SCHOOL OF SOFTWARE, ZHEJIANG UNIVERSITY SUMMER CAMP REPORT-TASK II: RESEARCH ON THE HALLUCINATION PROBLEM OF MULTIMODAL LARGE MODELS

A PREPRINT

Gaoyun Lin
School of Computer Science
Zhejiang university of Technology
202103340109@zjut.edu.cn

SUMMARY

During a 10-day-plus summer camp, I conducted research on Task II: Investigating the hallucination problem in multimodal large models. Initially, I studied the background knowledge of Transformers and large language models (LLMs), followed by a focus on methods to mitigate the hallucination problem, specifically on decoding methods and post-hoc approaches that do not require additional training, data, or external knowledge.

I replicated ***eight*** methods presented at top conferences in the past year, such as ICML, ACL and CVPR. Due to time and equipment constraints, I was able to reproduce most of the experimental data for the OPERA method, while for other methods, I utilized 1,000 randomly selected images from the COCO val2014 dataset. I generated descriptions under fixed random seeds and token length constraints of 64 and 512 tokens, respectively, documenting their failure cases and visualizing the corresponding information, additionally, I conducted POPE tests under random settings.

Subsequently, I analyzed these failure cases from a theoretical perspective and proposed my own improvements. Finally, I summarized the cutting-edge foundations of mitigating hallucination in multimodal large models. All experiments were conducted on the AutoDL platform, utilizing the following rented equipment: GPU L20 (48GB) with 20 vCPU Intel(R) Xeon(R) Platinum 8457C and 100GB memory, and an additional GPU RTX 4090D (24GB) with 15 vCPU Intel(R) Xeon(R) Platinum 8474C and 80GB memory. The case analyses, experimental data, codes, and other related materials can be found in the folder uploaded.

Contents

1 Introduction

Large Language Models (LLMs) have revolutionized the field of artificial intelligence (AI) by demonstrating remarkable capabilities in understanding, generating, and interacting with human language. These models, exemplified by architectures such as GPT-3 OpenAI [2022] and BERT Devlin et al. [2018], leverage massive amounts of textual data and sophisticated deep learning techniques to achieve unprecedented levels of performance in natural language processing (NLP) tasks. LLMs can generate coherent and contextually relevant text, translate languages, summarize documents, and even engage in complex conversational interactions. However, despite their impressive abilities, LLMs are not without limitations, particularly in generating outputs that may occasionally be factually incorrect or contextually inappropriate, a phenomenon known as "hallucination Agrawal et al. [2023]."

Building on the success of LLMs, researchers have developed multimodal large models that integrate multiple types of data, such as text, images, and audio, to create more versatile and context-aware AI systems. These multimodal models, like CLIP Radford et al. [2021] and DALL-E, are designed to process and understand information from various modalities simultaneously, enabling applications such as image captioning, visual question answering, and cross-modal retrieval. By combining the strengths of different data types, multimodal models can provide more comprehensive and accurate insights, making them highly valuable in a wide range of fields, from healthcare to autonomous driving Song et al. [2023].

However, the integration of multiple modalities introduces new challenges, one of the most significant being the issue of hallucination. In the context of multimodal models, hallucination refers to the generation of outputs that are either factually incorrect, contextually irrelevant, or completely fabricated based on the input data. This problem can manifest in various ways, such as generating an incorrect description of an image, providing inaccurate answers to questions based on visual input, or creating content that does not align with the provided data. The complexity of processing and fusing diverse data sources increases the risk of these erroneous outputs, making it a critical area of research Montagnese et al. [2020].

This report aims to conduct cutting-edge research on the hallucination of multimodal large models, including replication and improvement. Initially, I introduce the eight papers provided in the project tasks, categorizing them into two groups: studies on hallucination mitigation and studies on hallucination evaluation. I focus on the research on hallucination mitigation, presenting experimental data and analyses of the replicated methods. I particularly emphasize decoding methods and post-hoc approaches that do not require additional training, data, or external knowledge, providing corresponding experimental results and analyses. Based on my own observations and motivations, I propose my own improvement methods, detailing their implementation and experimental results. Finally, I explore more advanced research directions in this field.

2 Listed Papers

2.1 Overview

In this section, I analyze the eight papers listed in the project. I categorize these papers into two groups: Hallucination Evaluation and Hallucination Mitigation, with three papers in the former category and five in the latter. I replicated three papers from the latter category, with a particular focus on the OPERA Huang et al. [2024] method, where I reproduced most of the data, including the CHAIR experimental data for three models (LLaVA-1.5 7B, InstructBLIP 7B, MiniGPT-4 7B) using a random selection of 1,000 images from COCOval2014, as well as the POPE data for four models (LLaVA-1.5 7B, InstructBLIP 7B, MiniGPT-4 7B, Shikra) under Random, Popular and Adversarial setting.

For the other two papers, which discuss the Activation Decoding Chen et al. [2024a] and Residual Visual Decoding Zhong et al. [2024] methods, I used the LLaVA-1.5 7B model with a fixed set of 1,000 images from COCOval2014, generating descriptions for these images and obtained the corresponding CHAIR data. I then analyzed these descriptions, including metrics such as those described below, and evaluated the POPE performance of the corresponding models and methods under random settings.

2.2 Evaluation Metric

The CHAIR metrics are used to assess the occurrence of hallucinated objects in generated captions. The CHAIR metrics consist of two parts:

- **CHAIR_S** (Caption Hallucination Assessment In Retrieval - Sentence Level):

$$\text{CHAIR\_S} = \frac{\text{hallucinated\_caption\_count}}{\text{total\_captions}} \tag{1}$$

where *hallucinated_caption_count* is the number of captions containing hallucinated objects, and *total_captions* is the total number of captions.

- **CHAIR_I** (Caption Hallucination Assessment In Retrieval - Object Level):

$$\text{CHAIR\_I} = \frac{\text{total\_hallucinated\_objects}}{\text{total\_mentioned\_objects}} \tag{2}$$

where *total_hallucinated_objects* is the total number of hallucinated objects across all captions, and *total_mentioned_objects* is the total number of mentioned objects across all captions.

- **Position Ratio of Hallucinated Objects**:

$$\text{Position Ratio} = \frac{\text{Index of Hallucinated Object}}{\text{Total Number of Words in Description}} \tag{3}$$

where *Index of Hallucinated Object* is the position of the hallucinated object within the description, and *Total Number of Words in Description* is the total number of words in the generated description.

- **Co-occurrence Matrix**:

$$\text{Co-occurrence Count} = \sum_{i,j} \text{Pairs of Objects (i, j) appearing together} \tag{4}$$

where *Pairs of Objects (i, j)* are pairs of objects that appear together in the generated descriptions.

- **Coverage**:

$$\text{Coverage} = \frac{\text{Number of Correctly Mentioned Objects}}{\text{Total Number of Actual Objects}} \tag{5}$$

where *Number of Correctly Mentioned Objects* is the count of objects correctly mentioned in the description, and *Total Number of Actual Objects* is the total count of objects present in the image.

- **Hallucinated Object Counts**:

$$\text{Hallucinated Object Count} = \sum \text{Number of Hallucinated Objects in Each Description} \tag{6}$$

where *Number of Hallucinated Objects in Each Description* is the count of objects that were mentioned in the description but not present in the actual image.

- **Hallucination Ratio**:

$$\text{Hallucination Ratio} = \frac{\text{Number of Hallucinated Objects}}{\text{Total Number of Words in Description}} \tag{7}$$

where *Number of Hallucinated Objects* is the count of hallucinated objects in the description, and *Total Number of Words in Description* is the total number of words in the generated description.

2.3 Hallucination Evaluation

2.3.1 HaELMWang et al. [2023]

Motivation. Previous works have primarily focused on hallucinations in language models, but the unique characteristics of LVLMs necessitate specialized evaluation methods. This study aims to develop a framework for systematically evaluating hallucination in LVLMs, considering both the complexity of multimodal data and the need for accurate, practical assessments.

Method. The study presents a framework called Hallucination Evaluation based on Large Language Models (HaELM) to evaluate hallucinations in LVLMs. The approach includes several key components:

First, data collection involves gathering responses from LVLMs, including both hallucinatory and non-hallucinatory outputs. Realistic hallucination responses are manually annotated from descriptions generated by LVLMs when describing images. Second, synthetic data generation is achieved using ChatGPT to produce additional hallucination data. Specific prompts are crafted based on reference captions to ensure that the synthetic data closely aligns with realistic hallucination patterns. Third, model training involves fine-tuning a language model (LLaMA) with the collected data. The training process focuses on enabling the model to distinguish between hallucination and accurate descriptions based on image captions. Next, the evaluation process uses the HaELM framework to assess the propensity for hallucination in different LVLMs. The framework compares the outputs of these models against reference descriptions to quantify the extent of hallucination. Finally, an analysis of contributing factors is conducted to understand the reasons behind hallucinations in LVLMs. This analysis examines the effects of various prompts, generation lengths, and sampling methods. Additionally, attention patterns within LVLMs are visualized to identify potential causes of hallucination.

2.3.2 ReEvalYu et al. [2024a]

Motivation. The motivation behind this study is to tackle the challenge of hallucinations in Retrieval-Augmented Large Language Models (LLMs). Hallucinations occur when models generate content not grounded in the provided evidence, which undermines their reliability. This issue is critical as LLMs are increasingly used for tasks requiring accurate and trustworthy information synthesis. The study focuses on evaluating the robustness of LLMs in using new evidence, addressing gaps where static benchmarks fail to ensure the model's reliance on the provided data. This evaluation is essential for applications involving sensitive or dynamic information.

Method. The study introduces ReEval, a framework designed to evaluate the propensity and severity of hallucinations in LLMs when using retrieval-augmented evidence. The method involves several key steps: first, identifying seed test cases from datasets like Natural Questions and RealtimeQA, where LLMs can correctly answer questions in both closed-book and open-book settings. Next, synthetic data is generated using prompt chaining to perturb the original evidence through techniques such as answer swapping and context enriching. Finally, the generated test cases are used to evaluate LLMs by measuring their accuracy in maintaining consistency with the perturbed evidence, thus assessing their susceptibility to hallucinations. This approach leverages adversarial machine learning to ensure LLMs reliably use new evidence, providing a cost-effective and transferable evaluation method.

2.3.3 R-BenchWu et al.

Motivation. The motivation behind this study is to address the problem of relationship hallucinations in Large Vision-Language Models (LVLMs). While previous efforts have focused on object hallucinations, which can be mitigated with object detectors, the hallucinations related to inter-object relationships remain underexplored. Accurate understanding of these relationships is crucial for comprehensive visual comprehension. The study aims to evaluate how well LVLMs understand and describe these relationships, considering the long-tail distribution in visual instruction tuning datasets and the tendency of current LVLMs to rely excessively on common sense knowledge rather than actual visual content.
Method. The study introduces R-Bench, a benchmark designed to evaluate relationship hallucinations in LVLMs.

R-Bench includes image-level questions assessing relationship existence and instance-level questions evaluating local visual comprehension. First, COCO captions are parsed to extract relationship triplets, matched with nocaps dataset captions to form relationship seeds. Using GroundingDINO, significant objects are annotated with bounding boxes. Next, prompts are generated based on these relationships and bounding boxes, and fed into a Large Language Model (LLM) to create questions. These questions are manually filtered to ensure accuracy. Finally, the benchmark evaluates various LVLMs, measuring their accuracy in identifying and reasoning about inter-object relationships. This method highlights models' tendencies to overlook visual content in favor of common sense knowledge and their difficulties with spatial reasoning based on contextual information.

2.4 Hallucination Mitigation

2.4.1 OPERAHuang et al. [2024]

Motivation. The motivation for OPERA stems from the observation that hallucinations in multi-modal large language models (MLLMs) are closely tied to the knowledge aggregation patterns in the self-attention mechanism. Specifically, MLLMs tend to focus excessively on a few summary tokens during token generation, often neglecting image tokens, which leads to inaccurate or nonsensical descriptions of the image content. This over-trust in summary tokens can cause the model to generate hallucinations, such as describing objects that are not present in the image or misinterpreting the image content. OPERA aims to mitigate this issue by introducing an Over-Trust Penalty and a Retrospection-Allocation strategy to adjust the token selection process during decoding without the need for additional training data or external knowledge.

Method. The method introduced by OPERA includes two main components: the Over-Trust Logit Penalty and the Retrospection-Allocation strategy.

1. **Over-Trust Logit Penalty:** OPERA identifies that hallucinations are often linked to the model's tendency to over-trust certain tokens, particularly summary tokens that aggregate previous knowledge. This over-trust can lead to hallucinations when these tokens overshadow the visual tokens from the image. To counteract this, OPERA introduces a penalty on the model logits during the beam search decoding process.

The penalty term $\alpha \cdot \varphi(w_{\leq t})$ is computed using a column-wise metric on the self-attention weights. Specifically, the self-attention weights ω are calculated as:

$$\omega = \text{SoftMax}\left(\frac{QK^T}{\sqrt{D}}\right)$$

where Q and K are the query and key features, and D is the feature dimension. For a given local window of attention weights W_{t-1}^k, the metric $\varphi(w_{\leq t})$ is defined as the maximum column-wise product:

$$\varphi(w_{\leq t}) = \max_j \prod_{i=j}^{t-1} \sigma\omega_{i,j}$$

where σ is a scaling factor.

2. **Retrospection-Allocation Strategy:** This strategy involves rolling back the decoding process when a strong over-trust pattern is detected. If the maximum value of the penalty metric is observed repeatedly in recent tokens, OPERA will retrospectively reallocate the token selection to avoid patterns that lead to hallucination.

The retrospection condition is met when the overlap count N_{overlap} of the maximum penalty locations C exceeds a threshold r:

$$N_{\text{overlap}} = \sum_{c \in C} \mathbf{1}(c = s), \quad s = \text{Mode}(C)$$

If $N_{\text{overlap}} \geq r$, the decoding rolls back to the summary token and selects a new token from the complementary set.

Through these mechanisms, OPERA effectively reduces hallucinations by penalizing over-trust patterns and reallocating token selection during decoding, demonstrating significant improvements in accuracy and reliability across various MLLM models and tasks.

2.4.2 Residual Visual DecodingZhong et al. [2024]

Motivation. The motivation for MMHalSnowball arises from the observation that hallucinations in large vision-language models (LVLMs) can accumulate over time, a phenomenon termed Multimodal Hallucination Snowballing. When LVLMs generate incorrect or nonsensical descriptions, these hallucinations can mislead the model in subsequent interactions, causing a significant drop in performance. The study investigates the extent to which accumulated hallucinations influence LVLMs' subsequent responses, especially when these models are required to answer specific visual questions within a hallucinated conversation context. To address this issue, MMHalSnowball proposes a framework to evaluate and mitigate the snowballing effect without requiring additional training data.

Method. The MMHalSnowball method consists of two primary components: the evaluation framework and the mitigation strategy using Residual Visual Decoding (RVD).

1. **Evaluation Framework:** The evaluation framework simulates hallucinatory conversations to test LVLMs' behavior. It consists of the following steps:

- **Hallucination Allocation:** Categorize hallucinations into existence, attribute, relation, and imagination types.
- **Hallucination Creation:** Generate hallucinatory descriptions using ChatGPT to rewrite fact sentences and modify regional descriptions.
- **Conversation Construction:** Construct conversations that include both factual and hallucinatory contexts.
- **Evaluation Metrics:** Measure the model's performance using accuracy (Acc), flip rate (FR), and weak flip rate (WFR).

The model's response accuracy is defined as:

$$\text{Acc} = \frac{\sum_{i=1}^{n} \text{Score}(y_i, \hat{y}_i)}{n}$$

where y_i is the expected answer, $\hat{y}_i$ is the generated response, and $\text{Score}(y_i, \hat{y}_i)$ evaluates if y_i is entailed in $\hat{y}_i$.

2. **Residual Visual Decoding (RVD):** RVD aims to emphasize visual information during the inference process to counteract hallucinations. It includes:

- **Residual Visual Predictions:** Construct input that connects the visual input v with the current text query x to derive an output distribution focusing on visual information.
- **Revised Output Distribution:**

$$p_{\text{RVD}}(y|v,h,x) = \text{softmax}(\alpha \cdot \text{logit}_\theta(y|v,x) + (1-\alpha) \cdot \text{logit}_\theta(y|v,h,x))$$

  where α is a scaling factor that adjusts the emphasis on visual information.
- **Adaptive Distribution Blending:** Adjust the scaling parameter α dynamically based on the Jensen-Shannon divergence (JSD) between the output distributions from visual and text context:

$$\tau = \text{JSD}(p_\theta(y|v,x) \| p_\theta(y|x)), \quad \alpha = \min(\beta \cdot \tau, 1)$$

The RVD process ensures that the model maintains its contextual abilities while reducing the influence of hallucinations by adjusting the weight given to visual information dynamically.

2.4.3 Activation DecodingChen et al. [2024a]

Motivation. Hallucinations occur when LLMs generate incorrect information that seems factual. This study found that correct responses have sharper (more focused) activations in their internal states compared to incorrect responses. By using this observation, the study aims to reduce hallucinations by measuring and utilizing this "sharpness."

Method. The method involves using an entropy-based metric to measure the sharpness of activations in LLMs. This process is broken down into several steps:

1. **Activation Analysis:** Measure how sharply the model's internal states (hidden states) respond to context tokens. Sharper activations indicate higher confidence and correctness.

2. **Entropy Calculation:** Calculate the entropy, which is a measure of uncertainty, for the activations. Lower entropy means sharper, more focused activations. The entropy $E(v_p, v_{1:t})$ for a token v_p is calculated as:

$$E(v_p, v_{1:t}) = -\sum_{i=1}^{t} P(i|v_p, v_{1:t}) \log P(i|v_p, v_{1:t})$$

where $P(i|v_p, v_{1:t})$ is the normalized activation probability.

3. **Incorporation into Decoding:** Adjust the probabilities of generating each token based on their entropy values. Tokens with lower entropy (sharper activations) are favored. The adjusted probability is:

$$P(v_p|v_{1:p-1}) \propto e^{-\lambda E(v_p, v_{1:t})} P(v_p|v_{1:p-1})$$

where λ controls the impact of entropy on the token probability.

4. **Decoding Process:** Use the adjusted probabilities in standard decoding algorithms (like greedy decoding or beam search) to generate the final output.

2.4.4 VOLCANOLee et al. [2023]

Motivation. The problem of hallucinations often arises because the model's vision component fails to correctly interpret the image. VOLCANO aims to fix this by using a self-feedback mechanism to improve the initial responses based on visual input.

Method. VOLCANO's method involves three main steps: critique, revise, and decide.

1. **Critique**: VOLCANO first generates an answer to a question using the image. Then, it critiques this answer by producing feedback that highlights any mistakes or inconsistencies with the image.

2. **Revise**: Using the feedback, VOLCANO revises the initial answer to correct the errors. This step uses detailed visual information from the feedback to improve the accuracy of the response.

3. **Decide**: Finally, VOLCANO checks if the revised answer is better than the original. If the revised answer is an improvement, it is accepted. If not, the process may repeat to further refine the answer.

This process helps VOLCANO reduce errors and provide more accurate answers that align with the visual content. It has shown significant improvements in performance on various tests, making it effective in addressing the issue of multimodal hallucinations.

2.4.5 HalluciDoctorYu et al. [2024b]

Motivation. HalluciDoctor aims to solve the problem of hallucinations in visual instruction data. Hallucinations happen when large language models (LLMs) generate incorrect or nonsensical information that seems factual. The goal is to reduce these hallucinations and improve the accuracy of the generated content, especially in important fields like medicine where precision is critical.

Method. The method includes several steps to identify and reduce hallucinations in LLM outputs:

1. **Data Collection**: Collect a large dataset of visual instruction data, including both correct and incorrect examples. This helps train the model to distinguish between accurate and hallucinatory content.

2. **Model Training**: Train the LLM on this dataset, using techniques to penalize incorrect outputs and reinforce correct ones. This helps the model learn to avoid generating hallucinations.

3. **Evaluation**: Regularly evaluate the model's performance on a separate validation set to ensure it is improving and reducing the frequency of hallucinations. Adjustments are made based on these evaluations to further fine-tune the model.

2.5 Evaluation and Analysis

2.5.1 Performance

I will give the results of my reproduced OPERA, Residual Visual Decoding and Activation Decoding, as described in 2.1

Model	Accuracy	Precision	Recall	F1 score	Yes ratio
InstructBLIP 7B	90.0	93.5	86.7	90.0	47.5
MiniGPT-4 7B	79.5	89.4	68.5	77.5	39.2
LLaVA-1.5 7B	90.1	93.6	86.8	90.1	50.3

Table 1: POPE (OPERA): Result on 'Random' split

Model	Accuracy	Precision	Recall	F1 score	Yes ratio
InstructBLIP 7B	83.1	80.9	86.7	83.7	53.3
MiniGPT-4 7B	73.3	75.6	68.7	72.0	45.1
LLaVA-1.5 7B	85.7	83.8	88.5	86.1	52.5

Table 2: POPE (OPERA):Result on 'Popular' split

Model	Accuracy	Precision	Recall	F1 score	Yes ratio
InstructBLIP 7B	80.4	77.0	86.7	81.6	56.0
MiniGPT-4 7B	71.3	72.6	68.6	70.5	47.0
LLaVA-1.5 7B	78.8	74.1	88.5	80.7	59.4

Table 3: POPE (OPERA): Result on 'Adversarial' split

Method	CHAIR_S	CHAIR_I
Activation_Decoding	0.146	0.132
Residual_Visual_Decoding	0.123	0.123
OPERA	0.103	0.101

Table 4: CHAIR Scores (64)

2.5.2 Analysis

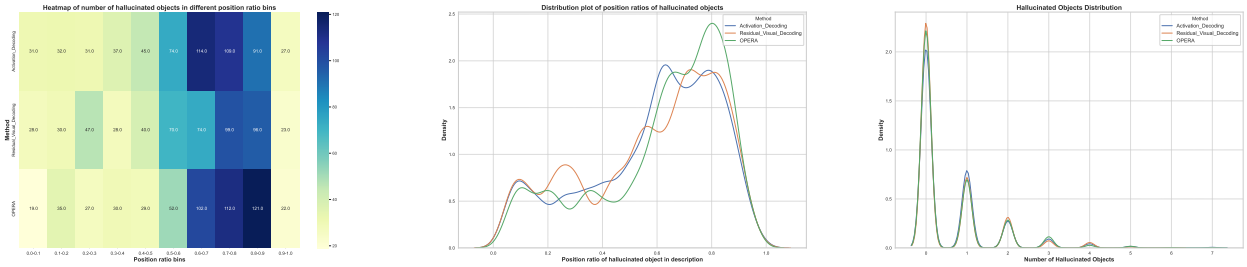
OPERA. OPERA shows several failure cases (see the uploaded folder) that illustrate its limitations. For example, in the image COCO_val2014_000000578385.jpg, the model erroneously mentions a "bowl" and a "spoon" on the table. This suggests that the backtracking strategy fails to correct errors effectively in multiple backtracking

Method	CHAIR_S	CHAIR_I
Activation_Decoding	0.375	0.251
Residual_Visual_Decoding	0.338	0.234
OPERA	0.341	0.217

Table 5: CHAIR Scores (512)

Method	Accuracy	Precision	Recall	F1 score
OPERA	90.1	93.6	86.8	90.1
Activation_Decoding	89.4	92.6	86.1	89.1
Residual_Visual_Decoding	88.6	91.9	85.3	88.4

Table 6: POPE Scores on Random Split

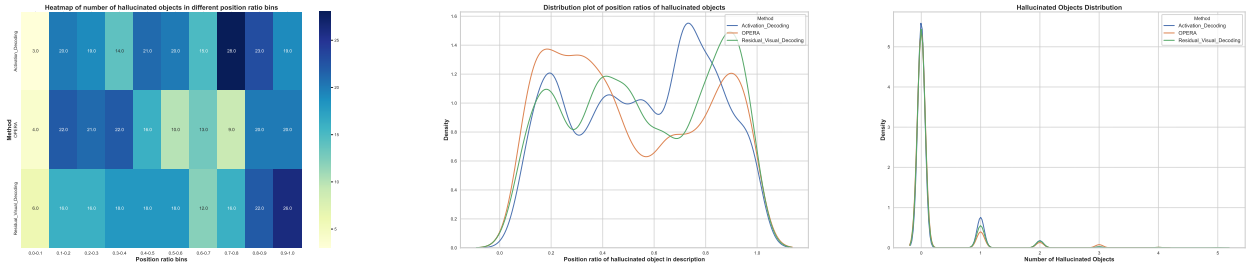


(a) Heatmap (b) Distribution Plot (c) Hallucinated Objects Distribution

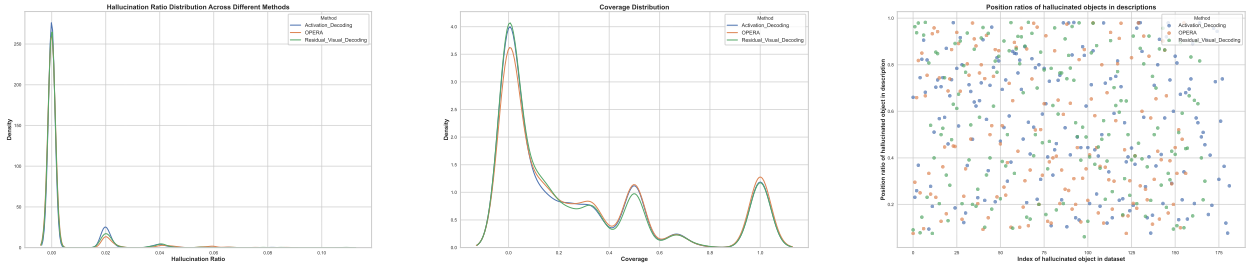
(d) Hallucination Ratio Distribution (e) Coverage Distribution (f) Scatter Plot

Figure 1: Comparison of Activation Decoding, OPERA, and Residual Visual Decoding Methods (512 tokens)

scenarios. This failure can be attributed to the model's reliance on previous tokens which might contain erroneous context. The over-trust penalty might not be strong enough to counteract the accumulated context errors. In COCO_val2014_000000202843.jpg, the model generates a non-existent "bench", indicating poor performance of the over-trust penalty in complex scenes. Complex scenes with multiple objects and intricate details challenge the model's ability to correctly allocate attention, leading to persistent hallucinations. The penalty mechanism might not sufficiently discourage the generation of plausible but incorrect objects when the visual context is highly detailed. Another failure case, COCO_val2014_000000375430.jpg, shows the model incorrectly mentioning a "vase" and a "cup", highlighting multiple hallucinations in complex scenes where the strategy fails to mitigate effectively. This scenario underscores the difficulty in balancing the penalization of over-trusted tokens with the need to maintain coherent and accurate descriptions. When the scene complexity increases, the backtracking mechanism struggles to identify and correct the initial point of error, leading to compounding mistakes. In COCO_val2014_000000530619.jpg, the model mistakenly identifies a "backpack", demonstrating the backtracking strategy's failure in multiple backtracking scenarios. The iterative nature of backtracking can lead to recursive errors, where each backtracking step introduces new potential points of failure, making it challenging to converge on a correct description. This case shows that the model's error correction mechanism might be too localized, missing the broader context required to accurately describe the scene. Lastly, in COCO_val2014_000000408049.jpg, the model generates a non-existent "backpack", showing the over-trust penalty's inability to completely eliminate hallucinations in complex scenes. This indicates that the penalty applied might not be proportionate to the degree of hallucination risk present in intricate visual contexts. The balance



(a) Heatmap (b) Distribution Plot (c) Hallucinated Objects Distribution



(d) Hallucination Ratio Distribution (e) Coverage Distribution (f) Scatter Plot

Figure 2: Comparison of Activation Decoding, OPERA, and Residual Visual Decoding Methods (64 tokens)

between penalizing incorrect trust and preserving the model's descriptive capabilities is delicate and difficult to achieve consistently.

Chart comparisons indicate a more evenly distributed position ratio for hallucinated objects but with a higher frequency above 0.6, suggesting that hallucinations often occur later in the descriptions. This pattern suggests that the initial part of the description sets a context that might be increasingly misinterpreted as the description progresses. The coverage distribution shows that OPERA performs well in low coverage areas but is slightly lacking in medium coverage (0.4-0.6). This indicates that while the strategy manages to limit the breadth of hallucinations, it struggles to maintain this control as the description lengthens and complexity increases. The number of hallucinated objects is higher in single-object scenarios compared to multi-object scenes, indicating that OPERA is somewhat better at handling scenes with fewer items. This might be because simpler scenes reduce the cognitive load on the model, making it easier to apply the over-trust penalty effectively. The hallucination ratio is more concentrated and stable, indicating that while hallucinations are present, they are fewer and more predictable in their occurrence. This suggests that OPERA introduces a degree of consistency in its hallucination patterns, even if it does not entirely eliminate them. From my perspective OPERA shows good overall control of hallucinations, especially in low coverage and high position ratio areas, but is slightly lacking in medium coverage control and still faces multiple hallucination issues in complex scenes.

Activation Decoding. Activation Decoding also exhibits several notable failure cases. For example, in the image COCO_val2014_000000024712.jpg, the model incorrectly identifies a non-existent "apple". This suggests that the model fails to distinguish background noise, leading to hallucinations when handling complex backgrounds. Another case, COCO_val2014_000000378904.jpg, shows the model mistakenly recognizing background clutter as a specific object "cup", indicating that context sharpness does not effectively filter out background noise. In COCO_val2014_000000289036.jpg, the model generates a non-existent "sofa", which can be attributed to context distortion during long text generation. Similarly, in COCO_val2014_000000517490.jpg, the model incorrectly identifies a "bookshelf", showing insufficient context sensitivity to capture nuances in the image. Finally, in COCO_val2014_000000372514.jpg, the model generates a non-existent "plant", due to increased dependency on context accumulating misleading information, resulting in hallucinations. Chart comparisons for In-Context Sharpness show that hallucinated objects often appear in the latter part of the description (high position ratios), with the highest frequency in low coverage (<0.2) and single hallucinated objects being most common. The hallucination ratio is concentrated in the low ratio region (<0.02), but more hallucinations appear in the tail. In conclusion, In-Context Sharpness performs well in low coverage and single-object scenes but is prone to hallucinations in the latter part of descriptions and multi-object scenes, with poor control over high hallucination ratios.

Residual Visual Decoding. Residual Visual Decoding method also has its share of failure cases. In the image COCO_val2014_000000370138.jpg, the model mentions multiple non-existent “bowls" and “bottles", reflecting insufficient handling of multiple similar objects, leading to quantity recognition errors. In COCO_val2014_000000327845.jpg, the model generates a non-existent “bench", indicating failure in accurately capturing the number of objects in the image. Another case, COCO_val2014_000000348792.jpg, shows the model incorrectly mentioning a “bottle", pointing to inaccurate object recognition in complex backgrounds. In COCO_val2014_000000292583.jpg, the model generates a non-existent “bicycle", demonstrating quantity recognition issues in multi-object scenes. Finally, in COCO_val2014_000000423821.jpg, the model mistakenly identifies a “backpack", highlighting inaccurate object recognition in complex scenes leading to hallucinations.

Chart comparisons for Residual Visual Decoding show higher numbers of hallucinated objects in the high position ratio region (>0.6) but overall more even distribution. It performs excellently in low coverage and single-object scenes. The frequency of single hallucinated objects is highest, indicating slightly better performance in multi-object scenes compared to the other methods. The hallucination ratio is concentrated in the low ratio region (<0.02), demonstrating stable performance. Based on the detailed failure case analysis and chart comparisons Residual Visual Decoding excels in low coverage and single-object scenes, with better control in early and middle positions, but needs improvement in controlling hallucinations in the latter part of descriptions and addressing quantity recognition issues in multi-object scenes.

3 Additional Works

3.1 Overview

In this section, I focus on **five** additional methods (not included in the reference list) for mitigating hallucinations in multimodal large models that do not require additional training, data, or external knowledge, and I have reproduced these methods. First, I will briefly introduce these five methods. Then, for each method, I conduct experiments similar to those for Activation Decoding2.1, present the experimental data, and provide a brief analysis.

3.2 DOLAChuang et al. [2023]

Motivation. Hallucinations occur when LLMs generate incorrect information, which is problematic in fields requiring precise information, like medicine or law. DOLA leverages the fact that higher layers in LLMs typically contain more factual details.

Method. DOLA contrasts outputs from higher and lower layers of the LLM. By comparing these layers, DOLA emphasizes the more reliable information from the higher layers. Additionally, DOLA dynamically selects the best layers to use during each step of text generation, ensuring the most accurate and relevant information is prioritized. This approach significantly improves the factual accuracy of the outputs, making the generated responses more truthful and reliable.

3.3 VCDLeng et al. [2024]

Motivation. The hallucinations occur when models generate descriptions of objects that are not present in the images, leading to incorrect outputs. This issue is particularly critical in applications requiring high accuracy, such as medical imaging or autonomous driving. The study aims to reduce these hallucinations by introducing a method called Visual Contrastive Decoding (VCD).

Method. Visual Contrastive Decoding (VCD) works by comparing the model’s outputs generated from original and distorted visual inputs. This approach helps in reducing the model’s reliance on statistical biases and language priors, which are common causes of hallucinations. VCD does not require additional training or external tools, making it an efficient solution. By contrasting the outputs, VCD ensures that the generated content is more closely aligned with the actual visual input, thus significantly reducing object hallucinations and improving the overall accuracy of the LVLMs.

3.4 WoodpeckerYin et al. [2023]

Motivation. Existing methods often require retraining the models with specific data, which is resource-intensive. Woodpecker proposes a novel, training-free approach to directly correct hallucinations without needing additional training data or computation.

Method. Woodpecker employs a five-stage process to correct hallucinations in MLLMs. First, it extracts key concepts from the generated text. Then, it formulates questions around these concepts to diagnose potential hallucinations. The answers to these questions are validated against visual information using pre-trained expert models. Next, it generates visual claims based on this validated information, creating a visual knowledge base. Finally, Woodpecker uses this knowledge base to correct the original text, ensuring that it accurately reflects the visual content. This method enhances the reliability and interpretability of MLLM outputs by providing clear evidence for the corrections made.

3.5 HALCChen et al. [2024b]

Motivation. Existing methods often rely on extensive retraining with additional data, which is resource-intensive. HALC aims to provide a more efficient solution by using a novel adaptive focal-contrast decoding technique that can be implemented without the need for retraining. **Method.** HALC employs adaptive focal-contrast decoding to reduce object hallucinations. This technique dynamically adjusts the focus on different regions of the image during the text generation process. By comparing and contrasting the outputs from these varied focal points, HALC can detect and mitigate hallucinations. This method ensures that the generated descriptions are more accurately aligned with the visual content, thereby improving the reliability and accuracy of LVLM outputs without requiring additional training data or significant computational resources.

3.6 LUREZhou et al. [2023]

Motivation. LURE (Learning to Understand and Reduce Hallucinations) seeks to provide a streamlined approach to identifying and mitigating these hallucinations by analyzing the underlying causes and implementing targeted adjustments.

Method. LURE introduces a multi-faceted approach to address object hallucination. First, it employs a comprehensive analysis of the hallucination patterns within LVLM outputs, identifying common triggers and contexts where hallucinations occur. Based on these insights, LURE utilizes a combination of fine-tuning on a curated dataset and a post-processing filtering mechanism. The fine-tuning process involves minimal additional data, carefully selected to cover typical hallucination scenarios, while the post-processing filter dynamically evaluates generated text, removing or correcting hallucinated objects based on contextual coherence and visual relevance. This dual approach not only reduces the incidence of hallucinations but also enhances the overall quality and accuracy of the model's outputs with minimal additional computational overhead.

3.7 Evaluation and Analysis

3.7.1 Performance

Method	CHAIR_S	CHAIR_I
DoLa	0.129	0.118
HALC	0.119	0.113
Woodpecker	0.149	0.131
LURE	0.126	0.113
VCD	0.131	0.116

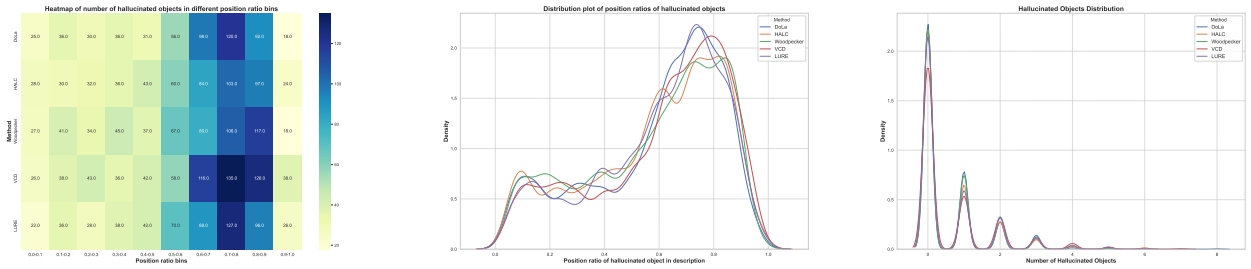
Table 7: CHAIR Scores (64)

Method	CHAIR_S	CHAIR_I
DoLa	0.351	0.219
HALC	0.327	0.225
Woodpecker	0.360	0.240
LURE	0.337	0.233
VCD	0.363	0.231

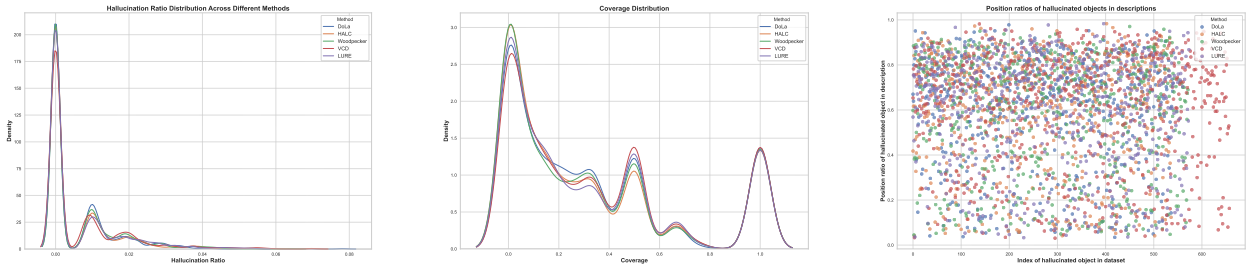
Table 8: CHAIR Scores (512)

Method	Accuracy	Precision	Recall	F1 score
DoLa	89.6	92.8	86.3	89.4
HALC	88.8	92.0	85.6	88.1
Woodpecker	90.0	93.3	86.9	90.2
LURE	88.3	91.6	84.9	87.9
VCD	89.8	93.1	86.7	89.9

Table 9: POPE Scores on Random Split



(a) Heatmap (b) Distribution Plot (c) Hallucinated Objects Distribution



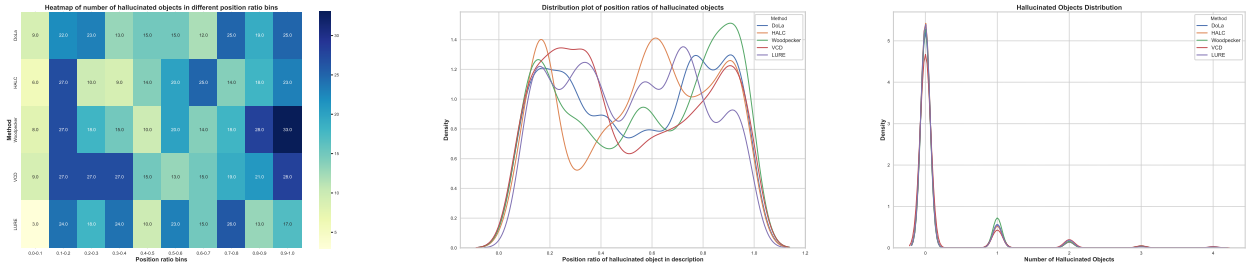
(d) Hallucination Ratio Distribution (e) Coverage Distribution (f) Scatter Plot

Figure 3: Comparison of DoLa HALC Woodpecker LURE VCD Methods (512 tokens)

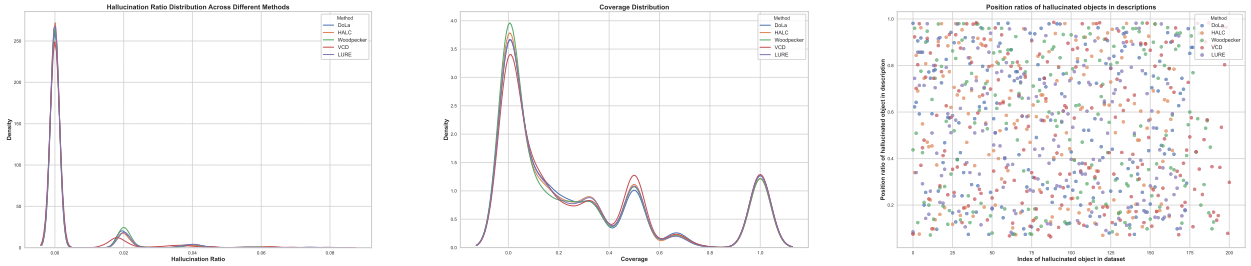
3.7.2 Analysis

DoLa. The DoLa method shows several notable failure cases, emphasizing its challenges in handling complex scenes. For instance, in the image COCO_val2014_000000298600.jpg, the model incorrectly identifies a non-existent “bowl". This suggests that the model fails to distinguish background noise, leading to hallucinations when handling complex backgrounds. Another case, COCO_val2014_000000286018.jpg, shows the model mistakenly recognizing background clutter as a specific object “bench", indicating that context sharpness does not effectively filter out background noise. Chart comparisons for DoLa show that the position ratio of hallucinated objects tends to be higher (above 0.6), suggesting that hallucinations are more likely to occur later in the descriptions. The coverage distribution shows that DoLa performs well in low coverage areas but struggles with medium coverage (0.4-0.6). The number of hallucinated objects is higher in single-object scenarios, indicating a general challenge in handling scenes with fewer items. The hallucination ratio distribution is concentrated in the low ratio region (<0.02), indicating fewer and more predictable hallucinations.

HALC. The HALC method also demonstrates several critical failure cases. For example, in the image COCO_val2014_000000375430.jpg, the model erroneously mentions a “vase" and a “cup" on the table, indicating that the hallucination mitigation strategy is ineffective in complex scenes with multiple objects. In COCO_val2014_000000432588.jpg, the model generates a non-existent “dog", highlighting the model’s struggle with differentiating between similar objects in detailed scenes. Chart comparisons for HALC show that the position ratio of hallucinated objects is higher (above 0.6), indicating a higher likelihood of hallucinations later in descriptions. The coverage distribution shows good performance in low coverage areas but struggles with medium coverage. HALC handles single-object scenarios better than multi-object scenes, and its hallucination ratio is concentrated in the low ratio region.



(a) Heatmap (b) Distribution Plot (c) Hallucinated Objects Distribution



(d) Hallucination Ratio Distribution (e) Coverage Distribution (f) Scatter Plot

Figure 4: Comparison of DoLa HALC Woodpecker LURE VCD Methods (64 tokens)

Woodpecker. The Woodpecker method exhibits several failure cases illustrating its limitations. For instance, in the image COCO_val2014_000000298600.jpg, the model mentions a “bowl" that does not exist in the scene, indicating poor performance in background noise reduction. In COCO_val2014_000000286018.jpg, the model incorrectly identifies a “bench", showing that the hallucination correction mechanism fails to handle complex backgrounds effectively. Chart comparisons for Woodpecker show a high position ratio for hallucinated objects (above 0.6), suggesting that hallucinations are more frequent later in descriptions. The coverage distribution reveals good performance in low coverage but difficulties in medium coverage. Woodpecker performs better in single-object scenarios, with hallucinations concentrated in the low ratio region.

LURE. The LURE method, designed to analyze and mitigate object hallucination in large vision-language models, has demonstrated several failure cases that highlight its limitations. For example, in the image COCO_val2014_000000375430.jpg, the model mentions a "vase" that does not exist in the scene, indicating challenges in accurately identifying objects in cluttered environments. Chart comparisons for LURE show a significant position ratio for hallucinated objects (above 0.6), suggesting that hallucinations tend to occur more frequently later in descriptions. The coverage distribution reveals that LURE performs well in low coverage but faces difficulties in medium coverage. Additionally, LURE performs better in scenarios involving fewer objects, with hallucinations concentrated in the low ratio region.

VCD. The VCD method exhibits several critical failure cases. In the image COCO_val2014_000000298600.jpg, the model incorrectly identifies a “bowl", suggesting challenges in background noise filtering. In COCO_val2014_000000286018.jpg, the model mistakenly mentions a “bench", showing ineffective context filtering in complex backgrounds. Chart comparisons for VCD show higher position ratios for hallucinated objects (above 0.6), indicating more hallucinations later in descriptions. The coverage distribution shows good performance in low coverage but struggles with medium coverage. VCD performs better in single-object scenarios, with hallucinations concentrated in the low ratio region.

Based on the detailed failure case analysis and chart comparisons, each method exhibits distinct strengths and weaknesses. DoLa, HALC, Woodpecker, LURE, and VCD all show a trend where hallucinations are more likely to occur later in descriptions, struggle with medium coverage, and handle single-object scenarios better than multi-object scenes.

4 My Method

4.1 Motivation

Effectiveness of OPERA Method. The OPERA method has proven effective in addressing this issue by introducing an Over-trust Penalty and a Retrospection-Allocation strategy. These techniques help mitigate the tendency of MLLMs to overly rely on summary tokens, which often leads to neglecting visual tokens and generating hallucinatory content. By penalizing this over-trust during decoding and reallocating token selection as needed, OPERA significantly reduces hallucinations without requiring additional data or training.

Addressing Hallucinations in Long Responses with LURE. Observations indicate that in long responses, hallucinations tend to appear predominantly in the latter parts of the description. To address this specific issue, the LURE method was introduced. LURE is designed to handle hallucinations that occur due to uncertainty, co-occurrence patterns, and positional factors within the text. It uses a revisor model trained to correct potentially hallucinatory descriptions, particularly effective in the latter segments of long responses. Given these capabilities, I have decided to incorporate the LUREZhou et al. [2023] method to ensure the accuracy and reliability of extended MLLM outputs.

Enhancing Visual Input Focus in MLLMs. According to literatureFavero et al. [2024], it has been found that as the length of the response increases, MLLMs tend to progressively reduce their focus on visual information. This reduction in attention to visual inputs over time exacerbates the hallucination problem, as models become more reliant on previously generated text rather than the visual content. To counteract this tendency and improve the model's accuracy in describing visual elements throughout the entire response, I aim to enhance the model's visual input processing capabilities. This enhancement will ensure that visual information remains a significant focus even as the response lengthens, thereby minimizing hallucinations and improving overall output fidelity.

Methods for Enhancing Visual Input Focus in MLLMs.

- **Saliency Detection**: Use a pre-trained DeepLabV3 model to generate a saliency map, highlighting significant regions in the image.
- **Combine Salient Regions with Image**: Merge the saliency map with the original image to ensure that the model focuses on the most important visual information during processing.

Due to time and device constraints, I did not combine LURE with OPERA successfully, but I combined the Enhancing Visual Input Focus and OPERA methods as the final result and analysis.

4.2 Evaluation

In this section I compare the results of my method to the results of all the methods described above.

Method	CHAIR_S	CHAIR_I
My_Method	0.115	0.102
DoLa	0.129	0.118
HALC	0.119	0.113
Woodpecker	0.149	0.131
LURE	0.126	0.113
Activation_Decoding	0.146	0.132
Residual_Visual_Decoding	0.123	0.123
OPERA	0.103	0.101
VCD	0.131	0.117

Table 10: CHAIR Scores (64)

5 Extra Research Progress Report

5.1 Hallucination Detecting

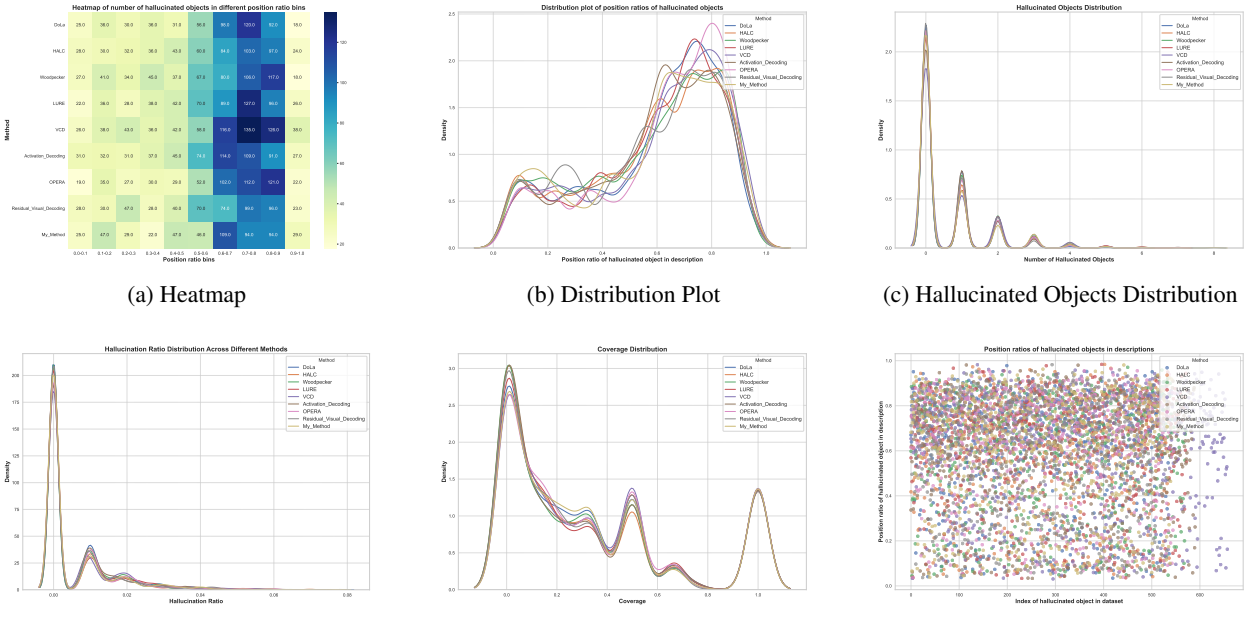
Recent research has focused on detecting and preventing hallucinations in large vision-language models (LVLMs). Gunjal et al introduce M-HalDetecGunjal et al. [2024], a multimodal hallucination detection dataset, and propose

Method	CHAIR_S	CHAIR_I
My_Method	0.317	0.212
DoLa	0.351	0.220
HALC	0.327	0.226
Woodpecker	0.360	0.241
LURE	0.337	0.233
Activation_Decoding	0.375	0.251
Residual_Visual_Decoding	0.338	0.234
OPERA	0.341	0.217
VCD	0.363	0.231

Table 11: CHAIR Scores (512)

Method	Accuracy	Precision	Recall	F1 score
OPERA	90.1	93.6	86.8	90.1
DoLa	89.6	92.8	86.3	89.4
HALC	88.8	92.0	85.6	88.1
Woodpecker	90.0	93.3	86.9	90.2
LURE	88.3	91.6	84.9	87.9
Activation_Decoding	89.4	92.6	86.1	89.1
Residual_Visual_Decoding	88.6	91.9	85.3	88.4
VCD	89.8	93.1	86.7	89.9
My_Method	90.0	93.3	86.8	90.1

Table 12: POPE Scores on Random Split

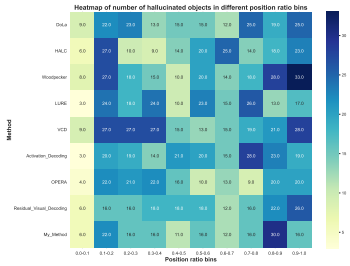


(a) Heatmap (b) Distribution Plot (c) Hallucinated Objects Distribution

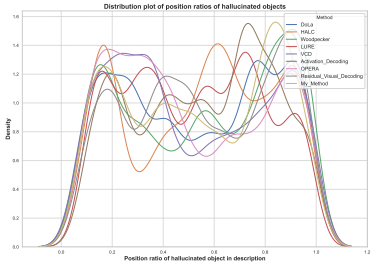
(d) Hallucination Ratio Distribution (e) Coverage Distribution (f) Scatter Plot

Figure 5: Comparison of DoLa HALC Woodpecker LURE VCD Activation_Decoding OPERAR esidual_Visual_Decoding My_Method (512 tokens)

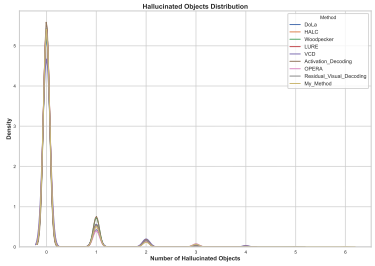
Fine-grained Direct Preference Optimization (FDPO), which reduces hallucination rates by 41% and 55% in different models through human evaluation. Zhai et al.introduce HallE-SwitchZhai et al. [2023] to control object existence hallucinations in LVLMs. This method reduces hallucination by 44% compared to other models while maintaining object coverage. Chen et al. propose a unified hallucination detection framework for multimodal large language modelsChen et al. [2024c], addressing the challenge of generating unfaithful content. This research emphasizes the need for robust detection methods to ensure the reliability of generated outputs.



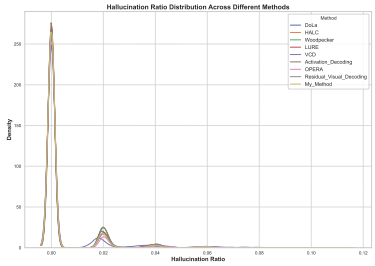
(a) Heatmap



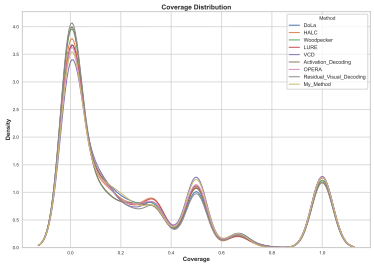
(b) Distribution Plot



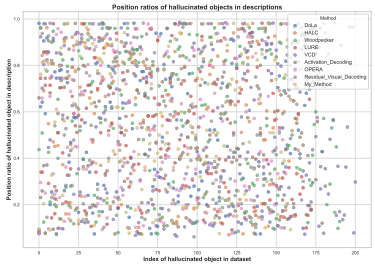
(c) Hallucinated Objects Distribution



(d) Hallucination Ratio Distribution



(e) Coverage Distribution



(f) Scatter Plot

Figure 6: Comparison of DoLa HALC Woodpecker LURE VCD Activation_Decoding OPERAR esidual_Visual_Decoding My_Method (64 tokens)

5.2 Hallucination Evaluating

Recent research has focused on evaluating hallucinations in large vision-language models (LVLMs). Li et al. introduce POPE, a polling-based object presence evaluation method that systematically evaluates object hallucinations in LVLMs and finds that objects frequently mentioned in visual instructions are prone to be hallucinated Li et al. [2023]. Liu et al. present HallusionBench, an image-context reasoning benchmark challenging for GPT-4V(ision), LLaVA-1.5, and other multi-modality models. This benchmark sheds light on the illusion or hallucination of LVLMs and provides insights on improving them Guan et al. [2024]. Lovenia et al. propose NOPE, a negative object presence evaluation to measure object hallucination in vision-language models, highlighting the vulnerabilities of current models to object hallucination Lovenia et al. [2023]. Cui et al. conduct a holistic analysis of hallucination in GPT-4V(ision), identifying bias and interference challenges Cui et al. [2023]. Jing et al. introduce FaithScore, a fine-grained evaluation metric that measures the faithfulness of the generated answers from LVLMs, demonstrating that current systems still generate hallucinated content Jing et al. [2023].

5.3 Hallucination Mitigating

Liu et al. introduce LRV-InstructionLiu et al. [2023], a large and diverse visual instruction tuning dataset designed to mitigate hallucination by including both positive and negative instructions, leading to more robust visual instruction tuning and significant reduction in hallucinations. Gunjal et al. propose HA-DPO (Hallucination-Aware Direct Preference Optimization)Zhao et al. [2023], which optimizes models to reduce hallucination by finetuning them with detailed preference data. This method reduces hallucinations in LVLMs like InstructBLIP and other multi-modal models significantly. Yan et al. present ViGoR (Visual Grounding with Fine-Grained Reward Modeling)Yan et al. [2024], which improves visual grounding in LVLMs by incorporating detailed visual annotations and a fine-grained reward model to reduce hallucinations. Another approachYue et al. [2024], "Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective," emphasizes reducing complexity in decision-making to prevent hallucinations by optimizing the decision points in the model's architecture. This method aims to streamline the decision-making process, thereby reducing the chances of generating hallucinated content. Cha et al. propose Pensieve, a new method which mitigates visual hallucinations through a retrospect-then-compare strategy, effectively reducing the risk of generating hallucinatory contentYang et al. [2024].

References

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1333–1351, 2024a.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In *Forty-first International Conference on Machine Learning*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, 2024.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*, 2024.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024a.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024b.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

TB OpenAI. Chatgpt: Optimizing language models for dialogue. openai, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huanmin Liu. Can knowledge graphs reduce hallucinations in llms? : A survey. *ArXiv*, abs/2311.07914, 2023. doi:10.48550/arXiv.2311.07914.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Shezheng Song, Xiaopeng Li, and Shasha Li. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *ArXiv*, abs/2311.07594, 2023. doi:10.48550/arXiv.2311.07594.

Marcella Montagnese, Pantelis Leptourgos, Charles Fernyhough, Flavie Waters, Frank Larøi, Renaud Jardri, Simon McCarthy-Jones, Neil Thomas, Rob Dudley, John-Paul Taylor, Daniel Collerton, and Prabitha Urwyler. A review of multimodal hallucinations: Categorization, assessment, theoretical perspectives, and clinical recommendations. *Schizophrenia Bulletin*, 2020. doi:10.1093/schbul/sbaa101.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*, 2024c.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pages 14375–14385, 2024.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*, 2023.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118*, 2024.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. Pensieve: Retrospect-then-compare mitigates visual hallucination. *arXiv preprint arXiv:2403.14401*, 2024.