution for the hallucinated tokens (e.g., ‘book’). The generated image tends to amplify the logits of tokens corresponding to the visualized hallucination.

where θ denotes the parameters of the MLLM, y_t represents the t -th token of response, and $y_{<t}$ is the sequence of tokens generated up to time t . f_{θ} denotes the logit distribution generated by the MLLM. Hallucination refers to the phenomenon where the output y generated by the MLLM does not correspond to the input image v . This study focuses on mitigating hallucinations while maintaining the overall performance of the MLLM as a language model.

Text-to-Image Generation. The core component of ConVis is the T2I model that generates images based on a given query. The goal of the T2I model is to create an image that accurately depicts the query. Among the recently proposed T2I models, we utilize Hyper-SDXL (Ren et al. 2024), an enhanced version of Stable Diffusion (Ho, Jain, and Abbeel 2020), which has demonstrated excellent T2I performance. The diffusion-based Hyper-SDXL model begins with a pure noise and progressively reconstructs it through an iterative reverse diffusion process which ultimately results in the generated image v'_0 .

Hallucination Visualization

We hypothesize that the T2I model can help mitigate hallucinations by providing visual contrast signals during the decoding process. If the T2I model receives a caption generated by the MLLM that contains hallucinations, it will faithfully visualize those hallucinations in the generated image. We refer to this process as *hallucination visualization*.

To implement this, ConVis first generates an initial caption c for the original image v using a simple instruction text that directs the MLLM to describe the image. This process is illustrated in Figure 2. The T2I model then takes the caption c as a query and generates an image v' based on it. If the caption contains hallucinations, these will be faithfully visualized in the generated image v' . Conversely, if the initial caption is accurate and free of hallucinations, the generated image will be semantically similar to the original image.

Diversity of Generated Images. Given that the current T2I model may not generate images that fully align with the captions, we address this limitation by increasing the diversity of the generated images using the following approaches: (1) We first generate a diverse set of n captions using Nucleus Decoding (Holtzman et al. 2020) instead of Greedy Decod-

ing. (2) Then, the T2I model uses these n captions to generate n corresponding images. This approach increases coverage of the various potential hallucinations that the MLLM might generate by diversifying the captions. Additionally, by using multiple images instead of a single one, we enhance the robustness of our method against the T2I model’s potential misalignment between the caption and the generated image due to its imperfect performance.

We have found these approaches to be effective, with detailed results available in the experiment section.

Contrastive Decoding

Hallucinations in captions cause visual differences between the original image v and the generated image v' . We mitigate these hallucinations by capturing the visual contrast signals from these differences. To achieve this, during the decoding process, we utilize both the original image v and the n generated images to produce the logit distribution for each image. The final contrastive logit distribution \hat{f}_{θ} is derived by averaging the contrastive logit distributions between the original image and each generated image as follows:

$$\hat{f}_{\theta} = \frac{1}{n} \sum_{i=1}^n \left((1 + \alpha) f_{\theta}(\cdot | v, x, y_{<t}) - \alpha f_{\theta}(\cdot | v'_i, x, y_{<t}) \right), \quad (2)$$

where α is a hyperparameter that controls the strength of the difference between the logit distributions from the original and generated images. The contrastive logit distribution \hat{f}_{θ} is used to generate the response y . For tokens associated with hallucinations, the contrastive logit distribution is significantly amplified compared to other tokens, allowing us to penalize these tokens and reduce the hallucinations.

Note that, Equation 2 is similar to the contrastive decoding methods used in VCD (Leng et al. 2024) and HALC (Chen et al. 2024). However, our method is distinguished from existing approaches by directly capturing visual contrastive signals from the hallucinations visualized by the T2I generative model.

Method	LLaVA-1.5		mPLUG-Owl2		MiniGPT-4	
	CHAIR _S ↓	CHAIR _I ↓	CHAIR _S ↓	CHAIR _I ↓	CHAIR _S ↓	CHAIR _I ↓
Greedy Search	22.4 ± 1.11	7.4 ± 0.27	22.2 ± 1.10	7.3 ± 0.24	34.0 ± 1.11	13.8 ± 0.85
Nucleus Sampling	26.0 ± 1.93	9.5 ± 0.76	25.2 ± 1.59	9.3 ± 0.34	30.1 ± 1.45	14.2 ± 0.90
Beam Search	19.5 ± 1.42	6.4 ± 0.09	18.3 ± 0.42	6.0 ± 0.34	31.1 ± 1.03	12.4 ± 0.59
VCD	23.7 ± 1.90	8.2 ± 0.80	25.7 ± 1.30	9.0 ± 0.28	31.6 ± 1.83	13.8 ± 0.83
OPERA	<u>18.5 ± 0.90</u>	6.6 ± 0.23	<u>18.2 ± 0.40</u>	6.2 ± 0.18	30.6 ± 1.06	12.5 ± 0.91
HALC	23.7 ± 2.66	9.1 ± 0.41	24.3 ± 1.22	9.4 ± 0.19	<u>24.2 ± 1.91</u>	<u>10.8 ± 0.53</u>
Ours	18.4 ± 0.53	<u>6.4 ± 0.37</u>	17.6 ± 3.54	<u>6.0 ± 0.89</u>	23.5 ± 0.31	10.0 ± 0.69

Table 1: Evaluation results on the CHAIR benchmark using the MSCOCO dataset (val2014 split). We conduct experiments with three different sets of 500 images, each selected by random seeds. The reported value is the mean of the results from the three different seeds, with the \pm symbol representing the standard deviation.

Experiments

Benchmarks. To evaluate the performance of our method, we conduct experiments on three benchmarks to evaluate the mitigation of hallucinations and two general-purpose benchmarks to assess the general performance of the MLLM:

- Hallucination: *CHAIR* (Rohrbach et al. 2018), *HallusionBench* (Guan et al. 2024), and *Polling-based Object Probing Evaluation (POPE)* (Li et al. 2023c)
- General-purpose: *MLLM Evaluation (MME)* (Fu et al. 2023) and *LLaVA-Bench* (Liu et al. 2024b)

Detailed information on these benchmarks can be found in the Appendix.

Backbones. To evaluate our method, we utilize three well-known MLLMs with publicly available checkpoint weights: LLaVA-1.5 (Liu et al. 2024a), mPLUG-Owl2 (Ye et al. 2024), and MiniGPT-4 (Zhu et al. 2024).

Compared Methods. Our method is designed to replace existing decoding methods used in the LLM component, and therefore, we compare it against baselines such as Greedy Search, Nucleus Sampling (Holtzman et al. 2020), and Beam Search (beam=5). We also evaluate our method’s effectiveness against other decoding methods in hallucination mitigation, including OPERA (Huang et al. 2024), VCD (Leng et al. 2024), and HALC (Chen et al. 2024). We use the same hyperparameters borrowed from the original papers of the compared methods to ensure a fair comparison.

Implementation Details. We utilize the Hyper-SDXL (Ren et al. 2024) T2I model for image generation. Specifically, in all experiments, unless otherwise noted, we use the Step 1 generation results of Hyper-SDXL model. The maximum length of text queries that the T2I model could accept is 77 tokens, which is too short to process the captions generated by MLLM. To address this, we leverage Compel (Stewart 2023), which allows for processing more than 77 tokens. We set the maximum token count for the caption generation to 256 and use Nucleus sampling with a temperature of 0.7 and a top- p of 0.9 to generate the images. The query used in this process is “Please describe this image in detail.” We set the number of generated images, n , to 4, producing four images based on distinct captions generated using different random seeds. For contrastive decoding, we follow (Li et al. 2023b)

using adaptive plausibility constraint to contrast only meaningful tokens. The plausibility constraint hyperparameter λ is set to 0.1. We also set α , which controls the degree of contrastive emphasis, to 1 for captioning-based metrics such as CHAIR and LLaVA-Bench, and to 0.1 for VQA metrics, including POPE, HallusionBench, and MME. To generate responses, we use a greedy decoding approach for all methods. For CHAIR, we sample three different sets of images using different random seeds and assess the performance using the mean and standard deviation of these results.

Experimental Results

Results on CHAIR. We report our evaluation results on the CHAIR (Rohrbach et al. 2018) benchmark in Table 1. Our assessment includes basic decoding strategies—Greedy search, Nucleus sampling, and Beam search—along with three state-of-the-art approaches—VCD (Leng et al. 2024), OPERA (Huang et al. 2024), and HALC (Chen et al. 2024). Our method achieves the best performance on the CHAIR_S metric across all three backbone models (LLaVA-1.5, mPLUG-Owl2, and MiniGPT-4). Remarkably, it significantly improves the CHAIR_S score compared to both the basic decoding strategies and the state-of-the-art methods, highlighting its superior ability to mitigate hallucinations. In terms of the CHAIR_I metric, our method consistently ranks either first or second across all backbone models. These results demonstrate that our method both excels in reducing the total number of hallucinations throughout entire sentences and minimizes the number of hallucinated objects across all evaluated image sets.

Results on HallusionBench. In Table 2, we present the evaluation results for the visual dependent category of the HallusionBench (Guan et al. 2024) benchmark. HallusionBench is evaluated with the assistance of GPT-4V, which incurs significant costs; therefore, we conduct experiments using only the LLaVA-1.5 (Liu et al. 2024a) backbone. Our method demonstrates superior performance in Figure Accuracy (fAcc), outperforming all baseline decoding strategies (Greedy Search, Nucleus Sampling, Beam Search) as well as state-of-the-art techniques (VCD, OPERA, HALC). This indicates that our model effectively interprets the visual details of images when responding to visually dependent questions,

Method	Figure Acc (fAcc)	All Acc (aAcc)
Greedy Search	<u>22.2</u>	50.1
Nucleus Sampling	17.8	46.2
Beam Search	19.1	48.4
VCD	21.7	47.5
OPERA	20.9	49.9
HALC	21.7	<u>50.6</u>
Ours	23.5	50.8

Table 2: Evaluation results on HallusionBench. We report Figure Acc and All Acc using LLaVA-1.5.

Method	LLaVA-1.5	mPLUG-Owl2	MiniGPT-4	Average
VCD	<u>82.8</u>	81.6	59.8	74.7
OPERA	83.0	83.3	66.1	<u>77.4</u>
HALC	50.6	83.4	<u>69.7</u>	67.9
Ours	83.0	83.0	69.9	78.6

Table 3: Evaluation results on the POPE benchmark using the MSCOCO dataset (val2014 split).

indicating its ability to mitigate hallucinations by providing responses that closely align with the given visual content. Furthermore, our method achieves the highest performance on the All Accuracy (aAcc) metric, which measures overall accuracy across all questions within the visual dependent category, demonstrating its effectiveness in handling a wide range of visually dependent queries.

Results on POPE. Table 3 reports the evaluation results on the POPE (Li et al. 2023c) benchmark using the MSCOCO (Lin et al. 2014) dataset (val2014 split). We present the average F1-scores across the three POPE question splits—Random, Popular, and Adversarial—for three different backbone models. Detailed performances on each POPE question split are in the Appendix.

Our method achieve a new SOTA performance on MiniGPT-4, and demonstrate performance comparable to existing techniques on LLaVA-1.5 and mPLUG-Owl2. In terms of average performance across all backbones, our method outperforms previous techniques. This indicates that our approach consistently delivers strong performance across various backbones.

While we achieves overall strong performance on this benchmark, the performance improvements across different backbone models are relatively modest. This might be because the POPE question split does not fully align with the types of hallucinations that T2I models generate. POPE questions, which ask, “Is this [object] in this image?” sample objects randomly, popularly, or adversarially. Meanwhile, our method visualizes hallucinations in captions generated by prompts like “Please describe this image in detail.” As a result, T2I model may visualize the objects unrelated to the actual POPE questions which limits our method’s effectiveness. This limitation will be explored further through a qualitative analysis of POPE samples later in this section.

Results on MME. In Table 4, we present the evaluation re-

Method	Category		Total
	Perception	Cognition	
Greedy Search	1472.5	303.9	<u>1776.4</u>
Nucleus Sampling	1203.4	311.1	1514.5
Beam Search	<u>1478.0</u>	287.5	1765.5
VCD	1326.7	374.6	1701.3
OPERA	1456.9	306.4	1763.3
HALC	887.7	269.6	1157.3
Ours	1487.6	306.1	1793.7

Table 4: Evaluation results on the MME using LLaVA-1.5.

Method	Complex	Conv	Detail	All
Greedy Search	82.0	47.3	64.1	67.0
Nucleus Sampling	76.2	41.2	52.6	59.9
Beam Search	83.9	58.7	58.8	70.0
VCD	79.9	53.5	56.3	66.2
OPERA	78.7	53.0	58.3	66.0
HALC	55.8	31.1	50.4	47.1
Ours	84.2	63.5	64.8	73.3

Table 5: Evaluation results on LLaVA-Bench using LLaVA-1.5.

sults on the MME benchmark using the LLaVA-1.5 backbone. Due to space limitations, we focus on the performance in the two main categories of the MME benchmark: Perception and Cognition. Scores for the subcategories are provided in the Appendix. Our method outperforms all others in the Perception category, demonstrating its effectiveness in accurately interpreting and processing visual information across various tasks. This strong performance indicates that our model is particularly well-suited for visual tasks, making it highly effective for applications that require precise visual understanding. In the Cognition category, our method demonstrates competitive performance, comparable to OPERA and superior to HALC, further underscoring the versatility and robustness of our approach. While VCD excels in cognitive tasks, our method achieves stronger overall performance when both the Perception and Cognition categories are considered together. This suggests that our model provides a more comprehensive and effective solution across diverse tasks. Its balanced and reliable performance in both visual and cognitive challenges makes it an adaptable solution for a wide range of applications.

Results on LLaVA-Bench. Table 5 shows the experimental results on the LLaVA-Bench, which verify whether the language model capabilities are preserved. For this evaluation, we uses the LLaVA-1.5 backbone. Our method outperforms existing techniques across all categories: complex reasoning, conversation, and detailed description. These results demonstrate that our method effectively mitigates hallucinations while also enhancing the performance of the MLLM.

T2I Model	CLIPScore \uparrow	LLaVA-1.5		mPLUG-Owl2		MiniGPT-4	
		CHAIR _S \downarrow	CHAIR _I \downarrow	CHAIR _S \downarrow	CHAIR _I \downarrow	CHAIR _S \downarrow	CHAIR _I \downarrow
Hyper-SD1.5	30.87	20.2	6.6	19.4	6.4	28.2	11.8
SDXL-Turbo	32.33	18.8	6.6	20.2	6.68	25.2	9.9
Hyper-SDXL	32.85	17	5.6	17	5.3	24.4	10.0

Table 6: Our performance when differentiating the T2I models for visualizing hallucinations. We generate captions with nucleus sampling and set max new token for 64 and generate the image with those captions. Inference step for diffusion set to be all 1.

Captioning by	LLaVA-1.5	mPLUG-Owl2	MiniGPT-4
CHAIR _S \downarrow			
Greedy Search	19.4	19.4	27.2
Nucleus Sampling	18.8	15.2	24.4
CHAIR _I \downarrow			
Greedy Search	6.6	6.4	11.6
Nucleus Sampling	6.7	5.1	10.3

Table 7: Comparison of performance using a single image ($n = 1$) generated by two different decoding strategies, Greedy search and Nucleus sampling.

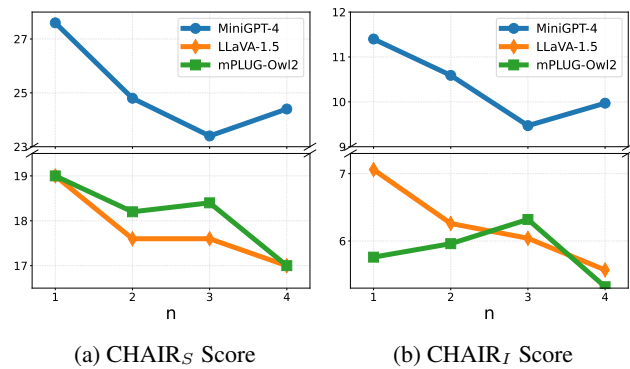


Figure 4: Effect of the number of images with different captions.

Analysis and Discussion

Diversity of Generated Captions and Images. Although T2I models have made significant advancements, they still struggle to generate images that perfectly align with the given captions (Ruiz et al. 2023). To address these limitations, we increase the coverage of hallucination visualization by generating diverse images. Specifically, we use Nucleus sampling, which is known for producing more varied responses than Greedy search, to generate multiple captions. These captions are then utilized to generate images.

To evaluate the effectiveness of this strategy, we analyze how caption diversity impacts hallucination reduction. First, we compare the CHAIR scores of the final responses when using Greedy search and Nucleus sampling during the image generation stage. In this experiment, we limit the number of generated images to one and compare which decoding strategy performs better. As shown in Table 7, Nucleus sampling outperforms Greedy search, demonstrating its po-

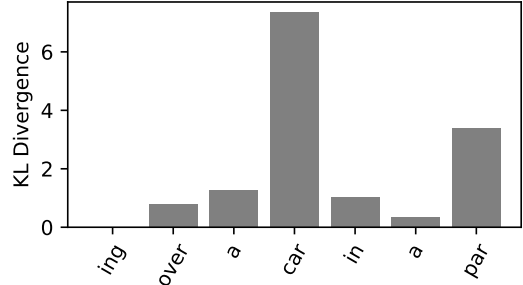


Figure 5: KL divergence between output distributions across each decoding step when the MLLM is provided with the images and caption from Figure 6 (a). The KL divergence is significantly elevated for the hallucinated token “car”.

tential to generate more diverse captions. Furthermore, in Figure 4, we investigate how the number of generated images from different captions using Nucleus sampling affects CHAIR scores. We observe that the number of images n increases, both CHAIR_S and CHAIR_I scores improve, confirming that using multiple reconstructed images, rather than a single image, is more effective for improving performance. These findings validate our design choice of utilizing Nucleus sampling and multiple captions for image generation.

Impacts of Image Generation Quality. To investigate the impact of generated image quality on hallucination mitigation, we evaluate the performance of our method using various text-to-image (T2I) models. Table 6 presents the generation quality (CLIPScore) of the T2I models alongside their corresponding CHAIR scores. We compare three T2I models: Hyper-SD1.5 (Ren et al. 2024), SDXL-Turbo (Sauer et al. 2023), and Hyper-SDXL (Ren et al. 2024), with the inference step fixed at 1.

The results indicate a clear trend: as the CLIPScore improves, so does the CHAIR score. Notably, SDXL-Turbo consistently outperforms Hyper-SD1.5 across all backbones, except for mPLUG-Owl2. Moreover, Hyper-SDXL significantly outperforms Hyper-SD1.5 in all cases. These findings suggest that using higher-quality T2I models, which are better aligned with the original captions, can more effectively mitigate hallucination issues. Consequently, we believe that as more advanced T2I models are developed, the performance of our method will continue to improve.

Qualitative Analysis. Figure 5 shows the KL divergence between output distributions at each decoding step when the

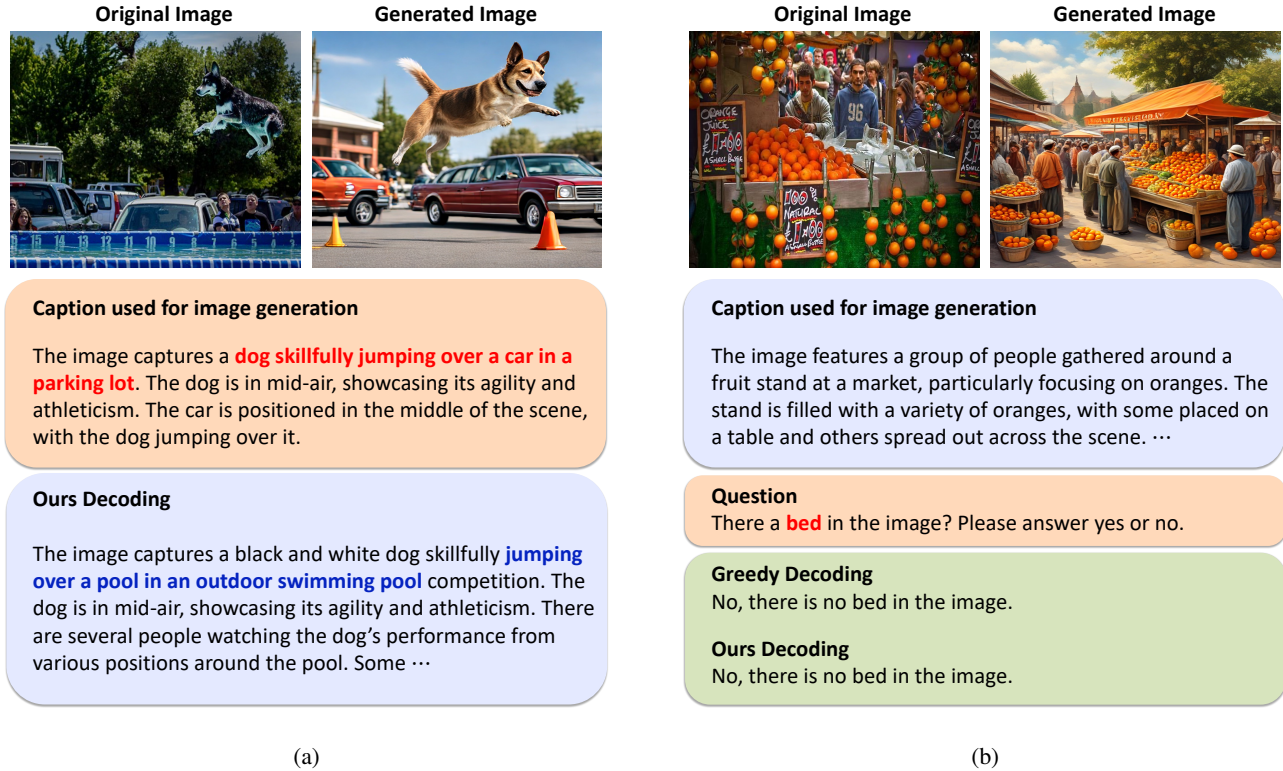


Figure 6: Qualitative samples using LLaVA-1.5 for backbone model. (a) shows an example that T2I model faithfully generate the images that depicts the hallucination in the caption. (e.g., jumping over a car) (b) is an example of our limitation in VQA tasks, which there can be a misalignment between visualized hallucination and actual main subject of question.

images and caption from Figure 6 (a) are provided to the MLLM. We observe that the KL divergence is high for the hallucinated token *car*, while non-hallucinated tokens exhibit lower KL divergence. This indicates that the generated image can produce visual contrastive signals for hallucinated tokens when compared to the original image. This supports our argument that the differences between the original and generated images are primarily influenced by the hallucinated tokens.

To more clearly demonstrate how our method mitigates multimodal hallucinations, we present an example in Figure 6 (a), illustrating the process from the initial hallucinated caption to the generated image, followed by the contrastive decoding result. Specifically, for an image of a dog jumping into a pool, the MLLM incorrectly describes the scene as “a dog jumping over a car in a parking lot.” Using this caption, the T2I model generates a reconstructed image that faithfully visualized the hallucinated content. By contrasting the distributions of the reconstructed and original images during decoding, our method effectively reduces hallucinations.

Limitations. One of the key limitations of our approach is its strong dependence on T2I generation models. This reliance may hinder effectiveness in tasks like VQA, where the generated captions can sometimes contain hallucinations that deviate significantly from the specific question. This limitation is particularly evident in our experiments with the

POPE benchmark, where the performance gain is not as significant as expected. Regarding questions about the presence of specific objects, if the object in question is not related to the hallucinations generated by the caption, visualizing with a T2I model may not sufficiently reflect the information needed for the VQA task. In Figure 6 (b), a question about the presence of a bed in an original image where people are looking at fruits might not be well served by the reconstructed image. This indicates the effectiveness of our method may decrease for certain type of questions.

Currently, our technique employs a fixed prompt for image captioning. However, we believe that adapting the prompt to respond more specifically to the given question could mitigate this issue. We plan to explore this adaptive approach in future work.

Conclusion

In this paper, we presented ConVis, a novel contrastive decoding method designed to mitigate hallucinations in MLLMs. By utilizing a T2I generation model, our approach effectively visualizes hallucinations and contrasts probability distributions between the original and reconstructed images. This process allows for the penalization of hallucinated content during the decoding phase, all without the need for additional data or model retraining.

Our extensive experiments across five benchmarks,

including CHAIR, HallusionBench, and LLaVA-Bench, demonstrated that ConVis consistently reduces hallucinations while preserving the core language model capabilities of MLLMs. The method achieves competitive or superior performance compared to existing techniques in various categories, validating its effectiveness in enhancing the reliability of MLLM outputs.

References

- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33. **2**
- Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In *International conference on machine learning*. **1, 2, 3, 4**
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 37. **1**
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*. **2, 4, 10**
- Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385. **2, 4, 10**
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33. **3**
- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. **3, 4**
- Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427. **1, 2, 4**
- Kim, M.; Kim, M.; Bae, J.; Choi, S.; Kim, S.; and Chang, B. 2024. Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. **1**
- Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589. **2**
- Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. 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*, 13872–13882. **1, 2, 3, 4**
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. **2**
- Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. **4**
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; and Wen, J.-R. 2023c. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. **2, 4, 5, 10**
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. **5, 10**
- Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*. **1, 2**
- Liu, F.; Zhu, T.; Wu, X.; Yang, B.; You, C.; Wang, C.; Lu, L.; Liu, Z.; Zheng, Y.; Sun, X.; et al. 2023b. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1): 226. **1**
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306. **2, 4**
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36. **1, 2, 4, 10**
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. **2**
- Ren, Y.; Xia, X.; Lu, Y.; Zhang, J.; Wu, J.; Xie, P.; Wang, X.; and Xiao, X. 2024. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *arXiv preprint arXiv:2404.13686*. **1, 3, 4, 6**
- Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *The 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. **2, 4, 10**
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510. **6**
- Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*. **6**
- Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130. **1**
- Stewart, D. 2023. Compel. <https://github.com/damian0815/compel>. **4**
- Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*. **1**
- Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal

large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051. [2](#), [4](#)

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*. [1](#), [2](#)

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816. [1](#), [2](#)

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *International Conference on Learning Representations*. [1](#), [2](#)

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning Representations*. [2](#), [4](#)

Appendix

Benchmarks

In this appendix, we provide additional details into the benchmarks referenced in the main paper. To evaluate hallucinations, we employ the following five benchmarks:

CHAIR (Rohrbach et al. 2018) evaluates how well the generated captions align with the content of the given image. CHAIR consists of two versions: CHAIR_S, which measures the inaccuracies at the sentence level, and CHAIR_I, which evaluates at the object level within the sentence by comparing the number of false objects to the total number of objects. For evaluation, we use the val2014 split of the MSCOCO (Lin et al. 2014) dataset, which includes annotations for 80 object categories. We randomly select 500 images from the entire dataset and used the prompt “Please describe this image in detail.” for the MLLM.

HallusionBench (Guan et al. 2024) is a hallucination evaluation benchmark designed to assess whether a model ignores visual context and relies solely on language priors (Language Hallucination) or exhibits the opposite phenomenon (Visual Illusion). The questions in HallusionBench are divided into two main categories, one of which is the Visual Dependent (VD) category. In this category, pairs of similar but different images are presented, and the same question is asked for each pair. The questions are presented in a VQA format with binary ground truth (GT) answers. Accuracy is calculated using GPT-4V by determining whether the model’s responses are similar to, different from, or difficult to compare with the answers generated by GPT-4V. Since this paper focuses on preventing MLLMs from generating hallucinated information based on a given image, we specifically conduct experiments on the Visual Dependent category.

Polling based Object Probing Evaluation (POPE) (Li et al. 2023c) is a VQA-based metric proposed to assess hallucinations in MLLMs. This metric evaluates the MLLM’s response to the prompt “Is [object] is in this image?” To emphasize that this is a binary VQA task, we appended the prompt with “Please answer yes or no.” To select objects referenced in the question prompt, we followed three different sampling options: random, popular, and adversarial. We evaluated performance across all sampling options.

MLLM Evaluation (MME) (Fu et al. 2023) evaluates the capabilities of MLLMs, dividing the evaluation into two major categories: perception and cognition. The perception category includes fine-grained tasks such as existence, count, location, rough color, poster, celebrity, scene, landmark, artwork identification, and OCR. The cognition category includes tasks like commonsense reasoning, numerical calculations, text translation, and code reasoning. All questions in this benchmark are structured to be answered with a simple yes or no.

Using the **LLaVA-Bench** (Liu et al. 2024b), we further demonstrated how well our proposed method maintains the language model performance. This benchmark involves posing various situational questions, such as dialogue, detailed descriptions, and complex reasoning, to randomly selected images from the MSCOCO val2014 dataset. A total of 60

Benchmark	Max New Tokens
CHAIR	64
HallusionBench	64
POPE	16
MME	128
LLaVA-Bench	512

Table A1: Maximum number of generated tokens utilized in the response generation for each benchmark experiment.

questions are used to assess whether the model faithfully follows the instructions. The generated answers are evaluated by comparing them to the responses of a text-only GPT-4 model.

Additional Implementation Details and Experimental Results

We present further implementation details and experimental results that were omitted from the main paper due to space limitations. Table A1 outlines the maximum lengths set for response generation. Additionally, Table A2 provides the complete evaluation results on the POPE benchmark using the MSCOCO dataset, including analyses of Random, Popular, and Adversarial scenarios across three MLLM backbones. Finally, Table A3 offers a full comparison of category performance for the MME benchmark in LLaVA-1.5.

Method	LLaVA-1.5			mPLUG-Owl2			MiniGPT-4			Average
	Random	Popular	Adversarial	Random	Popular	Adversarial	Random	Popular	Adversarial	
Greedy Sample (Nucleus) Beam (n=5)	84.6	83.4	81.3	85.8	83.5	80.3	74.1	68.2	67.1	78.7
	78.7	77.0	76.2	82.5	79.5	76.9	60.5	61.6	57.3	72.24
	<u>85.0</u>	83.7	81.5	85.5	<u>83.5</u>	80.7	71.0	67.6	64.6	78.12
VCD	85.3	82.9	80.1	84.7	81.8	78.4	61.5	59.3	58.7	74.74
OPERA	84.4	<u>83.4</u>	81.2	<u>85.8</u>	83.5	<u>80.5</u>	69.3	65.7	63.2	77.44
HALC	50.8	50.6	50.4	86.0	83.6	<u>80.5</u>	74.3	68.1	<u>66.8</u>	67.90
Ours	84.7	83.2	81.1	85.6	83.1	80.2	74.3	68.3	67.1	<u>78.62</u>

Table A2: Full report of evaluation on the POPE benchmark using the MSCOCO dataset (val2014 split).

Method	Existence	Count	Position	Color	Posters	Celebrity	Scene	Landmark	Artwork	OCR	Common	Numerical	Text Trans	Code Reas
Greedy Sample (Nucleus) Beam (n=5)	195.0	133.3	<u>133.3</u>	155.0	<u>138.0</u>	128.5	<u>153.5</u>	153.2	125.0	132.5	<u>121.4</u>	37.5	82.5	62.5
	<u>180.0</u>	136.6	116.6	138.3	122.7	<u>131.4</u>	146.2	145.2	109.2	100.0	122.1	85.0	92.5	75.0
	195.0	153.3	<u>133.3</u>	155.0	135.7	126.7	152.7	152.5	122.5	<u>130.0</u>	116.4	40.0	<u>87.5</u>	62.5
VCD	195.0	163.3	138.3	155.0	<u>138.0</u>	128.5	152.7	<u>154.0</u>	<u>123.0</u>	<u>130.0</u>	115.0	45.0	65.0	62.5
OPERA	165.0	101.6	98.3	<u>153.3</u>	116.6	107.9	135.2	125.7	109.5	90.0	108.5	<u>67.5</u>	67.5	<u>67.5</u>
HALC	110.0	78.3	90.0	100.0	60.2	69.4	109.5	98.5	101.7	70.0	92.1	50.0	77.5	50.0
Ours	195.0	<u>158.3</u>	<u>133.3</u>	155.0	143.2	139.7	153.8	155.3	121.5	132.5	118.6	45.0	<u>87.5</u>	55

Table A3: Evaluation results on the MME benchmark using LLaVA-1.5 for MLLM backbone, conducted across 10 subcategories focused on perception and 4 subcategories focused on cognition.