Hallucination of Multimodal Large Language Models: A Survey

ZECHEN BAI, Show Lab, National University of Singapore, Singapore
PICHAO WANG, Amazon Prime Video, USA
TIANJUN XIAO, AWS Shanghai AI Lab, China
TONG HE, AWS Shanghai AI Lab, China
ZONGBO HAN, Show Lab, National University of Singapore, Singapore
ZHENG ZHANG, AWS Shanghai AI Lab, China
MIKE ZHENG SHOU*, Show Lab, National University of Singapore, Singapore

This survey presents a comprehensive analysis of the phenomenon of hallucination in multimodal large language models (MLLMs), also known as Large Vision-Language Models (LVLMs), which have demonstrated significant advancements and remarkable abilities in multimodal tasks. Despite these promising developments, MLLMs often generate outputs that are inconsistent with the visual content, a challenge known as hallucination, which poses substantial obstacles to their practical deployment and raises concerns regarding their reliability in real-world applications. This problem has attracted increasing attention, prompting efforts to detect and mitigate such inaccuracies. We review recent advances in identifying, evaluating, and mitigating these hallucinations, offering a detailed overview of the underlying causes, evaluation benchmarks, metrics, and strategies developed to address this issue. Additionally, we analyze the current challenges and limitations, formulating open questions that delineate potential pathways for future research. By drawing the granular classification and landscapes of hallucination causes, evaluation benchmarks, and mitigation methods, this survey aims to deepen the understanding of hallucinations in MLLMs and inspire further advancements in the field. Through our thorough and in-depth review, we contribute to the ongoing dialogue on enhancing the robustness and reliability of MLLMs, providing valuable insights and resources for researchers and practitioners alike. Resources are available at: https://github.com/showlab/Awesome-MLLM-Hallucination.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Natural language processing**; *Machine learning*.

Additional Key Words and Phrases: Hallucination, Multimodal, Large Language Models, Vision-Language Models.

ACM Reference Format:
Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of Multimodal Large Language Models: A Survey. *Preprint* 1, 1 (April 2024), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding Author

Authors' addresses: Zechen Bai, Show Lab, National University of Singapore, 4 Engineering Drive 3, Singapore, Singapore, zechenbai@u.nus.edu; Pichao Wang, Amazon Prime Video, Washington, USA, pichaowang@gmail.com; Tianjun Xiao, AWS Shanghai AI Lab, Shanghai, China, tianjux@amazon.com; Tong He, AWS Shanghai AI Lab, Shanghai, China, htong@amazon.com; Zongbo Han, Show Lab, National University of Singapore, 4 Engineering Drive 3, Singapore, Singapore, hanzb1997@gmail.com; Zheng Zhang, AWS Shanghai AI Lab, Shanghai, China, zhaz@amazon.com; Mike Zheng Shou, Show Lab, National University of Singapore, 4 Engineering Drive 3, Singapore, Singapore, mike.zheng.shou@gmail.com.

ACM 0000-0000/2024/4-ART
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recently, the emergence of large language models (LLMs) [29, 81, 85, 99, 132] has dominated a wide range of tasks in natural language processing (NLP), achieving unprecedented progress in language understanding [39, 47], generation [128, 140] and reasoning [20, 58, 87, 107, 115]. Leveraging the capabilities of robust LLMs, multimodal large language models (MLLMs) [22, 75, 111, 138], sometimes referred to as large vision-language models (LVLMs), are attracting increasing attention. MLLMs show promising ability in multimodal tasks, such as image captioning [66], visual question answering [22, 75], etc. However, there is a concerning trend associated with the rapid advancement in MLLMs. These models exhibit an inclination to generate hallucinations [69, 76, 137], resulting in seemingly plausible yet factually spurious content.

The problem of hallucination originates from LLMs themselves. In the NLP community, the hallucination problem is empirically categorized into two types [44]: 1) *factuality hallucination* emphasizes the discrepancy between generated content and verifiable real-world facts, typically manifesting as factual inconsistency or fabrication; 2) *faithfulness hallucination* refers to the divergence of generated content from user instructions or the context provided by the input, as well as self-consistency within generated content. In contrast to pure LLMs, research efforts of hallucination in MLLMs mainly focus on the discrepancy between generated **text response** and provided **visual content** [69, 76, 137], *i.e.*, cross-modal inconsistency. This difference suggests that studies in LLMs cannot be seemingly transferred to MLLMs. Therefore, there is a growing need to comprehensively survey recent advancements in MLLMs' hallucination phenomena to inspire new ideas and foster the field's development.

In the realm of computer vision, object recognition is the core task, including sub-tasks such as object classification [60], detection [27], and segmentation [37], etc. Similarly, studies on hallucination in MLLMs primarily focus on object hallucination. In pre-MLLM era, there is a pioneering work on object hallucination in image captioning [90], evaluating object existence by comparing captions and image content. In MLLMs, object hallucination has been empirically categorized into three categories: 1) *category*, which identifies nonexistent or incorrect object categories in the given image; 2) *attribute*, which emphasizes descriptions of the objects' attributes, such as color, shape, material, etc; and 3) *relation*, which assesses the relationships among objects, such as human-object interactions or relative positions. Note that some literature may consider objects counting, objects event, etc., as independent hallucination categories; however, in this work, we include them into *attribute* category.

As numerous studies exist on the underlying causes of hallucinations in LLMs, the unique challenges posed by cutting-edge MLLMs warrant an in-depth investigation. Our analysis specifically targets the unique origins of hallucinations in MLLMs, spanning a spectrum of contributing factors from data, model, training, to the inference stage. In addition, we provide a comprehensive overview of benchmarks and metrics designed specifically for evaluating hallucinations in MLLMs. Then, we review and discuss recent works tailored to mitigate the problem of hallucination from the viewpoints of the identified causes.

Through our comprehensive survey, we aim to contribute to advancing the field of MLLMs and offer valuable insights that deepen understanding of the opportunities and challenges associated with hallucinations in MLLMs. This exploration not only enhances our understanding of the limitations of current MLLMs but also offers essential guidance for future research and the development of more robust and trustworthy MLLMs.

Comparison with existing surveys. In pursuit of reliable generative AI, hallucination stands out as a major challenge, leading to a series of survey papers on its recent advancements. For pure LLMs, there are several surveys [44, 129], describing the landscape of hallucination in LLMs. In

contrast, there are very few surveys on hallucination in the field of MLLMs. To the best of our knowledge, there is only one concurrent work [76], a short survey on the hallucination problem of LVLMs. However, our survey distinguishes itself in terms of both taxonomy and scope. We present a layered and granular classification of hallucinations, as shown in Fig. 1, drawing a clearer landscape of this field. Additionally, our approach does not limit itself to specific model architectures as prescribed in the work of [76], but rather dissects the causes of hallucinations by tracing back to various affecting factors. We cover a larger range of literature both in terms of paper number and taxonomy structure. Furthermore, our mitigation strategies are intricately linked to the underlying causes, ensuring a cohesive and targeted approach.

Organization of this survey. In this paper, we present a comprehensive survey of the latest developments regarding hallucinations in MLLMs. The survey is organized as follows: We begin by providing sufficient context and defining concepts related to LLMs, MLLMs, hallucination, etc. Next, we delve into an in-depth analysis of the factors contributing to hallucinations in MLLMs. Following this, we present a set of metrics and benchmarks employed for evaluating hallucinations in MLLMs. We then elaborate on a range of approaches designed to mitigate hallucinations in MLLMs. Finally, we delve into the challenges and open questions that frame the current limitations and future prospects of this field, offering insights and delineating potential pathways for forthcoming research.

2 DEFINITIONS

2.1 Large Language Models

Before moving to multimodal large language models, it is essential to introduce the concept of large language models. Typically, LLMs encompass a range of transformer-based models that are extensively trained on vast textual datasets. Prominent examples include GPT-3 [8], PaLM [18], LLaMA [99], and GPT-4 [82]. Through scaling both data volume and model capacity, LLMs demonstrate notable emergent capabilities, including In-Context Learning[8], Chain-of-Thought prompting[107] and instruction following[86], among others.

The characteristics and behaviors of LLMs are intricately linked to their training processes. LLMs typically undergo three primary training stages: pre-training, Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF). Below, we provide a concise overview of each stage to facilitate comprehension.

Pre-trianing. Pre-training serves as a fundamental phase in the learning process of LLMs [134]. During this stage, language models engage in autoregressive prediction, wherein they predict the subsequent token in a sequence. By undergoing self-supervised training on vast textual datasets, these models develop an understanding of language syntax, gain access to world knowledge, and enhance their reasoning capabilities. This pre-training process establishes a solid groundwork for the models to undertake subsequent fine-tuning tasks effectively.

Supervised Fine-Tuning. Although pre-training equips LLMs with substantial knowledge and skills, it's important to acknowledge that its primary focus is on optimizing for completion. Consequently, pre-trained LLMs essentially function as completion machines, which may create a misalignment between the objective of predicting the next word within LLMs and the user's objective of obtaining desired responses. To address this disparity, the concept of Supervised Fine-Tuning (SFT) [125] has been introduced. SFT involves further training LLMs using a meticulously annotated set of (instruction, response) pairs, thereby enhancing the capabilities and controllability of LLMs.

Reinforcement Learning from Human Feedback. Although SFT has made strides in enabling LLMs to adhere to user instructions, there remains a need for further alignment with human preferences. Among the various methods, Reinforcement Learning from Human Feedback

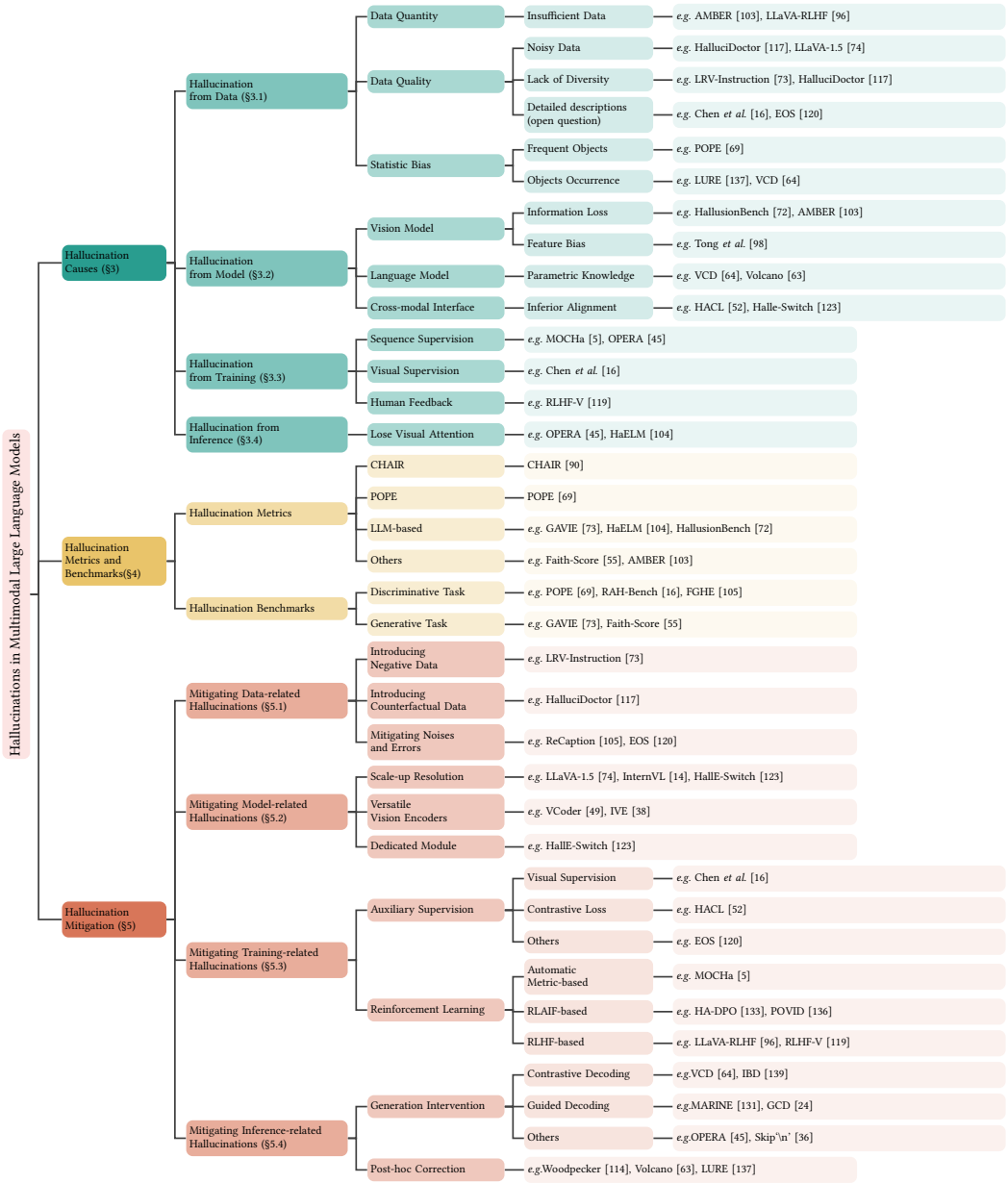

Fig. 1. The main content flow and categorization of this survey.

(RLHF) [19, 84, 94] emerges as a notable approach for achieving alignment through reinforcement learning. RLHF typically employs a preference model [7], trained to predict preference rankings based on prompts and human-labeled responses. To better align with human preferences, RLHF optimizes the LLM to generate outputs that maximize rewards provided by the trained preference model, often utilizing reinforcement learning algorithms like Proximal Policy Optimization (PPO) [93]. This

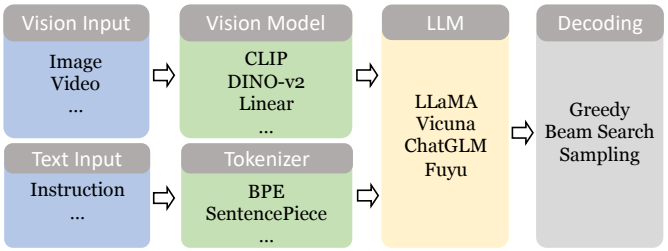

Fig. 2. Popular architecture of multimodal large language model.

integration of human feedback into the training loop has demonstrated effectiveness in enhancing the alignment of LLMs.

2.2 Multimodal Large Language Models

MLLMs [22, 75, 111, 138] typically refers to a series of models that enable LLMs to perceive and comprehend data from various modalities. Among them, vision+LLM is particularly prominent, owing to the extensive research on vision-language models (VLMs) [51, 88, 116] prior to LLMs. As a result, MLLMs are sometimes referred to as vision-LLMs (VLLMs) or large vision language models (LVLMs). The goal of MLLMs is to activate the visual capabilities of LLMs, enabling them to "see" the world via images or videos. Combined with strong reasoning and language generation abilities, MLLMs trigger a series of downstream tasks in multimodal domains, such as image/video captioning and visual question answering. Additionally, MLLMs serve as the foundation for applications in other fields, such as AI assistants, embodied agents, and robotics.

Integrating the two modalities of vision and language involves primarily two types of approaches. The first line of work is built upon off-the-shelf pre-trained uni-modal models. Specifically, these MLLMs usually incorporate a learnable interface between pre-trained visual encoders and LLMs. The interface extracts and integrates information from visual modalities. Such interfaces can be further categorized into 1) learnable query-based and 2) projection layer based. Learnable query-based methods, exemplified by Q-Former [66], as used in MiniGPT-4 [138] and Instruct-BLIP [22], utilize a set of learnable query tokens to capture visual signals via cross-attention. Projection layer-based methods, as widely applied in LLaVA [75], Shikra [12], etc., involve training a linear projection layer or a Multi-Layer Perceptron (MLP) module to transform extracted visual features. Both types of interfaces aim to transform pre-trained visual features into the input space of pre-trained LLMs.

Another line of work is represented by Fuyu-8B [4] and Gemini [97]. Unlike previous methods that leverage pre-trained uni-modal models, these works employ end-to-end training from scratch. Taking Fuyu-8B as an example, it does not employ any pre-trained vision encoder. Instead, it directly inputs image patches and employs a linear projection to transform the raw pixels of each patch into embeddings.

The abstracted pipeline is depicted in Fig. 2. MLLMs take input from both visual and textual modalities, learning from multimodal instructions and responses, which leads to remarkable performance across various multimodal tasks. Regarding the training of MLLMs, we provide a concise overview of the training process for interface-based MLLMs. Given that end-to-end models are closed-source, the training details are unknown. Typically, the training of interface-based MLLMs consists of two stages: 1) pre-training, 2) instruction tuning.

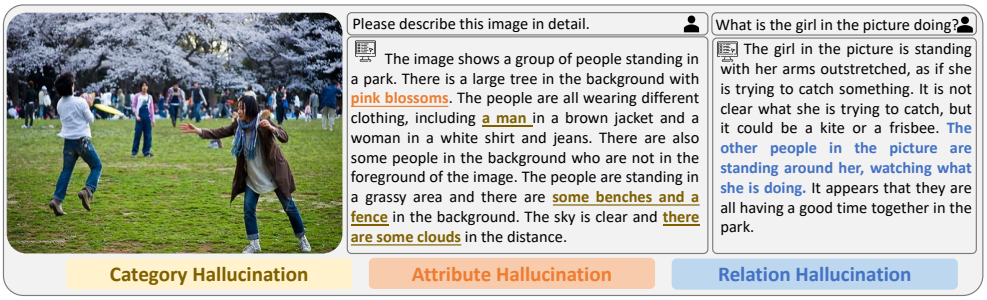

Fig. 3. Three types of typical hallucination.

Pre-training. Given that models from each modality are pre-trained on their respective data, the objective of this pre-training phase is to achieve cross-modal feature alignment. During training, both the pre-trained visual encoder and LLM remain frozen, with only the cross-modal interface being trained. Similar to traditional VLMs training, as exemplified by CLIP [88], web-scale image-text pairs [92] are utilized for training. Given that the final output is at the LLM side, the most widely used loss function in this stage is the text generation loss, typically cross-entropy loss, which aligns with the pre-training of LLMs. Certain studies (e.g., [22, 66]) explore the incorporation of contrastive loss and image-text matching loss to further enhance alignment. After training, the interface module maps the visual features into the input embedding space of the LLM.

Instruction Tuning. Similar to LLMs, after pre-training, the current model still lacks instruction following ability in the multimodal context. During the instruction tuning stage, both machine-generated datasets [75] and human-annotated QA datasets [48, 59, 80] are utilized to enhance the model's ability to comprehend and follow multimodal instructions. Unlike pre-training data, the format and quality of instruction tuning data significantly impact the model's performance. It is usually in the format of *visual content - instruction - response*. Empirical studies also demonstrate that high-quality data significantly enhances the performance of MLLMs. During this stage, there are various options for training, such as fine-tuning LLM parameters in full [75], or using techniques like LoRA [41] to tune specific LLM parameters.

2.3 Hallucinations in Multimodal Large Language Models

Hallucination of MLLM generally refers to the phenomenon where the generated text response does not align with the corresponding visual content. State-of-the-art studies in this field primarily focus on object hallucination, given that objects are central to research in computer vision and multimodal contexts. Regarding inconsistency, two typical failure modes are: 1) missing objects, and 2) describing objects that are not present in the image or with incorrect statements. Empirically, the second mode has been shown to be less preferable to humans. For example, the LSMDC challenge [91] shows that correctness is more important to human judges than specificity. In contrast, the coverage of objects is less perceptible to humans. Thus, object coverage is not a primary focus in studies of object hallucination. Empirically, object hallucination can be categorized into three types: *object category*, *object attribute*, and *object relation*. An example of the three types of hallucination is shown in Fig. 3.

- Category. MLLMs identify nonexistent object categories or incorrect categories in the given image. For example, in Fig. 3, "some benches and a fence", "some clouds", described in the text response do not exist in the given image.

- Attribute. The object categories identified by MLLMs are accurate, while the descriptions of these objects' attributes (such as color, shape, material, content, counting, action, etc.) are wrong. In Fig. 3, "pink blossoms" is hallucinated by the MLLM as the color is inaccurate.
- Relation. All objects and their attributes are described correctly, but the relationships among them (such as human-object interactions or relative positions) do not align with the actual image content. In Fig. 3, "...standing around her, watching..." is a typical example of relation hallucination, as the objects are presented in the image but the relation is inaccurate.

It's worth noting that some literature may categorize objects counting, objects event, etc., as independent hallucination categories. In this work, we classify them under the *attribute* category. The definition of hallucination types aligns well with the domain of compositional generalization [79, 121] of VLMs, which investigates visio-linguistic generalization and reasoning abilities.

3 HALLUCINATION CAUSES

Hallucinations have multifaceted origins, spanning the entire spectrum of MLLMs' capability acquisition process. In this section, we delve into the root causes of hallucinations in MLLMs, primarily categorized into four aspects: *Data*, *Model*, *Training*, and *Inference*.

3.1 Data

Data stands as the bedrock for MLLMs, enabling them to gain cross-modal understanding and instruction-following capabilities. However, it can inadvertently become the source of MLLM hallucinations. This mainly manifests in three aspects: quantity, quality, and statistical bias.

3.1.1 Quantity. Deep learning models are data-hungry, especially large models like MLLMs. The amount of data plays an important role in building robust and reliable MLLMs. Currently, image-text pair datasets [92] and visual QA [48, 80] data are used for training MLLMs. Although these datasets are usually larger than typical datasets in computer vision, they are still far less abundant than the text-only data used for training LLMs in terms of quantity. Insufficient data could potentially lead to problematic cross-modal alignment, resulting in hallucinations [96, 103].

3.1.2 Quality. Given the increasing demand for large-scale training data, heuristic data collection methods are employed to efficiently gather vast volumes of data. While these methods provide extensive data, they offer no guarantee of quality, thereby increasing the risk of hallucinations. Data quality relevant to hallucinations can be further categorized into the following three facets.

- **Noisy data.** As mentioned in the definition section, training MLLMs involves two stages. The pre-training stage employs image-text pairs crawled from the web, which contain inaccurate, misaligned, or corrupted data samples. The noisy data would limit the cross-modal feature alignment [117, 120], which serves as the foundation of MLLMs. As for the instruction tuning data, prevalent methods, such as LLaVA [75], utilize the advanced GPT-4 [82] model to generate instructions. However, ChatGPT is a language model that cannot interpret visual content, leading to the risk of noisy data. Moreover, language models themselves suffer from the issue of hallucination [44], further increasing the risk. LLaVA-1.5 [74] adds human annotated QA data into instruction following and shows improved results, revealing the effect of noisy data.
- **Lack of diversity.** Recent works [73, 117] reveal that the diversity of data also plays a crucial role. For the data used in the two training stages, instruction tuning data are more likely to have this issue since it is usually in a relatively small amount. One prominent property is that most instruction following data samples are composed of conversations regarding the image content. We regard this type of data as *positive instruction*, as it always faithfully reflects the

image content. In contrast, *negative instruction* data [73] and *reject answering* responses [11] are rare in the datasets. Given such training data, one potential drawback observed by recent studies [69, 73] is that current models tend to answer "*Yes*" for any instructions presented to the model, even when a proper answer should be "*No*", leading to hallucination. This phenomenon indicates the effect of data diversity.

- **Detailed descriptions (open question)** The impact of the level of detail in textual descriptions on this matter remains an open question. As discussed in Sec. 2.2, the texts in pre-training data, such as LAION [92], usually describe the salient objects' overall content. While the texts in the instructing tuning stage, such as LLaVA-150K [75], consist of more detailed descriptions. This LLaVA-150K dataset is generated by GPT-4 based on objects recognized by vision models. One recent work [16] argues that within the training data, detailed descriptions related to object position, attributes, and non-salient objects are usually absent. This property results in incomplete cross-modal alignment and deprives the model of grounding ability [62, 126]. However, another work [120] hypothesizes that the text descriptions in the instruction tuning data contain too much details, exceeding the perception limit of MLLMs. When trained with such detailed data, in an attempt to fit the detail level and length distribution of ground truth captions, the model may risk expressing details that it cannot discern from the image, and therefore exhibit hallucinations. The detail level of the training data remains an open question.

3.1.3 Statistic bias. Neural networks, especially large language models, possess an intrinsic tendency to memorize training data, as noted in [23]. The nous (*e.g.*, objects) distribution in the training dataset has strong effects on the behavior of the model. Frequently appeared objects and object co-occurrence are two prominent types of statistical bias, as discussed in [69, 90, 137]. For example, '*person*' might be one of the most frequently appearing objects in the training data. During inference, even if the given image does not contain a person, the model still tends to predict the presence of a person. On the other hand, object co-occurrence refers to the phenomenon that the model will remember which two objects usually 'go together' [90]. For instance, given an image of a kitchen with a refrigerator, MLLMs are prone to answer '*Yes*' when asked about a microwave, as refrigerators and microwaves frequently appear together in kitchen scenes. Bias exists in most datasets. Increasing the scale of data may alleviate the effect, but cannot fully resolve it, given the long-tail distribution of the real world.

3.2 Model

Currently, the architecture of popular MLLMs is composed of several components, usually including pre-trained vision model, pre-trained LLM, and alignment module as we discussed above. Since these models are connected together, instead of end-to-end training from scratch, the error of each module can be accumulated. Inferior and problematic output from each module may lead to hallucinations.

- **Weak vision model.** As mentioned in related works [31, 90, 103], a primary potential reason for hallucination is a weak vision model, which can lead to misclassification or misinterpretation of visual concepts. Even the most powerful vision model may still experience information loss during the encoding process. Weak vision model implies weak perception, which fundamentally undermines the multimodal understanding.
- **Language model prior.** The modern architecture of MLLMs is imbalanced. Usually, the language model is much larger and stronger than the vision model, leading to a tendency to prioritize language-based information [31, 63, 64, 73, 90]. A typical phenomenon is that the knowledge entailed in the language model, also termed as parametric knowledge, can

override the visual content. For example, given an image showing a red banana, which is counter-intuitive in the real world, an MLLM may still respond with "yellow banana", as "banana is yellow" is a deep-rooted knowledge in the LLM. Such language/knowledge prior makes the model overlook the visual content and response with hallucination.

- **Weak alignment interface.** The alignment interface plays an essential role in MLLMs, as it serves as the bridge between the two modalities. A weak alignment interface can easily cause hallucinations. One potential cause of a weak alignment interface is data, as discussed in earlier sections. Apart from that, the interface architecture itself and training loss design also matter [52, 77, 123]. Recent work [52] argues that the LLaVA-like linear projection interface preserves most of the information, but lacks supervision on the projected feature. Visualization in [52] reveals that the features after the projection layer remain distinct from the language embeddings. The distribution gap causes trouble in cross-modal interaction, leading to hallucination. On the other hand, Q-former-like [66] architecture has diverse supervision on the extracted visual feature, aligning it to the language embedding space. However, the use of learnable queries inevitably results in the loss of fine-grained visual information.

3.3 Training

The training objective of MLLMs is basically the same as LLMs, *i.e,* auto-regressive next token prediction loss. This loss is straightforward yet effective and easy to scale up, showing promising performance in language modeling. However, some studies in the field of MLLMs have suggested that the next-token prediction loss might not be suitable for learning visual content due to its complex spatial structure [5, 16]. Additionally, the loss optimizes at the token level, while lacking supervision at the sequence level [5]. Another perspective is that, unlike training LLMs, the RLHF stage is absent in training procedure of MLLMs [96, 119], becoming a potential cause of hallucination.

3.4 Inference

As for inference, some works also argues a potential issue in the auto-regressive generation. During generation, as the sequence length grows, the self-attention will focus more on the previously generated text tokens, *i.e.,* the attention on the visual content is diluted [45, 102–104]. Through visualizing the attention map during generation [45, 104], it can be observed that the generated content focuses more on previous special tokens, such as punctuation, rather than visual content tokens. The issue of 'losing attention' would also lead to the model's output response being irrelevant to the visual content.

4 HALLUCINATION METRICS AND BENCHMARKS

In this section, we present a comprehensive overview of existing hallucination metrics and benchmarks, which are designed to assess the extent of hallucinations generated by existing cutting-edge MLLMs. Currently, the primary focus of these benchmarks is on evaluating the object hallucination of MLLM-generated content. Tab. 1 illustrates a summary of related benchmarks.

CHAIR [90]. As one of the early works, the metric of CHAIR was proposed to evaluate object hallucination in the traditional image captioning task. This is achieved by computing what proportion of words generated are actually in the image according to the ground truth sentences and object segmentations. The computation of the CHAIR metric is straightforward and easy to understand. The metric has two variants: per-instance (denoted as CHAIR_i) and per-sentence

Table 1. Summary of most relevant benchmarks and metrics of object hallucination in MLLMs. The order is based on chronological order on arxiv. In the metric column, Acc/P/R/F1 denotes Accuracy/Precision/Recall/F1-Score.

Benchmark	Venue	Underlying Data Source	Size	Task Type	Metric	Hallucination Type			
						Category	Attribute	Relation	Others
CHAIR [90]	EMNLP'18	MSCOCO [70]	5,000	Gen	CHAIR	✓	✗	✗	✗
POPE [69]	EMNLP'23	MSCOCO [70]	3,000	Dis	Acc/P/R/F1	✓	✗	✗	✗
MME [113]	arXiv'23 Jun	MSCOCO [70]	1457	Dis	Acc/Score	✓	✓	✗	✓
CIEM [42]	NeurIPS-W'23	MSCOCO [70]	78120	Dis	Acc	✓	✗	✗	✗
M-HalDetect [32]	arXiv'23 Aug.	MSCOCO [70]	4,000	Dis	Reward Model Score	✓	✗	✗	✗
MMHal-Bench [96]	arXiv'23 Sep.	Open-Images [61]	96	Gen	LLM Assessment	✓	✗	✗	✓
GAVIE [73]	ICLR'24	Visual-Genome [59]	1,000	Gen	LLM Assessment	Not Explicitly Stated			
NOPE [77]	arXiv'23 Oct.	Open-Images [61]	36,000	Dis	Acc/METEOR [3]	✓	✗	✗	✗
HaELM [104]	arXiv'23 Oct.	MSCOCO [70]	5,000	Gen	LLM Assessment	Not Explicitly Stated			
FaithScore [55]	arXiv'23 Nov.	MSCOCO [70]	2,000	Gen	FaithScore	✓	✓	✓	Obj. Counting
Bingo [21]	arXiv'23 Nov.	Unknown	370	Gen	Human Assessment	✗	✗	✗	Model Bias
AMBER [103]	arXiv'23 Nov.	Web	15,202	Dis & Gen	AMBER Score	✓	✓	✓	✗
RAH-Bench [16]	arXiv'23 Nov.	MSCOCO [70]	3,000	Dis	False Positive Rate	✓	✓	✓	✗
HallusionBench [72]	CVPR'24	Unknown	1,129	Gen	LLM Assessment	✗	✗	✗	Model Diagnose
CCEval [123]	arXiv'23 Dec.	Visual-Genome [59]	100	Gen	LLM-based CHAIR	✓	✗	✗	✗
MERLIM [100]	arXiv'23 Dec.	MSCOCO [70]	31,373	Dis	Accuracy	✓	✗	✓	Obj. Counting
FGHE [105]	arXiv'23 Dec.	MSCOCO [70]	200	Dis	Acc/P/R/F	✓	✓	✓	Obj. Behavior
MOCHa [5]	arXiv'23 Dec.	Synthetic	2,000	Gen	OpenCHAIR [5]	✓	✓	✗	✗
CorrelationQA [35]	arXiv'24 Feb.	Synthetic	7,308	Dis	Acc/AccDrop	✗	✗	✗	Model Bias
VQAv2-IDK [11]	arXiv'24 Feb.	VQAv2 [30]	6,624	Dis	Acc	✗	✗	✗	IK [11]
MHaluBench [13]	arXiv'24 Feb.	MSCOCO [70]	1,860	Gen	Acc/P/R/F	✓	✓	✗	T2I
VHTest [46]	arXiv'24 Feb.	MSCOCO [70]	1,200	Dis & Gen	Acc	✓	✓	✗	✓
Hal-Eavl [53]	arXiv'24 Feb.	MSCOCO [70] & LAION [92]	10,000	Dis & Gen	Acc/P/R/F & LLM Assessment	✓	✓	✓	Obj. Event

(denoted as CHAIR_s):

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|},$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}.$$

In the paper of CHAIR [90], the range of objects is restricted to the 80 MSCOCO objects. Sentence tokenization and synonyms mapping are applied to determine whether a generated sentence contains hallucinated objects. Ground-truth caption and object segmentations both serve as ground-truth objects in the computation. In the MLLM era, this metric is still widely used for assessing the response of MLLMs.

POPE [69]. When used in MLLMs, the work of [69] argues that the CHAIR metric can be affected by the instruction designs and the length of generated captions. Therefore, it proposes a new evaluation metric as well as a benchmark, called Pooling-based Object Probing Evaluation (POPE). The basic idea is to convert the evaluation of hallucination into a binary classification task by prompting MLLMs with simple *Yes*-or-*No* short questions about the probing objects (*e.g.*, Is there a *car* in the image?) Compared to CHAIR, POPE offers increased stability and flexibility. Based on this metric design, it further proposed an evaluation benchmark, drawing 500 images from the MSCOCO dataset. The questions in the benchmark consist of both positive and negative questions. The positive questions are formed based on the ground-truth objects, while the negative questions are built from sampling nonexistent objects. The benchmark is divided into three subsets according to different negative sampling strategy: random, popular, and adversarial. Popular and adversarial sampling are specifically designed to assess frequently appeared objects and object co-occurrence. As an early representative work, POPE serves as a foundation of object hallucination evaluation.

MME [113]. MME is a comprehensive evaluation benchmark for MLLMs. It covers the examination of perception and cognition abilities, encompassing 14 subtasks. Regarding object hallucination,

there are four popular object related subtasks in its perception evaluation, including object existence, count, position, color. Similar to POPE, these tasks are formulated as *Yes*-or-*No* tasks.

CIEM [42] CIEM is a benchmark to evaluate hallucination of MLLMs. Unlike previous works utilize human annotated objects, CIEM is generated using an automatic pipeline. The pipeline takes the text description of a specific image as input and utilize advanced LLMs to generate QA pairs. Although the LLM-based data generation pipeline is not completely reliable, empirical result shows that the generated data has low error rate, around 5%.

MMHal-Bench [96] Comprising 96 image-question pairs, ranging in 8 question categories × 12 object topics, MMHal-Bench is a dedicated benchmark for evaluating hallucination in MLLMs. The 8 question categories cover various types of hallucination, including object attributes, counting, spatial relations, etc. During the evaluation of MMHal-Bench, the GPT-4 model is employed to analyze and rate the responses.

GAVIE [73] GPT4-Assisted Visual Instruction Evaluation (GAVIE) is proposed to assess the LMM output in two different aspects: *Relevancy* to evaluate the instruction-following performance and *Accuracy* to measure the visual hallucination in the LMM output. It comprises a benchmark with 1,000 samples and an evaluation approach. GAVIE evaluates the output of MLLMs in an open-ended manner and does not require human-annotated ground-truth answers. The core idea is to ask the advanced GPT-4 to work as a smart teacher and score the answer by taking image content, human instruction, and model response as input.

NOPE [77] This paper proposes to establish a distinction between object hallucination and incorrectness. a) Object hallucination refers to a phenomenon in VQA where a VL model's response includes a non-existent object, despite the ground truth answer being a negative indefinite pronoun (e.g., "none", "no one", etc). This is denoted as NegP. b) Incorrectness occurs when a VL model fails to accurately respond to a question with a ground truth answer that is anything other than NegP, denoted as Others. This paper argues that the existing VQA datasets have a significantly imbalanced distribution, containing too little NegP data. Therefore, NOPE (Negative Object Presence Evaluation) is proposed in this paper to complement the absent NegP data. During evaluation, traditional metrics, including Accuracy and METEOR, are employed.

HaELM [104] Most LLM-based evaluation benchmarks employ advanced ChatGPT or GPT-4 models to assess the quality of the MLLM response. In contrast, the work of Hallucination Evaluation based on Large Language Models (HaELM) proposes to train a specialized LLM for hallucination detection. It collects a set of hallucination data generated by a wide range of MLLMs, simulates data using ChatGPT, and trains an LLM based on LLaMA [99]. After that, the HaELM model becomes proficient in hallucination evaluation, leveraging reference descriptions of images as the basis of assessment.

FaithScore [55] Considering the natural forms of interaction between humans and MLLMs, FaithScore aims to evaluate free-form responses to open-ended questions. Different from LLM-based overall assessment, FaithScore designs an automatic pipeline to decompose the response, evaluate, and analyze the elements in detail. Specifically, it includes three steps: descriptive sub-sentence identification, atomic fact generation, and fact verification. The evaluation metric involves fine-grained object hallucination categories, including entity, count, color, relation, and other attributes. The final computation of FaithScore is the ratio of hallucinated content.

Bingo [21] Bingo (Bias and Interference Challenges in Visual Language Models) is a benchmark specifically designed for assessing and analyzing the limitations of current popular MLLMs, such as GPT-4V [83]. It comprises 190 failure instances, along with 131 success instances as a comparison. This benchmark reveals that state-of-the-art MLLMs show the phenomenon of bias and interference. Bias refers to the model's susceptibility to generating hallucinatory outputs on specific types of examples, such as OCR bias, region bias, etc. Interference refers to scenarios in which the judgment

of the model can be disrupted, making it more susceptible to hallucination. Due to the small amount of data in this benchmark, the assessment and analysis are mostly conducted by humans.

AMBER [103] Upon the application and evaluation of MLLMs, the tasks can be roughly divided into generative tasks and discriminative tasks. For generative tasks, this paper argues that most existing works rely on additional LLMs, suffering from computational cost. As for discriminative tasks, the most popular evaluation suite is POPE [69]. However, POPE lacks fine-grained hallucination types such as attributes and relations. AMBER (An LLM-free Multi-dimensional Benchmark) is proposed to support the evaluation of generative tasks and discriminative tasks, including object existence hallucination, attribute hallucination, and relation hallucination. It further combines the **CHAIR** [90] metric in generative tasks and **F1** in discriminative tasks to form the AMBER Score as follows:

$$\textbf{AMBER Score} = Avg(1 - \textbf{CHAIR}, \textbf{F1}). \tag{1}$$

RAH-Bench [16] Relation-Associated Hallucination Benchmark (RAH-Bench) can be regarded as an upgraded version of POPE, containing 3,000 yes-or-no questions with their corresponding images. Different from POPE, RAH-Bench further divides the negative questions into three subsets. Each subset contains 500 questions with misleading statements in the different aspects, including: 1) categorical hallucination, 2) attribute hallucination, 3) relation hallucination.

HallusionBench [72] To diagnose and analyze the potential failure modes of MLLMs, HallusionBench evaluates hallucination from a different perspective. It consists of 455 visual-question control pairs, with 346 different figures and a total of 1129 questions covering diverse topics and formats. The questions are divided into two categories: *Visual Dependent* and *Visual Supplement*. The *Visual Dependent* questions are defined as questions that do not have an affirmative answer without the visual context. This setting aims to evaluate visual commonsense knowledge and visual reasoning skills. The *Visual Supplement* questions can be answered without the visual input; the visual component merely provides supplemental information or corrections. This setting is designed to evaluate visual reasoning ability and the balance between parametric memory (language prior) and image context. This division provides a new perspective for understanding and diagnosing MLLMs.

CCEval [123] CCEval focuses on the hallucination evaluation of detailed captions. Traditional caption-based evaluation benchmarks and metrics, like CHAIR, are known to favor short captions. However, short captions often lack detail and contain less information. To address this issue, CCEval randomly samples 100 images from Visual Genome to form a benchmark. In evaluation, GPT-4 is utilized to parse the captions generated by MLLMs and extract objects. Additionally, this work introduces the "coverage" metric on top of CHAIR to ensure that the captions are detailed enough. This metric computes the ratio of objects in the caption that match the ground truth to the total number of ground truth objects. It additionally records the average number of objects as well as the average length of captions as auxiliary metric. Compared with CHAIR, CCEval employs more diverse objects, as reflected in the source of ground truth (Visual Genome vs. COCO) and caption parsing (GPT-4 vs. rule-based tool).

MERLIM [100] MERLIM (**M**ulti-modal **E**valuation benchma**R**k for **L**arge **I**mage-language **M**odels) is a test-bed aimed at empirically evaluating MLLMs on core computer vision tasks, including object recognition, instance counting, and identifying object-to-object relationships. MERLIM contains over 279K image-question pairs, and has a strong focus on detecting cross-modal hallucinations. Interestingly, when organizing the data, a set of edited images is intentionally added. Based on the original image, an inpainting strategy is employed to remove one object instance in the image. With this original-edited image pair, one can compare the output of the target MLLM and identify the hallucinated objects that lack visual grounding.

FGHE [105] Fine-Grained Object Hallucination Evaluation (FGHE) follows a binary classification approach similar to POPE to evaluate MLLMs. However, unlike POPE, FGHE requires a different set of binary questions to measure fine-grained hallucination. The FGHE dataset consists of 50 images and 200 binary questions divided into three categories: (a) multiple-object questions, which verify the relationships between multiple objects in the image; (b) attribute questions, which verify attributes of objects in the image; and (c) behavior questions, which verify behaviors or objects in the image. The questions are manually defined by human annotators on a subset of 50 images from the validation set of the MSCOCO dataset. Similar to POPE, the Accuracy, Precision, Recall, and F1 score are employed as the evaluation metrics.

OpenCHAIR [5] The traditional CHAIR metric relies on the closed list of 80 objects in the MS-COCO dataset, limiting its application. To measure object hallucination in the open-vocabulary settings, *OpenCHAIR* expands CHAIR by relaxing the strong reliance on the closed vocabulary. The 'open-vocabulary' manifests in two ways. Firstly, when building the benchmark, it organizes a dataset consisting of synthetic images with corresponding captions, which include diverse, open-vocabulary objects using a text-to-image diffusion model. Secondly, during computing the metric, CHAIR checks if words or their synonyms (as given by fixed vocabulary lists) are found in ground-truth annotations. In contrast, OpenCHAIR extracts concrete objects from a predicted caption and identifies hallucinated objects from this list by querying an LLM. Similar to CHAIR, the final metric computation is based on the hallucination rate.

Hal-Eval [53] The work of Hal-Eval [53] identifies another type of object hallucination: event hallucination. This type of hallucination fabricates a fictional target and constructing an entire narrative around it, including its attributes, relationships, and actions. This effort further completes the definition of hallucination types. In addition, this work proposes an evaluation benchmark, which encompasses both discriminative and generative evaluation methods. This is achieved by collecting two evaluation subsets, each tailored to the discriminative and generative evaluation methods, respectively.

CorrelationQA [35] CorrelationQA is a dedicated benchmark to quantify the effect of hallucination induced by the spurious visual input. This type of hallucination usually occurs when providing the MLLM with images that are highly relevant but inconsistent with the answers, causing MLLMs to suffer from hallucination. Such visual inputs are defined as 'spurious visual inputs'. This benchmark reveals that most of mainstream MLLMs, including GPT-4V, suffer from hallucination when presented with such spurious visual inputs. This phenomenon indicates that an image can induce MLLMs to instinctively focus on visual content, resulting in responses that are predominantly based on visual information without proper reasoning and thinking.

VQAv2-IDK [11] It has been widely discussed that in the binary QA scenario, MLLMs generally have a bias on answering 'Yes-or-No,' leading to hallucination. In a more detailed question and answer scenario, MLLMs generally tend to respond to the user's question plausibly, even if the desired answer is 'I don't know'. The concept is defined as *'I Know (IK)' hallucination* in the work of [11]. Accordingly, a new benchmark, VQAv2-IDK, is proposed to specifically evaluate this type of hallucination. VQAv2-IDK is a subset of VQAv2 comprising unanswerable image-question pairs as determined by human annotators. In this benchmark, 'I Know (IK)' hallucination has been further categorized into four types:

- Unanswerable: no one can know.
- Don't know: human may not know, but robot might.
- False questions: refers non-existing.
- Not sure: ambiguous to answer.

This benchmark opens a new track for the study of hallucination in MLLMs.

MHaluBench [13] This benchmark does not aim to evaluate the MLLMs themselves. Instead, it is intentionally designed to evaluate the hallucination detection tools of MLLMs, *i.e.*, judge whether a tool can successfully detect the hallucination produced by an MLLM. Thus, the benchmark consists of hallucinatory examples. Specifically, the benchmark unifies image-to-text tasks and the text-to-image tasks into one evaluation suite: cross-modal consistency checking. The hallucinatory examples are generated using leading MLLMs and image generation models, such as LLaVA [75], MiniGPT-4 [138], DALL-E2 [89], and DALL-E3 [6]. During evaluation, the benchmark can be used to compare different hallucination detection methods based on their performance. So far, there are not many dedicated hallucination detection methods. This work serves as a basis for this direction.

VHTest [46] VHTest categorizes visual properties of objects in an image into 1) individual properties, such as existence, shape, color, orientation, and OCR; and 2) group properties, which emerge from comparisons across multiple objects, such as relative size, relative position, and counting. Based on such categorization, the authors further defined 8 visual hallucination modes, providing a very detailed evaluation of hallucination in MLLMs. Furthermore, the collected 1,200 evaluation instances are divided into two versions: "open-ended question" (OEQ) and "yes/no question" (YNQ). Such design enables this benchmark to evaluate both generative and discriminative tasks.

Comparison of mainstream models We compare the mainstream MLLMs on some representative benchmarks, providing a holistic overview of their performance from different dimensions. The results are shown in Table 2 for generative tasks and Table 3 for discriminative tasks. We observe that the MLLMs' performance is not always consistent across different benchmarks. It indicates that different benchmarks have different evaluation dimensions and emphases.

Table 2. Comparison of mainstream MLLMs on **generative** benchmarks. The numbers come from the original papers of these benchmarks.

Model	LLM Size	CHAIR (On AMBER) ↓	AMBER Score ↑	HallusionBench All-Acc ↑	FaithScore (LLaVA-1k) ↑	FaithScore (COCO-Cap) ↑	Hal-Eval In-domain Gen. Acc ↑	Hal-Eval Out-of-domain Gen. Acc ↑
mPLUG-Owl [111]	7B	23.1	54.1	43.93	0.7167	0.8546	27.3	29.5
Multimodal-GPT [28]	7B	-	-	-	0.5335	0.5440	-	-
InstructBLIP [22]	7B	10.3	86.2	45.26	0.8091	0.9392	35.5	41.3
GPT-4V [83]	-	4.3	92.7	65.28	-	-	-	-
LLaVA (7B) [75]	7B	13.5	69.3	-	-	-	23.3	26.3
LLaVA (13B) [75]	13B	-	-	-	0.8360	0.8729	-	-
MiniGPT-4 (7B) [138]	7B	-	-	35.78	0.5713	0.6359	61.4	50.1
MiniGPT-4 (13B) [138]	13B	15.9	76.7	-	-	-	-	-
mPLUG-Owl2 [112]	7B	10.6	84.0	47.30	-	-	-	-
LLaVA-1.5 (7B) [74]	7B	8.6	82.9	-	-	-	44.6	46.4
LLaVA-1.5 (13B) [74]	13B	-	-	46.94	0.8566	0.9425	-	-
CogVLM [106]	7B	7.9	86.1	-	-	-	-	-
Qwen-VL-Chat [2]	7B	-	-	39.15	-	-	-	-
Open-Flamingo [1]	9B	-	-	38.44	-	-	-	-
LRV-Instruction [73]	-	-	-	42.78	-	-	-	-

5 HALLUCINATION MITIGATION

In this section, we present a comprehensive review of contemporary methods aimed at mitigating hallucinations in MLLMs. Based on the properties and perspectives of these methods, we systematically categorize them into four groups. Specifically, we investigate approaches addressing hallucination from Data, Model, Training, and Inference.

Table 3. Comparison of mainstream MLLMs on **discriminative** benchmarks. The numbers come from the original papers of these benchmarks.

Model	LLM Size	MME Existence Score ↑	MME Count Score ↑	MME Position Score ↑	MME Color Score ↑	POPE Random F1-Score ↑	POPE Random F1-Score ↑	POPE Adversarial F1-Score ↑	RAH-Bench F1 Score ↑	AMBER Dis. F1-Score ↑	AMBER Score ↑	Hal-Eval In-domain Event. F1 ↑	Hal-Eval Out-of-domain Event. F1 ↑
mPLUG-Owl [111]	7B	120.00	50.00	50.00	55.00	68.06	66.79	66.82	69.3	31.2	54.1	47	46.6
ImageBind-LLM [34]	7B	128.33	60.00	46.67	73.33	-	-	-	-	-	-	-	-
InstructBLIP [22] (7B)	7B	-	-	-	-	-	-	-	89.1	82.6	86.2	66.2	66.6
InstructBLIP [22] (13B)	13B	185.00	143.33	66.67	153.33	89.29	83.45	78.45	84.7	-	-	-	-
VisualGLM-6B [25]	6B	85.00	50.00	48.33	55.00	-	-	-	-	-	-	-	-
Multimodal-GPT [28]	7B	61.67	55.00	58.33	68.33	66.68	66.67	66.67	-	-	-	-	-
PandaGPT [95]	7B	70.00	50.00	50.00	50.00	-	-	-	-	-	-	-	-
LaVIN [78]	13B	185.00	88.33	63.33	75.00	-	-	-	-	-	-	-	-
Cheetor [67]	7B	180.00	96.67	80.00	116.67	-	-	-	-	-	-	-	-
GPT-4V [83]	-	190.00	160.00	95.00	150.00	-	-	-	-	89.6	92.7	-	-
LLaVA [75] (7B)	7B	-	-	-	-	-	-	-	73.3	32.0	69.3	35.1	14.0
LLaVA [75] (13B)	13B	185.00	155.00	133.33	170.00	68.65	67.72	66.98	71.8	-	-	-	-
LRV-Instruction [73]	7B	165.00	111.67	86.67	165.00	-	-	-	-	-	-	-	-
Lynx [122]	7B	195.00	151.67	90.00	170.00	-	-	-	-	-	-	-	-
MMICL [130]	11B	170.00	160.00	81.67	156.67	-	-	-	-	-	-	-	-
Muffin [118]	13B	195.00	163.33	66.67	165.00	-	-	-	-	-	-	-	-
Otter [65]	7B	195.00	88.33	86.67	113.33	-	-	-	-	-	-	-	-
Qwen-VL-Chat [2]	7B	158.33	150.00	128.33	170.00	-	-	-	-	-	-	-	-
SPHINX [71]	13B	195.00	160.00	153.33	160.00	-	-	-	-	-	-	-	-
VPGTrans [124]	7B	70.00	85.00	63.33	73.33	-	-	-	-	-	-	-	-
BLIVA [43]	11B	180.00	138.33	81.67	180.00	-	-	-	-	-	-	-	-
InfMLLM [135]	13B	195.00	145.00	170.00	195.00	-	-	-	-	-	-	-	-
LLaMA-Adapter V2 [26]	7B	185.00	133.33	56.67	118.33	-	-	-	-	-	-	-	-
MiniGPT-4 [138]	13B	68.33	55.00	43.33	75.00	78.86	72.21	71.37	-	69.3	76.7	48.2	53.0
mPLUG-Owl2 [112]	7B	185.00	155.00	88.33	150.00	-	-	-	-	78.5	84.0	-	-
LLaVA-1.5 [75]	7B	-	-	-	-	-	-	-	-	74.4	82.9	48.9	34.2
CogVLM [106]	7B	195.00	165.00	103.33	160.00	-	-	-	-	80	86.1	-	-

5.1 Data

As discussed in the section on hallucination causes 3, data is one of the primary factors inducing hallucination in MLLMs. For mitigating hallucination, recent works make attempts on data, including introducing negative data [73], introducing counterfactual data [117], and reducing noise and errors in existing dataset [105, 120].

LRV-Instruction [73] LRV-Instruction is proposed to address the issue that existing instruction tuning data primarily focus on positive instruction samples, leading the model to consistently answer 'Yes'. LRV-Instruction is designed to include both positive and negative instructions for more robust visual instruction tuning, where the negative instructions include: 1) 'Nonexistent Object Manipulation': introducing nonexistent objects, activities, attributes, and interactions; 2) 'Existent Object Manipulation': manipulating existent objects with inconsistent attributes; 3) 'Knowledge Manipulation': manipulating knowledge in instructions.

HalluciDoctor [117] This paper addresses the object hallucination problem in MLLMs by calibrating the instruction-tuning dataset. The calibration is conducted from two perspectives. Firstly, it develops a hallucination detection pipeline via consistency cross-checking of multiple MLLMs. Based on the detection result, the hallucinated content can be eliminated. Secondly, this work observes that long-tail distribution and object co-occurrence in the training data are two primary factors of hallucination. Thus, a counterfactual visual instruction generation strategy is proposed to expand the dataset. Using the proposed methods, the instruction tuning data can be balanced and experience reduced hallucination. MLLMs trained on the calibrated dataset are shown to be less prone to hallucination.

ReCaption [105] This work proposes a framework called ReCaption to rewrite the text captions of existing image-text pairs in datasets. The framework comprises two steps: 1) keyword extraction, which extracts verbs, nouns, and adjectives from the caption; and 2) caption generation, which employs an LLM to generate sentences based on the extracted keywords. Ultimately, the framework produces a set of high-quality image-caption pairs. Experiment results show that the model trained

on the rewritten caption dataset has higher accuracy on certain benchmarks, such as the POPE benchmark [69]. Despite the performance improvement, the question of why rewritten captions can reduce hallucination remains an open problem.

EOS Decision [120] Previous work [137] provides an observation that hallucination tends to occur with objects positioned later in the generated descriptions. Intuitively, an ideal scenario is that the MLLM can terminate the generation process in a timely manner. This idea is thoroughly explored in the work of [120] from the perspective of end-of-sequence (EOS) decision. The key insight is that the training data may exceed the perception limit of the MLLM. When trained with such data, the model may attempt to fit the detail level and length distribution of ground truth captions. However, it may risk expressing details that it cannot discern from the image, and therefore exhibit hallucinations. Thus, the authors explored approaches to enhance the model's end-of-sequence (EOS) decision-making process, ensuring timely termination when it reaches the perception limit. Regarding data, this work proposes a data filtering strategy to eliminate harmful training data that could impair the model's ability to end sequences.

5.2 Model

5.2.1 Scale-up Resolution. Enhancing the perception ability of MLLMs has been shown to improve their overall performance and reduce hallucination [14, 74, 75, 123]. One important update when upgrading from LLaVA [75] to LLaVA-1.5 [74] is to scale up the CLIP ViT vision encoder from CLIP-ViT-L-224 to CLIP-ViT-L-336, resulting in considerable performance improvement. Qwen-VL [2] has shown the effectiveness of gradually enlarging image resolution from 224×224 to 448×448. InternVL [2] scales up the vision encoder to 6 billion parameters, enabling processing of high-resolution images. Regarding hallucination, HallE-Switch [123] has investigated the impact of vision encoder resolution on its proposed CCEval benchmark. Among the three studied vision encoders (CLIP-ViT-L-112, CLIP-ViT-L-224, CLIP-ViT-L-336), higher resolution generally results in lower degrees of hallucination. These works indicate that scaling up vision resolution is a straightforward yet effective solution.

5.2.2 Versatile Vision Encoders. Several studies [38, 49, 98] have investigated vision encoders for MLLMs. Typically, the CLIP ViT image encoder is used as the vision encoder in most MLLMs thanks to its remarkable ability to extract semantic-rich features. However, CLIP has been shown to lose some visual details compared to pure vision models like DINO ViT [10]. Therefore, recent studies have proposed complementing this information loss by incorporating visual features from other vision encoders. The work of [98] proposes mixing features from CLIP ViT and DINO ViT. Specifically, it experimented with additive and interleaved features. Both settings show that there is a trade-off between the two types of features. A more dedicated mechanism is needed.

Concurrently, a visual expert-based model proposed in [38] aims to mitigate the information loss caused by the CLIP image encoder. Instead of merely mixing features, this paper enhances the visual perception ability of MLLMs by focusing on knowledge enhancement, relying on two pivotal modules: multi-task encoders and the structural knowledge enhancement module. The multi-task encoders are dedicated to integrating various types of latent visual information extracted by multiple visual encoders. Additionally, the structural knowledge enhancement module is designed to utilize visual tools, such as OCR tools and object detectors, to extract prior knowledge from visual inputs.

Following the approach of the structural knowledge enhancement module in [38], another line of research investigates the utilization of vision tool models to enhance the perception of MLLMs. VCoder [49] utilizes additional perception formats, such as segmentation masks and depth maps, to enhance the object identification ability of the MLLM. Another work [54] ensembles additional

object detection and optical-character recognition models into the MLLM architecture. It also explores various ways to integrate this information, including training-free infusion, LoRA [41] augmented retraining, and LoRA augmented finetuning.

5.2.3 Dedicated Module. Following our previous discussion, the parametric knowledge embedded in the LLM is identified as a significant factor leading to hallucination, directing the generation to be based on language knowledge instead of visual content. To address this issue, the work of [123] proposes training a dedicated "*switch*" module, termed *HallE-Switch*, which controls the extent of parametric knowledge within detailed captions. The detailed implementation is inspired by LM-switch [33], which involves adding a control parameter ϵ serving as a "switching value". The switch module is trained using contrastive training data from both contextual (visual content-related) and parametric datasets. During inference, addressing hallucination can be attempted by tuning the control parameter ϵ.

5.3 Training

5.3.1 Auxiliary supervision. The primary supervision signal of training MLLMs is language modeling loss (implemented as *CrossEntropyLoss*) in both pre-training and finetuning stage. However, such supervision may not be sufficient to process the rich information encoded in the visual content.

Accordingly, the work of [16] constructs a fine-grained vision instruction dataset based on Panoptic Scene Graph (PSG), called Relation-Associated Instruction (RAI-30k). In addition to standard dialogues, each instruction in RAI-30k is associated with a relation annotation in PSG, which includes mask annotations for related instances. With these additional annotations, it further supervises MLLMs with mask prediction loss using a state-of-the-art expert vision model, SAM [57], guiding MLLMs to focus on highly-related image content. With the additional supervision from the mask prediction loss, MLLMs are encouraged to extract features that can better represent these crucial instances, thus generating more accurate responses and mitigating vision hallucination. The intuitive idea of supervising MLLMs with grounding shows promising performance in mitigating hallucination.

Another line of work analyzes the training loss from the perspective of embedding space distribution. As introduced earlier, popular MLLMs typically project the encoded vision features into the input space of a specific LLM. A recent work, HACL [52], argues that an ideal projection should blend the distribution of visual and textual embeddings. However, despite visual projection, a significant modality gap exists between textual and visual tokens, suggesting that the current learned interfaces are not effective in mapping visual representations into the textual representation space of LLMs. This issue potentially exacerbates the tendency for MLLMs to generate more hallucinations. Therefore, HACL proposes enhancing the alignment between visual and textual representations through contrastive loss. Texts with hallucinations are used as hard negative examples for image anchors. The loss pulls representations of non-hallucinating text and visual samples closer while pushing representations of non-hallucinating and hallucinative text apart. Experiment results show that this method not only reduces hallucination but also enhances performance on other popular benchmarks.

Recalling the work of EOS Decision [120], to teach the model to terminate the generation process properly, this work also designs a learning objective, termed Selective EOS Supervision, in addition to the data filtering strategy. This is achieved by simply modifying the Maximum Likelihood Estimation (MLE), enabling the model to mitigate hallucination through learning from regular instruction data.

5.3.2 Reinforcement Learning. Reinforcement learning (RL) is introduced to train MLLMs for mitigating hallucinations by conducting the following perspectives: 1) Automatic Metric-based

Optimization, 2) Reinforcement Learning from AI Feedback, 3) Reinforcement Learning from Human Feedback.

Automatic Metric-based Optimization. Motivated by the limitation of LLMs (and MLLMs) training, which is unable to optimize at the sequence level, the MOCHa [5] framework is proposed to apply reinforcement learning. This work aims to improve the accuracy and relevance of image captioning, thereby reducing hallucination. The framework introduces three metric-based objectives to guide the reinforcement learning process for image captioning: 1) Natural Language Inference (NLI) for fidelity, focusing on the accuracy of the caption in describing the image content; 2) BERTScore [127] for semantic adequacy, assessing the relevance and richness of the description; and 3) Kullback–Leibler (KL) divergence for regularization, which constrains the model to stay close to its initial policy. The framework incorporates these objectives into a multi-objective reward function for reinforcement learning. Subsequently, the proximal policy optimization reinforcement learning algorithm is employed to maximize the expected reward. By promoting the creation of accurate, contextually appropriate, and varied descriptions, the hallucination of MLLM can be mitigated.

Reinforcement Learning from AI Feedback (RLAIF). HA-DPO [133] addresses hallucination as a preference selection problem by training models to prioritize accurate responses over hallucinatory ones. To achieve this goal, HA-DPO initially constructs a high-quality dataset. Specifically, it first utilizes MLLMs to generate descriptions corresponding to images, then employs GPT-4 to detect whether these descriptions contain hallucinations. If hallucinations are detected, the descriptions are rewritten. Thus, HA-DPO constructs a dataset that includes both accurate descriptions (positive samples) and hallucinatory descriptions (negative samples). HA-DPO then trains the model using these sample pairs, enabling it to distinguish between accurate and hallucinatory descriptions. This goal is achieved through direction preference optimization (DPO), which optimizes a specific loss function designed to maximize the model's preference for positive samples while minimizing its preference for negative samples.

A concurrent work, Silkie [68], introduces a similar approach of utilizing preference-based reinforcement learning to enhance the faithfulness of MLLMs. Specifically, it emphasizes the concept of reinforcement learning from AI feedback (RLAIF) by distilling preferences from a more robust MLLM, *i.e.*, GPT-4V [83]. Responses are first generated by models from 12 MLLMs, and then assessed by GPT-4V. The constructed dataset, termed as VLFeedback, contains preferences distilled from GPT-4V and is utilized to train other MLLMs through direct preference optimization.

A more recent work, POVID [136], challenges the assumption underlying previous DPO-based methods. These methods rely on the traditional preference data generation process in LLMs, where both preferred and dispreferred responses may potentially be incorrect. Therefore, this work proposes the Preference Optimization in VLLM with AI-Generated Dispreferences (POVID) framework, aiming to exclusively generate dispreferred feedback data using AI models. The dispreferred data is generated by: 1) utilizing GPT-4V to introduce plausible hallucinations into the answer, and 2) provoking inherent hallucination by introducing noise into MLLMs. In the DPO optimization framework, the ground-truth multimodal instructions serves as the preferred answers.

Reinforcement Learning from Human Feedback (RLHF). HalDetect [32] first introduces the M-HalDetect dataset for detecting hallucinations, which covers a wide range of hallucinatory content, including non-existent objects, unfaithful descriptions, and inaccurate relationships. It then proposes a multimodal reward model to detect hallucinations generated by MLLMs. The reward model is trained on the M-HalDetect dataset to identify hallucinations in the generated text. To utilize the trained reward model to reduce hallucinations, the authors introduced Fine-grained Direct

Preference Optimization (FDPO). FDPO uses fine-grained preferences from individual examples to directly reduce hallucinations in generated text by enhancing the model's ability to distinguish between accurate and inaccurate descriptions.

LLaVA-RLHF [96] also try to involve human feedback to mitigate hallucination. It extends the RLHF paradigm from the text domain to the task of vision-language alignment, where human annotators were asked to compare two responses and pinpoint the hallucinated one. The MLLM is trained to maximize the human reward simulated by an reward model. To address the potential issue of *reward hacking*, *i.e.,* achieving high scores from the reward model does not necessarily lead to improvement in human judgements, it proposes an algorithm named Factually Augmented RLHF. This algorithm calibrates the reward signals by augmenting them with additional information such as image captions.

Similarly, RLHF-V [119] also employs the RLHF paradigm to enhance the pre-trained MLLM. Specifically, this work emphasizes two improvements: 1) at the data level, it proposes to collect human feedback in the form of fine-grained segment-level corrections, providing a clear, dense, and fine-grained human preference. 2) at the method level, it proposes dense direct preference optimization (DDPO) that directly optimizes the policy model against dense and fine-grained segment-level preference.

Another similar work, ViGoR [110], also designs a fine-grained reward model to update pre-trained MLLMs, aiming to improve visual grounding and reduce hallucination. The reward modeling in this work encompasses both human preferences and automatic metrics. Specifically, it collects human judgment and preferences for the responses generated by MLLMs by asking crowd-workers to provide fine-grained feedback at the sentence level. The collected human preference data is used to train a reward model. Additionally, it leverages advanced vision perception models to automatically score the grounding and fidelity of the text generated by an MLLM. Both sources are combined into a single reward score during the reinforcement learning procedure.

5.3.3 Unlearning. Unlearning refers to a technique designed to induce a model to 'forget' specific behaviors or data, primarily through the application of gradient ascent methods [9]. Recently, unlearning for LLMs has been receiving increasing attention [50], effectively eliminating privacy vulnerabilities in LLMs. In the context of MLLMs, a recent work [109] introduces the Efficient Fine-grained Unlearning Framework (EFUF), applying an unlearning framework to address the hallucination problem. Specifically, it utilizes the CLIP model to construct a dataset comprised of both positive samples and negative (hallucinated) samples. The training loss is applied separately for positive and negative at the sub-sentence level. To the best of our knowledge, EFUF [109] is the first and only work that applies the unlearning framework to the task of hallucination mitigation, opening up a new path for future research.

5.4 Inference

5.4.1 Generation Intervention.

Contrastive Decoding. VCD (Visual Contrastive Decoding) [64] is designed to suppress the statistical biases and language priors in MLLMs during the decoding phase. The main assumption of VCD is that a distorted visual input would lead to text responses with more biases and priors. Thus, by contrasting output distributions derived from original and distorted visual inputs, VCD aims to effectively reduce the over-reliance on statistical bias and language priors. Specifically, the decoding probability distribution is calibrated using the reference (distorted) distribution.

Following the same idea of contrastive decoding, IBD [139] proposes an image-biased decoding strategy. Specifically, IBD involves computing a more reliable next-token probability distribution by contrasting the predictions of the original model with those of an image-biased model, which

focuses more on the image information. The image-based model is created by modifying the attention weight matrix structure within the original model, without altering its parameters. This approach emphasizes the knowledge of the image-biased model and diminishes that of the original model, which may be text-biased. Thus, it encourages the extraction of correct content while suppressing hallucinations resulting from textual over-reliance.

Guided Decoding. MARINE [131] proposes a training-free approach. It employs an additional vision encoder for object grounding and utilizes the grounded objects to guide the decoding process. Specifically, it innovatively adapts the classifier-free guidance [40] technique to implement guided decoding, showing promising performance in emphasizing the detected objects while reducing hallucination in the text response.

Similarly, GCD [24] devises a CLIP-Guided Decoding (GCD) approach. It first verifies that CLIPScore [88] can effectively distinguish between hallucinated and non-hallucinated sentences through a series of studies across different models and datasets. Based on this conclusion, it further recalibrates the decoding process of MLLMs, including two steps: 1) reliability scoring, which designs a (CLIP-based) scoring function aiming to assign higher scores to candidate responses that are less likely to be hallucinated, and 2) guided sentence generation, which generates responses based on this scoring. This is implemented in a similar way to beam search but at the sentence level.

HALC [15] provides a key insight that when decoding a specific token in the MLLM, identifying a token-wise optimal *visual context* to provide the most informative visual grounding can effectively reduce hallucination. *Visual context* refers to the visual tokens that can be grounded from the generated text response. An oracle study showed that decoding from the provided optimal visual contexts eliminates over 84.5% of hallucinations. Based on the insight and observation, the authors designed mechanisms to locate the fine-grained visual information to correct each generated token that might be hallucinating. This is essentially a visual content-guided decoding strategy. In addition to token-level correction, HALC also incorporates a *matching-based beam search* that utilizes a visual matching score to steer the generation of the final outputs, balancing both object hallucination mitigation and text generation quality.

Others. The work of OPEAR [45] makes an interesting observation that most hallucinations are closely tied to the knowledge aggregation patterns manifested in the self-attention matrix, *i.e.,* MLLMs tend to generate new tokens by focusing on a few summary tokens rather than all the previous tokens. Such a partial over-trust inclination results in neglecting image tokens and describing the image content with hallucination. Based on this observation, a decoding method for MLLMs grounded in an **O**ver-trust **P**enalty and a **R**etrospection-**A**llocation strategy is proposed. First, a penalty term on the model logits is introduced during the MLLM beam-search decoding process to mitigate the over-trust issue. Additionally, to handle the hard cases that cannot be addressed by the penalty term, a more aggressive strategy called the rollback strategy is proposed to retrospect the presence of summary tokens in the previously generated tokens and reallocate the token selection if necessary.

Another interesting study observes that the hallucination of MLLMs seems to be easily triggered by paragraph break '\n\n' [36]. Based on this observation, this work proposes two simple methods to reduce hallucination by avoiding generating '\n' during generation. First, intuitively, users can design the prompt to instruct the model to output responses within one paragraph, avoiding '\n'. Besides, the authors tried to alter the output logits during generation by manually lowering the probability of generating '\n'. Experimental results show that this simple strategy can alleviate hallucination on popular benchmarks.

5.4.2 Post-hoc Correction. Post-hoc correction refers to first allowing the MLLM to generate a text response and then identifying and eliminating hallucinating content, resulting in less hallucinated output. This is usually achieved by grounding on visual content [114], pre-trained revisior [137], and self-revision [63].

Woodpecker [114] is an early attempt on hallucination detection and correction. Similar to how a woodpecker heals trees, Woodpecker picks out and corrects hallucinations from the generated text. The key idea of Woodpecker is to extract key concepts from the generated text and validate them using visual content. Subsequently, the hallucinated concepts can be detected and corrected accordingly. Specifically, it consists of five stages: 1) *Key concept extraction* identifies the main objects mentioned in the generated sentences; 2) *Question formulation* asks questions around the extracted objects; 3) *Visual knowledge validation* answers the formulated questions via expert models; 4) *Visual claim generation* converts the above Question-Answer (QA) pairs into a visual knowledge base; 5) *Hallucination correction* modifies the hallucinations and adds the corresponding evidence under the guidance of the visual knowledge base. Woodpecker is a training-free method, where each component can be implemented using either hand-crafted rules or off-the-shelf pre-trained models.

Another line of work rectifies the generated text using a dedicatedly trained revisor model. Specifically, inspired by denoising autoencoders [101], which are designed to reconstruct clean data from corrupted input, LURE [137] employs a hallucination revisor that aims to transform potentially hallucinatory descriptions into accurate ones. To train such a revisor model, a dataset has been constructed. Each example in this dataset consists of an image accompanied by a hallucinatory description, with the correct description serving as the target output. The hallucinatory descriptions are generated by modifying the accurate descriptions using GPT-3.5. These adjustments are guided by factors related to object hallucination, including co-occurrence, object uncertainty, and object position. After that, the authors fine-tuned an MLLM using this dataset to serve as a revisor, which is used as an additional step for rectifying the output of an MLLM during generation.

Similar idea has also been explored in Volcano [63]. It introduces a self-revising mechanism to reduce hallucination. It consists of four stages: 1) generate initial response; 2) generate feedback for the initial response; 3) revise the response using this feedback; 4) compare the responses before and after revision to decide which one is better. Stages 2-4 are repeated iteratively. To provide better feedback and decision-making, the model is fine-tuned on a curated dataset. The dataset is organized using ChatGPT.

LogicCheckGPT [108] is a more recent self-revising-based hallucination mitigation method. Unlike Volcano [63], which revises the generated response with the help of *general* feedback, LogicCheckGPT delves into the *logical consistency* of MLLMs' responses. Specifically, the approach can be formulated into two stages: the first stage involves inquiring attributes of objects, followed by inquiring objects based on attributes. Whether their responses can form a logical closed loop serves an indicator of object hallucination. If the ratio of closed loops to the total number of questions exceeds a certain threshold, rectify the hallucinated objects by prompting the MLLM.

6 CHALLENGES AND FUTURE DIRECTIONS

The research of hallucination in MLLMs is still at early stage, remaining a variety of research problems to be explored. In this section, we delve into the challenges and future directions of this pivotal domain.

6.1 Data-centric Challenges and Innovations

The reliance of MLLMs on large volumes of data presents significant challenges in terms of data quality, diversity, and bias. In Sec. 3.1, previous works have identified several core issues that may

cause hallucination. In order to improve the accuracy and reliability of hallucinated content, it is crucial to ensure that MLLMs have access to high-quality and diverse training data. Future research should focus on developing techniques for data collection, augmentation, and calibration. Firstly, collecting enough data at the initial stage is crucial to address the data scarcity issue and increase data diversity. Secondly, data augmentation is an effective solution to further expand the size of data. Finally, exploring methods for re-calibrating existing datasets is crucial. This includes eliminating biases, promoting diversity and inclusivity, and mitigating other potential issues that may induce hallucinations.

6.2 Cross-modal Alignment and Consistency

The key challenge of multimodal hallucination is the cross-modal consistency issue. Ensuring that generated content remains consistent and contextually relevant to the input modality requires sophisticated techniques for capturing and modeling cross-modal relationships. The direction of cross-modal alignment encompasses both MLLMs training and hallucination evaluation. Regarding training, future research should explore methods for aligning representations between different modalities. Achieving this goal may involve designing more advanced architectures, introducing additional learning objectives [52], or incorporating diverse supervision signals [16]. Regarding evaluation, cross-modal consistency checking has been a long-standing topic, ranging from multi-modal understanding [66, 88] to text-to-image generation [13, 17]. Drawing on proven experiences from these domains to improve the assessment of MLLM hallucination, or unifying them into an overall framework, may be promising research directions.

6.3 Advancements in Model Architecture

Despite recent advancements in model architectures of LLMs and MLLMs, designing effective architectures specifically tailored to hallucination remains a challenge. Developing advanced model architectures capable of capturing complex linguistic structures and generating coherent and contextually relevant output based on input visual content is essential for improving the performance of MLLMs. Future research can explore innovative architectural designs based on identified causes of hallucination. This includes developing stronger visual perception models, innovative cross-modal interaction modules capable of transferring cross-modal information seamlessly, and novel large language model architectures faithful to input visual content and text instructions, etc.

6.4 Establishing Standardized Benchmarks

The lack of standardized benchmarks and evaluation metrics poses significant challenges in assessing the degree of hallucination in MLLMs. In Table 1, it can be observed that there is a variety of evaluation benchmarks, but a lack of unified standards. Among them, one of the most popular benchmarks might be POPE [69], which employs a 'Yes-or-No' evaluation protocol. However, this binary-QA manner does not align with how humans use MLLMs. Accordingly, some benchmarks specifically evaluate the hallucination of MLLMs in the (free-form) generative context. Yet, they often rely on external models, such as vision expert models or other LLMs, which limits their widespread application. Moving forward, future research can investigate standardized benchmarks that are theoretically sound and easy to use. Otherwise, research on methods to mitigate hallucinations may be built on an incorrect foundation.

6.5 Reframing Hallucination as a Feature

Recently, discussions on social media [56] have suggested that hallucination can be regarded as an inherent feature of LLMs and MLLMs. The models are like dream machines. Human users direct their dreams with prompts. The prompts start the dream, and based on the model's hazy

recollection of its training documents, most of the time the result goes someplace useful. It's only when the dreams enter deemed factually incorrect territory that we label them as 'hallucinations'. From this perspective, leveraging hallucination capabilities as a feature in downstream applications presents exciting opportunities for enhancing user experiences and enabling new use cases. As humans are the end-users of these models, the primary goal is to enrich human user experiences. Future research may switch the optimization objective from specific cross-modal benchmarks to human experience. For example, Some content may cause hallucinations but will not affect the user experience, while some content may. Alternatively, integrating hallucination to inspire more creative ideas in real-world applications could also be intriguing.

6.6 Enhancing Interpretability and Trust

Existing methods for hallucination mitigation are primarily based on empirical observations of specific patterns, such as skipping the '\n' token and penalizing over-trust tokens. However, despite the impressive improvements achieved on specific benchmarks, understanding the underlying mechanisms and decision-making processes remains challenging. Future research should focus on developing techniques for interpreting and explaining the generation process of MLLMs, thereby providing insights into the factors influencing hallucinated content. This includes investigating methods for visualizing model internals, identifying salient features and linguistic patterns, and tracing the generation process from input to output. Enhancing the interpretability of MLLMs will not only improve our understanding of model behavior but also enable users to better assess hallucinated content in practical applications.

6.7 Navigating the Ethical Landscape

As MLLMs become increasingly proficient at generating realistic text, ethical considerations surrounding the use of generated content become paramount. Especially in the context of hallucination, the generated response may contain severely concerning ethical content, amplifying the importance of the problem. Addressing ethical concerns related to misinformation, bias, privacy, and societal impact is crucial for promoting responsible AI practices in the development and deployment of MLLMs. In addition to addressing typical object hallucination, future research on MLLM hallucinations should prioritize ethical considerations throughout the entire lifecycle of MLLM development, from data collection and model training to deployment and evaluation.

7 CONCLUSION

Based on powerful large language models, multimodal large language models demonstrate remarkable performance across various multimodal tasks. However, the phenomenon of hallucination presents a significant challenge to the practical applications of MLLMs, giving rise to undeniable concerns about safety, reliability, and trustworthiness. In this comprehensive survey, we conducted a thorough examination of hallucinations within multimodal large language models, focusing on their underlying causes, evaluation metrics, benchmarks, and mitigation methods. Despite considerable progress, hallucination remains a complex and persistent concern that warrants ongoing investigation. The challenge of hallucination in multimodal large language models remains compelling, requiring continuous scrutiny and innovation. In light of these challenges, we have outlined several promising future directions in this burgeoning domain. Through navigating the intricate landscape of hallucinations, we aim for this survey to serve as a foundational resource for addressing the complexities of hallucination phenomena in MLLMs. We envision this survey empowering researchers and practitioners to dedicate efforts to advancing research and developing robust solutions in this vital area of study.