Building More Accountable Multi-Modal LLMs Through Spatially-Informed Visual Reasoning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Recent research has demonstrated that debate mechanisms among Large Language Models (LLMs) show remarkable potential for enhancing reasoning capabilities and promoting responsible text generation. However, it remains an open question whether debate strategies can effectively generalize to Multi-Modal Large Language Models (MLLMs). In this paper, we address this challenge by proposing a location-aware debate framework specifically designed for MLLMs to mitigate hallucination without requiring additional external knowledge. Our approach introduces an asymmetric debate structure across both textual and visual modalities. For textual processing, one MLLM instance generates a comprehensive image description while identifying object locations, while a second instance "zooms in" on specific regions of interest to evaluate and refine the initial descriptions. For visual processing, we introduce a novel hybrid attention module that fuses visual self-attention with cross-modal attention between textual and visual information, effectively highlighting critical content regions. The framework incorporates a judge component that evaluates the complete debate process and selects the most reliable output between the two debating instances. Our experimental results demonstrate that this approach substantially reduces hallucination across diverse MLLMs and evaluation metrics. Moreover, the framework serves as a readily integrable complement to existing hallucination mitigation methods. By employing consistent procedures and standardized prompts across all investigated tasks, our framework proves both effective and highly adaptable, enabling direct application to a broad range of black-box MLLMs without architectural modifications.

1 Introduction

Recent advancements in multi-modal large language models (MLLMs) have demonstrated significant progress, achieving outstanding performance across various vision-language tasks Bai et al. (2023); Alayrac et al. (2022); Li et al. (2023); Zhu et al. (2023); Liu et al. (2024b,a); Peng et al. (2023b); Team et al. (2023); Dai et al. (2023). With the ability to process both image and text inputs, these general-purpose foundation models are versatile and can be adapted to a wide range of tasks, including image generation Black et al. (2023), biomedical applications Li et al. (2024), text-to-video generation Cai et al. (2023), and reasoning Lai et al. (2024).

While the remarkable performance and versatility of MLLMs are highly favorable, they are plagued by a well-known issue called “hallucination." Specifically, MLLMs often generate incorrect responses regarding the existence of objects, their color, quantity, orientation, and spatial relationships. Moreover, some of their responses are entirely irrelevant to the input images. These flaws pose significant challenges to the development of responsible and robust multi-modal intelligence agents, particularly

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

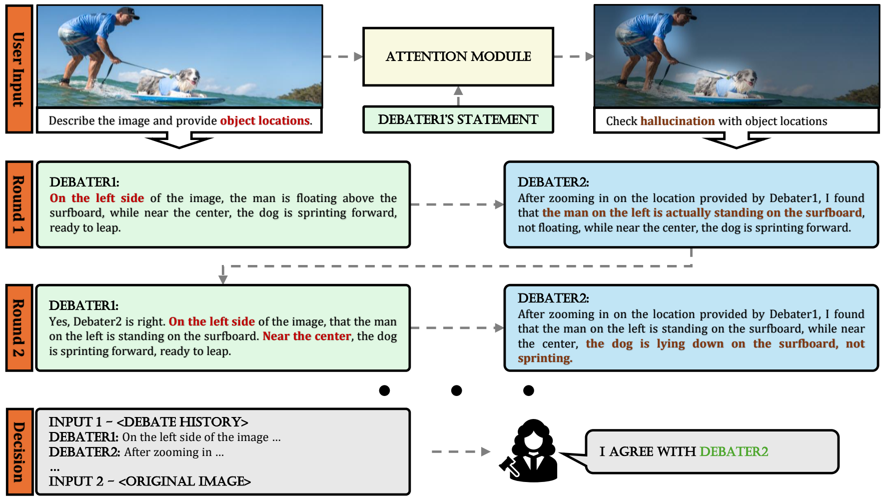

Figure 1: The Overall Debating Pipeline of the Proposed Location-Aware Framework. Debater 1 is tasked with generating general descriptions and identifying object locations within the image, while Debater 2 focuses on providing detailed descriptions of specific regions of interest, guided by both textual input and the hybrid attention module. The judge evaluates and selects between the debaters' statements rather than modifying the final description.

in critical domains such as healthcare Li et al. (2024), autonomous driving Wei et al. (2024), and military applications Rivera et al. (2024).

To address the challenge of "hallucination," various approaches have been proposed, including instruction tuning Liu et al. (2023), over-trust penalty Huang et al. (2024), instruction correlation Wang et al. (2024), the replacement of uncertain objects Zhou et al. (2023), and multi-agent debate Lin et al. (2024); Khan et al. (2024); Du et al. (2023). While all these methods have demonstrated effectiveness, the multi-agent debate strategy is particularly appealing, as it does not rely on costly external knowledge, such as additional instruction data for training, and offers an intuitively designed solution Liu et al. (2023).

Building on this idea, debate mechanisms have been explored in LLM to enhance reasoning and factual accuracy in text generation Khan et al. (2024); Du et al. (2023); Liang et al. (2023). A similar framework was then directly extended to the domain of MLLMs Lin et al. (2024).

In this paper, we argue that the debate framework should differ from the general LLM framework due to the presence of multi-modal inputs, i.e., text and images. The spatial information from images is often underutilized, leading to suboptimal results in the debating framework. To address this, we propose a simple yet highly effective asymmetric debate framework for MLLMs. Specifically, one MLLM is tasked with describing objects along with their corresponding spatial locations in the image, as illustrated in Figure 1. Another MLLM instance then reviews and critiques the responses from the first debater. Importantly, we emphasize spatial information in both modalities. While the textual description allows the second MLLM instance to infer object locations, we further enhance spatial awareness by utilizing a hybrid attention module that dims unrelated areas while highlighting the described regions in the image. This design enables the second debater to focus on key regions of interest through both textual and visual guidance. The process is repeated over multiple rounds. Finally, the debate history, along with the input image, is presented to a judge, who determines the winner and provides the final response to the query.

To comprehensively assess the effectiveness of our proposed framework, we evaluate it from three key perspectives: object-level hallucination, object-existence hallucination, and overall text quality. These aspects are quantified using four evaluation metrics: Caption Hallucination Assessment with Image Relevance (CHAIR), GPT-4-assisted evaluation, and Polling-based Object Probing Evaluation (POPE). Through extensive experiments on benchmarks and hallucination metrics, we conclude the findings and contributions as follows:

1. The proposed location-aware debate fosters more responsible responses compared to single-modal debate. Previous debate frameworks often overlook the spatial information inherent in objects within images. As a result, debaters tend to distribute their attention uniformly across regions of interest rather than focusing on the most relevant areas, as guided by the input text and image. To address this, we first introduce a debater specifically tasked with clarifying the locations of recognized objects. This simple yet effective design significantly enhances response accuracy, reducing CHAIR scores by an average of **7.02%**. Building on this, we further integrate spatial information into the image using an attention module, which mitigates hallucinations even further, reducing CHAIR scores by an average of **9.52%**.

2. The proposed location-aware debates between MLLMs help generate more responsible content and are widely adaptable. We conduct experiments across various decoding methods, including greedy decoding, nucleus sampling Holtzman et al. (2019), beam search decoding Sutskever (2014), DoLa Chuang et al. (2023), and OPERA Huang et al. (2024), as well as different types of MLLMs, including InstructBLIP Dai et al. (2023), MiniGPT-4 Zhu et al. (2023), LLaVA-1.5 Li et al. (2024), and Shikra Chen et al. (2023). While some of these baselines are specifically designed to reduce hallucinations, the location-aware debate framework continuously enhances their effectiveness as a general, readily integrable approach. Notably, we observe a **2%** to **35.56%** reduction in hallucination rates, consistently reflected in the CHAIR metric.

3. The judge should choose the right statement among debaters rather than providing a summary. Providing the debating history and input data to the LLM judge allows for a more comprehensive response to queries. However, this process may inevitably introduce new hallucinations if the judge is tasked with summarizing and refining the debaters' statements. Therefore, we instruct the judge to select the most accurate statement rather than synthesizing or reinterpreting the debaters' responses. This design consistently reduces hallucinations across various evaluation metrics.

2 Related Work

2.1 Hallucination in Large Foundation Models

Recent advancements in computational resources have significantly accelerated research on large-scale foundational models. MLLMs, such as LLaVA Liu et al. (2024b), Vicuna Chiang et al. (2023), Shikra Chen et al. (2023), MiniGPT-4 Zhu et al. (2023), and others Bai et al. (2023); Dai et al. (2023); Li et al. (2022, 2023), enhance content understanding and generation by leveraging information from multiple modalities. However, these models can sometimes generate text that is inaccurate or fails to address the given query Zhang et al. (2023a). Such limitations arise from various factors, including overfitting, training data biases, and insufficient response validation mechanisms. To address these challenges, previous research has explored various approaches, including data augmentation Lee et al. (2022), fine-tuning techniques Ouyang et al. (2022); Lee et al. (2023), debating Khan et al. (2024) and self-refinement strategies Manakul et al. (2023); Peng et al. (2023a). Extending to multi-modal foundation models, some efforts have been dedicated to instruction tuning Liu et al. (2023) and statistical analysis-based error correction Zhou et al. (2023). More recently, researchers have introduced a nearly cost-free approach that mitigates hallucinations by penalizing over-confident tokens Huang et al. (2024).

2.2 Debate Strategies

While numerous approaches have been proposed to reduce hallucinations using a single LLM agent Wei et al. (2022,?); Yao et al. (2024); Shinn et al. (2024), there is a growing trend of leveraging multiple agents working collaboratively to enhance generation quality through post-training refinement. The initial efforts focused on communicative agents for thought exploration, which later evolved into the concept of multi-agent debate, designed to mimic human-like discourse to improve factual accuracy and reasoning Du et al. (2023). Building on this foundation, the framework was extended to interactive debates, incorporating LLM judges to facilitate the selection of more truthful responses Khan et al. (2024). In parallel, the multi-agent debate (MAD) framework introduced divergent chain-of-thought exploration, demonstrating promising results in translation tasks Liang et al. (2023). Ultimately, this debate paradigm was further extended to the multi-modal LLM (MLLM) domain Lin et al. (2024).

2.3 Region-Level Image Attention

In vision-related research, identifying key regions for fine-grained analysis is a widely adopted strategy. This approach plays a crucial role in object detection Ren et al. (2016); Redmon (2016);

Zang et al. (2024), where it helps localize target objects. Beyond detection, large foundation models have applied similar techniques to open-vocabulary object recognition Kamath et al. (2021); Zhou et al. (2022); Liu et al. (2024c), enabling more flexible and adaptive visual understanding. Region-level attention has also been leveraged in related tasks such as image captioning Yang et al. (2017); Wu et al. (2024) and graph generation Tang et al. (2019); Yang et al. (2022), demonstrating its versatility in structured representation learning. More recently, this concept has been incorporated into instruction tuning to enhance model performance across a broader range of applications Zhang et al. (2023b). Building on these advancements, we extend region-aware mechanisms to multi-modal LLM debates, promoting hierarchical evaluation and improving the reliability of generated responses.

3 Methodology

In this section, we present our proposed multi-agent debate framework for MLLMs. First, we provide an overview of the framework in Section 3.1. Next, we detail the technical implementation of the Hybrid Attention Module in Section 3.2.

3.1 The Overall Debating Pipeline

We adopt an asymmetric multi-agent debate framework with two MLLMs acting as debaters, as illustrated in Figure 1. Unlike previous debate methods, we assign two tasks to the first debater: (1) answering a standard query and (2) identifying the locations of recognized objects within the image. The input image for Debater1 remains unaltered to ensure that no pre-assigned attention influences its response.

Debater2 receives the same query as Debater1 but is also provided with textual descriptions of object locations, making it location-aware. To further enhance object-level attention, we introduce a hybrid attention module shown in Figure 2, which enables more fine-grained articulation of objects of interest. Specifically, the hybrid attention mechanism consists of a visual self-attention block and a cross-attention block, which operate between the raw input image and Debater1's statement. This module helps highlight critical details for Debater2 and improves the overall debate process. A detailed explanation of the hybrid attention module is provided in Section 3.2.

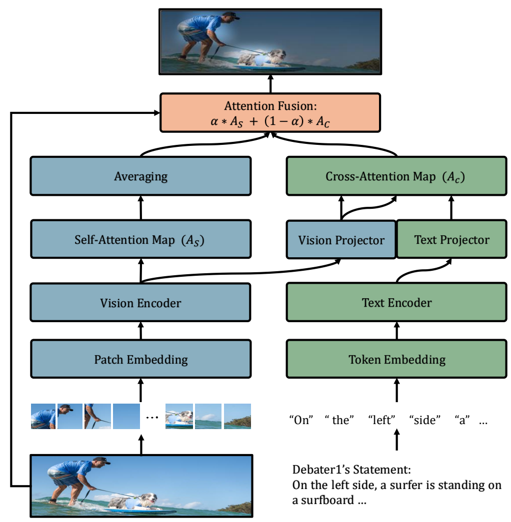

Figure 2: The design of the proposed hybrid attention module. The final attention map is composed of visual self-attention and cross-attention between Debater 1's statements and visual content, ensuring no critical objects are overlooked.

The attention module allows Debater2 to focus on key regions for further inspection and discussion, significantly reducing potential hallucinations. Importantly, the assigned attention from the hybrid attention module is dynamically adjusted based on Debater1's description, meaning that the region of interest may shift accordingly.

Finally, the debating history, along with the raw input image, is fed into the judge, a third MLLM instance. Crucially, instead of summarizing the debaters' statements, the judge is tasked with selecting the most accurate description of the image. Additionally, we provide the judge with the unprocessed image to ensure a fair evaluation and prevent any potential bias introduced by the attention module.

3.2 Hybrid Attention Module

To enable fine-grained visual-linguistic understanding, we propose a hybrid attention mechanism that combines CLIP's Radford et al. (2021) intrinsic self-attention with cross-modal attention and refines the resulting maps via advanced post-processing and adaptive fusion strategies. The overall design of this module is shown in Figure 2. More specifically, given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first employ CLIP's Vision Transformer (ViT-B/32) to partition it into $N = HW/P^2$ non-overlapping patches (with $P = 32$). These patches are encoded into patch embeddings:

$$X = \{x_1, \ldots, x_N\} \in \mathbb{R}^{N \times d},$$

where $d = 768$ is the embedding dimension. N is a perfect square so that the patches can be arranged into a square grid of dimensions $\sqrt{N} \times \sqrt{N}$. Simultaneously, a text description D, is processed by CLIP's text encoder to yield token embeddings, which is formally defined as:

$$T = \{t_1, \ldots, t_M\} \in \mathbb{R}^{M \times d},$$

with M denoting the sequence length.

To capture spatial relationships within the image, we extract self-attention features from the final three layers (of the total $L = 12$ layers) of the Vision Transformer. For each layer l among the last three, the attention matrix $A^{(l)}$, with a shape $[1,\ \text{num\_heads},\ N+1,\ N+1]$) is averaged over heads, and the attention corresponding to the [CLS] token is removed. The resulting map is then reshaped into a $\sqrt{N} \times \sqrt{N}$ grid:

$$A_S^{(l)} = \text{reshape}\Big(\text{mean}\left(A^{(l)}_{:,1:,1:}\right), \sqrt{N}, \sqrt{N}\Big).$$

These layer-specific maps are averaged to produce a preliminary self-attention map A_S, which is subsequently refined via Gaussian smoothing, normalization, and contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE) Reza (2004), followed by percentile-based boosting.

While self-attention captures most of the critical objects within the image, it may not always align precisely with the descriptions provided by Debater1. In some cases, certain objects may be overlooked, even when they are of interest to the MLLM instances and central to the discussion. To address this issue, we incorporate information from Debater 1's descriptions to refine the attention mechanism. Specifically, by leveraging these spatial features, we compute cross-modal attention between visual and textual representations to improve alignment and ensure a more accurate focus. Therefore, in parallel, cross-modal attention is computed by aligning the visual and textual representations. Specifically, we extract the text [CLS] token from the text encoder, denoted as $t_{\text{cls}} = T[:, 0]$, and project it into the common embedding space via CLIP's text projection layer defined as $T_{proj} = Projector_{\text{text}}(t_{\text{cls}})$. For the image, we discard the [CLS] token from the patch embeddings to form $X_{patch} = \{x_2, \ldots, x_N\}$, and project these using the vision projection layer $V_{proj} = Projector_{vision}(X_{patch})$. The cross-attention map is then computed as:

$$A_C = \text{softmax}\big(V_{proj}\, T_{proj}^{\top} / \tau\big),$$

where $\tau = 0.07$, is a temperature parameter that scales the similarity scores. The map A_C is reshaped into a $\sqrt{N} \times \sqrt{N}$ grid and refined with Gaussian smoothing, CLAHE-based contrast enhancement, and percentile boosting.

Finally, the two refined attention maps are fused to yield the final attention map:

$$A = (1 - \alpha)A_S + \alpha A_C, \qquad s.t. \quad \alpha >= 0,$$

where the fusion weight α is set to a fixed value, e.g., $\alpha = 0.3$ for self-attention and $1 - \alpha = 0.7$ for cross-attention. Finally, the integrated attention map is then thresholded at the 70th percentile using the operator $\Phi(\cdot)$ and normalized via a sigmoid activation: $M = \sigma(\Phi(A))$, ensuring that the final mask M robustly highlights the image regions most relevant to the text description.

4 Experiments and Results

In this section, we present the experiments and related results. Specifically, in Section 4.1, we outline the experimental setup, including baseline MLLMs, generation and decoding methods, evaluation metrics, and implementation details. This setup ensures a comprehensive assessment of the proposed framework and provides the necessary information for reproducing the experimental results. In Section 4.2, we report the evaluation results with and without the proposed location-aware debate across multiple metrics, including CHAIR, GPT-4-assisted evaluation, and POPE. We provide an in-depth analysis of hallucination reduction and text quality improvements to demonstrate the effectiveness of the proposed framework. Lastly, in Section A, we conduct ablation studies to highlight the importance of location-aware debate and reveal the influence of critical hyperparameters.

4.1 Experiments Setup

Baseline Models. Following the previous paradigm Huang et al. (2024), we select four representative MLLMs: InstructBLIP Li et al. (2022), MiniGPT-4 Zhu et al. (2023), LLaVA-1.5 Li et al. (2024), and Shikra Chen et al. (2023). These models are chosen to represent different strategies for vision-text alignment, including linear projection layers and Q-Former Li et al. (2023). To ensure consistency throughout the debate process, all debaters use the same 7B-parameter MLLM. Additionally, for the hybrid attention module, we employ CLIP ViT-B/32 Radford et al. (2021) across all experiments.

Baselines Methods. We evaluate the proposed debate framework against various baseline methods, ranging from standard greedy decoding and nucleus sampling Holtzman et al. (2019) to beam search decoding Sutskever (2014), DoLA Chuang et al. (2023), and the more recent OPERA Huang et al. (2024).

To further assess the robustness of our framework, we deliberately include two techniques specifically designed to mitigate hallucination: DoLA and OPERA. Despite their hallucination-reducing mechanisms, we find that the proposed debate framework still provides additional benefits. DoLA refines token selection by contrasting differences in logits between earlier and later transformer layers, leveraging the observation that factual knowledge in LLMs is often localized to specific layers. Building on this, OPERA introduces a penalty term on model logits during beam search decoding to address overconfidence, along with a rollback strategy that detects and re-evaluates summary tokens in previously generated outputs, enabling a more reliable token allocation.

Method	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Greedy	58.8	23.7	31.8	9.9	45.0	14.7	55.8	15.4
Greedy + Debate	53.9	20.4	27.5	8.8	40.2	12.9	51.6	14.2
Nucleus	54.6	23.8	31.8	11.2	46.8	14.0	55.3	15.2
Nucleus + Debate	50.1	21.5	28.3	10.5	43.6	13.4	51.0	14.1
Beam Search	55.8	16.0	31.2	9.5	47.2	13.4	52.4	14.2
Beam Search + Debate	51.0	13.5	27.7	8.9	43.2	12.8	49.0	13.5
DoLa	48.8	15.7	32.2	10.0	47.3	14.5	54.5	14.8
DoLa + Debate	45.6	14.5	29.0	9.9	42.2	13.9	50.2	12.7
OPERA	47.8	14.1	27.0	9.8	46.6	12.8	39.8	12.5
OPERA + Debate	42.5	12.7	17.4	9.6	41.4	11.8	35.1	10.4

Table 1: CHAIR hallucination evaluation results on sentence ($C_S \downarrow$) and image level ($C_I \downarrow$) with and without the proposed debate framework. The max new tokens is set to 512.

Method	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Greedy	30.4	14.8	24.4	8.2	20.6	6.4	22.2	7.1
Greedy + Debate	28.3	13.3	21.7	7.5	18.2	6.0	19.6	6.6
Nucleus	30.4	15.8	23.8	8.6	26.4	8.6	22.5	7.8
Nucleus + Debate	28.2	14.2	21.1	7.8	23.2	7.8	19.6	6.8
Beam Search	21.5	7.2	23.4	7.8	19.0	6.0	21.2	6.6
Beam Search + Debate	19.5	7.0	22.1	7.6	15.8	5.6	18.8	5.8
DoLa	22.5	7.2	24.2	8.2	20.2	6.3	20.6	6.5
DoLa + Debate	20.9	7.0	21.7	8.0	18.6	5.8	18.4	6.0
OPERA	16.8	7.1	22.6	8.4	14.5	5.6	14.2	6.3
OPERA + Debate	15.2	6.5	20.2	7.3	12.6	5.2	12.7	5.8

Table 2: CHAIR hallucination evaluation results on sentence ($C_S \downarrow$) and image level ($C_I \downarrow$) with and without the proposed debate framework. The max new tokens is set to 64.

4.2 Experimental Results

4.2.1 CHAIR Evaluation

The Caption Hallucination Assessment with Image Relevance (CHAIR) Rohrbach et al. (2018) is an evaluation metric designed specifically to assess object hallucination and object-existence-level hallucination in image captioning tasks. Given descriptions of images, CHAIR quantifies the degree of object hallucination with high accuracy. The metric measures the ratio of objects mentioned in the description that are not present in the ground-truth label set. More specifically, CHAIR evaluates hallucination in both textual and visual contexts, distinguishing between sentence-level hallucination, i.e., CHAIR_S (C_S) and image-level hallucination $\text{CHAIR}_I(C_I)$. Formally, these two metrics are defined as follows:

$$\text{CHAIR}_S = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|},$$
$$\text{CHAIR}_I = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$

We use the MSCOCO dataset Lin et al. (2014) for our experiments and CHAIR evaluation. Specifically, we randomly select 1,000 images from the validation set of COCO2014 to compute the average CHAIR value. The prompt consists of a system message that contains a user question and context information. Concretely, the user question follows a standard prompt:

$$\langle \textit{User Question} \rangle = \textit{“Please describe this image in detail”},$$

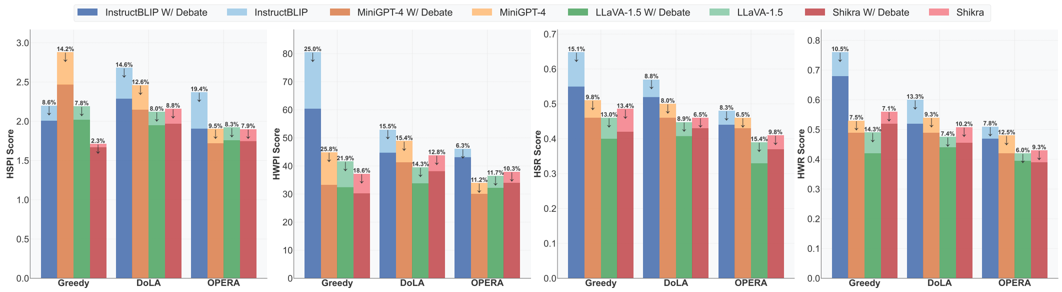

Figure 3: The reduced hallucination ratio from GPT-4-assisted evaluation on the VG-100K dataset. The numbers above the downward arrows of each bar represent relative decrease ratios. We report four aspects of hallucination.

following previous research Huang et al. (2024); Chuang et al. (2023); Sutskever (2014). The system message is designed to ensure that the MLLMs recognize their role as debaters and are tasked with providing object locations. Formally, it is defined as:

"You are participating in a debate about an image."
"Answer ⟨User Question⟩ *"*
"Describe the location of detected objects."
"Here is the previous context: ⟨Context⟩ *"*,

where ⟨Context⟩ represents the statement from the other debater and remains empty if the debate is at the initialization stage.

For the judge's prompt, we only task it with selecting the better statement between the two debaters. Formally, given the final statements from the two debaters, denoted as ⟨Context1⟩ and ⟨Context2⟩, respectively, we apply the following prompt:

"Debater 1's Statement: ⟨Context1⟩ *"*
"Debater 2's Statement: ⟨Context2⟩ *"*
"As a Judge, choose the best statement from two debaters."

To ensure a comprehensive evaluation, we set the *max new tokens* to 64 and 512 for the generation tasks of MLLMs. The results are presented in Table 1 and Table 2. We observe a notable reduction in sentence-level hallucinations, ranging from **6.56% to 35.56%**. For image-level hallucinations, the reduction remains favorable, varying from **2.00% to 16.80%**.

We also report results with a *64 max token*. In this setting, the reduction rate for CHAIR$_S$ ranges from **5.56% to 16.84%**, while for image-level hallucinations, the reduction rate for CHAIR$_I$ varies from **2.44% to 13.10%**. The average reduction ratio is lower compared to the *512 max token* setting. However, this observation is expected, as longer generations are more prone to severe hallucinations Huang et al. (2024).

4.2.2 GPT4-assisted Evaluation

Beyond object and object-existence hallucination, additional evaluation aspects for the debate framework would be beneficial. In particular, attributes, locations, and spatial relationships of objects have not been systematically quantified or assessed. To address this gap, we further evaluate our framework on HalluBench Zhao et al. (2023), one of the most widely used benchmarks for hallucination assessment. For ground-truth references, we use descriptions from the Visual Genome (VG) dataset Krishna et al. (2017). To assess hallucinations in generated descriptions, we rely on GPT-4 for detailed analysis. Specifically, the collected descriptions are fed directly into GPT-4, which is prompted to analyze hallucinations on a sentence-by-sentence basis. For MLLM prompting, we maintain the exact setup used in the CHAIR evaluation, setting the maximum token length to 512

More specifically, we report four aspects of hallucination: the number of hallucinated sentences per image (HSPI), the number of hallucinated words per image (HWPI), the ratio of hallucinated sentences (HSR), and the ratio of hallucinated words (HWR). The three decoding methods, Greedy, DoLA, and OPERA, are presented in detail in Figure 3.

Method	InstructBLIP	MiniGPT-4	LLaVA-1.5	Shikra
Greedy	80.2	58.5	82.2	81.1
Greedy + Debate	83.1	60.2	83.6	83.8
Nucleus	80.2	57.8	82.5	81.2
Nucleus + Debate	83.4	59.5	83.9	83.6
Beam Search	84.4	70.3	84.9	82.5
Beam Search + Debate	85.3	71.8	86.5	84.2
DoLa	83.4	72.8	83.2	82.1
DoLa + Debate	85.1	74.9	84.4	83.9
OPERA	84.8	73.3	85.4	82.7
OPERA + Debate	85.4	75.1	85.8	84.2

Table 3: POPE (↑) hallucination evaluation results on four MLLM models. We report the average F1-score computed on *random*, *popular*, and *adversarial* splits of POPE.

We observe that the debate framework consistently helps MLLMs generate more reliable content across various perspectives and evaluation metrics. Specifically, the average HSPI decreased from 2.27 to 2.06, representing a relative reduction of 9.35% on average difference decoding methods. Similarly, averaged HWPI was significantly reduced from 47.09 to 38.89, corresponding to an average improvement of 17.42%. In terms of sentence-level hallucination, HSR dropped from 0.50 to 0.46, yielding an average reduction of 9.55%, while averaged HWR decreased from 0.54 to 0.50, resulting in an average decrease of 8.97%. These findings highlight the effectiveness of the proposed debate framework in mitigating hallucination across different models and decoding methods. Concrete numerical results and additional findings on Beam Search and Nucleus Sampling are provided in the Supplementary Materials.

4.2.3 POPE Evaluation.

More recently, the POPE evaluation has been introduced to assess MLLMs in terms of object-level hallucination. It has gained widespread adoption in recent research Huang et al. (2024); Lin et al. (2024). To evaluate our framework on this benchmark, we maintain the same prompt design used in the CHAIR evaluation but modify the user question as follows:

$$\langle \textit{User Question} \rangle = \textit{“Please describe this image in detail”},$$

which is a standardized query specifically designed to determine whether the model can accurately identify the presence of a given object in an image.

The POPE evaluation consists of three distinct settings: *Random*, *Popular*, and *Adversarial*. Under the *Random* setting, objects are randomly sampled from the entire dataset to assess the model's ability to recognize general objects. In the *Popular* setting, evaluation is conducted on the most frequently described objects in the dataset, focusing on the model's capability to verify common object occurrences. Finally, the *Adversarial* setting evaluates the model's ability to distinguish objects that are visually or semantically relevant to those present in the image, measuring its robustness against misleading cues.

Consistent with previous evaluate settings, we report the results over four MLLMs with their averaged F1 scores with and without debate framework. We notice more obvious improvement especially over naive decoding method such as Greedy and Nucleus Search.

5 Conclusion

In this work, we introduce a novel and effective multi-modal debate framework to pursue more responsible generation and reduced hallucination. The location-aware debate differs significantly from traditional single-modal debate frameworks by incorporating location awareness in visual content. This is achieved through both textual descriptions and a hybrid attention module to encourage fine-grained attention in visual contexts. Extensive experiments demonstrate that the proposed framework effectively reduces object-level hallucination and object-existence hallucination while simultaneously enhancing overall text quality under various metrics. More importantly, the framework generalizes effectively across different MLLMs and decoding methods. We hope this work inspires further research on debate frameworks for MLLMs.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. (2023). Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Cai, Y., Mao, S., Wu, W., Wang, Z., Liang, Y., Ge, T., Wu, C., You, W., Song, T., Xia, Y., et al. (2023). Low-code llm: Visual programming over llms. *arXiv preprint arXiv:2304.08103*, 2.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. (2023). Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. (2023). Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

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. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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*, pages 13418–13427.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. (2021). Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. (2024). Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

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*, pages 9579–9589.

Lee, H., Phatale, S., Mansoor, H., Lu, K. R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., and Rastogi, A. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M., and Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2024). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

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 *European conference on computer vision*, pages 740–755. Springer.

Lin, Z., Niu, Z., Wang, Z., and Xu, Y. (2024). Interpreting and mitigating hallucination in mllms through multi-agent debate. *arXiv preprint arXiv:2407.20505*.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. (2023). Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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*, pages 26296–26306.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2024b). Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2024c). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Manakul, P., Liusie, A., and Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. (2023a). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. (2023b). Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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*, pages 8748–8763. PMLR.

Redmon, J. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. (2024). Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 836–898.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. (2018). Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Sutskever, I. (2014). Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Tang, K., Zhang, H., Wu, B., Luo, W., and Liu, W. (2019). Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al. (2024). Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., and Wang, Y. (2024). Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087.

Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., and Wang, L. (2024). Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer.

Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., and Liu, Z. (2022). Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer.

Yang, L., Tang, K., Yang, J., and Li, L.-J. (2017). Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zang, Y., Li, W., Han, J., Zhou, K., and Loy, C. C. (2024). Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, pages 1–19.

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. (2023a). How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Liu, Y., Chen, K., and Luo, P. (2023b). Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., and He, C. (2023). Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Textual Locations	Visual Locations	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra		Avg.
		C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I	
✗	✗	2.43%	4.31%	6.12%	3.77%	3.02%	2.23%	2.72%	3.41%	3.50%
✓	✗	6.44%	7.94%	13.23%	4.20%	7.32%	3.71%	6.98%	6.32%	7.02%
✗	✓	3.35%	6.77%	7.86%	4.10%	4.03%	3.89%	3.47%	5.42%	4.86%
✓	✓	8.56%	11.36%	16.25%	5.34%	9.58%	6.59%	8.30%	10.19%	9.52%

Table 4: Average decrease rates of CHAIR values for sentence-level ($C_S \downarrow$) and image-level ($C_I \downarrow$) across four models. Each data point is the average over 5 decoding methods, including Greedy, Nucleus, Beam Search, DoLA and OPERA.

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. (2022). Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. (2023). Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Ablation Study

We conduct detailed ablation studies on critical hyperparameters, including the number of debate rounds, Judge's role and the fusion weight α. Additionally, we analyze the impact of incorporating location information in both the text and vision branches by comparing results with and without this crucial information. These experiments are conducted across different decoding methods and MLLMs shown in Table 1. We then report the average reduction rate in the CHAIR evaluation with the maximum number of new tokens set to 512.

Debate Rounds We report the average reduction rate in the CHAIR evaluation as the number of debate rounds varies from 0 to 4. The results are shown in left sub-plot Figure 4. While additional debate rounds consistently improve performance, we observe that the benefits become marginal beyond two rounds. Considering the trade-off between efficiency and model performance, we set the number of debate rounds to 2 in the previous experiments of this paper.

The Judge's Role We set the Judge with two settings: one that requests it to naively choose the better statement from two debaters, and another that refines the statements from debaters with further summarization. The results are shown by the red curve in the left sub-plot in Figure 4. We found that the judge should choose the right statement among debaters rather than providing a summary, especially as the number of debate rounds increases. We believe this observation is expected, as the quality of generated content improves with further debate among debaters. However, the judge's statement is not evaluated, thus may potentially introduce some additional, but marginal, hallucination.

Fusion Weight Similarly, we explore the range of fusion weight from 0 to 1 and reported the decreased rate of the averaged CHAIR value. We set $\alpha = 0.3$ in previous experiments given it's best results. The details are in right sub-plot Figure 4.

Importance of Location-Aware Debate We ablate the most critical components that enable MLLMs to be location-aware of different detected objects and facilitate fine-grained discussion with appropriate attention. The results are shown in Table 4. We observe that textual location descriptions contribute the most to the overall performance of the debate framework. While hybrid visual attention also improves the framework's effectiveness, it is additive to textual descriptions and further helps reduce hallucinations.

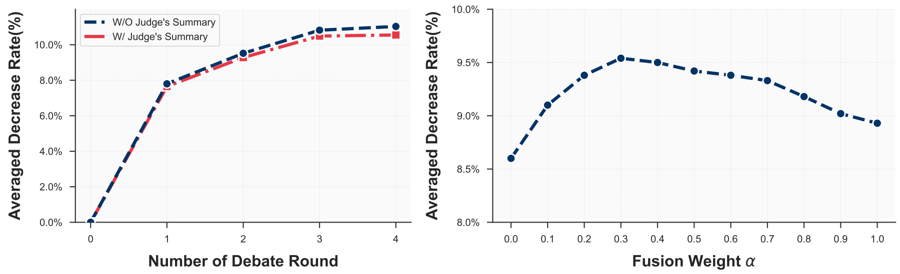

Figure 4: Left: Impact of debate rounds on CHAIR reduction rate with and without further summary from the Judge. Right: Impact of fusion weight (α) on CHAIR reduction rate.

B Implementation details.

We set all the debaters and the judge to be the same type of MLLM during the debate process. The debate proceeds through two rounds. For the attention module, we build upon CLIP's ViT-B/32 backbone, which relies on a 12-layer Vision Transformer and a matching 12-layer text transformer, both pre-trained on large-scale image-text data. The input image is partitioned into non-overlapping 32×32 patches. Similarly, each input text sequence is tokenized and embedded to produce 768-dimensional token representations. The final attention map is generated through a weighted fusion of the self- and cross-attention maps, where $\alpha = 0.7$ by default, followed by a thresholding operator that retains values above the 70th percentile. For OPERA and beam search, we set $N_{\text{beam}} = 5$. For nucleus sampling, we set $p = 9$. The indices of candidate pre-mature layers are set to "0,2,4,6,8,10,12,14," while the mature layer index is set to 32 for DoLa.

In the greedy decoding approach, the next token is simply selected based on the highest probability. Beam search decoding, on the other hand, maintains a set of candidate sequences, i.e., beams, to optimize the final generation. Unlike these deterministic methods, nucleus sampling dynamically truncates the probability distribution, filtering out low-probability tokens and sampling from a concentrated set of high-confidence candidates.