Object hallucinations in Multimodal large language models: A survey

Vangmay Sachan*[1, 2]

[1] *National University of Singapore*

Abstract

This survey presents a comprehensive analysis of hallucinations in multimodal large language models (MLLMS) also known as Large Vision Language Models (LVLMs) focusing on techniques related to the mitigation of these hallucinations. MLLMs have demonstrated significant advancements and remarkable abilities in multimodal tasks, outperforming traditional models in areas such as image recognition, natural language understanding. However, despite their potential, MLLMs often suffer from a critical limitation which causes them to generate outputs inconsistent with the visual content as a result which has raised questions about their reliability in real world applications. We begin by introducing concepts crucial for the understanding of MLLMs and hallucinations, this involves an overview of components within LLMs and MLLMs. The paper then explores the different causes of hallucinations, categorizing them based on the reason behind their occurrence. We then review the recent advances in approaches involving mitigating such hallucinations and how effective they are.

This survey aims to provide researchers with a comprehensive understanding of the current research landscape concerning hallucinations in multimodal large language models (MLLMs), by integrating existing knowledge and identifying promising directions for future exploration. By critically examining key challenges and opportunities, we seek researchers with valuable insights, driving innovative solutions to address the complex issue of hallucinations in MLLMs. This work aspires to catalyze advancements in the field, paving the way for more reliable and robust MLLM-based systems across diverse applications.

1 Introduction

1.1 Motivation

The development of the transformer model [1] has sparked a significant surge in the development of large language models (LLMs) [2], which have achieved remarkable advancements across various tasks in natural language processing (NLP). These models excel in language understanding [3], reasoning [4], and generation [5], often reaching near-human performance levels. By Harnessing the capabilities of LLMs, Multimodal large language models (MLLMs) integrate multiple forms of information, such as images and videos , enabling modern AI systems to comprehend both textual and visual data. MLLMs have demonstrated promising outcomes in tasks such as image captioning, visual question answering, proving their high potential in real-world applications. However, a critical challenge persists: these models often exhibit hallucinations [6] that result in responses being either factually inaccurate or nonsensical responses to user prompts. Hallucinations refer to the generation of responses that are either factually incorrect, nonsensical, or based on non-existent patterns or objects perceived by the model. This issue undermines the deployment of MLLMs in real-world applications, where accuracy and reliability are paramount. The occurrence of hallucinations raises concerns about the safety, trustworthiness, and overall usability of these models, especially in critical domains like healthcare, autonomous systems, and decision-making processes. This survey discusses the recent advances in identification and mitigation of such hallucinations. This paper starts from the basics, giving an introduction about MLLMs then discussing what hallucinations are, why they occur, what benchmarks are used to measure them and then we look at several mitigation strategies.

*vangmay.sachan@u.nus.edu

1.2 Organization of this survey

This survey addresses recent advancements in identifying and mitigating hallucinations within MLLMs. It begins with a foundational introduction to MLLMs, followed by an exploration of what hallucinations are, their underlying causes, and the benchmarks used to evaluate them. We examine several strategies aimed at reducing these hallucinations, providing a comprehensive overview of current research efforts in this critical area. Finally, we outline the challenges and different avenues of improvement, offering potential pathways for future research.

1.3 Scope of the Survey

This survey focuses on the problem of hallucinations within AI based models, particularly multimodal large language models that are capable of coming from different modalities of information such as text, images, audio and video. The survey primarily covers research aimed at mitigating object hallucinations which can be categorized in three categories.

- **Object Category:** This type involves MLLMs erroneously identifying incorrect objects or categories within a given image. For instance, an MLLM might incorrectly assert the presence of "a bench" in a park image that contains no such object.

- **Object Attribute:** In this scenario, while the identified object categories are correct, the descriptions of these objects do not align with the visual content. Examples include inaccuracies in color, position, or pose.

- **Object Relation:** Here, all objects and their attributes are accurately described; however, the interactions between these objects are misrepresented. For example, a description may incorrectly portray a person's actions in an image.

Drawing upon the taxonomy of a previous survey 7, we explore the mitigation of hallucinations arising during four key stages of MLLM training.

- **Data:** Hallucinations that stems from inadequate visual datasets, poor quality within datasets, and statistical biases present in those datasets.

- **Model:** These hallucinations arise from weaknesses within the MLLM itself, such as ineffective language model, vision encoder or a flawed alignment module.
- **Training:** Adjustments to the training task of MLLMS can help mitigate such hallucinations
- **Inference:** Hallucinations which can occur while the model is generating output, this can occur when the length of generated output is long and leads to diluted attention on the visual content provided or because of occurrences of special character like "
n"

2 Definitions and Preliminary information

2.1 Large Language Models

Since multimodal large language models are built upon large language models, it is crucial that we understand the concept of large language models. Some examples of LLMs include GPT-3 [8], LLaMA [9], and GPT-4 [2]. LLMs are transformer based models that are trained on extremely text based large datasets. LLMs have been able to show promising results in several language tasks such as information extraction, summarization and translation.

Pretraining: Initially, LLMs are pre-trained on vast amounts of unlabeled data using self-supervised learning in an autoregressive. This phase helps the model learn general language patterns, grammar, facts, and some reasoning capabilities by predicting the next word in a sequence or filling in missing words (e.g., masked language modeling). However, pre-trained models are general and may not perform optimally on specific tasks (e.g., translation, summarization, or answering specific domain-related questions).

Supervised-fine-tuning: Involves adapting the pre-trained Large Language Model (LLM) to perform a specific task by training it further on an annotated dataset, which contains input-output pairs that demonstrate the desired behavior. Example, translation, sentiment analysis.

2.2 Multimodal Large Language Models

MLLMs are a branch of models that are able to allow LLMs to combine data from multiple modalities and allow it to "perceive" images, the ultimate goal is to provide means for an LLM to "see" visual information in the form of videos/images and enable downstream tasks such as image captioning. In order to merge both modalities the MLLMs contain a pre-trained LLM and also a vision encoder, together they are combined via a learnable interface which is trained such that the visual content feeded into the model is interpreted by the LLMs input space. Essentially the task of MLLMs is to combine two kinds of knowledge, parametric knowledge which is the knowledge received from the language model and visual knowledge which is received from the vision model and then generate an output. Training of MLLMs consists of 2 steps.

Pretraining : This is done to allow the model to achieve cross-modal feature alignment. The visual encoder and LLM remain frozen, while the cross-modal interface is being trained.

Instruction tuning: This stage is similar to fine-tuning a large language model, in this stage QA datasets are used to enhance the model's ability to follow multi-modal instructions and tasks, mostly this is done to enhance a MLLM's performance in a specific set of domain specific tasks including image-captioning, visual question answering.

2.3 Hallucinations

Hallucinations refer to the phenomenon when the output generated by the MLLM as a response to the user's prompt does not align with the visual content provided to the model and provides responses that are either nonsensical or inaccurate and contain details that might not be present in the visual content provided 7. The main type of hallucinations this survey discusses in Object hallucinations, where the model is unable to see objects in the visual content provided. There are three primary types of object hallucinations.

- **Object Category:** This type involves MLLMs erroneously identifying incorrect objects or categories within a given image. For instance, an MLLM might incorrectly assert the presence of "a bench" in a park image that contains no such object.
- **Object Attribute:** In this scenario, while the identified object categories are correct, the descriptions of these objects do not align with the visual content. Examples include inaccuracies in color, position, or pose.
- **Object Relation:** Here, all objects and their attributes are accurately described; however, the interactions between these objects are misrepresented. For example, a description may incorrectly portray a person's actions in an image.

3 Causes of Hallucinations

In this section, we look at the causes of hallucinations in MLLMs which can occur primarily within four aspects of an MLLMs pipeline: Data, Model, Training, Inference. In this section we look at the main causes of hallucinations.

3.1 Data

Data is the foundation of such strong and capable models and allows them to gain such high accuracy on a varying levels of tasks and most importantly, provide LLMs with the understanding of language. However it can also become a source of hallucinations. Deep learning models, particularly large language models (LLMs) rely heavily on extensive datasets to achieve reliable performance. The amount of data available plays a crucial role in developing robust models, Currently image-text datasets like visual QA are used to train powerful MLLMs. However, a significant challenge arises from the fact that visual datasets are considerably less abundant than text only datasets. The scarcity of visual data can lead to unreliable cross-modal alignment which is essential for interpreting visual content. When the vision model is undertrained due to insufficient data, it can lead to hallucination. Thus, addressing the imbalance of dataset availability is vital for enhancing the reliability of MLLMs. Moreover, Current data collection methods are quite efficient and they promise quantity, however they do not give guarantee on the quality of data. Oftentimes, noisy data present in the training datasets lead to limitations in the cross-modal alignment of the MLLMs. Moreover, a lack of diversity in the data can lead to potential biases within the MLLMs. Besides that statistical bias can also play a role in making the data become the cause of hallucinations in a model. The distribution of nouns in the training datasets often has strong effects on the model. This is because neural networks have a strong tendency to memorize the dataset, so for example "person", might be something that appears often

in the dataset so even if the image does not contain a person the model might predict the presence of a person. Similarly, models also learn objects that "go together", given an image of a kitchen with a fridge, the model might hallucinate and predict the presence of a microwave. This is because in the training datasets the images of a kitchen would often have a microwave and fridge.

3.2 Model

The architecture of modern MLLMs consist of 3 key components: a pre-trained vision model, a pre-trained language model and an alignment interface module. Since these models are connected together, any error present in these modules can accumulate and lead to hallucinations. [7]

Weak-Vision model: A weak vision model can lead to misinterpretation of visual details and also a loss of visual features. A weak perception in the model means it would be able to capture less features from the provided image and hence lesser multimodal understanding.

Stronger Language Model: Due to the boom in research related to large language models, vision models are unable to catch up to the capability of LLMs as a result, MLLMs often have a stronger language model, this can lead to a situation where the model relies more on the language based information and tends to override the visual content.

Weak alignment interface: In order to combine the visual encoder and the large language model, MLLMs have a third component called an alignment interface, this module task is to learn the representations of the visual content and project this representation onto the language model's representation space. This is the part that basically allows the model to "see". If the alignment interface hasn't been trained properly, or has been trained by using insufficient data this can cause the model to fail in perceiving the visual information in sufficient details leading to a loss of visual information and leading to weak perception. Having a weak perception means that the model is unable to "see" the visual content with enough details hence it has less information about it. This could lead to a significant increase in missing object hallucinations.

3.3 Training

The training objective for multimodal large language models (MLLMs) largely mirrors that of traditional large language models (LLMs), specifically utilizing an auto-regressive next-token prediction loss. Although the method is proven to be useful for LLMs, However, some studies indicate that this next-token prediction loss may not be ideal for learning visual content due to the complex spatial structures involved.

3.4 Inference

Some studies argue that auto-regressive generation can led to attenuation of visual information as the length of output sequence groups because the self attention mechanism will focus more on the previously generated tokens and because of which the attention on visual content gets diluted. Hence the issue of losing attention could also lead to the model generating responses that are not related to the visual content as output size grows.

4 Hallucination Benchmarks

GAVIE [10] GPT4-Assisted Visual Instruction Evaluation assesses two key factors: Relevancy, which evaluates instruction-following performance and Accuracy to measure the visual hallucination in the output. It comprises a benchmark with 1,000 samples and an and does not require human-annotated ground-truth answers.

CHAIR 11 (Caption Hallucination Assessment with Image relevance) metric is designed for evaluating hallucinations in object captioning tasks. It calculates what proportion of words generated are actually in the image using the ground truth sentences and object segmentation as a bases.

FAITHSCORE 12 is designed to gauge responses to open-ended questions by evaluating the natural interactions between humans and multimodal large language models (MLLMs). FaithScore leverages an automated process that breaks down responses into components for detailed evaluation and analysis. The process is a three step pipeline where the segments within the response that contain descriptions are identified, then the specific atomic facts are extracted from these segments and they are checked for consistency against the original input. The metric focuses on fine-grained categories of object hallucination, which include attributes such as entity, count, color, relation, and other characteristics. Ultimately, FaithScore calculates a ratio representing the amount of hallucinated content in the responses.

POPE 5 Introduces a new evaluation metric and benchmark known as Pooling-based Object Probing Evaluation (POPE). The core concept behind POPE is to transform the assessment of hallucinations into a binary classification task by asking MLLMs straightforward Yes-or-No questions about specific objects in images (e.g., "Is there a car in the image?"). By sampling objects that LVLMs are prone to hallucinate they construct a set of hard questions to poll LVLMs. As standard answers to these questions are just Yes or No they easily identify them without complex parsing rules and avoid the influence of instruction designs and caption lengths.

5 Hallcination Mitigation

5.1 Hallucinations caused by Data

Hallucidoctor 13 addresses the object hallucination problem in MLLMs by augmenting the instruction tuning dataset used to train the MLLMs. The augmentation is a 2 step process.It includes a hallucination detection pipeline via consistency cross checking using multiple MLLMs. Based on those results the hallucinating content is removed. A key observation is that long tail distribution and object co-occurrences in the training data are two primary causes of hallucination. Following this observation, a counterfactual visual instruction strategy is deployed to expand the dataset. As a result, MLLMs trained on the augmented datasets are shown to be less prone to hallucinations.

LRV Instruction 10 Observed that without sufficient negative instructions, MLLMs are more likely to generate hallucinations in the form of false or inconsistent descriptions that do not align with the provided images or instructions. This occurs because the models may over-rely on positive examples and fail to learn the boundaries of what should not be included in their outputs. As a result of this finding, LRV Instruction is proposed which is a large containing 400k visual, instructions, covering 16 vision and language tasks with both positive and negative instructions in different semantic levels and style, the dataset includes both positive

and negative instructions for a more robust and unbiased instruction tuning.

5.2 Hallucinations caused by Model

HallESwitch 14 Parametric knowledge in the LLM is identified as a key factor leading to hallucinations, especially when the MLLM focuses more on the parametric knowledge and lesser on the visual content. [HallESwitch] identifies that when language descriptions include finer details than what the vision module can verify, hallucinations are likely to be induced. To address this issue, a novel approach is introduced which involves using a specialized module called HallESwitch. What sets this approach apart is that it can condition the captioning to switch between exclusively depicting contextual knowledge and blending it with parametric knowledge to image inferred objects. This is achieved using a control parameter(ϵ) that serves as a switching value. The module is trained via contrastive training datasets that cover both contextual and parametric datasets. For contextual datasets they set the $\epsilon =$ -1 and for data with both $\epsilon = 1$. During inference, they change the control value to any value in [-1, 1] to change how much the model relies towards parametric knowledge. As a result the method reduced hallucinations by 44% as compared to LLaVA7B

InternVL To bridge vision models with LLMs, existing VLLMs commonly employ lightweight glue layers such as QFormer or linear projection to align features of vision and language models. Such alignment contains several limitations. **[InternVL]** 15 reveals a large gap in both parameter scale and feature representation ability between vision encoders and LLMs. In effort to bridge the gap, the paper presents a large scale vision language foundation model which aligns the large scale vision encoders with LLMs and introduces a progressive image-text alignment strategy for efficient training of large scale vision language foundation models. Results showed that the model demonstrated strong performance in perception tasks, vision language tasks and also multimodal dialogue and achieved SoTA performance on 32 generic visual linguistic benchmarks.

VCoder 16 [VCoder] found that MLLMs are capable of answering details about an image however when prompted to for example "Counting the number of people in the image" (Perception tasks), they often deliver sub-par performance. They hypothesize that one of the main reasons for this is the absence of conversations covering object identification for not just salient objects but also the objects in the background of the image. Aiming to improve multimodal LLMs in the simple yet fundamental object level perception skills including counting the paper introduces a novel approach to enhance MLLMs by integrating vision encoders. VCoder functions as an adapter to allow MLLMs to leverage visual information through the use of auxiliary perception modalities as control inputs by utilizing additional input formats in the form of segmentation masks and depth maps to enhance the object identification ability of the MLLM.

5.3 Hallucinations caused by Training

HACL 17 Finds that there is a gap between textual and visual representations which indicate unsatisfactory cross-modal representation of visual and textual embeddings of LLMs. This issue accelerates the tendency of MLLMs to produce hallucinatory responses. They propose enhancing this alignment via contrastive learning. Hallucinatory texts are used as hard negative examples. The method pulls representations of non-hallucinatory text and visual samples closer together while representations of non-hallucinating and hallucinative texts are pushed away. This yields results that improve performance on popular benchmarks.

[MOCHA] 18 Mocha framework proposes to apply state of the art Reinforcement learning methods to the accuracy and relevance of image captioning and reduce hallucinations. A multi objective reward function is designed which consists of 2 parts

1. NLI based model for fidelity to measure the accuracy of captioning.

2. BERTScore to measure whether the output caption contains sufficient details.

3. KL Regularization penalty to reward the model which constrains the model to stay close to its initial policy.

Together these 3 objectives form the total reward function for the model. After which, in order to train the model they make use of a Proximal Policy optimization algorithm that has been used by recent works on text generation.

[EOS DECISION] 19 Often training data for MLLMs consists of rich visual semantics from multiple human annotations or vision expert models and is rewritten in lengthy paragraphs. This training data might exceed the visual perception capability of the MLLMs, especially for subtle image features that are small or can get easily mistaken. During training with such data, the model may attempt to fit the detail level and length distribution of ground truth caption and risk expressing details that it cannot discern from the image leading to hallucinations this incident is even accelerated because the MLLMs decides to terminate generation by assessing the completeness of the text through comparing the generated text with the image. Using these findings, the authors propose two techniques 1) A learning objective, 2) A data filtering based approach called Scoring EOS Supervision which aims to eliminate such harmful training data that can impair the model's ability to end sequences.

Noiseboost 20 discovered that one origin of hallucination is from the summarization mechanism of large language models, leading to excessive dependence on linguistic tokens while neglecting vision information. As a result, NoiseBoost was proposed which is a broadly applicable method for alleviating hallucinations. NoiseBoost introduces noise perturbations, which are controlled random variations added during the training process. This method plays the part of a regularizer and promotes a balanced distribution of attention between visual and linguistic tokens, thereby enhancing the model's robustness by preventing the dilution of visual knowledge within the MLLM. As a result, NoiseBoost improves dense caption accuracy by 8.1% in human evaluations and allows for effective use of unlabeled data, achieving comparable results with only 50% of labeled data.

RLHF-V 21 uses the [RLHF] paradigm to enhance MLLMs, specifically it identifies and addresses 2 issues via using 2 solutions. 1) They identify that responses about rich image content can be long and complex making it difficult for the model to know which response is preferable. 2) This coarse grained ranking feedback makes it difficult to accurately allocate credit to the desirable behaviors. These challenges are addressed via the following methods: At the data level, they propose to collect human feedback in the form of fine grained segment level corrections. At the method level, They propose dense direct preference op- timization (DDPO), a new variant of DPO [42] that addresses the traditional RLHF objective in an equivalent but efficient supervised fashion.

HA-DPO 22 frames hallucinations as a preference selection problem. This is done by training the model to prioritize accurate responses instead of hallucinatory ones. This effect is achieved via a 3 step process first, randomly selected images from the VG dataset are fed into the MLLM to generate descriptions corresponding to image then GPT 4 is used to detect hallucinations within these descriptions. If hallucination is present then it is prompted to re-write an accurate response. Through this, a new dataset is constructed that contains both hallucinatory and non hallucinatory responses . The MLLM model is then trained via DPO (Direct Preference optimization) which enables it to recognize hallucinatory and non hallucinatory responses and optimize the model in such a manner that it always prefers the non-hallucinatory response using a special kind of loss function.

5.4 Hallucinations caused by Inference

RBD 23 The approach aims to balance the amount of attention the model gives to parametric and visual knowledge during generation which helps to reduce the occurence of hallucination. The method employs 2 branches

- **Textual Branch**: Introduces noise into the textual input to prevent the model from over-relying on parametric knowledge and more on visual knowledge.
- **Visual Branch**: Focuses on selecting significant visual tokens from the image, refines the attention mechanism and ensures that the model also focuses on the visual knwoldge so it doesn't get diluted.

The idea behind this approach is to recalibrate how attention gets distributed among the textual and visual inputs, by balancing the attention RBD helps the model to lose vision as the size of output grows.

DENTIST 24 In order to tackle various types of hallucinations, DENTIST is a unified framework for hallucination mitigation. The core idea is to classify the queries then perform a classification step and take a different course of actions depending on the type of hallucination caused. The framework consists of three main steps Potential Hallucination Classification : The framework first involves classifying queries into two main categories 1) Perception Query: Which requires the model to have vision ability and 2) Reasoning: Which tests the model's reasoning and logical ability . This classification helps to identify the type of hallucination that can be present in the generated answer. The classification is done by prompting Chat-GPT to classify the result into a perception query or a reasoning query.

- **Divide and Conquer Treatment:** Based on the type of query that has been classified different strategy is deployed.
- Perception: The response is verified by asking sub-questions based on the query that focus on specific aspects of the visual content provided, then the LVLM answers the sub-questions to produce sub-answers which are then aggregated to form the output with fewer hallucinations.
- **Reasoning queries:** Sometimes, LVLMs only generate the results of logical reasoning and not the logical thinking process behind them which might be important to the user, hence a Chain-Of-Thought based method is used to enhance the response.
- **Validation Loop:** The framework then leverage an iterative validation loop which enhances the responses until no significant semantic changes occur between iterations.

The framework achieved a 13.44%/10.2%/15.8% improvement in accuracy on Image Quality, a Coarse Perception visual question answering (VQA) task in the MMbench benchmark, over the baseline InstructBLIP/LLaVA/VisualGLM

HELPD 25 The method employs Hierarchical feedback learning that integrates feedback mechanisms at both object and sentence level. The mechanism assesses whether the objects mentioned in the generated text are present in the corresponding images, it involves extracting object sets from the generated sentences and the labeled reference sentences. Then GPT-4 is used to evaluate the coherence of the generated output based on the visual content provided to the MLLM. A supplemental method Vision-enhanced Penalty Decoding is also incorporated, which utilizes visual attention to adjust penalty scores during the decoding process.

MemVR 26 [MemVR] believes that the capabilities of LLMs to comprehend and memorize different modalities are distant. Since an image possesses a higher information density as compared to textual data they assume that LLMs struggle to understand and memorize vision tokens compared to text tokens, which leads to hallucinations. They come up with a metric to quantify the uncertainty of the MLLM by first computing the probability of the next token and then using an entropy based metric to define the uncertainty. This leads to the observation that in the context of tokens involving objects, attributes or relations uncertainty is high. Based upon this finding they develop a method aimed at reducing hallucinations in MLLMs. Inspired by human cognitive behavior, MemVR allows models to revisit visual information when they show uncertainty, enhancing response reliability. Notably, it operates without additional training and leverages dynamic layer injection of visual features. Experimental results demonstrate that MemVR significantly mitigates hallucinations and improves overall model performance across various benchmarks, offering a promising approach for enhancing the accuracy of AI systems that process multimodal data. Experimental results demonstrate that MemVR enhances comprehensive performance on diverse tasks and mitigates hallucinations improving overall accuracy across various scenarios.

[VCD] 27 VCD is a technique deployed at the decoding stage and is used to mitigate hallucinations induced by language priors and statistical biases. The key observation made by authors is that distorted input leads to more hallucinated outputs containing priors and statistical biases. Hence they proposed a technique which compares the output of an Image V by making two inferences, first is the regular image V and second is a distorted version of this image by adding some predefined augmentations called V‘. By contrasting the outputs by the original image V and distorted image V‘ VCD aims to reduce the risk of hallucinations caused by statistical biases and unimodal priors.

[SKIP /" n"] 28 The work observes that hallucinations of LLMs is often trigger by a paragraph break character. Inorder to mitigate such hallucinations triggered by “” they propose two approach. 1) Design prompts in such a way that the model doesn't need to use “” 2) The output of “” is avoided by modifying the output decoding and reducing the logits corresponding to the ‘’ token.

MARINE 29 leverages an additional pretrained object grounding vision encoder for object grounding called DEtection TRansformer which provides supplementary visual information as a guide to the decoding process. It also strikes a balance between efficiency, instruction following ability and effectiveness in reducing

object hallucinations.

CGD 30 The technique is based on the key observation that CLIPScore can effectively differentiate between output that is hallucinated and non-hallucinated, this hypothesis is verified through a series of studies. Based on the observation, a two step algorithm is proposed.

1. Reliability scoring: Which designs a scoring function aiming to prioritize candidate responses which are less likely to be hallucinated

2. Guided sentence generation: Which generates the responses based on the scoring function designed in the previous step. The approach effectively mitigates hallucination while preserving the quality of generation .

OPERA[31] Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation **(OPERA)** introduces a novel decoding method aimed at reducing hallucinations in multi-modal large language models (MLLMs). It combines two key components: the Over-Trust Penalty (OTP), which penalizes excessive confidence in predictions, and the Retrospection-Allocation Strategy (RAS), which revisits and reassesses previous outputs to improve factual accuracy. Through extensive experiments, OPERA demonstrates a significant reduction in hallucination rates while maintaining or enhancing overall model performance, all without the need for additional training data or knowledge sources. This approach offers a practical solution to enhance the reliability of MLLMs, making it valuable for applications requiring high precision.

6 Future Directions

Future works aiming to mitigate 7 hallucinations include some key areas of focus. To enhance the performance of MLLMs, it is important to develop more advanced architectures that are capable of capturing complexities of language and able to generate logically sound and contextually relevant output based on the visual content provided. Future research should explore creating model architectures based on recent findings about hallucinations and embed hallucination mitigation techniques within the models design as this will enable a family of lesser mitigation prone models. Moreover, future work should also look towards the direction of cross modal alignment and aligning the representations between different modalities as it ensures that the generated content remains relevant to the input modality. Although this requires a lot of complex techniques, we believe its worthwhile to explore this avenue because a strong cross modal alignment can lead to consistent and hallucination free outputs without the need of external techniques.

7 Conclusion

In this paper, we have surveyed recent advancements in hallucinations within multimodal large language models, focusing on their causes, benchmarks and mitigation methods, moreover we also provide some future directions in the domain.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need; 2023. Available from: https://arxiv.org/abs/1706.03762.

[2] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.. GPT-4 Technical Report; 2024. Available from: https://arxiv.org/abs/2303.08774.

[3] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding. In: International Conference on Learning Representations; 2021. Available from: https://openreview.net/forum?id=d7KBjmI3GmQ.

[4] Qiao S, Ou Y, Zhang N, Chen X, Yao Y, Deng S, et al.. Reasoning with Language Model Prompting: A Survey; 2023. Available from: https://arxiv.org/abs/2212.09597.

[5] Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen JR. Evaluating Object Hallucination in Large Vision-Language Models; 2023. Available from: https://arxiv.org/abs/2305.10355.

[6] Xu Z, Jain S, Kankanhalli M. Hallucination is Inevitable: An Innate Limitation of Large Language Models; 2024. Available from: https://arxiv.org/abs/2401.11817.

[7] Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, et al.. Hallucination of Multimodal Large Language Models: A Survey; 2024. Available from: https://arxiv.org/abs/2404.18930.

[8] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.. Language Models are Few-Shot Learners; 2020. Available from: https://arxiv.org/abs/2005.14165.

[9] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al.. LLaMA: Open and Efficient Foundation Language Models; 2023. Available from: https://arxiv.org/abs/2302.13971.

[10] Liu F, Lin K, Li L, Wang J, Yacoob Y, Wang L. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning; 2024. Available from: https://arxiv.org/abs/2306.14565.

[11] Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object Hallucination in Image Captioning; 2019. Available from: https://arxiv.org/abs/1809.02156.

[12] Jing L, Li R, Chen Y, Du X. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models; 2024. Available from: https://arxiv.org/abs/2311.01477.

[13] Yu Q, Li J, Wei L, Pang L, Ye W, Qin B, et al.. HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data; 2024. Available from: https://arxiv.org/abs/2311.13614.

[14] Zhai B, Yang S, Xu C, Shen S, Keutzer K, Li C, et al.. HallE-Control: Controlling Object Hallucination in Large Multimodal Models; 2024. Available from: https://arxiv.org/abs/2310.01779.

[15] Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, et al.. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks; 2024. Available from: https://arxiv.org/abs/2312.14238.

[16] Jain J, Yang J, Shi H. VCoder: Versatile Vision Encoders for Multimodal Large Language Models; 2023. Available from: https://arxiv.org/abs/2312.14233.

[17] Jiang C, Xu H, Dong M, Chen J, Ye W, Yan M, et al.. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model; 2024. Available from: https://arxiv.org/abs/2312.06968.

[18] Ben-Kish A, Yanuka M, Alper M, Giryes R, Averbuch-Elor H. Mitigating Open-Vocabulary Caption Hallucinations; 2024. Available from: https://arxiv.org/abs/2312.03631.

[19] Yue Z, Zhang L, Jin Q. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective; 2024. Available from: https://arxiv.org/abs/2402.14545.

[20] Wu K, Jiang B, Jiang Z, He Q, Luo D, Wang S, et al.. NoiseBoost: Alleviating Hallucination with Noise Perturbation for Multimodal Large Language Models; 2024. Available from: https://arxiv.org/abs/2405.20081.

[21] Yu T, Yao Y, Zhang H, He T, Han Y, Cui G, et al.. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback; 2024. Available from: https://arxiv.org/abs/2312.00849.

[22] Zhao Z, Wang B, Ouyang L, Dong X, Wang J, He C. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization; 2024. Available from: https://arxiv.org/abs/2311.16839.

[23] Liang X, Yu J, Mu L, Zhuang J, Hu J, Yang Y, et al.. Mitigating Hallucination in Visual-Language Models via Re-Balancing Contrastive Decoding; 2024. Available from: https://arxiv.org/abs/2409.06485.

[24] Chang Y, Jing L, Zhang X, Zhang Y. A Unified Hallucination Mitigation Framework for Large Vision-Language Models; 2024. Available from: https://arxiv.org/abs/2409.16494.

[25] Yuan F, Qin C, Xu X, Li P. HELPD: Mitigating Hallucination of LVLMs by Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding; 2024. Available from: https://arxiv.org/abs/2409.20429.

[26] Zou X, Wang Y, Yan Y, Huang S, Zheng K, Chen J, et al.. Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models; 2024. Available from: https://arxiv.org/abs/2410.03577.

[27] Leng S, Zhang H, Chen G, Li X, Lu S, Miao C, et al.. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding; 2023. Available from: https://arxiv.org/abs/2311.16922.

[28] Han Z, Bai Z, Mei H, Xu Q, Zhang C, Shou MZ. Skip : A Simple Method to Reduce Hallucination in Large Vision-Language Models; 2024. Available from: https://arxiv.org/abs/2402.01345.

[29] Zhao L, Deng Y, Zhang W, Gu Q. Mitigating Object Hallucination in Large Vision-Language Models via Classifier-Free Guidance; 2024. Available from: https://arxiv.org/abs/2402.08680.

[30] Deng A, Chen Z, Hooi B. Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding; 2024. Available from: https://arxiv.org/abs/2402.15300.

[31] Huang Q, Dong X, Zhang P, Wang B, He C, Wang J, et al.. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation; 2024. Available from: https://arxiv.org/abs/2311.17911.