A Survey on Hallucination in Large Vision-Language Models

Hanchao Liu* , **Wenyuan Xue*** , **Yifei Chen** , **Dapeng Chen** , **Xiutian Zhao** ,
Ke Wang , **Liping Hou** , **Rongjun Li** , **Wei Peng**

IT Innovation and Research Center, Huawei Technologies

{liuhanchao2, xuewenyuan1, chenyifei14, chendapeng8, zhaoxiutian,
wangke215, houliping1, lirongjun3, peng.wei1}@huawei.com

Abstract

Recent development of Large Vision-Language Models (LVLMs) has attracted growing attention within the AI landscape for its practical implementation potential. However, "hallucination", or more specifically, the misalignment between factual visual content and corresponding textual generation, poses a significant challenge of utilizing LVLMs. In this comprehensive survey, we dissect LVLM-related hallucinations in an attempt to establish an overview and facilitate future mitigation. Our scrutiny starts with a clarification of the concept of hallucinations in LVLMs, presenting a variety of hallucination symptoms and highlighting the unique challenges inherent in LVLM hallucinations. Subsequently, we outline the benchmarks and methodologies tailored specifically for evaluating hallucinations unique to LVLMs. Additionally, we delve into an investigation of the root causes of these hallucinations, encompassing insights from the training data and model components. We also critically review existing methods for mitigating hallucinations. The open questions and future directions pertaining to hallucinations within LVLMs are discussed to conclude this survey.

1 Introduction

In the rapidly evolving realm of artificial intelligence, large language models (LLMs) such as GPT-4 [OpenAI, 2023], LLaMA [Touvron *et al.*, 2023a], and LLaMA2 [Touvron *et al.*, 2023b] have made remarkable strides in natural language understanding (NLU) and generation (NLG). To harness the NLU and NLG capabilities of LLMs for vision-language tasks, a prevalent approach is inserting visual features as supplementary inputs to LLMs and aligning them with textual features. This method has been adapted in several large vision-language models (LVLMs) such as MiniGPT-4 [Zhu *et al.*, 2023], LLaVA [Liu *et al.*, 2023c], and LLaVA-1.5 [Liu *et al.*, 2023b]. Despite the promising results demonstrated by existing LVLMs, an unignorable issue has been impeding their practical application: *hallucination*. Hallucination

---

*Equal Contribution

Figure 1: Hallucination examples in LVLMs. Hallucination symptoms may manifest as deficiencies in various vision-language tasks like *judgment* and *description*, or factual errors in different visual semantics, such as objects, attributes, and relations.

in LVLM denotes a disagreement between factual content of images and corresponding generated textual content, akin to the solely textual hallucination encountered in large language models [Huang *et al.*, 2023a].

Existing studies [Rohrbach *et al.*, 2018; Li *et al.*, 2023b; Hu *et al.*, 2023; Zhai *et al.*, 2023] have addressed the issue of hallucination in image captioning models with a predominant focus on the "presence of objects", specifically whether the objects depicted in a given image are accurately described by the text generated by the model. In contrast to image captioning models trained within closed domains, LVLMs leverage the robust understanding and expressive capabilities of LLMs to obtain more detailed and explicable generated descriptions. However, these enhanced abilities also diversify and potentially exacerbate hallucination, which is not only limited to object existence, but also manifested in the descriptive errors such as attribute and relation errors. We focus on visual hallucinations, referring to all inconsistencies between the semantic content conveyed by the image and the textual content generated by the model.

Hallucination symptoms in LVLMs are multifaceted. From a cognitive perspective, hallucinations can manifest as errors in true/false judgments and inaccuracies in the descrip-

tion of visual information. For instance, as illustrated in the first example of Figure 1, the model ill responds to questions such as "Is there a cat in the image?" and "Are there four birds in the image?", demonstrating flawed factual discernment. Additionally, the second example shows an misalignment of the generated description and the visual facts. Meanwhile, a visual semantics perspective provides us with a ternary taxonomy: hallucination on object, attribute, and relationship. For example, the model generates non-existent objects in the image such as "laptop" and "small dog", provides incorrect attribute descriptions such as describing the man being "long-haired", and makes inaccurate assertions about the relationships between objects such as claiming the bicycle is "in front of" the man. Current methods evaluate these hallucination in LVLMs based on the models' cognitive performance, focusing on two main aspect: non-hallucinatory generation and hallucination discrimination. The former involves a detailed analysis of the hallucinatory elements in the model's response and quantifying their proportion. The latter, on the other hand, only requires a binary judgment of whether the response comprises any hallucinatory content. These approaches are comprehensively discussed in §3.

While the LLM community has extensively discussed the causes of hallucinations in LLMs, LVLMs' visual modality introduces unique challenges in analyzing these occurrences. We conduct a thorough analysis of hallucinations in LVLMs, focusing on the training data and model characteristics. Our analysis indicates that hallucinations in LVLMs are caused not only by the generative nature of LLMs but also by biased training data, the inability of vision encoders to accurately ground images, misalignment among different modalities, insufficient context attention, and many other factors. Following this, we provide a comprehensive overview of the existing hallucination mitigation methods. Corresponding to the causes, current mitigation approaches predominantly focus on the optimization of training data, refinement of various modules within LVLMs, and post-processing of generated outputs. These methods are employed to diminish the occurrence of hallucinations, thereby yielding more faithful responses. In the end, we list several considerable directions for developing the research of hallucinations in LVLMs.

In summary, this study aims to provide insights for the development of LVLMs and explore the opportunities and challenges related to LVLM hallucinations. This exploration not only helps us understand the limitations of current LVLMs, but also offers important guidance for future research and the development of more reliable and efficient LVLMs.

2 Hallucination in the Era of LVLM

2.1 Large Vision-Language Models

LVLMs are advanced multimodal models that process visual and textual data to solve compound tasks involving both vision and natural language. Incorporating LLMs' capabilities, LVLMs are an evolution of previous Vision-Language Pretrained Models (VLPMs) [Long *et al.*, 2022].

LVLM architectures typically comprise three components: a visual encoder, a modality connection module, and a LLM. The visual encoder, often an adaption of the CLIP vision encoder [Radford *et al.*, 2021], transforms input images into visual tokens. The connection module is designed to align visual tokens with the word embedding space of the LLM, ensuring that the LLM can handle the visual information. There are various methods for modality alignment, including cross-attention [Alayrac *et al.*, 2022], adapters [Gao *et al.*, 2023], Q-Formers [Li *et al.*, 2023a; Dai *et al.*, 2023a; Zhu *et al.*, 2023], and simpler structures such as linear layers or multi-layer perceptrons (MLP) [Liu *et al.*, 2023c; Chen *et al.*, 2023b; Liu *et al.*, 2023b]. The LLM functions like the central processing unit within LVLMs, receiving aligned visual and textual information and subsequently synthesizing this information to produce responses.

The training of LVLMs involves two key stages: (1) pretraining, where LVLMs acquire vision-language knowledge from aligned image-text pairs, and (2) instruction-tuning, during which LVLMs learn to follow human instructions using a varied task dataset. Following these stages, LVLMs can efficiently process and interpret visual and textual data, enabling them to reason about concepts in compound multimodal tasks like Visual Question Answering (VQA).

2.2 Hallucination in LVLMs

Hallucination in LVLMs refer to contradictions between the visual input (taken as 'fact') and the textual output of a LVLM. Through the lens of visual-language tasks, LVLM hallucination symptoms can be interpreted as deficiencies in judgement or description.

A judgement hallucination occurs when the model's response to a user's query or statement is in disagreement with the actual visual data. For example, as shown in Figure 1, when being presented with an image depicting three birds and inquired whether there is a cat in the picture, the model wrongly affirms with a "Yes". A description hallucination, on the other hand, is a failure to faithfully depict the visual information. As exemplified in lower part of Figure 1, where the model inaccurately describes the man's hair, the quantity and color of cups, the bicycle's location, and fabricates non-existent objects such as laptop and dog.

From a semantic perspective, such misalignment can be characterized by claiming nonexistent objects, incorrect object attributes, or inaccurate object relations, as highlighted in different colors.

2.3 Unique Challenges regarding Hallucination in LVLMs

LVLMs handle vision-language tasks by incorporating both visual and language modules. However, the integration also pose unique challenges in hallucination detection, causal inference, and mitigation methods.

Hallucination Detection Difficulties   The multimodal nature of LVLM impedes detecting hallucinations. LVLM Hallucinations may manifest across various semantic dimensions, including, but not limited to the object, attribute, and relation [Zhai *et al.*, 2023; You *et al.*, 2023]. To detect these hallucinations comprehensively, the model is tasked not only on natural language understanding but also employing fine-grained visual annotations and aligning them precisely with the generated text.

Intertwined Causes The causes of hallucinations in LVLMs are often multifaceted. On one hand, there are data-related issues shared by both LLMs and LVLMs [Hu *et al.*, 2023] such as misinformation, biases, and knowledge boundary limits. However, LVLMs uniquely suffer from their incorporation of visual data. For instance, visual uncertainty, such as unclear or distorted images, can exacerbate the language priors and statistical biases in LVLMs, leading to severer hallucinations [Liu *et al.*, 2023a].

Compound Mitigation Methods Besides adapting LLM-targeted hallucination mitigation methods, such as data quality enhancement, encoding optimization, and aligning with human preferences, LVLM-specific methods also include refining visual representations and improving the multi-modal alignment. For instance, enlarging visual resolution has been suggested to effectively reduce hallucinations [Bai *et al.*, 2023]. Nonetheless, training high-resolution visual encoders with extensive data can be highly resource-demanding. Therefore, it is imperative to investigate more cost-effective strategies for augmenting visual representations. Moreover, the significant gap between visual and textual tokens suggest that improving the vision-language token alignment can potentially lower the incidence of hallucinations [Jiang *et al.*, 2023].

3 Evaluation Methods and Benchmarks

Establishing the concepts of hallucination in LVLM, we then turn to examine existing LVLM hallucination evaluation methods and benchmarks. Corresponding to the hallucination symptoms in description and judgement tasks mentioned in Figure 1, current evaluation methods can be classified into two major approaches: (1) assessing the model's ability of non-hallucinatory content generation, and (2) evaluating the model's ability of hallucination discrimination, as shown in Figure 2. Similarly, the benchmarks can also be categorized into discriminative and generative ones on the basis of evaluation tasks, as demonstrated in Table 1.

3.1 Evaluation on Non-Hallucinatory Generation

Evaluating non-hallucinatory generation is to measure the proportion of hallucinated content in the outputs. Currently, there are primarily two main types: *handcrafted pipeline* and *model-based end-to-end* methods.

Handcrafted Pipeline Methods Handcrafted pipeline methods feature strong interpretability by manually designing multiple steps with specific and clear objectives. CHAIR [Rohrbach *et al.*, 2018] targets evaluating object hallucinations in image captioning by quantifying differences of objects between model generation and ground-truth captions. However, such conventional methods struggle with LVLMs' vast object categories [Lu *et al.*, 2018]. To address this, CCEval [Zhai *et al.*, 2023] employs object alignment module powered by GPT-4 before applying CHAIR. FAITH-SCORE [Jing *et al.*, 2023] offers a reference-free and fine-grained evaluation method for evaluating the faithfulness of free-form model responses, which comprises identifying descriptive sub-sentences, extracting atomic facts, and comparing extracted facts against the corresponding input image.

Model-based End-to-End Methods End-to-end methods directly assess the performance of LVLMs by evaluating their responses, such as scoring the responses or determining the presence of hallucinations in responses. Existing end-to-end evaluation methods can broadly be categorized into two types. The first type is *LLM-based evaluation* which implements an advanced LLM (e.g., GPT-4) to rate LVLM generated content based on hallucination [Liu *et al.*, 2023a; Sun *et al.*, 2023]. These methods leverage the LLM's robust natural language understanding and processing capabilities. By integrating visual information (such as dense captions and object bounding boxes), user instructions, and model responses as input, LLMs are prompted to evaluate and score these responses. The second type is *hallucination data driven model evaluation* which constructs labelled hallucination datasets for fine-tuning models to detect hallucinations. [Gunjal *et al.*, 2023] creates the M-HalDetect dataset with annotated LVLM image descriptions and fine-tune an InstructBLIP model on the dataset for hallucination identification. Similarly, [Wang *et al.*, 2023c] compiles a dataset of hallucinatory responses generated by ChatGPT and then fine-tune a LLaMA model with LoRA [Hu *et al.*, 2022] on this dataset to differentiate hallucinatory responses from non-hallucinatory ones.

3.2 Evaluation on Hallucination Discrimination

Hallucination discrimination evaluation approach aim to assess the hallucination discrimination ability of LVLMs. The methods follow this approach typically adopt a question-answering format, inquiring LVLMs with descriptions that agree or conflict with the content of the provided input image and assessing the models' responses [Li *et al.*, 2023b; Hu *et al.*, 2023; Lovenia *et al.*, 2023]. POPE [Li *et al.*, 2023b] designs binary (*Yes-or-No*) questions about object presence in images such as "*Is there a person in the image?*" to evaluate the hallucination discrimination ability of LVLMs. The objects asked in questions are selected under three distinct sampling strategies: random (selecting random absent objects), popular (choosing frequently absent objects), and adversarial (selecting absent objects often co-occurring with present ones). CIEM [Hu *et al.*, 2023] is akin to POPE, yet it automates object selection by prompting ChatGPT. NOPE [Lovenia *et al.*, 2023] is another VQA-based method that is designed to evaluate LVLMs' ability of discerning the absence of objects in visual queries, with correct responses being negative statements.

3.3 Evaluation Benchmarks

Distinct from standard LVLM benchmarks that assess general LVLM capabilities, LVLM hallucination benchmarks specifically target non-hallucinatory generation or hallucination discrimination. Many such benchmarks have been proposed along with the evaluation methods. Table 1 shows a synthesis of representative benchmarks. These benchmarks are categorized based on the type of evaluation approach: Discriminative (Dis) or Generative (Gen).

Discriminative Benchmarks POPE [Li *et al.*, 2023b], NOPE [Lovenia *et al.*, 2023], and CIEM [Hu *et al.*, 2023] are
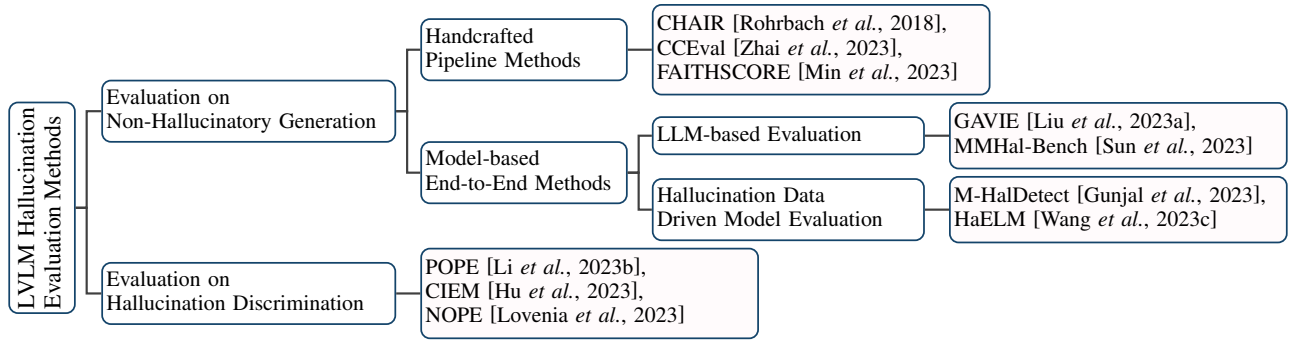
Figure 2: Taxonomy of Hallucination Evaluation Methods for LVLMs

Table 1: Hallucination Evaluation Benchmarks for LVLMs. 'Dis' and 'Gen' stand for 'Discriminative' and 'Generative', respectively.

Benchmark	Evaluation	Size	Hallucination Symptoms	Metrics
POPE [Li *et al.*, 2023b]	Dis	3000	Object	Accuracy
NOPE [Lovenia *et al.*, 2023]	Dis	17983	Object	Accuracy
CIEM [Hu *et al.*, 2023]	Dis	72941	Object	Accuracy
M-HalDetect [Gunjal *et al.*, 2023]	Gen	4000	Object & Attribute & Relation	Reward Model Score
GAVIE [Liu *et al.*, 2023a]	Gen	1000	Object & Attribute	Accuracy & Relevancy
FAITHScore [Jing *et al.*, 2023]	Gen	2000	Object & Attribute & Relation	FAITHScore
HaELM [Wang *et al.*, 2023c]	Gen	5000	Not Explicitly Stated	Accuracy
MMHal-Bench [Sun *et al.*, 2023]	Gen	96	Object & Attribute & Relation	Rating Score
AMBER [Wang *et al.*, 2023b]	Dis & Gen	15202	Object & Attribute & Relation	AMBER Score

discriminative benchmarks, and their dataset sizes are 3000, 17983, and 72941, respectively. All of these three benchmarks focus only on object hallucinations and adopt accuracy as the evaluation metric, which is obtained by inquiring whether objects are present in images and comparing model responses with ground-truth answers.

Generative Benchmarks As delineated in Table 1, current research emphasizes generative benchmarks over discriminative ones. While discriminative benchmarks primarily evaluate hallucinations on object level, generative benchmarks extend their scope to include a broader spectrum of hallucinations, including attribute and relation ones [Gunjal *et al.*, 2023; Liu *et al.*, 2023a; Jing *et al.*, 2023; Wang *et al.*, 2023c; Sun *et al.*, 2023]. Notably, AMBER [Wang *et al.*, 2023b] stands out as a comprehensive benchmark that integrates both generative and discriminative tasks. Furthermore, we observe that the metrics employed in generative benchmarks tend to be more complex and varied, in contrast to the reliance on accuracy metrics in discriminative benchmarks. This is probably due to that, distinct from discriminative methods that mostly compare model outputs with ground-truth answers, generative evaluation methods require tailored metrics to target specific hallucination categories in an attempt to analyze the responses generated by LVLMs.

4 Causes of LVLM Hallucinations

Hallucinations in LVLMs can arise from various factors. As depicted in Figure 3, in correspondence with LVLM pipeline. These causes can be categorized into the following aspects:

4.1 Hallucinations from Data

The quality of training data largely impacts training efficiency and model performance. However, existing LVLM training data suffer from several quality challenges, fostering certain types of hallucinations.

Data Bias One notable issue of data bias is the presence of *distribution imbalance* in existing training data, particularly in factual judgment QA pairs where the majority of the answers are "Yes" [Hu *et al.*, 2023; Liu *et al.*, 2023a]. LVLMs trained on such biased data tend to answer "Yes" consistently and provide explanations even for incorrect or misleading prompts. Another issue is *data homogeneity*, which impedes the model's ability to understand visual information and execute instructions under various environments [You *et al.*, 2023; Liu *et al.*, 2023a]. For example, MiniGPT-4 can only describe images irrespective of the specific question posed by users due to the lack of diversified instruction learning. Similarly, LLaVA struggles to accurately describe local visual relations due to the limited varieties in visual information it was trained on.

Annotation Irrelevance A large amount of instruction data is primarily synthesized from image-caption and detection data using LLMs. Nevertheless, due to the unreliability of these generative models, this approach encounters a challenge of annotation relevance: the generated long instructions often contains objects, attributes, and relationships that may not correspond to the fine-grained content depicted in the images [Liu *et al.*, 2023a]. Training on such data also catalyzes hallucinations.
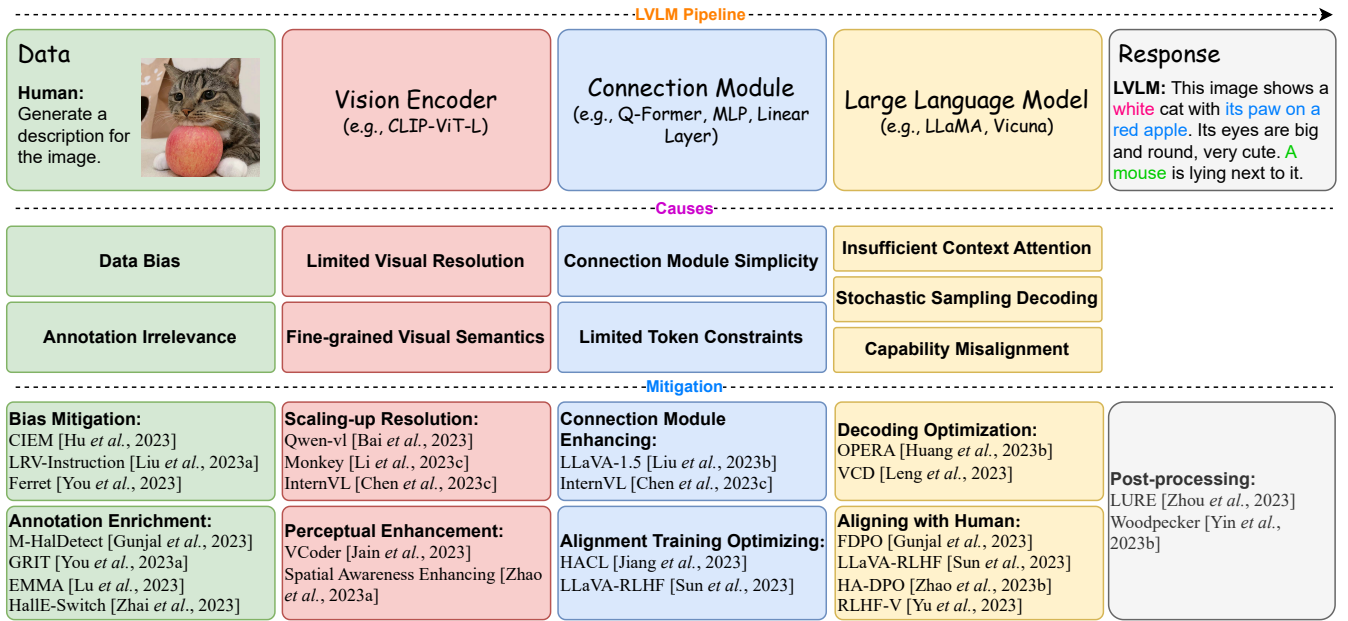
LVLM Pipeline

Data
Human: Generate a description for the image.

Vision Encoder (e.g., CLIP-ViT-L)

Connection Module (e.g., Q-Former, MLP, Linear Layer)

Large Language Model (e.g., LLaMA, Vicuna)

Response
LVLM: This image shows a white cat with its paw on a red apple. Its eyes are big and round, very cute. A mouse is lying next to it.

Causes

Data Bias	Limited Visual Resolution	Connection Module Simplicity	Insufficient Context Attention
Annotation Irrelevance	Fine-grained Visual Semantics	Limited Token Constraints	Stochastic Sampling Decoding
			Capability Misalignment

Mitigation

Bias Mitigation:
CIEM [Hu *et al.*, 2023]
LRV-Instruction [Liu *et al.*, 2023a]
Ferret [You *et al.*, 2023]

Annotation Enrichment:
M-HalDetect [Gunjal *et al.*, 2023]
GRIT [You *et al.*, 2023a]
EMMA [Lu *et al.*, 2023]
HallE-Switch [Zhai *et al.*, 2023]

Scaling-up Resolution:
Qwen-vl [Bai *et al.*, 2023]
Monkey [Li *et al.*, 2023c]
InternVL [Chen *et al.*, 2023c]

Perceptual Enhancement:
VCoder [Jain *et al.*, 2023]
Spatial Awareness Enhancing [Zhao *et al.*, 2023a]

Connection Module Enhancing:
LLaVA-1.5 [Liu *et al.*, 2023b]
InternVL [Chen *et al.*, 2023c]

Alignment Training Optimizing:
HACL [Jiang *et al.*, 2023]
LLaVA-RLHF [Sun *et al.*, 2023]

Decoding Optimization:
OPERA [Huang *et al.*, 2023b]
VCD [Leng *et al.*, 2023]

Aligning with Human:
FDPO [Gunjal *et al.*, 2023]
LLaVA-RLHF [Sun *et al.*, 2023]
HA-DPO [Zhao *et al.*, 2023b]
RLHF-V [Yu *et al.*, 2023]

Post-processing:
LURE [Zhou *et al.*, 2023]
Woodpecker [Yin *et al.*, 2023b]

Figure 3: The causes and mitigation methods of hallucinations in LVLMs.

4.2 Hallucinations from Vision Encoder

The widely adopted vision encoders in LVLMs are derived from CLIP, which maps visual and textual features to the same space by contrastive learning. Though CLIP achieves excellent performance on various visual understanding tasks, it is nonetheless limited in fully expressing visual information, such as restricted visual resolution and fine-grained visual semantics.

Limited Visual Resolution Higher image resolution can make visual encoder more accurate in object recognition and perceive more visual details, thus alleviating hallucinations [Zhai *et al.*, 2023; Li *et al.*, 2023c]. However, handling a wider range of image resolutions is computation-wise demanding, hence existing models usually use smaller image resolution, such as 224×224 and 336×336 pixels for LLaVA and LLaVA-1.5 respectively.

Fine-grained Visual Semantics CLIP aligns visual content and word embeddings through contrastive learning that majorly focuses only on salient objects, falling to capture fine-grained aspects of an image and resulting in hallucinations [Jain *et al.*, 2023; Cho *et al.*, 2022], such as background description, object counting, and object relation.

4.3 Hallucinations from Modality Aligning

The connection module projects visual features into the LLM's word embedding space, aligning visual and textual modalities. Consequently, misalignment could be a key factor in hallucination generation [Sun *et al.*, 2023; Zhao *et al.*, 2023b]. Furthermore, [Jiang *et al.*, 2023] reveal that even in advanced LVLMs like MiniGPT-4 and LLaVA, the gap between visual and textual features stands significant and lead to hallucinations. The simplicity of connection module and constraint of tokens are two main causes.

Connection Module Simplicity Simple structures, such as linear layers, are commonly employed as connection modules to align visual and textual modalities. Although cost-efficient, the simplicity of the connection modules hinder comprehensive multimodal connection [Liu *et al.*, 2023b], thereby increasing the risk of hallucinations.

Limited Token Constraints Q-Former is a widely used multimodal alignment module in LVLMs such as BLIP-2, InstructBLIP and MiniGPT-4. It is trained with various objectives to encode a predetermined number of random initialized tokens (e.g., 32) into visual features that are aligned with text. However, the restricted quantity of these tokens prevents encoding all the information present in images during the alignment process [Yin *et al.*, 2023a; Chen *et al.*, 2023c], and consequent information loss raises the risk of hallucinations.

4.4 Hallucinations from LLM

LLMs are pivotal in the architecture of LVLMs, as LLMs significantly enhance LVLMs ability to process complex multimodal tasks. However, this integration also introduce the inherent hallucination challenges of LLMs to LVLMs. A growing body of literature has explored various factors contributing to hallucinations in LLMs [Ji *et al.*, 2023; Zhang *et al.*, 2023b; Huang *et al.*, 2023a]. In this section, we briefly summarize potential causes of hallucinations originating from LLMs within the scope of vision-language.

Insufficient Context Attention Insufficient context attention is the phenomenon where the model, during the decoding process, focuses only on partial information from the context. The phenomenon includes focusing excessively on the current segment of generated content while ignoring input visual information [Wang *et al.*, 2023a; Lee *et al.*, 2023], prioritiz-

ing language patterns to produce fluent yet inaccurate content [Wang *et al.*, 2023c], and concentrating on partial summarizing tokens of the generated content, resulting in fabrication and hallucinatory content [Huang *et al.*, 2023b].

Stochastic Sampling Decoding Stochastic sampling introduces randomness into the decoding process and is a commonly used decoding strategy in LLMs. It helps prevent the generation of low-quality text often associated with sequences of high likelihood and serves to enrich the generated content [Holtzman *et al.*, 2020]. However, the incorporation of randomness in the model also amplify the risk of hallucinations [Chuang *et al.*, 2023; Dziri *et al.*, 2021].

Capability Misalignment Capability misalignment in LLMs is the disparity between the model's inherent capabilities that established during the pre-training phase, and the expanded requirements imposed during the instruction tuning. This misalignment leads LLMs to produce responses beyond their established knowledge limits, increasing the potential for hallucinatory content [Huang *et al.*, 2023a].

5 Mitigation of LVLM Hallucination

Various methods have been proposed to mitigate hallucinations in LVLMs. As depicted in Figure 3, these methods are tailored to address the identified causes. In this section, we elaborate on them from the following perspectives:

5.1 Mitigation for Data

Optimizing training data is a direct and effective way to alleviate hallucinations. Regarding the explored deficiency in current datasets (§4.1), some measures are taken to address the issues in the following two perspectives:

Bias Mitigation Considering the imbalance of positive and negative samples in the training data, CIEM [Hu *et al.*, 2023] utilizes off-shelf LLMs to generate contrastive question-answer pairs from annotated image-text datasets. These pairs are then leveraged in introduced Contrastive Instruction Tuning (CIT). LRV-Instruction [Liu *et al.*, 2023a] proposes a large and diverse dataset consisting of 400,000 visual instructions. This dataset not only includes positive visual instructions but also incorporates negative ones on three different semantic levels. Furthermore, delving into a fine-grained manner, Ferret [You *et al.*, 2023] proposes to mine 95,000 negative samples by replacing the original categories, attributes, or quantity information with analogous fake ones. This method effectively enhances the model robustness.

Annotation Enrichment Constructing richly annotated datasets helps mitigating hallucination by supervising LVLMs to extract visual content accurately and align modalities comprehensively. One notable dataset is M-HalDetect [Gunjal *et al.*, 2023]. It consists of 4,000 image-description pairs sourced from COCO labelled with object existence, relative positions, and attributes. Another valuable dataset is GRIT [You *et al.*, 2023]. GRIT features 1.1 million samples with refer-and-ground instructions over hierarchical spatial knowledge. In addition, [Zhai *et al.*, 2023] develops two datasets: one with RAM-system [Zhang *et al.*, 2023a] grounded objects and the other with objects grounded by

RAM or generated by LLMs. Lastly, [Lu *et al.*, 2023] generated 180,000 fine-grained instruction samples, covering multi-round conversations, vision-prompted recognition, and fact-checking.

5.2 Mitigation for Vision Encoder

Scaling-up Vision Resolution [Bai *et al.*, 2023] have shown the effectiveness of gradual enlargement of image resolution from 224×224 to 448×448. MONKEY [Li *et al.*, 2023c] processes high-resolution images by dividing them into patches with the size matched the vision encoder, and local features extracted from patches are processed by the LLM. [Chen *et al.*, 2023c] proposes InternVL that scales up the vision encoder to 6 billion parameters and can process images with widths ranging from 1,664 to 6,144 pixels. However, this approach is resource-demanding for pretraining with large-scale data.

Perceptual Enhancement Most existing LVLMs employ the ViT from CLIP as the vision encoder, which majorly focuses only on salient objects and inevitably omit some visual cues. To enhance the object-level perceptual capability, [Jain *et al.*, 2023] proposes VCoder that uses extra perception modalities, such as segmentation map, depth map, or both, as the control inputs through additional vision encoders. To enhance the spatial awareness capability, [Zhao *et al.*, 2023a] suggests introducing addition pre-trained models to acquire spatial position information and scene graph details, which are then used to guide the LVLMs in addressing user queries.

5.3 Mitigation for Connection Module

Connection Modules Enhancing As discussed in §4.3, the limited capacity of the connection module hinders perfect modality alignment, increasing hallucination risks. To better align the vision and language modalities, researchers have developed more capable connection modules recently. For instance, LLaVA-1.5 enhances the connection module in LLaVA by upgrading from a single linear layer to a MLP. Moreover, [Chen *et al.*, 2023c] utilizes LLaMA2 to construct QLLaMA, which significantly outperforms the Q-Former in aligning visual features with text.

Alignment Training Optimizing Strengthening the alignment training process is an effective method to enhance modality alignment and reduce the risk of hallucinations. [Jiang *et al.*, 2023] observe a significant gap between visual and textual tokens in LVLMs. By explicitly adding new learning objectives to bring visual and textual tokens closer, the hallucinations can be reduced. On the other hand, [Sun *et al.*, 2023] employ Reinforcement Learning from Human Feedback (RLHF) [Stiennon *et al.*, 2020] to align different modalities, achieving hallucination reduction.

5.4 Mitigation for LLM

Decoding Optimization In response to insufficient context attention, some recent research attempts to enable the model to focus on proper contexts during decoding, thereby mitigating hallucinations. [Huang *et al.*, 2023b] notice LVLMs may generate hallucinations by over-focusing on few summary tokens and neglecting image tokens. To mitigate hallucinations,

they develop a novel decoding strategy named OPERA that modifies the beam search process with a weighted scoring system to de-prioritize over-trusted candidates and a retrospection mechanism for re-evaluating decision points. [Leng *et al.*, 2023] focused on visual uncertainty and its impact on hallucinations caused by statistical bias and language priors. They proposed a visual contrastive decoding strategy, which contrasts outputs from original and altered visuals, helping correct the model's over-reliance on language priors and statistical biases.

Aligning with Human   Training LVLMs to align with human preferences can improve the quality of model responses. Recently, some efforts focus on preference-aligned training, where non-hallucinatory responses are favored, while hallucinatory contents are not preferred. LLaVA-RLHF [Sun *et al.*, 2023] pioneers the use of RLHF for training LVLMs to align with human preferences. The authors first train a reward model to mirror human preferences. Subsequently, they apply a fact augmented RLHF and the reward model to align the LVLMs with human preferences. Another category of methods [Gunjal *et al.*, 2023; Zhao *et al.*, 2023b; Yu *et al.*, 2023] are based on Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023] that trains LLMs directly from human preference data, avoiding the complexities of reward-model-based reinforcement learning methods like RLHF. After constructing hallucinatory labelled training data, these work present DPO variants, notably HA-DPO [Zhao *et al.*, 2023b] and FDPO [Gunjal *et al.*, 2023], to train the LVLMs better align with human and generate non-hallucinatory responses.

5.5   Mitigation via Post-processing

Beyond manipulating modules of LVLMs, hallucinations can be mitigated through post-processing or output editing via an additional module or operations. The post-processing methods input visual data, user instructions, and LVLM responses, and outputs hallucination-free, refined responses.

More recently, two post-processing methods were introduced to reduce hallucinations in generated image descriptions: LURE [Zhou *et al.*, 2023] and Woodpecker [Yin *et al.*, 2023b]. LURE leverages insights into the causative factors of object hallucination to train a revisor model based on MiniGPT-4. Woodpecker is a training-free post-processing method comprising five stages that extract key concepts, formulate questions, apply visual knowledge validation, generate a structured visual knowledge base, and refine the generated content by correcting inaccuracies and integrating evidence from the knowledge base.

6   Future Directions

The research on hallucinations in LVLMs is still in its early stage, leaving considerable room for development and improvement. We detail these areas of further research below:

Supervision Objective   Meticulous supervision objectives are crucial for models to acquire desired capabilities or suppress unsought symptoms. Current LVLMs predominantly employ coarse-grained (e.g., image-caption) [Li *et al.*, 2023a]

alignments and whole-image instruction tuning, which limits their ability to understand complex visual details. More detailed, spatially-aware supervision such as object masked language modeling [Dai *et al.*, 2023b] for alignment and position-enhanced instruction following [Chen *et al.*, 2023a] hold immense potentials for further advancements.

Enriching Modalities   Integrating multi-modal signals not only expands the capabilities of models but also notably enhances their performance on single-modal tasks [Girdhar *et al.*, 2023]. This can be attributed to high-quality representation learning from correlated and complementary information. Recent LVLMs research made strides in incorporating more modalities (e.g., videos) through methods including through cache retrieval [Han *et al.*, 2023] and pre-alignment [Lin *et al.*, 2023]. These studies show benefits in addressing the "existence" issue. Moving forward, exploration into designing advanced paradigms for coupling multiple modalities is a worthwhile avenue of investigation.

LVLMs as Agent   Empowered by LLMs, the understanding and reasoning capabilities of LVLMs are propelled to unprecedented levels. Yet, accurately perceiving and expressing specific visual details (like numbers in tables) is still challenging. Rather than using more advanced visual encoders with extensive training, a practical alternative is to teach LVLMs to utilize visual tools (e.g., detection and segmentation models) and reason with accurate and structured outputs. Recent research on "LVLMs as agent" [Liu *et al.*, 2023d] demonstrates impressive fidelity in tasks like OCR-VQA, object tagging, and grounding, validating the viability of this direction.

Delving into Interpretability   The post-processing methods (§5.5) address hallucinations in LVLMs by refining generated outputs, but require careful design and higher computational costs. Further study is needed on LVLMs' internal hallucination mechanisms and potential solutions. Several decoding strategies (§5.4) and model editing techniques [Cheng *et al.*, 2023] show good reliability and effectiveness against hallucination. Future in-depth research into LVLMs' defects could lead to promising interpretability-oriented solutions.

7   Conclusion

Equipped with advanced vision encoders, robust LLMs, and modality alignment modules, LVLMs excel in open-domain vision-language tasks. However, hallucinations significantly challenge LVLMs' practical application. In this survey, we conduct a meticulous investigation into the phenomenon of hallucinations in LVLMs. This exploration covers a detailed analysis of the fundamental causes behind these hallucinations, assesses innovative evaluation methods along with pertinent benchmarks, and discusses effective mitigation approaches. We also delve into the existing challenges and discuss potential directions. This survey aims to lay a foundation for tackle complexities of hallucinations in LVLMs and facilitate future research towards practical implementation of these models in various applications.

References

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, 2022.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[Chen *et al.*, 2023a] Chi Chen, Ruoyu Qin, Fuwen Luo, et al. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023.

[Chen *et al.*, 2023b] Jun Chen, Deyao Zhu, Xiaoqian Shen, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[Chen *et al.*, 2023c] Zhe Chen, Jiannan Wu, Wenhai Wang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[Cheng *et al.*, 2023] Siyuan Cheng, Bozhong Tian, Qingbin Liu, et al. Can we edit multimodal large language models? In *EMNLP*, 2023.

[Cho *et al.*, 2022] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, et al. Fine-grained image captioning with clip reward. In *Findings of NAACL*, 2022.

[Chuang *et al.*, 2023] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, et al. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[Dai *et al.*, 2023a] Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[Dai *et al.*, 2023b] Wenliang Dai, Zihan Liu, Ziwei Ji, et al. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *EACL*, 2023.

[Dziri *et al.*, 2021] Nouha Dziri, Andrea Madotto, Osmar R Zaiane, et al. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *EMNLP*, 2021.

[Gao *et al.*, 2023] Peng Gao, Jiaming Han, Renrui Zhang, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, et al. Imagebind one embedding space to bind them all. In *CVPR*, 2023.

[Gunjal *et al.*, 2023] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

[Han *et al.*, 2023] Jiaming Han, Renrui Zhang, Wenqi Shao, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[Holtzman *et al.*, 2020] Ari Holtzman, Jan Buys, Li Du, et al. The curious case of neural text degeneration. In *ICLR*, 2020.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[Hu *et al.*, 2023] Hongyu Hu, Jiyuan Zhang, Minyi Zhao, et al. Ciem: Contrastive instruction evaluation method for better instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[Huang *et al.*, 2023a] Lei Huang, Weijiang Yu, Weitao Ma, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[Huang *et al.*, 2023b] Qidong Huang, Xiaoyi Dong, Pan Zhang, et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

[Jain *et al.*, 2023] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*, 2023.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.

[Jiang *et al.*, 2023] Chaoya Jiang, Haiyang Xu, Mengfan Dong, et al. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint arXiv:2312.06968*, 2023.

[Jing *et al.*, 2023] Liqiang Jing, Ruosen Li, Yunmo Chen, et al. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.

[Lee *et al.*, 2023] Seongyun Lee, Sue Hyun Park, Yongrae Jo, et al. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023.

[Leng *et al.*, 2023] Sicong Leng, Hang Zhang, Guanzheng Chen, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[Li *et al.*, 2023b] Yifan Li, Yifan Du, Kun Zhou, et al. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

[Li *et al.*, 2023c] Zhang Li, Biao Yang, Qiang Liu, et al. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023.

[Lin *et al.*, 2023] Bin Lin, Bin Zhu, Yang Ye, et al. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[Liu *et al.*, 2023a] Fuxiao Liu, Kevin Lin, Linjie Li, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Yuheng Li, et al. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[Liu *et al.*, 2023c] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. Visual instruction tuning. In *NeurIPS*, 2023.

[Liu *et al.*, 2023d] Shilong Liu, Hao Cheng, Haotian Liu, et al. Llava-plus: Learning to use tools for creating multi-modal agents. *arXiv preprint arXiv:2311.05437*, 2023.

[Long *et al.*, 2022] Siqu Long, Feiqi Cao, Soyeon Caren Han, et al. Vision-and-language pretrained models: A survey. In *IJCAI*, 2022.

[Lovenia *et al.*, 2023] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, et al. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*, 2023.

[Lu *et al.*, 2018] Jiasen Lu, Jianwei Yang, Dhruv Batra, et al. Neural baby talk. In *CVPR*, 2018.

[Lu *et al.*, 2023] Jiaying Lu, Jinmeng Rao, Kezhen Chen, et al. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*, 2023.

[Min *et al.*, 2023] Sewon Min, Kalpesh Krishna, Xinxi Lyu, et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[Rohrbach *et al.*, 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, et al. Object hallucination in image captioning. In *EMNLP*, 2018.

[Stiennon *et al.*, 2020] Nisan Stiennon, Long Ouyang, Jeffrey Wu, et al. Learning to summarize with human feedback. In *NeurIPS*, volume 33, 2020.

[Sun *et al.*, 2023] Zhiqing Sun, Sheng Shen, Shengcao Cao, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[Touvron *et al.*, 2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Touvron *et al.*, 2023b] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2023a] Bin Wang, Fan Wu, Xiao Han, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023.

[Wang *et al.*, 2023b] Junyang Wang, Yuhang Wang, Guohai Xu, et al. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[Wang *et al.*, 2023c] Junyang Wang, Yiyang Zhou, Guohai Xu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

[Yin *et al.*, 2023a] Shukang Yin, Chaoyou Fu, Sirui Zhao, et al. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[Yin *et al.*, 2023b] Shukang Yin, Chaoyou Fu, Sirui Zhao, et al. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[You *et al.*, 2023] Haoxuan You, Haotian Zhang, Zhe Gan, et al. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[Yu *et al.*, 2023] Tianyu Yu, Yuan Yao, Haoye Zhang, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

[Zhai *et al.*, 2023] Bohan Zhai, Shijia Yang, Xiangchen Zhao, et al. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023.

[Zhang *et al.*, 2023a] Youcai Zhang, Xinyu Huang, Jinyu Ma, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[Zhang *et al.*, 2023b] Yue Zhang, Yafu Li, Leyang Cui, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[Zhao *et al.*, 2023a] Yongqiang Zhao, Zhenyu Li, Zhi Jin, et al. Enhancing the spatial awareness capability of multi-modal large language model. *arXiv preprint arXiv:2310.20357*, 2023.

[Zhao *et al.*, 2023b] Zhiyuan Zhao, Bin Wang, Linke Ouyang, et al. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

[Zhou *et al.*, 2023] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, et al. Analyzing and mitigating object hallucination in large vision-language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.