

A Survey of Multimodal Hallucination Evaluation and Detection

Zhiyuan Chen^{1,2}, Yuecong Min^{1,2}, Jie Zhang^{1,2}, Bei Yan^{1,2}, Jiahao Wang³, Xiaozhen Wang³, Shiguang Shan^{1,2}

¹State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, 100190, China.

²University of Chinese Academy of Sciences, Beijing, 100049, China.

³Trustworthy Technology and Engineering Laboratory, Huawei, Shenzhen, China.

Contributing authors: chenzhiyuan21@mail.ucas.ac.cn; minyuecong@ict.ac.cn;
zhangjie@ict.ac.cn; yanbei23s@ict.ac.cn; wangjiahao50@huawei.com;
jasmine.xwang@huawei.com; sgshan@ict.ac.cn;

Abstract

Multi-modal Large Language Models (MLLMs) have emerged as a powerful paradigm for integrating visual and textual information, supporting a wide range of multi-modal tasks. However, these models often suffer from hallucination, producing content that appears plausible but contradicts the input content or established world knowledge. This survey offers an in-depth review of hallucination evaluation benchmarks and detection methods across Image-to-Text (I2T) and Text-to-image (T2I) generation tasks. Specifically, we first propose a taxonomy of hallucination based on faithfulness and factuality, incorporating the common types of hallucinations observed in practice. Then we provide an overview of existing hallucination evaluation benchmarks for both T2I and I2T tasks, highlighting their construction process, evaluation objectives, and employed metrics. Furthermore, we summarize recent advances in hallucination detection methods, which aims to identify hallucinated content at the instance level and serve as a practical complement of benchmark-based evaluation. Finally, we highlight key limitations in current benchmarks and detection methods, and outline potential directions for future research.

Keywords: Multi-modal large language model, Vision-language model, Diffusion model, Hallucination evaluation, Hallucination detection

1 Introduction

Over the past few years, Multi-Modal Large Language Models (MLLMs) have demonstrated remarkable advancements in bridging visual and textual data, supporting a wide range of multi-modal understanding and generation

tasks. Image-to-Text (I2T) models such as GPT-4o ([Open AI, 2024](#)), Gemini ([Team et al., 2023](#)), and Qwen-VL ([Wang et al., 2024](#)) excel at tasks like visual question answering and image captioning, achieving robust image recognition and reasoning capabilities without relying on external tools. Conversely, Text-to-Image (T2I) models like Stable Diffusion ([Rombach et al., 2022](#)) and

DALL-E (Ramesh et al., 2021) have made strides in generating high-quality images that align with user-specified content or artistic styles through textual prompts. Despite recent progress, both I2T and T2I models continue to face several significant challenges, such as robustness to distribution shifts (Zhang et al., 2024; Guo et al., 2024; Aafaq et al., 2022; Liu et al., 2023), susceptibility to attacks (Vice et al., 2024; Zhang et al., 2025; Fan et al., 2024), and hallucination (Huang et al., 2023; Bai et al., 2024; Lan et al., 2024; Liu et al., 2024). In particular, hallucination arises when models generate outputs that seem plausible but deviate from the given input or factual knowledge.

The hallucination problem remains a significant challenge to the development of large models. Early studies primarily focus on hallucination within I2T models, emphasizing inconsistencies between generated textual outputs and their corresponding visual inputs (Rohrbach et al., 2018; Li et al., 2023). More recent works have extended this focus to T2I models (Chen et al., 2024; Hu et al., 2023), where hallucinations manifest as misalignments between generated visual outputs and textual prompts. This survey focuses on the hallucinations across both I2T and T2I paradigms of MLLMs, categorizing them into faithfulness hallucinations and factuality hallucinations. Faithfulness hallucination involves inconsistencies between model outputs and user inputs or prior output, such as misidentified objects or images that do not correspond to textual prompts. In contrast, factuality hallucination refers to contradictions between model outputs and established world knowledge, exemplified by incorrect landmark recognition or violations of medical plausibility in X-ray imaging tasks.

As shown in Fig. 1, hallucination evaluation and benchmark construction have received increasing attention, resulting in numerous benchmarks for assessing hallucinations in both I2T models (Li et al., 2023; Wu et al., 2024; Fu et al., 2023; Seth et al., 2024; Zhou et al., 2023) and T2I models (Hu et al., 2023; Feng et al., 2023; Gokhale et al., 2022; Huang et al., 2024) from various perspectives. This paper systematically reviews existing I2T and T2I benchmarks targeting faithfulness and factuality hallucinations. Specifically, we provide a detailed comparison across benchmarks in terms of data sources, evaluation tasks, image-text

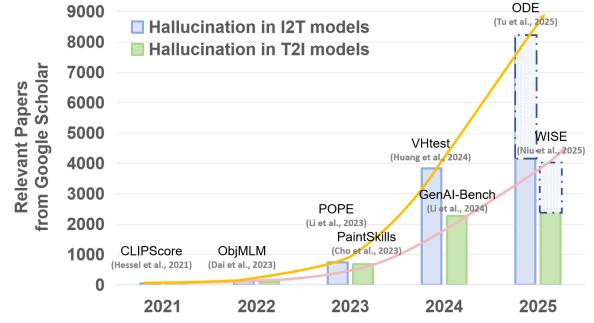


Fig. 1 Trends in the number of relevant papers on I2T and T2I hallucination evaluation from Google Scholar, highlighting the rapid growth in recent years. Dashed lines indicates approximate predictions.

pair construction, and hallucination types, highlighting the emerging trends towards automatic construction and fine-grained evaluation. Furthermore, we give a unified perspective on hallucination evaluation across T2I and I2T benchmarks, summarizing their commonalities and differences.

The flexible response formats of MLLMs present additional challenges for hallucination evaluation, particularly in free-form tasks such as image captioning and image generation. Therefore, hallucination detection methods play a crucial role in the construction of evaluation benchmarks by reducing reliance on costly human annotations and enabling scalable evaluation. Conversely, evaluation benchmarks provide annotated data and standardized metrics essential for the development and assessment of hallucination detection methods. To provide a more holistic perspective on the complementary relationship between evaluation and detection, we further provide a comprehensive summary of existing hallucination detection methods and discuss the feasibility of hallucination detection in I2T and T2I models from a unified perspective.

In a nutshell, this paper aims to advance hallucination evaluation in MLLMs through a comprehensive survey of benchmarks and detection methods for both I2T and T2I tasks. In contrast to existing surveys (Huang et al., 2023; Bai et al., 2024; Lan et al., 2024; Liu et al., 2024) that primarily focus on hallucinations arising in I2T tasks, we broaden the scope to include hallucinations in T2I tasks and provide a unified overview of both faithfulness and factuality hallucinations. Furthermore, we offer a systematic

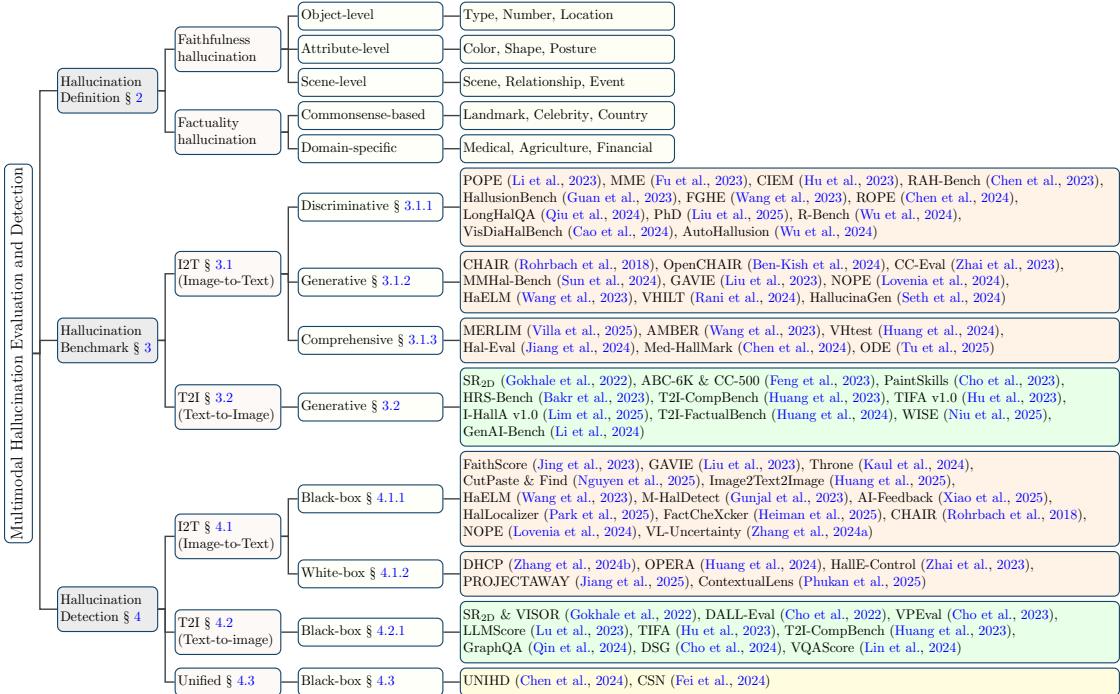


Fig. 2 Overview of the main structure and taxonomy presented in this survey.

summary of existing hallucination evaluation and detection methods and introduce a clear categorization framework. For example, we categorize detection methods into black-box and white-box approaches with different granularity. We also highlight recent efforts towards unified hallucination evaluation and detection across both I2T and T2I models, contributing to a more comprehensive understanding of hallucination phenomena.

Organization of this survey. To facilitate a clearer understanding of existing research, we organize this survey as follows. As shown in Fig. 2, we begin by defining hallucinations in both I2T and T2I tasks and highlight the relationships of different hallucination types observed between different tasks in Sec. 2. Next, we analyze and summarize existing hallucination benchmarks for I2T and T2I models in Sec. 3, focusing on aspects such as hallucination types, construction methodologies, and evaluation metrics, while highlighting recent trends in benchmark development. Building on this foundation, we categorize existing MLLMs hallucination detection methods and investigate the feasibility of unified detection approaches in Sec. 4. Finally, we discuss current limitations and challenges in hallucination evaluation, and suggest promising directions for future research.

2 Definition

Hallucination is a general psychological term (Bentall, 1990) referring to a percept-like experience that occurs without an external stimulus, possesses the vividness of genuine perception, and is not under voluntary control (Slade and Bentall, 1988). In the context of artificial intelligence, hallucination typically refers to the generation of unfaithful or counterfactual content. This phenomenon has garnered significant attention across multiple fields, including natural language generation (Ji et al., 2023), computer vision (Rohrbach et al., 2018; Yu et al., 2018), and multi-modal modeling (Bai et al., 2024).

In the context of computer vision, the term “hallucination” originated in image super-resolution (Baker and Kanade, 2000; Liu et al., 2007; Wang et al., 2014; Huang et al., 2019) task and has since been applied to other image generation tasks, such as inpainting (Yu et al., 2018; Quan et al., 2024) and novel sample generation (Hariharan and Girshick, 2017; Yang and Wang, 2025). In practice, hallucination is leveraged to generate realistic images or to enhance generalization capabilities. However, unlike its beneficial role in generative tasks, hallucination in

vision perception tasks is often undesirable. It typically refers to instances where the model produces false or misleading outputs that do not correspond to the input data. For example, in object detection (Rohrbach et al., 2018) or image caption (Kayhan et al., 2021; Piasco et al., 2021), a model may hallucinate objects that are not present in the scene, leading to reduced reliability and potential risks in safety-critical applications such as autonomous driving (You et al., 2024) or medical imaging (Ayaz et al., 2024).

With the rapid advancement of image generation techniques, the notion of hallucination has shifted to the alignment between generated content and the input prompt (Aithal et al., 2024; Kim et al., 2024; Wang et al., 2024). Moreover, hallucination in I2T, i.e., visual perception (Saleh and Tabatabaei, 2025), and T2I, i.e., visual generation (Agnese et al., 2020), tasks have become increasingly interrelated: generated images are used to evaluate the faithfulness of I2T models and vice versa. One interesting work attempts to detect hallucinations across both I2T and T2I tasks (Chen et al., 2024) within a unified framework. To provide a more consistent viewpoint between T2I and I2T tasks, we adopt a unified definition of hallucination applicable to both I2T and T2I settings: *any inconsistency between the generated content and either the input conditions or established world knowledge*. This definition is consistent with the psychological concept of hallucination as a perception without an external stimulus that contradicts reality and also provides a comprehensive viewpoint of recent works.

2.1 Hallucinations in I2T Tasks

Hallucination in Vision-Language Models (VLMs) refers to the generated text response being inconsistent with the input visual content or established world knowledge. Building on the taxonomy of hallucination in the context of LLM (Huang et al., 2025), we extend this framework to VLMs and categorize hallucinations into two types: faithfulness hallucination and factuality hallucination. The former captures inconsistencies between the generated contents and the visual inputs, while the latter refers to discrepancies between the generated contents and established world knowledge (Fu et al., 2023; Seth et al., 2024; Guan et al., 2023).

These kinds of hallucinations may arise from factors like noisy training data, strong language priors, cross-modal misalignment, and other related issues (Huang et al., 2025). To better assess the reliability of recent VLMs, numerous studies have proposed hallucination benchmarks that evaluate different aspects of this phenomenon, including object hallucination (Li et al., 2023), relationship hallucination (Wu et al., 2024), attribute hallucination (Fu et al., 2023) and so on. Fig. 3 and Fig. 4 present examples of different kinds of hallucinations in I2T tasks.

2.1.1 Faithfulness Hallucinations

Recent works on faithfulness hallucinations primarily focus on object hallucination (Bai et al., 2024; Rohrbach et al., 2018; Li et al., 2023; Ben-Kish et al., 2024; Zhai et al., 2023; Lovenia et al., 2024; Dai et al., 2023) and further categorize into category (Wang et al., 2023; Chen et al., 2024; Liu et al., 2025; Wang et al., 2023), attribute (Fu et al., 2023; Hu et al., 2023; Chen et al., 2023; Qiu et al., 2024; Wang et al., 2023), and relationship (Wu et al., 2024; Hu et al., 2023; Chen et al., 2023; Wang et al., 2023; Qiu et al., 2024; Villa et al., 2025; Wang et al., 2023) hallucination. These types form a fine-to-fine-grained taxonomy. Category hallucination refers to errors in object classification. Attribute hallucination refers to the errors in attribute recognition of specific objects. Relationship hallucination requires recognizing object category and attributes, as well as objects' interrelations. However, hallucinations such as scene hallucination cannot be captured by this taxonomy. Inspired by the hierarchical structure of computer vision tasks, ranging from image-level to pixel-level based on supervision strength (Nesteruk et al., 2024), we propose a thoughtful taxonomy of faithfulness hallucination based on the granularity of visual content: object-level, attribute-level and scene-level.

- *Object-level* hallucinations involve the basic information of objects within the image, such as object type, object number, object localization and so on. For example, in Fig. 3(a), the generated description is inconsistent with the image content, as apples are described as peaches.
- *Attribute-level* hallucinations refer to the condition that the basic information of objects

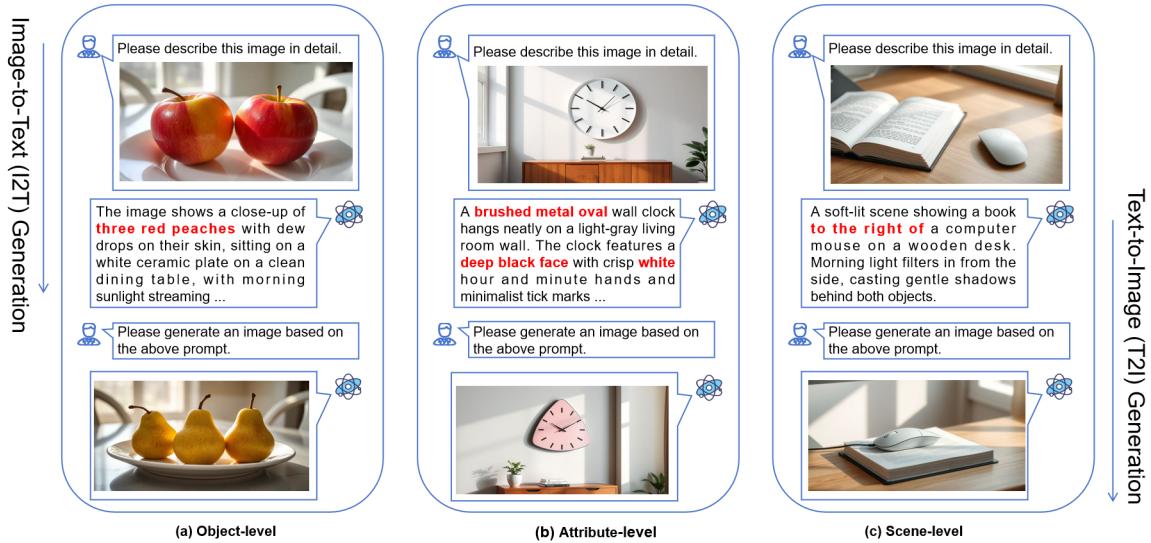


Fig. 3 Examples of object-level (a), attribute-level (b), and scene-level (c) hallucinations (from left to right) in image-to-text (top) and text-to-image (bottom) tasks. Hallucinated responses are highlighted in red.

identified by VLMs is accurate, but the fine-grained intra-object content is wrong, which can be categorized into color, shape, posture, direction and so on. In Fig. 3(b), the description of the clock is inconsistent with the image, particularly in terms of its material, shape, and color.

- *Scene-level* hallucinations refer to the inconsistencies between global visual inputs and textual outputs. VLMs correctly identify the basic information of objects, but hallucinate in inter-object information such as object relationships, object events, scene and so on. In Fig. 3(c), the spatial relation is described as “to the right of”, which contradicts the actual positioning in the image (“to the left of”).

2.1.2 Factuality Hallucinations

With the advancement of in-context learning and visual understanding capabilities of VLMs, they have been increasingly applied in practical scenarios including medical image analysis (Ayaz et al., 2024), autonomous driving (You et al., 2024) and smart agriculture (Arshad et al., 2025). However, VLMs may produce hallucinations due to the insufficient grounding in established world knowledge (Seth et al., 2024; Chen et al., 2024). Existing benchmarks primarily focus on common-sense knowledge (Fu et al., 2023; Guan et al.,

2023), while some recent works concerning specific domain knowledge (Seth et al., 2024; Chen et al., 2024) have emerged. Based on the source of established world knowledge, we categorized factuality hallucinations into commonsense-based and domain-specific hallucinations.

- *Commonsense-based* hallucinations refer to VLMs violating the general knowledge which is universally accepted by humans but usually implicitly stated, such as the appearance of landmarks or features of celebrities. For example, in Fig. 4(a), VLM identifies the landmark in the image as the Egyptian pyramids, while it is actually the Louvre Museum in France.
 - *Domain-specific* hallucinations refer to inconsistencies between the outputs of VLMs and factual information when processing images from specific domains. In Fig. 4(b), the VLM fails to recognize the refraction of the laser beam within the glass and instead hallucinates an additional laser beam inside the glass. In Fig. 4(c), the VLM identifies the lungs affected by COVID-19 in the X-ray image as healthy lungs.

2.2 Hallucinations in T2I Tasks

Hallucinations in T2I tasks refer to the generated image content being inconsistent with the

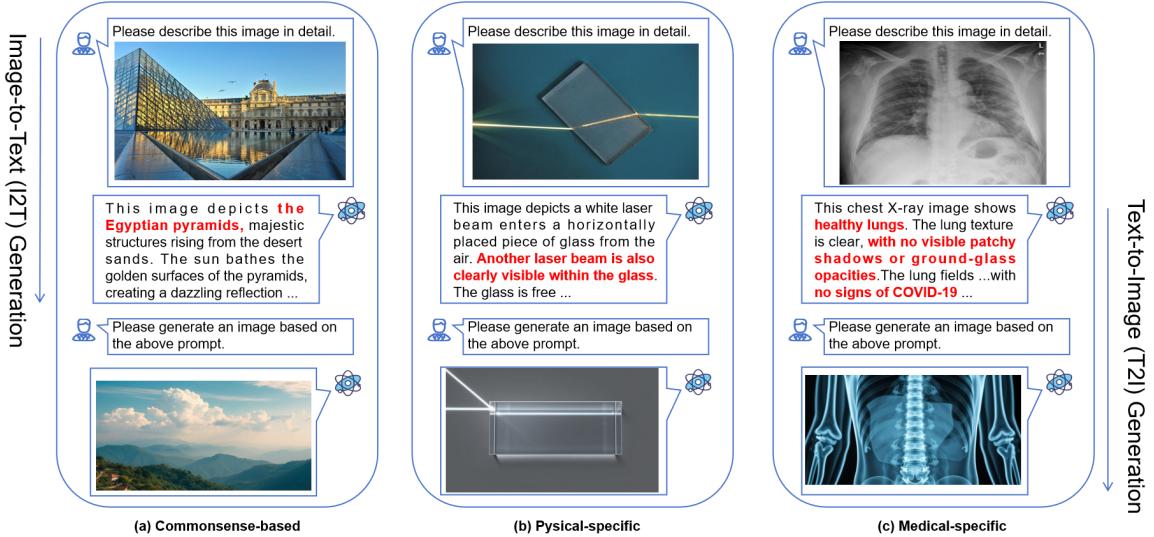


Fig. 4 Examples of Commonsense-based (a), Physical-specific (b), and Medical-specific (c) hallucinations (from left to right) in image-to-text (top) and text-to-image (bottom) tasks. Hallucinated responses are highlighted in red.

input text prompt or established world knowledge. By summarizing existing T2I hallucination benchmarks (Chen et al., 2024; Feng et al., 2023; Gokhale et al., 2022; Cho et al., 2023; Bakr et al., 2023; Fei et al., 2024; Saharia et al., 2022), we categorize hallucinations in T2I tasks into faithfulness hallucination and factuality hallucination, similar to the VLMs. Faithfulness hallucination refers to discrepancies between generated image content and input text prompt, while factuality hallucination refers to discrepancies between established world knowledge and input text prompt. Numerous studies have proposed benchmarks to evaluate faithfulness hallucination in T2I models (Feng et al., 2023; Gokhale et al., 2022; Cho et al., 2023; Bakr et al., 2023; Saharia et al., 2022), and some recent works (Huang et al., 2024; Meng et al., 2024) have also recognized the issue of factuality hallucination and introduce some factuality benchmarks. Fig. 3 and Fig. 4 present examples of different kinds of hallucinations in T2I tasks.

2.2.1 Faithfulness Hallucinations

We conduct the same taxonomy of faithfulness hallucinations in T2I tasks like I2T tasks: object-level, attribute-level and scene-level. Recent works primarily focus on object-level (Cho et al., 2023) and attribute-level (Hu et al., 2023; Feng et al., 2023) hallucination, while some studies have

explored scene-level hallucination, such as relationship hallucination (Huang et al., 2023, 2024).

- *Object-level* hallucinations refer to cases where T2I models generate incorrect basic object information during image synthesis, including object type, object quantity and so on. In Fig. 3(a), the fruits in the generated image are pears rather than peaches in the text prompt.
- *Attribute-level* hallucinations refer to cases where T2I models properly generate the object-level content, but the attributes of objects are wrong, including color, shape, and direction, which have been explicitly specified in the text prompt. For example, in Fig. 3(b), the generated image depicts a triangular clock, which does not match the “oval” shape specified in the text prompt.
- *Scene-level* hallucinations refer to T2I models incorrectly generating large-scale image content or inter-object information, including scene, object relationship and so on. In Fig. 3(c), the spatial relationship between computer mouse and book in the generated image is “above” rather than “to the right of”.

2.2.2 Factuality Hallucinations

T2I models can generate realistic images according to text prompts, but the image content often contradicts established world knowledge. Recent

works primarily focus on commonsense knowledge (Huang et al., 2024; Niu et al., 2025), such as the appearance of well-known landmarks or specific types of animals. Hallucinations involving domain-specific knowledge have received relatively little attention. Similar to VLMs, factuality hallucinations in T2I models can also be categorized into commonsense-based and domain-specific hallucinations.

- *Commonsense-based* hallucinations refer to cases where T2I models generate image content that conflicts with basic commonsense knowledge. In Fig. 4(a), the generated image depicts a mountain landscape, whereas the prompt is expected to evoke an image of the pyramids.
- *Domain-specific* hallucinations arise when the image content produced by T2I models deviates from established domain knowledge, such as physical laws, scientific facts, or medical imaging standards. In Fig. 4(b), the image generated by the T2I model not only fails to depict the refraction of the laser beam within the glass, but also shows an incorrect reflection angle on the glass surface. In Fig. 4(c), the image presents an abdominal X-ray instead of the expected chest (lung) X-ray, indicating a mismatch with domain-specific medical understanding.

2.3 Relationship between Hallucination Problems in I2T and T2I tasks

Based on the proposed definitional and classificatory framework of hallucinations in I2T and T2I models, we can gain a clearer understanding of the relationships and distinctions between current studies on hallucinations in I2T and T2I tasks.

Under a unified definition of hallucination, both faithfulness and factuality hallucinations commonly arise in I2T and T2I tasks. These phenomena largely arise from shared challenges such as cross-modal misalignment and insufficient or biased training data. In terms of hallucination categories, since faithfulness hallucinations are primarily linked to image-text alignment, both I2T and T2I models tend to exhibit similar patterns of hallucination across different granularities, namely the object-level, attribute-level, and scene-level.

Regarding factuality hallucinations, existing research has explored commonsense hallucinations in both I2T and T2I tasks, encompassing errors related to everyday knowledge or general scientific principles. However, domain-specific hallucinations show more significant differences across the two model types. For I2T models, recent studies have begun to examine hallucinations in specialized domains such as medicine (Seth et al., 2024; Chen et al., 2024; Hartsock and Rasool, 2024) and smart agriculture (Arshad et al., 2025), aided by the availability of domain-specific benchmarks and evaluation metrics. For instance, many I2T models have been pretrained on large-scale medical imaging datasets (Ye and Tang, 2025), and their hallucination performance is often evaluated using specialized medical benchmarks. These evaluations adopt coarse-grained generative metrics (e.g., BERTScore (Zhang et al., 2020)) or fine-grained task-specific metrics (e.g., accuracy, precision) in discriminative tasks (Seth et al., 2024; Hartsock and Rasool, 2024). Research on domain-specific hallucinations in T2I models remains relatively limited. In the medical domain, T2I models for medical image synthesis typically employ GAN-based (Uzunova et al., 2019) or diffusion-based methods (Pinaya et al., 2022; Polamreddy et al., 2024), and are mainly evaluated using general image quality metrics such as FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021). These metrics, however, fall short in capturing hallucination-related discrepancies at a fine-grained level. As a result, there remains a lack of dedicated hallucination benchmarks tailored to T2I models in domain-specific contexts. In summary, domain-specific hallucination evaluation in both I2T and T2I tasks has received growing attention, and fine-grained evaluation methods for T2I models indicates an important direction for future research.

3 Hallucination Evaluation Benchmarks and Metrics

In this chapter, we provide a systematic overview of the evolution of existing benchmarks. Tab. 1 summarizes metrics for hallucination evaluation, including their formula and derivatives.

Table 1 Summary of representative evaluation metrics for hallucination in MLLMs.

Metric	Formula	Usage	Derivatives
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	Accuracy measures the overall correctness of answers to YNQs or MCQs	Accuracy+ (Fu et al., 2023)
Precision	$\frac{TP}{TP + FP}$	Precision measures the proportion of predicted positives that are actually correct.	—
Recall	$\frac{TP}{TP + FN}$	Recall measures the proportion of actual positives that are correctly identified.	—
F1 Score	$\frac{2 \times Precision \cdot Recall}{Precision + Recall}$	F1 Score measures the harmonic mean of precision and recall to balance false positives and false negatives.	$F1_{PhD} = 2 \cdot \frac{\text{Yes Recall} \times \text{No Recall}}{\text{Yes Recall} + \text{No Recall}}$ (Liu et al., 2025) Composite = MAE/F1 Score (Heiman et al., 2025)
CHAIR _s (Rohrbach et al., 2018)	$\frac{ \{\text{hall. sents.}\} }{ \{\text{all sentences}\} }$	CHAIR _s measures the proportion of sentences containing any hallucinated objects.	Attack Success Rate (ASR) (Wu et al., 2024) OpenCHAIR (Ben-Kish et al., 2024)
CHAIR _i (Rohrbach et al., 2018)	$\frac{ \{\text{hall. obj.}\} }{ \{\text{all obj.}\} }$	CHAIR _i measures the proportion of hallucinated objects among all mentioned objects.	$\text{AMBER} = \frac{1}{2} \times (1 - \text{CHAIR} + \text{F1})$ (Wang et al., 2023)
GPT Rate (Sun et al., 2024)	0–6 rating (6 = no hallucination)	GPT Rate uses GPT as an evaluator to assess the degree of hallucination.	GAVIE (Liu et al., 2023), Fine-tuned model rate (Wang et al., 2023) $\text{FaithScore} = \frac{\sum_{i=1}^C \sum_{j=1}^{n_i} w_j^i \cot(\epsilon_j^i, I)}{\sum_{i=1}^C n_i}$ (Jing et al., 2023)
Visor (Gokhale et al., 2022)	Correct images Total images	Visor measures the proportion of images with correct spatial relationships.	UniDet (Huang et al., 2023)
VQAScore (Li et al., 2024)	Alignment score	VQAScore evaluates the alignment between image and text using a VQA model.	TIFA Score (Hu et al., 2023) I-HallA Score (Lim et al., 2025) WiScore (Niu et al., 2025) LLMScore (Lu et al., 2023)

Tab. 2 summarizes key details of these benchmarks, including their data sources, construction paradigms, and evaluation criteria. Hallucination benchmarks for I2T and T2I models are introduced in Sec.3.1 and Sec. 3.2 respectively, with their interrelation explored in Sec. 3.3.

3.1 Hallucination Benchmarks for I2T Models

Effectively evaluating hallucination in VLMs remains an open challenge. Recent I2T benchmarks assess hallucination across diverse task formats, including Yes or No Questions (YNQs), Multiple-Choice Questions (MCQs), Image Captioning (IC), and Visual Question Answering (VQA). These tasks are typically categorized by response type into two groups: discriminative (YNQs and MCQs) and generative (IC and VQA) (Lan et al., 2024; Bai et al., 2024). Following this convention, we categorize hallucination

benchmarks based on whether they constrain the output space (discriminative) or allow open-ended generation (generative), reflecting the different evaluation goals and challenges posed by each setting.

In addition, some benchmarks integrate both discriminative and generative tasks to analyze the inconsistencies in hallucination behavior across different evaluation tasks (Wang et al., 2023; Jiang et al., 2024), which has not been sufficiently addressed in previous surveys (Lan et al., 2024; Bai et al., 2024). To capture this feature, we introduce a third category, namely comprehensive benchmarks. This classification provides a clearer framework for understanding the design and focus of existing hallucination evaluations. We present representative benchmarks within each category in the following sections.

Table 2 Summary of representative evaluation benchmarks for hallucination in MLLMs, with task types classified as discriminative (Dis), generative (Gen), and comprehensive (Comp) tasks.

Task	Benchmark	Benchmark Type	Data Source	Size	Hallucination Type		Metric
					Faithfulness	Factuality	
I2T	POPE (Li et al., 2023)	Dis	MSCOCO (Lin et al., 2014)	3000	✓	✗	Acc/P/R/F1
	MME (Fu et al., 2023)	Dis	MSCOCO (Lin et al., 2014)	1277	✓	✓	Acc
	CIEM (Hu et al., 2023)	Dis	MSCOCO (Lin et al., 2014)	72941	✓	✗	Acc/P/R/F1
	RAH-Bench (Chen et al., 2023)	Dis	MSCOCO (Lin et al., 2014)	3000	✓	✗	FP
	HallusionBench (Guan et al., 2023)	Dis	Website	1129	✓	✓	Acc
	FGHE (Wang et al., 2023)	Dis	MSCOCO (Lin et al., 2014)	200	✓	✗	Acc/P/R/F1
	ROPE (Chen et al., 2024)	Dis	MSCOCO (Lin et al., 2014) & ADE20K (Zhou et al., 2017)	5000	✓	✗	Acc
	LongHalQA (Qiu et al., 2024)	Dis	Visual Genome (Krishna et al., 2016) & Object365 (Shao et al., 2019)	6485	✓	✗	Acc
	R-Bench (Wu et al., 2024)	Dis	MSCOCO (Lin et al., 2014)	11651	✓	✗	Acc/P/R/F1
	VisDiaHalBench (Cao et al., 2024)	Dis	GQA (Hudson and Manning, 2019)	25000	✓	✗	F1/EM
	AutoHallusion (Wu et al., 2024)	Dis	MSCOCO (Lin et al., 2014) & DALL-E-2	1000	✓	✗	ASR/MASR/CASR
	PhD (Liu et al., 2025)	Dis	TDIUC (Kafle and Kanan, 2017)	102564	✓	✓	F1 _{PhD}
	CHAIR (Rohrbach et al., 2018)	Gen	MSCOCO (Lin et al., 2014)	5000	✓	✗	CHAIR
	CC-Eval (Zhai et al., 2023)	Gen	Visual Genome (Krishna et al., 2016)	100	✓	✗	CHAIR/Cover
	HaELM (Wang et al., 2023)	Gen	MSCOCO (Lin et al., 2014)	5000	✓	✗	LLM Rate
	GAVIE (Liu et al., 2023)	Gen	Visual Genome (Krishna et al., 2016)	1000	✓	✗	GPT Rate
	OpenCHAIR (Ben-Kish et al., 2024)	Gen	Stable Diffusion	5000	✓	✗	OpenCHAIR
	MMHal-Bench (Sun et al., 2024)	Gen	OpenImages (Kuznetsova et al., 2020)	96	✓	✗	GPT Rate
	NOPE (Lovenia et al., 2024)	Gen	OpenImages (Kuznetsova et al., 2020)	36000	✓	✗	Acc/METEOR
	VHILT (Rani et al., 2024)	Gen	Website	2000	✓	✓	Acc
	HallucinaGen (Seth et al., 2024)	Gen	MSCOCO (Lin et al., 2014) & NIH Chest X-ray (Wang et al., 2017)	96000	✓	✓	Acc
T2I	AMBER (Wang et al., 2023)	Comp	Website	15202	✓	✗	Acc/CHAIR
	VHtest (Huang et al., 2024)	Comp	MSCOCO (Lin et al., 2014) & DALL-E-3	1200	✓	✗	Acc
	Hal-Eval (Jiang et al., 2024)	Comp	MSCOCO (Lin et al., 2014)	10000	✓	✗	Acc/Score
	Med-HallMark (Chen et al., 2024)	Comp	Slake (Liu et al., 2021), etc.	7341	✓	✓	MediHall Score
	MERLIM (Villa et al., 2025)	Comp	MSCOCO (Lin et al., 2014)	31373	✓	✗	Acc/F1
	ODE (Tu et al., 2025)	Comp	Stable Diffusion	8786	✓	✓	AMBER/Acc
	SR _{2D} (Gokhale et al., 2022)	Gen	MSCOCO (Lin et al., 2014)	25280	✓	✗	Visor Score
	DrawBench (Saharia et al., 2022)	Gen	Human, DALL-E	200	✓	✗	Human
	ABC-6K & CC-500 (Feng et al., 2023)	Gen	MSCOCO (Lin et al., 2014)	6900	✓	✗	Human/CLIP-R
	PaintSkills (Cho et al., 2023)	Gen	Template	7330	✓	✗	DETR Score
	HRS-Bench (Bakr et al., 2023)	Gen	GPT	15000	✓	✗	P/AC-T2I
T2I	T2I-CompBench (Huang et al., 2023)	Gen	MSCOCO (Lin et al., 2014) & Template & GPT	6000	✓	✗	BLIP-VQA/UniDet
	TIFA v1.0 (Hu et al., 2023)	Gen	MSCOCO (Lin et al., 2014)	4000	✓	✗	TIFA Score
	T2I-FactualBench (Huang et al., 2024)	Gen	GPT	3000	✓	✓	VQA Score
	GenAI-Bench (Li et al., 2024)	Gen	Human	1600	✓	✗	Human/VQAScore
	I-HallA v1.0 (Lim et al., 2025)	Gen	Textbook	200	✓	✓	I-HallA Score
WISE	WISE (Niu et al., 2025)	Gen	LLM-Constructed	1000	✓	✓	WiScore

3.1.1 Discriminative Benchmarks

There have been a large number of benchmarks utilizing discriminative tasks for hallucination evaluation. Compared to generative tasks, VLMs exhibit more stable performance in discriminative tasks (Li et al., 2023), which also allow a more fine-grained analysis of model weakness across various hallucination types by constructing different kinds of YNQs or MCQs. As more types of hallucinations are identified and LLMs demonstrate increasingly strong in-context learning capabilities, discriminative benchmarks have evolved beyond merely evaluating object hallucinations, with their construction processes becoming more automated. Considering the development of benchmarks, we provide an overview of discriminative benchmarks along two key dimensions: the inclusion of more diverse hallucination types and the adoption of more efficient data generation

methods. We highlight each benchmark’s data sources, construction methodologies, evaluation metrics and so on.

POPE (Li et al., 2023) is one of the early benchmarks using YNQs for evaluating object hallucinations of VLMs. It selects 500 images from the MSCOCO (Lin et al., 2014) validation set, ensuring that each image contains at least three annotated objects, and constructs YNQs in the form of “Is there a {object} in the image”. To generate these questions, different strategies are employed based on the expected answer. For “Yes” questions, real objects that are annotated in the image are used; for “No” questions, objects that do not appear in the image are selected using one of the following three sampling strategies: randomly selecting non-existent objects (random sampling), selecting common objects in the dataset (popular sampling), or selecting objects that frequently co-occur with other objects

(adversarial sampling). Evaluation metrics include Accuracy, Precision, Recall and so on.

ROPE (Chen et al., 2024) extends the evaluation of hallucinations in VLMs from single-object to multi-object scenarios by leveraging MCQs for evaluation. In this benchmark, VLMs are required to identify and select the correct object names from a provided list of objects in multiple bounding boxes, making it more challenging compared to POPE. The images are selected from the MSCOCO (Lin et al., 2014) and ADE20K (Zhou et al., 2017) datasets, with a focus on those containing more than five annotated objects and an Intersection-over-Union (IoU) of bounding boxes lower than 0.1. The benchmark is divided into four subsets: homogeneous, heterogeneous, adversarial, and in-the-wild. This work shows that the probability of hallucination errors increases as the number of recognized object categories in the image grows.

FGHE (Wang et al., 2023) builds upon the POPE framework and extends the scope of hallucination evaluation to include not only object hallucination, but also relationship, attribute, and behavior hallucinations using YNQs. This benchmark selects 50 images from MSCOCO (Lin et al., 2014), and human annotators construct 200 binary YNQs based on these images. Evaluation is performed using standard metrics such as Accuracy, Precision, F1 score, and other relevant indicators, enabling a more comprehensive analysis of fine-grained hallucination types in VLMs.

MME (Fu et al., 2023) provides a comprehensive evaluation of VLMs from both perceptual and cognitive perspectives, and involves various types of hallucinations. Perception tasks include coarse-grained recognition tasks such as object, quantity, color, and location, as well as fine-grained recognition tasks requiring external knowledge, such as movie posters, celebrity photos, and Optical Character Recognition (OCR). Cognition tasks evaluate abilities like commonsense reasoning, math, translation, and code reasoning. Images are sampled from public datasets such as MSCOCO (Lin et al., 2014), and YNQs are manually constructed for each category. This benchmark also proposes an Accuracy+ metric, the accuracy based on Yes-No paired questions, preventing potential Yes/No bias for a more reliable assessment.

HallusionBench (Guan et al., 2023) is a comprehensive evaluation benchmark that covers multiple types of hallucinations. HallusionBench categorizes hallucinations into two types: Visual Dependent hallucinations and Visual Supplement hallucinations, depending on the necessity of visual input. The former are based on visual image information, such as illusions and math, assessing the visual understanding and reasoning abilities of VLMs. The latter rely on external knowledge and can be answered without visual input, testing the models' ability to balance parameter memory with image context. Furthermore, HallusionBench incorporates manually edited images to evaluate model robustness. The evaluation employs YNQs and Accuracy metric, and also introduces a Yes/No bias test to quantitatively analyze the models' preferences. Experimental results show that large models generally perform poorly on HallusionBench, with GPT-4V achieving only 31.42% accuracy.

CIEM (Hu et al., 2023) evaluates hallucinations related to object existence, attributes, and relationships. It introduces an automated benchmark construction process. Using image annotations as references, CIEM leverages GPT-3.5 to automatically generate YNQs. After sampling 4,929 images from MSCOCO (Lin et al., 2014), GPT produces 37,193 “Yes” questions and 35,748 “No” questions. This scale greatly exceeds that of benchmarks relying on manual construction, demonstrating the efficiency of incorporating LLMs in question generation. However, the construction depends on annotated multi-modal datasets and cannot be applied to raw, unannotated images, and the quality of annotations may affect the accuracy of the generated question-answer pairs. Utilizing MSCOCO (Lin et al., 2014) dataset, the error rate is 5% under human verification, which demonstrates the effectiveness of using LLMs in question construction.

RAH-Bench (Chen et al., 2023) evaluates three types of hallucinations: object classification, attributes, and relationships. This work collects images from the MSCOCO validation set and utilizes GPT-4 to generate 3000 YNQs based on annotations. Unlike works (Li et al., 2023; Hu et al., 2023) evaluating all YNQs using the Accuracy metric, RAH-Bench introduces False Positive (FP) Rates to assess the model's performance across different hallucination types. While

Accuracy reflects overall correctness, False Positive Rates provide a more fine-grained diagnostic signal. Experimental results show that VLMs perform worse in attribute and relationship hallucinations, suggesting the need for targeted improvements in these areas.

R-Bench (Wu et al., 2024) is a benchmark designed to evaluate relationship hallucinations, including relationship existence, spatial relationship and actions. While previous benchmarks pay limited attention to relationship hallucination, R-Bench is a large-scale benchmark (11,651 instances) specifically focused on relationship hallucinations. The benchmark is divided into two levels: image-level and instance-level. The former focuses on the existence of relationships, and the latter evaluates specific types of relationships. To construct the benchmark, a set of candidate relationships are extracted from MSCOCO (Lin et al., 2014) using a scene graph parser. The images and corresponding descriptions are sourced from Nocaps (Agrawal et al., 2019). Based on the extracted relationships and Nocaps descriptions, LLM is used to generate YNQs. Experimental results reveal that VLMs are more prone to generating relationship hallucinations than object hallucinations, highlighting the need for targeted improvements in modeling relational reasoning.

AutoHallusion (Wu et al., 2024) is an automated benchmark construction pipeline, designed to evaluate object existence and spatial relationship hallucinations. To induce object hallucinations in VLMs, AutoHallusion employs text prompts to probe the model’s language priors (e.g., “We have {object1} in the image. What is the object to insert/remove under the strategy X where we wish to...”), and then performs targeted image editing based on the model’s responses. AutoHallusion primarily utilizes three image editing induction strategies: abnormal object insertion, paired object insertion, and correlated object removal. Hallucination detection is performed by GPT-4V-Turbo, achieving an accuracy rate of 92.6%. Experimental results show that AutoHallusion achieves hallucination induction success rates of 97.7% on synthetic datasets and 98.7% on real-world datasets. These results demonstrate that modifying image content based on VLM’s language priors could increase hallucination rates.

PhD (Liu et al., 2025) evaluates various types of hallucinations in VLMs, including object existence, attributes, and counting. It collects images and annotations from TDIUC (Kafle and Kanan, 2017), and creates word sets for different question types. For example, for color-related questions, given the color annotation of an object, GPT generates a set of alternative color words that exclude the correct one. The CLIP model then measures the similarity between the image and descriptions with the replaced keywords. Based on the keyword with the highest similarity, GPT generates binary YNQ questions. This process increases the challenge of the questions compared to previous methods that directly rely on GPT to generate questions. PhD also explores hallucination induction methods, including inaccurate prefix and counterfactual images, both of which successfully induce hallucinations in VLMs. Given that original F1 Score might be biased due to the models’ “Yes” or “No” preference, a new $F1_{\text{PhD}}$ score is adopted as the evaluation metric, which mitigates such bias, shown in Tab. 1.

LongHalQA (Qiu et al., 2024) is a benchmark containing 6,000 long texts with rich context, covering various hallucination types including object existence, attributes, counting, and relationships. Compared to short-context questions in previous benchmarks like “Is there an {object} in the image?”, LongHalQA introduces three types of long-context data: object-level description, image-level description, and multi-round conversation. These formats not only provide rich and open-ended contexts but also reflect real-world interaction scenarios, proposing new challenges to existing models’ reasoning and grounding capabilities. To construct the benchmark, GPT-4V is leveraged to generate two types of MCQs unifying both discriminative and generative tasks: description selection and continuation selection. Experimental results show that model rankings in LongHalQA are largely consistent with free-generation settings, indicating its effectiveness and consistency. Unified MCQs provide a new perspective for LLM-free evaluation.

Summary of Discriminative Benchmarks. As shown in Tab. 2, discriminative benchmarks exhibit clear trends in both evaluation focus and construction methods. Specifically, the

focus has expanded from basic faithfulness hallucinations to factuality hallucinations involving commonsense and domain-specific knowledge, reflecting a transition from evaluating merely perceptual abilities to assessing cognitive capabilities. Regarding construction methods, early benchmarks rely heavily on manually designed questions. More recent efforts such as CIEM (Hu et al., 2023) and RAH-Bench (Chen et al., 2023), leverage LLMs for automated question generation. Additionally, frameworks like AutoHallusion (Wu et al., 2024) and PhD (Liu et al., 2025) establish automated pipelines for benchmark construction and explore diverse methods for hallucination induction, such as adversarial attacks, cross-modal inputs, and counterfactual images. Notably, Long-HalQA (Qiu et al., 2024) innovatively integrates MCQ to unify generative and discriminative tasks, representing a promising direction for the evolution of future discriminative benchmarks.

Currently, research on factuality hallucinations in discriminative benchmarks remains limited, with a predominant focus mainly on commonsense knowledge and insufficient coverage of specialized domains. To better align with real-world application scenarios, future discriminative benchmark construction can incorporate richer contextual formats such as multi-turn dialogues and long-text, enabling a more comprehensive assessment of VLMs’ hallucination. Additionally, exploring more hallucination induction methods similar to PhD (Liu et al., 2025) could also enhance the complexity of discriminative benchmarks.

3.1.2 Generative Benchmarks

Compared to discriminative tasks, generative tasks typically elicit longer and open-ended responses, posing challenges for traditional accuracy-based metrics and necessitating dedicated hallucination detection methods. Early efforts in this area primarily targeted object hallucinations in conventional image captioning (Rohrbach et al., 2018). With the growing capabilities of large language models in contextual understanding and long-form generation, recent benchmarks increasingly incorporate these models to facilitate hallucination detection in generative settings (Ben-Kish et al., 2024; Zhai et al., 2023). In addition, concerns about potential data

leakage have raised questions about the reliability of benchmark-based evaluations (Sun et al., 2024; Rani et al., 2024). As a response, newer benchmarks have begun to explore alternative data sources, such as website-crawled (Rani et al., 2024) or synthetically generated images (Ben-Kish et al., 2024), to improve the robustness of evaluations. In this section, we present a comprehensive overview of existing generative hallucination benchmarks. Our discussion is structured around three major developments in this area: the evaluation of more diverse hallucination types, the expansion of task formats, and the adoption of automated detection methods. For each benchmark, we highlight its core design elements, including question types, data sources, and hallucination detection methodologies.

CHAIR (Caption Hallucination Assessment with Image Relevance) (Rohrbach et al., 2018) is one of the early works evaluating object hallucination in the image captioning task. CHAIR is proposed to calculate the proportion of hallucinated objects in VLM responses, including two variants: per-instance and per-sentence. As shown in Tab. 1, the former calculates what fraction of object instances are hallucinated (denoted as CHAIR_i), and the latter calculates what fraction of sentences include a hallucinated object (denoted as CHAIR_s). The hallucination detection is achieved by sentence tokenization and synonyms mapping, and for easier analysis, the scope of object type is restricted to the 80 MSCOCO (Lin et al., 2014) objects.

OpenCHAIR (Ben-Kish et al., 2024) is an open-vocabulary benchmark designed to evaluate object hallucination in the image captioning task. Compared to CHAIR (Rohrbach et al., 2018), OpenCHAIR expands the evaluation scope to include a wider range of rare and diverse objects, reflecting the variability of real-world scenarios. To construct the benchmark, LLM is used to generate diverse captions, and then a diffusion model generates the corresponding images. Additionally, OpenCHAIR replaces CHAIR’s fixed synonym lists with a dynamic semantic matching method based on an LLM, enabling a more flexible and accurate assessment of object hallucination in the open-vocabulary setting.

CC-Eval (Zhai et al., 2023) evaluates object hallucination in detailed image captions using GPT-4 assistance. It uses 100 randomly sampled

images from Visual Genome (Krishna et al., 2016), and requires VLMs to generate captions. GPT-4 extracts mentioned objects and compares them with the ground-truth annotations from Visual Genome. The benchmark adopts CHAIR to measure hallucination rate and introduces coverage to assess how well real objects are captured. These metrics jointly offer a comprehensive assessment of object-level accuracy.

NOPE (Lovenia et al., 2024) is designed to evaluate object hallucinations, specifically targeting cases where the answer should be “None” or “Nothing” (NegP data). Existing benchmarks contain very few such instances; therefore, NOPE constructs 36,000 NegP examples using two LLM-based methods: generate-from-scratch and list-then-rewrite. The former prompts LLM to directly generate questions based on the question type and image caption, while the latter decomposes the task into multi-turn dialogue for step-by-step question generation. Human evaluation shows that 50% of generate-from-scratch questions are correct, whereas list-then-rewrite achieved 92% accuracy, highlighting the benefits of a structured approach. NOPE uses Accuracy and METEOR as evaluation metrics. Experimental results reveal that existing VLMs are prone to hallucinations when asked about non-existent objects.

VHILT (Rani et al., 2024) collects images from the New York Times Twitter account. All the VQA questions and Caption tasks are manually designed through Amazon Mechanical Turk. The benchmark includes 2,000 instances and proposes 8 types of hallucinations, including contextual guessing (when the model generates unrelated elements that bear no resemblance to the subject at hand), identity incongruity, and geographic erratum. Final evaluation is conducted using human annotations. Results show that models frequently hallucinate on contextual guessing and identity incongruity, reflecting their limitations in visual cognition and reasoning. Given VHILT’s heavy reliance on manual annotation, future work could explore leveraging large models to automate parts of the process.

MMHal-Bench (Sun et al., 2024) collects images from the validation set of OpenImages (Kuznetsova et al., 2020) to prevent data leakage. The benchmark contains 96 handcrafted VQA questions, and covers 8 types of hallucinations, including object existence, attributes,

counting, and environment. Evaluation leverages GPT-4 to analyze and score the degree of hallucination based on the consistency with actual image content. Results show that GPT-4’s hallucination scores align with human annotations in 94% of cases, demonstrating the effectiveness of using LLMs for evaluation.

HaELM (Wang et al., 2023) evaluates VLMs’ performance on image captioning tasks utilizing a fine-tuned LLM. To construct the training dataset for fine-tuning, GPT is prompted to create 10,000 hallucinatory and 10,000 non-hallucinatory captions based on the MSCOCO (Lin et al., 2014) training set. LLaMA is then fine-tuned with LoRA to create a hallucination detector. For evaluation, 5,000 samples from the MSCOCO (Lin et al., 2014) 2014 test set are used. Results show that HaELM achieves 95% of GPT’s accuracy in detecting hallucinations. HaELM provides a scalable and cost-efficient solution that can be deployed locally without dependence on external APIs like GPT.

HallucinaGen (Seth et al., 2024) evaluates VLMs’ object recognition and reasoning capabilities through challenging visual-language tasks with increasing difficulty: Localization (LOC), Visual Context (VC), and Counterfactual (CF). Images are sourced from MSCOCO (Lin et al., 2014) and NIH Chest X-ray (Wang et al., 2017). Similar to NOPE (Lovenia et al., 2024), it employs implicit hallucination induction attacks, including locating absent entities and imagining non-existent objects. Evaluation combines word-matching algorithms with GPT-4o, using five prompts for robust assessment. Results show that as task difficulty increases, VLMs exhibit higher hallucination rates. In tasks involving domain-specific knowledge, even specialized models like LLaVA-Med struggle to perform better than random guessing. This highlights the vulnerability of VLMs in high-stakes scenarios like healthcare, where accuracy and reliability are critical.

Summary of Generative Benchmarks. As illustrated by the works above, current generative benchmarks have made significant progress in diversifying hallucination types, expanding task formats, and introducing automation in benchmark construction. Nevertheless, several critical

challenges remain. First, many existing benchmarks still rely on coarse metrics such as hallucination rates and LLM scores, lacking fine-grained and diagnostic assessments. Second, despite recent efforts to mitigate data leakage by using web-crawled or validation-set data, these sources may still overlap with the pretraining data of VLMs. Additionally, current automation methods are mostly limited to question generation, and the quality of generated content still heavily depends on human verification. Developing scalable and robust benchmark construction pipelines that integrate reliable automation with enhanced diagnostic capabilities remains a promising direction.

3.1.3 Comprehensive Benchmarks

Considering the limited expressiveness of discriminative benchmarks and the evaluation difficulty of generative benchmarks, comprehensive benchmarks attempt to take advantage of both and provide a more holistic evaluation result. We review existing comprehensive benchmarks within three key aspects: the expansion to more types of hallucinations, the adoption of more automated and scalable evaluation methods, and the efforts to address data leakage risks. As shown in Tab. 1 and Tab. 2, for each benchmark, we summarize its hallucination types, evaluation metrics, data sources, and other core designs.

MERLIM ([Villa et al., 2025](#)) evaluates existence, relation, and counting hallucinations using a large number of both edited and original images. GPT is used for automatic question generation. To construct the benchmark, the original images are edited by removing a selected object and seamlessly inpainting the missing region to preserve visual realism, creating paired edited versions for controlled comparison. By comparing model responses between the original and edited images, hallucinations lacking visual grounding can be effectively identified. Relationship and counting hallucinations are evaluated via YNQs. Object existence hallucinations are evaluated via open-ended questions, and model responses are processed with the spaCy library and GPT to extract object nouns, followed by calculating metrics such as Precision.

AMBER ([Wang et al., 2023](#)) evaluates hallucinations in object existence, attributes, and

relations through both discriminative and generative tasks. Images are manually collected from websites to prevent data leakage. Discriminative questions assess model accuracy and Yes/No bias, revealing that models like GPT-4V and Qwen-VL tend to favor “No” answers. The generative task involves extracting target objects from captions using linguistic tools and computing metrics such as CHAIR. This approach, however, is limited by errors in object extraction and mainly focuses on existence hallucinations. Unlike MERLIM, which reports separate results by hallucination type, AMBER introduces a unified AMBER Score to measure a model’s overall performance across both discriminative and generative tasks, shown in Tab. 1.

VHtest ([Huang et al., 2024](#)) covers eight types of hallucinations, including object existence, shape, color, and OCR, with 1,200 hallucination instances. To address issues of limited image diversity and potential data leakage, VHtest constructs a synthetic dataset. It first identifies initial hallucination instances from datasets like MSCOCO ([Lin et al., 2014](#)) using CLIP and DINO, then generates additional images with T2I models based on annotations. Question construction combines templates and manual editing to create YNQs and image captioning tasks. Yes/No responses are automatically evaluated against reference answers, while captions are manually assessed for hallucinations. Experiment results show that the benchmark, built by expanding initial hallucination instances, reduces the accuracy of models such as GPT-4V and LLaVA-1.5.

VisDiaHalBench ([Cao et al., 2024](#)) contains 25,000 multi-turn questions and 5,000 edited images, covering hallucinations like object existence, attribute, relation, and various verification questions. Based on GQA ([Hudson and Manning, 2019](#)) dataset, VisDiaHalBench edits both the images and corresponding scene graph by removing, replacing, or altering the color of objects. The edited scene graphs are then provided as prompts to GPT-4 to generate five-turn dialogue questions. The five-turn dialogue structure allows for in-depth evaluation of VLMs’ visual understanding and consistency. Evaluation uses keyword-filtered outputs to compute macro-average F1 and Exact Match (EM) scores. Experiment results show that GPT-4’s performance on VisDiaHalBench is significantly lower than its performance on GQA,

highlighting the new challenges posed by VisDia-HalBench for current VLMs.

Hal-Eval (Jiang et al., 2024) presents a systematic quantitative analysis of data leakage issues in VLMs. It constructs evaluation datasets by collecting images from both MSCOCO (Lin et al., 2014) (in-domain) and various web-sourced datasets (out-of-domain), and generates questions using GPT-4. Hal-Eval designs questions for both discriminative and generative tasks, covering hallucination types including existence, attribute, relation, and the newly introduced event hallucination. Experimental results show that most models tend to answer “Yes” in discriminative tasks, indicating a bias towards positive responses. Some models, such as LLaVA, exhibit stronger hallucination tendencies in out-of-domain scenarios compared to in-domain, possibly due to the model’s extensive use of in-domain images during instruction fine-tuning. In generative task evaluation, Hal-Eval employs a fine-tuned Hall-Evaluator. Its performance is comparable to GPT-4, demonstrating its effectiveness in identifying hallucinations across diverse datasets.

Med-HallMark (Chen et al., 2024) is a hallucination detection benchmark specifically developed for the multi-modal medical domain, supporting two tasks: Med-VQA and Imaging Report Generation (IRG). To evaluate model robustness, it includes four types of questions: conventional medical questions, confidence-weakening questions, counterfactual questions, and image depiction tasks. These are constructed through a combination of manual annotation, model-generated responses from LLaVA-Med and GPT, and prompt engineering techniques. To better characterize medical hallucinations, the benchmark introduces a hierarchical categorization consisting of five levels: catastrophic, critical, attribute, prompt-induced, and minor hallucinations. These categories are defined based on their clinical impact. Building on this categorization, the MediHall Score serves as a fine-grained metric. It quantifies hallucination severity at different levels: at the answer level for Med-VQA and the sentence level for IRG. With its diverse data sources, structured question types, and clinically grounded evaluation framework, Med-HallMark provides a comprehensive foundation for hallucination detection in medical multi-modal learning.

ODE (Open-set Dynamic Evaluation) (Tu et al., 2025) is a dynamic benchmark that evaluates hallucinations in VLMs by generating novel image-text pairs covering diverse object concepts and attributes. It constructs a weighted graph to represent objects and their co-occurrence distributions. The benchmark generates test samples across multiple distribution types, including standard, long-tail, random, and fictional pairs. This dynamic design reduces data leakage risks in static benchmarks and allows iterative updates to broaden evaluation scope. ODE uses generative and discriminative tasks with specific inquiry templates. Evaluation metrics include accuracy, CHAIR, and AMBER. Experiments show ODE effectively reveals hallucination patterns and supports fine-tuning to improve model reliability.

Summary of Comprehensive Benchmarks. As shown in the analysis and Tab. 2, comprehensive benchmarks remain relatively scarce. While efforts like MERLIM (Villa et al., 2025) and Hal-Eval (Jiang et al., 2024) incorporate both task types, fully unified evaluation frameworks that consistently integrate results across different formats are still in progress. Benchmarks such as AMBER (Wang et al., 2023) introduce unified metrics, though achieving balanced coverage of hallucination types across tasks remains a challenge. Nevertheless, several benchmarks provide valuable strategies for mitigating data leakage. For instance, Hal-Eval (Jiang et al., 2024) includes both in-domain and out-of-domain datasets to assess potential data leakage. VHtest (Huang et al., 2024) and ODE (Tu et al., 2025) utilize T2I generation models to construct synthetic datasets, improving scalability and reducing the risk of data leakage. These developments underscore the need for more comprehensive benchmarks that combine task diversity with diagnostic precision and robust data practices.

Summary of I2T Hallucination Benchmarks. Based on our review of the three types of I2T hallucination benchmarks, several key trends have emerged. First, the range of hallucination types is steadily expanding. Second, LLMs are increasingly involved in benchmark construction. Third, concerns about data leakage are gaining more attention.

Despite these advancements, notable imbalances remain. Faithfulness-focused benchmarks

have developed more extensively. In contrast, benchmarks targeting factuality hallucinations remain relatively underexplored. Current research on VLM hallucinations continues to prioritize image-text consistency, leaving broader factuality issues insufficiently addressed.

As shown in Tab. 2, there is an imbalance between discriminative and generative benchmarks. A greater number of benchmarks are discriminative, as they rely on predefined answer sets and allow straightforward hallucination detection. Due to the challenges in detecting hallucinations in open-ended outputs, generative benchmarks are relatively less common. Further progress in detection methods may help support the development of more generative benchmarks. A similar imbalance exists between single-task and comprehensive benchmarks. Most benchmarks are designed around a single task type, while fewer support both discriminative and generative tasks. The development of comprehensive benchmarks constitutes an important research direction.

3.2 Hallucination Benchmarks for T2I Models

Although benchmarks for I2T models have advanced rapidly, the hallucination evaluation in T2I models has also made steady progress in three key aspects: hallucination types, benchmark construction, and evaluation methods. Early works focus on assessing image-text alignment in aspects such as color (Feng et al., 2023) and spatial relationships (Gokhale et al., 2022), which are typically categorized as faithfulness hallucinations. Recent studies (Huang et al., 2024) extend the evaluation scope to factuality hallucinations by introducing prompts involving established world knowledge. Along with the diversification of hallucination types, the construction of T2I benchmarks is becoming more automated. Prompt design has evolved from manual selection and dataset sampling (Feng et al., 2023), to template-based or LLM-assisted generation (Huang et al., 2023; Bakr et al., 2023). As for evaluation methods, early approaches rely on manual annotation (Saharia et al., 2022) and object detection (Gokhale et al., 2022), while recent works adopt automated methods based on VLMs, VQA tasks, and caption tasks (Huang et al., 2023; Hu

et al., 2023; Lu et al., 2023). These developments enable a more fine-grained and scalable assessment of hallucinations in T2I models.

In the following section, we provide a comprehensive overview of existing T2I hallucination benchmarks, highlighting each benchmark’s hallucination types, data sources, and evaluation metrics. As shown in Tab. 2, T2I benchmarks are typically categorized as generative benchmarks, as they assess model performance based on the generation of images conditioned on textual prompts.

SR_{2D} (Gokhale et al., 2022) is a benchmark designed to evaluate spatial hallucination in T2I models. Based on 80 object categories from MSCOCO (Lin et al., 2014), the benchmark contains a total of 25,280 text prompts. The VISOR metric is used to calculate the accuracy of the generated images. It extracts spatial relationships from generated images using object detection and centroid determination algorithms. Experimental results show that T2I models exhibit poor performance in understanding and rendering spatial relationships, as indicated by low VISOR scores.

ABC-6K & CC-500 (Feng et al., 2023) are two benchmarks targeting color hallucination in T2I models. ABC-6K constructs 6400 text prompts by sampling descriptions from MSCOCO (Lin et al., 2014) and swapping color words in those that contain at least two color terms. CC-500 focuses on simple parallel text prompts, typically in the format of “a red apple and a yellow banana.” The evaluation includes human evaluation, automatic evaluation using the phrase alignment model GLIP, and system-level metrics such as FID (Heusel et al., 2017). Experimental results reveal three main challenges in generating images from combined prompts: attribute leakage, interchanged attributes, and missing objects.

DrawBench (Saharia et al., 2022) is a comprehensive benchmark including a diverse set of 200 prompts across 11 categories. These categories test various aspects of model performance, including color, quantity, spatial relationships, scene text, and unusual object interactions, as well as complex prompts with long descriptions, rare words, and misspellings. Human evaluations are conducted by presenting images from different models and asking raters to compare them in terms of image fidelity and image-text alignment.

PaintSkills (Cho et al., 2023) assesses image-text alignment in T2I tasks using 7330 template-generated text prompts. It evaluates hallucinations such as object existence, counting, and spatial relationships. Text prompts are generated through templates, and over 50000 images are created using Unity engine to train an object detector (DETR). The evaluation employs DETR to automatically verify object presence in the generated outputs, and determines spatial relationships through algorithmic analysis. Results show that T2I models generally perform poorly in object counting and spatial relationships.

HRS-Bench (Bakr et al., 2023) provides a comprehensive and scalable evaluation benchmark for T2I models. It assesses performance across five key dimensions: accuracy, robustness, generalization, fairness, and bias. The benchmark spans 50 diverse scenes encompassing domains such as fashion, animals, transportation, and objects. Departing from prior methods reliant on sampling or manual prompt design, HRS-Bench first constructs specific hallucination templates (e.g., “Describe a scene about 3 apples and 3 dining tables”). These templates are then expanded using GPT to generate more complex prompts (e.g., “Three apples are sitting on top of three dining tables in a cozy dining room”). For evaluation, HRS-Bench introduces the AC-T2I metric. This metric employs an image description model to caption the generated image and calculates the alignment score between this caption and the original text prompt.

T2I-CompBench (Huang et al., 2023) is a comprehensive open-world benchmark for evaluating compositional T2I generation. It includes 6,000 compositional text prompts, categorized into three main categories (attributes, relations, complex compositions) and six subcategories (color, shape, texture, spatial and so on). The prompts are derived from fixed sentence templates, existing datasets, or ChatGPT generation. Evaluation in T2I-CompBench employs a multi-tool strategy. Specifically, BLIP-VQA assesses object attributes (color, shape, texture); UniDet evaluates spatial relationships by detecting objects; a mean metric (averaging CLIPScore, BLIP-VQA, and UniDet scores) measures complex compositions. Additionally, VLMs (e.g., MiniGPT-4) evaluate overall image-text alignment via image description. This

multi-tool approach significantly enhances the precision of hallucination detection in T2I models.

TIFA v1.0 (Hu et al., 2023) constructs a text prompt dataset for evaluation by sampling from the image descriptions of MSCOCO (Lin et al., 2014), containing 4K text inputs. It utilizes a language model to pre-generate corresponding question-answer pairs for each prompt, covering diverse elements. This process yields 25K questions spanning 4.5K element types. Moving beyond coarse-grained evaluation, TIFA employs VQA tools for fine-grained assessment across 11 hallucination categories (e.g., object type, count, attributes). The resulting VQA accuracy is reported as the TIFA Score. Experiments show that TIFA Score has a higher correlation with human judgments than CLIPScore.

I-HallA v1.0 (Lim et al., 2025) constructs a dataset of real image-text pairs sourced from scientific and historical textbooks, with a focus on factuality hallucinations. The texts are enriched with external knowledge and hallucination reasoning using GPT-4’s pre-trained capabilities. An LLM then generates MCQs based on the augmented texts, while T2I models are tasked with producing corresponding images. During evaluation, a vision-language model answers MCQs associated with each image-text pair, and the I-HallA Score is calculated based on its accuracy. This MCQ-based framework offers great evaluation stability compared to caption.

T2I-FactualBench (Huang et al., 2024) is a benchmark specifically designed to evaluate factuality hallucinations in generated images. T2I-FactualBench collects 1,600 knowledge concepts across eight diverse domains. It defines three difficulty levels of factual prompts: (1) basic factual combinations (e.g., an image of a soccer ball and a baseball where the sizes of the objects should align with real-world facts); (2) instantiated factual combinations (e.g., a table with a half-eaten T-bone steak and a steaming plate of Mapo tofu); (3) instantiated knowledge concepts with interaction (e.g., a white West Highland Terrier playing fetch with a green hunting cap). For evaluation, T2I-FactualBench employs multiple-round VQA. It assigns three distinct scores corresponding to these difficulty levels: the Concept Factuality Score, Task Completeness Score, and Composition Factuality Score. This

systematic multi-dimensional assessment enables granular evaluation of factuality hallucinations.

WISE (World Knowledge-Informed Semantic Evaluation) (Niu et al., 2025) is a benchmark designed to evaluate factuality hallucinations in T2I tasks through complex prompts. It addresses factuality hallucinations across three core domains: natural sciences, spatiotemporal reasoning, and cultural knowledge. The benchmark comprises 1,000 evaluation prompts spanning 25 subfields. These prompts originate from diverse sources: educational materials, encyclopedias, common sense Q&A datasets, and LLM synthetic data. Human annotators optimize and extend the prompts to ensure clarity, complexity, and unambiguous real-world answers. For evaluation, WISE introduces the WiScore metric, which calculates a weighted sum of three GPT-4o-assigned dimension scores: consistency, realism, and aesthetic quality. This multi-dimensional approach provides a semantically rich assessment of image generation.

GenAI-Bench (Li et al., 2024) is a T2I benchmark with 1,600 manually constructed text prompts, covering common hallucination types such as attributes, spatial relations, and actions. GenAI-Bench also introduces complex real-world scenarios involving counting, comparisons, and textual negation. All prompts are refined with the assistance of GPT. Evaluation combines human scoring with VQAScore metric. VQAScore quantifies image-text consistency by computing the probability of a “Yes” response from a VQA model to the question “Does this figure show ‘{text}’?”. VQAScore exhibits higher correlation with human than CLIPScore on GenAI-Bench, validating its effectiveness for hallucination assessment.

Summary of T2I Hallucination Benchmarks. Based on the overview of existing T2I hallucination benchmarks, several key development trends can be identified. First, benchmark development has shifted from focusing solely on faithfulness hallucinations to include factuality hallucinations. For example, T2I-FactualBench (Huang et al., 2024) covers hallucinations involving both visual realism and established world knowledge. Second, the involvement of LLMs enables automated benchmark construction. In HRS-Bench (Bakr et al., 2023), text prompts are generated based on templates first and then expanded

by GPT. Third, leveraging VLMs facilitates fine-grained evaluation. TIFA v1.0 (Hu et al., 2023) employs VQA models to detect object-level and attribute-level hallucinations.

Despite recent progress, several challenges remain. Although some pretrained T2I models in specialized domains have emerged (Abaid et al., 2024), existing hallucination benchmarks still pay insufficient attention to domain-specific hallucinations. Moreover, many benchmarks still rely on coarse-grained scoring provided by VLMs (Niu et al., 2025; Li et al., 2024). Exploring more fine-grained evaluation methods represents a promising direction to overcome this limitation.

3.3 Relationship between Hallucination Benchmarks for I2T and T2I Models

The above overview of I2T and T2I hallucination benchmarks helps clarify the current landscape of both research areas. While the definition and evaluation of hallucinations originates earlier in I2T tasks, and related benchmarks are generally more comprehensive, T2I benchmarks focusing on image-text consistency have also received growing attention and are developing rapidly. Both directions share similar trends, including the broadening of hallucination categories, increasing automation in benchmark construction, the use of model-based evaluation methods, and a stronger focus on fine-grained analysis.

Regarding hallucination types, Regarding hallucination types, both faithfulness and factuality hallucinations have received increasing attention in I2T and T2I tasks. However, in T2I settings, factual content remains relatively under-represented, with most evaluations focusing on prompt consistency. Expanding the scope of T2I benchmarks and introducing established world knowledge during evaluation remain important directions for future research.

In terms of construction and evaluation strategies, there is a growing overlap in the use of models across both tasks. For instance, I2T benchmarks such as VHtest (Huang et al., 2024) and PhD (Liu et al., 2025) employ T2I models to generate or modify images for building datasets. Conversely,

Table 3 Summary of Hallucination Detection Methods in I2T, T2I, and Unified Scenarios

Task	Method	Access Type	Faithfulness Hallucination Detection Granularity			Detection Method
			Object-level	Scene-level	Attribute-level	
I2T	GAVIE (Liu et al., 2023)	Black-box	✓	✓	✓	LLM
	Throne (Kaul et al., 2024)	Black-box	✓	✗	✗	LLM
	HaELM (Wang et al., 2023)	Black-box	✓	✓	✓	LLM
	FaithScore (Jing et al., 2023)	Black-box	✓	✓	✓	VQA
	CutPaste & Find (Nguyen et al., 2025)	Black-box	✓	✓	✓	VQA
	FactCheXcker (Heiman et al., 2025)	Black-box	✓	✓	✗	VQA
	M-HalDetect (Gunjal et al., 2023)	Black-box	✓	✓	✓	Caption
	AI-Feedback (Xiao et al., 2025)	Black-box	✓	✓	✓	Caption
	HalLocalizer (Park et al., 2025)	Black-box	✓	✓	✓	Caption
	CHAIR (Rohrbach et al., 2018)	Black-box	✓	✗	✗	Word Matching
	NOPE (Lovenia et al., 2024)	Black-box	✓	✗	✗	Word Matching
	VL-Uncertainty (Zhang et al., 2024a)	Black-box	✓	✓	✓	Uncertainty
	Image2Text2Image (Huang et al., 2025)	Black-box	✓	✓	✓	Similarity
	DHCP (Zhang et al., 2024b)	White-box	✓	✗	✗	Attention
	OPERA (Huang et al., 2024)	White-box	✓	✗	✗	Attention
	HallIE-Control (Zhai et al., 2023)	White-box	✓	✓	✓	Feature
	PROJECTAWAY (Jiang et al., 2025)	White-box	✓	✗	✗	Logits
	ContextualLens (Plukhan et al., 2025)	White-box	✓	✓	✓	Logits
T2I	SR _{2D} & VISOR (Gokhale et al., 2022)	Black-box	✓	✓	✗	Detector
	DALL-Eval (Cho et al., 2022)	Black-box	✓	✗	✗	Detector
	VPEEval (Cho et al., 2023)	Black-box	✓	✓	✗	Detector & VQA
	TIFA (Hu et al., 2023)	Black-box	✓	✓	✓	VQA
	T2I-CompBench (Huang et al., 2023)	Black-box	✓	✓	✓	VQA & Detector & Caption
	GraphQA (Qin et al., 2024)	Black-box	✓	✓	✓	VQA
Unified	DSG (Cho et al., 2024)	Black-box	✓	✓	✓	VQA
	VQAScore (Lin et al., 2024)	Black-box	✓	✓	✓	VQA
	LLMScore (Lu et al., 2023)	Black-box	✓	✓	✓	Detector & Caption
Unified	UNIHID (Chen et al., 2024)	Black-box	✓	✗	✓	VQA
	CSN (Fei et al., 2024)	Black-box	✓	✓	✓	VQA

T2I benchmarks like TIFA v1.0 (Hu et al., 2023) and T2I-FactualBench (Huang et al., 2024) use VLMs to evaluate hallucinations. These practices suggest two complementary developments: the use of synthetic image data in I2T benchmarks, and the application of VLM-based evaluation in T2I benchmarks. They highlight the potential for mutual enhancement between I2T and T2I research in advancing hallucination evaluation.

4 Hallucination Detection Methods

In this chapter, we systematically review the existing hallucination detection methods for both I2T and T2I models. Tab. 3 summarizes the key details of the methods, including detection type, granularity, and technical approach. Fig. 5 summarizes three common hallucination detection methods for I2T and T2I models, including detector-based, caption-based and VQA-based methods. The specific detection methods for I2T and T2I models are presented in Sec. 4.1 and Sec. 4.2 respectively, and unified methods in Sec. 4.3, with their relationship in Sec. 4.4.

4.1 Hallucination Detection Methods for I2T Models

Hallucination detection has become a critical component in assessing the reliability of VLMs, especially in I2T tasks such as image captioning and visual question answering. Recent advances in this area have introduced a variety of detection strategies, broadly categorized into black-box and white-box approaches depending on the accessibility of model internals (Chakraborty et al., 2025). Given an input image, VLMs typically produce a sequence of output tokens, token-level probabilities, and internal layer representations. Black-box methods operate solely on the textual outputs, and often leverage external evaluators or models for hallucination identification. White-box methods assume access to internal model signals such as attention weights or intermediate embeddings. This classification serves both practical and analytical purposes: it separates methods by their access requirements, which not only reflects real-world deployment constraints but also helps

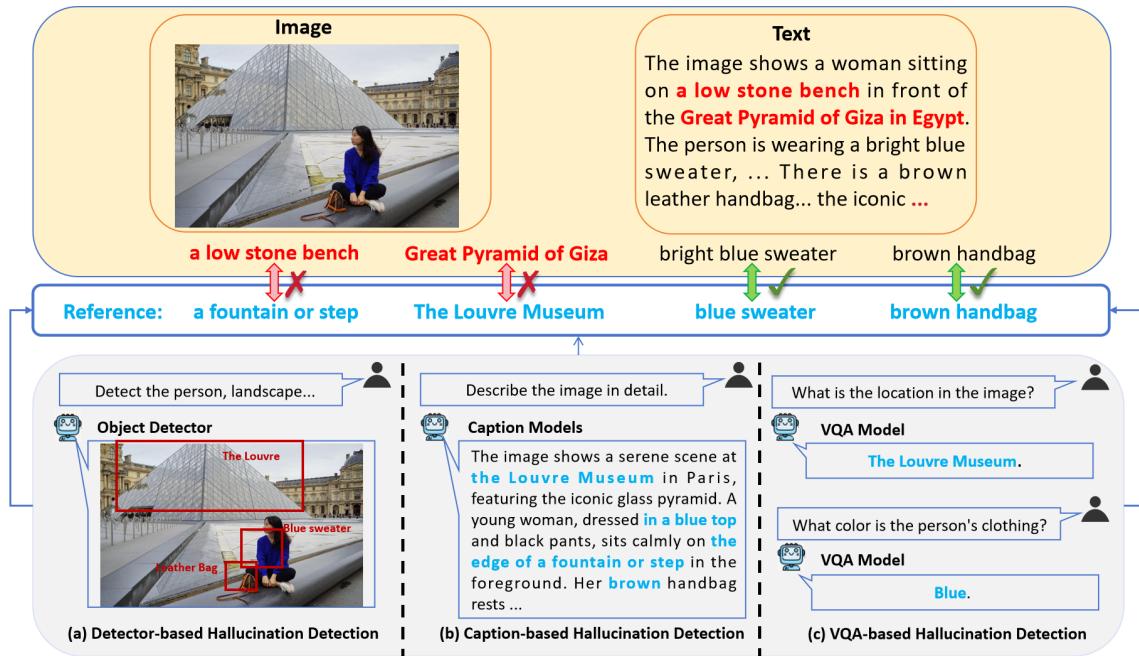


Fig. 5 Examples of detection methods. At the top is a mismatched image-text pair, and hallucinated text contents are highlighted in red. At the bottom are representative (a) Detector-based, (b) Caption-based, and (c) VQA-based hallucination detection methods arranged from left to right. The reference labels are highlighted in blue.

clarify the trade-offs between applicability and interpretability.

In the following sections, we organize existing VLM hallucination detection techniques according to this black-box/white-box taxonomy, highlighting their assumptions, mechanisms, and key applications.

4.1.1 Black-Box

Compared to white-box detection, black-box methods lack access to the internal information of VLMs and thus rely on external tools for hallucination detection. Current black-box hallucination detection methods for VLMs can be broadly categorized into these types: external model-based methods, word matching, uncertainty-based detection and similarity-based detection. External model-based methods leverage powerful models to analyze the outputs generated by the target VLMs, including LLM, VQA models, and caption models. Word matching applies text-matching paradigms to compare the generated outputs against reference data or input content. Uncertainty estimation detects hallucinations by computing the entropy or confidence

levels of the model’s responses, under the assumption that hallucinations are often associated with higher uncertainty. Similarity-based detection utilizes I2T models to generate reference images and calculate image-image similarity.

VQA-based Detection. FaithScore (Jing et al., 2023) evaluates hallucinations in generative questions generated by VLMs. It first leverages GPT to identify descriptive content in the model’s response, then breaks down this content into fine-grained atomic facts. A visual entailment model (VEM) is used to verify the consistency of these atomic facts with the visual information, helping to detect hallucinations in the response. The work utilizes FaithScore as the evaluation metric. The function $s(e_j^i, I)$ represents the VEM, where an output of 1 indicates that the image covers the content of the atomic fact. The final hallucination score is calculated by weighting these binary scores, shown in Tab. 1.

CutPaste & Find (Nguyen et al., 2025) is a training-free framework for detecting hallucinations in VLMs, leveraging a visual knowledge base and a VQA model. The knowledge base, built from

Visual Genome (Krishna et al., 2016), contains object crops, attributes, and relationships, serving as a structured visual reference. The detection pipeline consists of three main steps: (1) extracting a scene graph from the model-generated caption; (2) locating the mentioned objects via object detection; and (3) verifying the consistency of the caption through image similarity queries and VQA-based checks. Hallucinations are identified when inconsistencies arise between the caption, detection results, and the visual knowledge base. Experiments on POPE and R-Bench show that CutPaste & Find achieves competitive performance compared to prior methods.

FactCheXcker (Heiman et al., 2025) proposes an extensible and modular pipeline to detect and mitigate object and measurement hallucinations in radiology reports generated by medical VLMs. It constructs a benchmark focused on EndoTracheal Tube (ETT) placement, leveraging the MIMIC-CXR dataset and multiple report-generation models. The framework follows a query-code-update paradigm, where measurement queries are automatically generated from original reports, solved via domain-specific code modules, and incorporated back into updated reports. Evaluation metrics include precision for existence hallucination, Mean Absolute Error (MAE) for measurement hallucination, and composite scores combining both MAE and F1-score (shown in Tab. 1). Experiment results demonstrate that FactCheXcker significantly reduces hallucinations and improves both measurement accuracy and placement correctness across diverse models. Without requiring retraining, it serves as a valuable tool for enhancing clinical reliability in medical imaging AI systems.

LLM-based Detection. GAVIE (Liu et al., 2023) employs GPT to detect hallucinations in model responses. Specifically, the method inputs dense image descriptions with bounding box coordinates into GPT, then compares these descriptions to model responses for scoring. Evaluation criteria assess two dimensions: (1) Accuracy, measuring response consistency with image content; and (2) Relevance, determining whether responses directly address instructions. Hallucination scores derived from this dual-aspect analysis identify hallucinated content. Similar strategies are adopted

by CC-Eval (Zhai et al., 2023) and MMHal-Bench (Sun et al., 2024), which also rely on GPT to identify or score hallucinated contents.

Throne (Kaul et al., 2024) introduces an LLM-based framework for evaluating object-level hallucinations in free-form generative responses of VLMs. Throne employs public language models to analyze VLM-generated image descriptions prompted neutrally with "Describe this image in detail." Through Abstractive Question Answering (AQA), it systematically queries the presence of each object class in ground-truth annotations, generating precision-recall metrics that directly quantify hallucination severity in generative tasks. Experiments show that compared to CHAIR (Rohrbach et al., 2018) relying on word matching, Throne achieves lower hallucination misjudgment rates.

HaELM (Wang et al., 2023) uses a fine-tuned specialized LLM to serve as a hallucination detector for evaluating image descriptions. To construct the training corpus, it collects hallucination instances generated by VLMs and simulated via GPT models. The detector is based on LLaMA and trained using this curated dataset. Experimental results show that HaELM achieves hallucination detection performance comparable to GPT-3.5. Moreover, HaELM can be deployed locally, offering advantages in reduced computational cost.

Caption-based Detection. M-HalDetect (Gunjal et al., 2023) introduces a hallucination detection framework based on a fine-tuned multi-modal reward model. It constructs a training set using approximately 4,000 MSCOCO (Lin et al., 2014) images paired with captions generated by InstructBLIP, annotated at the sub-sentence level to capture fine-grained hallucinations. The reward model is initialized from InstructBLIP by replacing its final embedding layer with a classification head, and is trained to predict hallucination labels at both the sentence and segment levels. To support flexible detection, the model is trained under two settings: binary (accurate vs. hallucinated) and ternary (accurate, hallucinated, and analytical). The ternary setup allows the model to distinguish factual errors from subjective analysis, reducing confusion in ambiguous cases. Experiments show that the model effectively detects

hallucinations in generated captions and generalizes well to outputs from other LVLMs such as LLaVA and mPLUG-OWL.

AI-Feedback (Xiao et al., 2025) introduces a fine-grained hallucination detection framework for VLMs by constructing a detailed AI-generated feedback dataset. The model is trained to identify hallucinated content and assess its severity across various types, including object, attribute, and relationship. To build the dataset, 5000 images are sampled from Visual Genome (Krishna et al., 2016), and GPT-4 is leveraged to analyze the VLM-generated caption, which detects hallucinations and provides structured feedback. The detection model is trained on triplets consisting of the image, the VLM outputs, and the corresponding GPT feedback. In addition to enabling fine-grained hallucination detection, this work proposes Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO), a mitigation strategy that leverages severity labels to guide preference-based fine-tuning.

HalLocalizer (Park et al., 2025) proposes a lightweight framework for token-level hallucination detection in the process of VLM generating text responses. Trained on the HalLoc dataset (150000 token-annotated samples across VQA, instruction-following, and captioning tasks), it provides probabilistic confidence scores rather than binary decisions, addressing ambiguity in real-world scenarios. The architecture offers dual input modalities: processing either the VLMs’ final hidden states or generated text through a VisualBERT encoder with four linear classification heads (for object/attribute/relationship/scene hallucination types). Designed for minimal computational overhead, HalLocalizer achieves efficient inference with low latency while generating per-token hallucination probabilities. Its plug-and-play integration preserves base model efficiency, enabling seamless deployment in existing VLMs to enhance reliability without significant resource demands.

Word Matching. CHAIR (Rohrbach et al., 2018) and NOPE (Lovenia et al., 2024) employ word matching paradigms for hallucination detection while implementing specialized rules for distinct hallucination types. CHAIR targets object-level hallucinations through a synonym mapping system for 80 MSCOCO (Lin et al., 2014) objects

(e.g., “player” → “person”), involving text tokenization, singularization, and compound noun disambiguation (e.g., distinguishing “hot dog” from “dog”). Its innovation lies in fusing dual annotation sources: segmentation labels verify object presence, while human captions accommodate referential variations (e.g., semantic equivalence between “dining table” and “coffee table”). NOPE addresses negative semantic hallucinations via a dual evaluation framework. For Negative Pronoun (NegP) tasks, it applies rule-based accuracy to detect predefined negative pronouns ($A_{\text{NegP}} = \text{none, nothing, nobody}$); for non-negation tasks, it leverages METEOR metrics to assess unigram matches through weighted precision, recall, and alignment scores. Although relatively coarse, these methods offer simple and effective solutions; subsequent work such as Open-CHAIRR (Ben-Kish et al., 2024) builds upon this foundation by employing large language models as advanced text aligners.

Uncertainty-based Detection. Zhang et al. (2024a) proposes VL-Uncertainty, a black-box hallucination detection method based on model uncertainty estimation. They observe that vision-language models tend to produce consistent outputs for semantically similar inputs when confident, but show variance when hallucinations occur. To measure this, it generates paraphrased queries and perturbed images (e.g., Gaussian blur) and computes the entropy of model predictions across them. High entropy indicates uncertainty and potential hallucination. This method avoids manual annotations and demonstrates strong performance across 10 models and 4 benchmarks, highlighting uncertainty as a meaningful hallucination indicator.

Similarity-based Detection. Huang et al. (2025) introduces Image2Text2Image, a novel framework for evaluating the quality of image captioning based on image-to-image consistency, thereby removing the dependence on human-annotated references. It leverages diffusion models (e.g., Stable Diffusion) to assess the fidelity of generated captions in a self-supervised manner. The proposed framework consists of four main components: (1) an image captioning module that generates textual descriptions from input images; (2) an image encoder that extracts visual features; (3) a diffusion model that synthesizes images

from the generated captions; and (4) a similarity module that compares the original and synthesized images in feature space. Low similarity scores suggest potential deficiencies in the captions, revealing discrepancies between the original visual content and its textual representation. This approach enables automatic evaluation without reliance on human labels.

4.1.2 White-Box

Unlike black-box methods that rely solely on model outputs, white-box approaches leverage internal representations for hallucination detection. These signals provide granular insights into the model’s generation process, enabling deep analysis of hallucination patterns. Based on the internal parameters accessed, white-box hallucination detection for VLMs falls into three categories: attention-based detection, feature-based detection and logits-based detection. These approaches build upon the observation that internal representations exhibit distinguishable patterns between hallucinated and faithful content. By quantifying these disparities, white-box methods aim to detect and mitigate hallucinations effectively.

Attention-based Detection. DHCP ([Zhang et al., 2024b](#)) presents a white-box hallucination detection framework that exploits cross-modal attention patterns in LVLMs. Specifically, it captures the attention weights from the language model to visual tokens during the generation of the first output token, forming a structured (token \times layer \times head) attention tensor. Empirical analysis on the POPE-Extended dataset reveals that despite being textually identical, hallucinated and non-hallucinated responses exhibit significantly different attention distributions, which vary by hallucination type. Building on this observation, DHCP trains a lightweight two-stage MLP-based classifier (DHCP-d): the first-stage detector prioritizes high recall to flag potential hallucinations, while the second-stage refines predictions for higher precision. This cascaded design leads to a substantial performance improvement, achieving over 93% accuracy on the POPE-Extended test set. DHCP demonstrates that internal attention dynamics provide reliable and interpretable signals for hallucination detection without requiring access to model retraining or output references.

OPERA ([Huang et al., 2024](#)) detects and alleviates hallucination in VLMs by mitigating over-reliance on sparse summary tokens during decoding. Experiments show that hallucinations correlate with knowledge aggregation patterns in self-attention, where models prioritize recent summary tokens over contextual image information. Based on this observation, OPERA introduces an over-trust penalty to suppress over-aggregation in beam-search logits and a retrospection-allocation strategy to trigger token re-selection when critical patterns emerge. Experiments show that OPERA generally improves the VLMs’ performances on popular benchmarks including POPE ([Li et al., 2023](#)) and MME ([Fu et al., 2023](#)).

Feature-based Detection. HallE-Control ([Zhai et al., 2023](#)) identifies that hallucinations occur when LVLM language responses contain finer-grained details than the visual module can verify, leading to unwarranted inferences. The method introduces a control module with a trainable projection layer. This layer transforms output embeddings in a way that reduces hallucinated content by modulating the model’s reliance on unverifiable details. Hallucinations are further reduced by adjusting control parameters to modulate knowledge integration. This parameter-efficient approach reduces object hallucination by 44% compared to LLaVA-7B while maintaining object coverage, enabling controllable generation of contextual versus parametric knowledge.

Logits-based Detection. PROJECTAWAY ([Jiang et al., 2025](#)) detects and alleviates object-level hallucination in VLMs through logits lens. This technique, originally introduced in the context of language models, involves directly mapping intermediate activations to the vocabulary space using the unembedding layer, allowing for an interpretable view of token predictions at different layers. By applying this technique to VLMs, PROJECTAWAY probes each image patch to determine the presence of objects. This approach reveals that real objects exhibit significantly higher output probabilities than hallucinated ones at the token level. These probability distributions further enable patch-level spatial localization of objects. Building on this granular analysis, the method applies knowledge erasure via linear orthogonalization of image features against hallucinated object features. Results show that targeted edits to a model’s latent representations

achieve up to 25.7% hallucination reduction on COCO2014 while preserving performance.

ContextualLens (Phukan et al., 2025) detects and mitigates hallucinations in VLMs through token-level analysis and patch/bounding box-level grounding. Unlike the logit lens method that relies on non-contextual unembedding layer features (Jiang et al., 2025), ContextualLens leverages contextual token embeddings from middle VLM layers to capture multi-token concepts and spatial relationships. This enables granular hallucination detection across complex categories (attributes, actions, OCR) by computing cosine similarity between averaged answer token embeddings and image patch embeddings. For grounding, it generates precise bounding boxes via contextual embedding alignment, advancing zero-shot segmentation to grounded VQA. Evaluated on HQH benchmark (Yan et al., 2024), the method achieves significant improvements in mAP for hallucination detection while maintaining the training-free efficiency.

Summary of I2T Hallucination Detection. As shown in Tab. 3, hallucination detection primarily relies on black-box methods. Most black-box methods utilize external models to assist in detection, reducing manual labeling efforts. However, the detection results are constrained by the accuracy and generalizability of the auxiliary models. To further improve the detection performance, some studies have constructed hallucination detection datasets and fine-tuned existing models as detectors for evaluation (Gunjal et al., 2023). While effective, these approaches incur high training costs and their performance is heavily dependent on the quality of the constructed datasets. In contrast, the uncertainty-based methods (Zhang et al., 2024a) do not require additional training, but such approaches are prone to misjudging complex cases.

In comparison, white-box detection methods depend on the model’s internal signals, such as attention and feature. By accessing these internal representations, white-box methods enable the tracing of hallucination sources and offer valuable insights into the model’s generation process. They help reveal the flow of information and decision-making across layers, which can aid in diagnosing and mitigating specific hallucination cases. However, due to the complexity of internal

mechanisms, current white-box methods are generally limited to coarse-grained or holistic detection. They often struggle to distinguish between fine-grained hallucination types within generated outputs.

Overall, while black-box methods tend to offer better generalizability and easier deployment, both approaches face significant limitations. Existing detection efforts are primarily centered on faithfulness hallucinations. The detection of factuality hallucinations, which require reasoning over external or commonsense knowledge, remains an underexplored challenge in the field.

4.2 Hallucination Detection Methods for T2I Models

Early assessments of image generation primarily focus on coarse-grained evaluations such as image quality, using metrics like FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021). Building on these evaluations, hallucination detection in T2I models emphasizes image-text alignment at a fine-grained level, including object types, attributes, and relationships. Current hallucination detection approaches for T2I models largely rely on black-box methods, including object detection and VLMs. This section provides a comprehensive review of existing black-box methods for hallucination detection in T2I models.

4.2.1 Black-Box

Black-box hallucination detection methods in T2I tasks typically rely on external tools or models to verify whether the image accurately reflects the text prompt, including object detector, VQA model and caption model.

Detector-based Detection. SR_{2D} & VISOR (Gokhale et al., 2022) leverages pretrained object detectors and spatial relationship recognition algorithms to evaluate spatial consistency in T2I tasks. Objects and their spatial relationships are automatically identified in generated images. These elements are compared against the spatial layout specified in the text prompt to assess alignment with the textual description. This automated metric reveals severe spatial relationship limitations in T2I models, while demonstrating strong alignment with human judgment.

DALL-Eval (Cho et al., 2022) employs object detection models to quantify consistency failures in T2I models, specifically evaluating object existence, counting, and spatial relationships. The framework utilizes synthetic training data, which is derived from MSCOCO (Lin et al., 2014) and procedurally generated in Unity to train a DETR-based detector, which subsequently analyzes generated images for hallucinatory content.

VQA-based Detection. VPEval (Cho et al., 2023) employs a programmatic evaluation framework utilizing specialized visual modules to detect hallucinations across distinct categories, including object presence, OCR accuracy, spatial relationships, and quantitative counting. This approach dynamically assembles skill-specific evaluation pipelines, where dedicated modules (e.g., object detection and counting systems) are invoked to verify prompt-image alignment for targeted attributes. As an interpretable visual programming paradigm, VPEval overcomes limitations of monolithic evaluation models by providing skill-optimized assessments with explanatory justifications, demonstrating significantly stronger correlation with human judgment than conventional single model-based evaluation.

TIFA (Hu et al., 2023) leverages LLMs and VQA models to assess T2I generation fidelity. For a given text prompt, question-answer pairs are automatically generated using an LLM and filtered by a question-answer model. These questions are then posed to a VQA model using the generated image. Faithfulness is evaluated by assessing consistency between the VQA answers and the ground-truth answers derived from the text prompt. This reference-free metric enables fine-grained, interpretable evaluations, demonstrates superior correlation with human judgments, and reveals significant model deficiencies in spatial relations and object composition.

T2I-CompBench (Huang et al., 2023) evaluates compositional fidelity in synthesized images across three dimensions: object attribute, spatial relationships, and scene coherence. To assess object and attribute hallucinations, binary questions derived from text prompts are evaluated using VQA models. For spatial relationship verification, object detection combined with spatial recognition algorithms is employed. Regarding scene-level coherence, MiniGPT-4-generated descriptions are scored against textual requirements. Results show

limitations of T2I models to generate complex scene compositions.

GraphQA (Qin et al., 2024) proposes a method for detecting hallucinations in T2I models by integrating scene graphs and VQA models. LLMs automatically generate evaluative questions from text prompts, while object detection algorithms (e.g., GroundedSAM) and VLMs extract visual attributes and relationships to construct image knowledge graphs. Graph Question Answering then scores response consistency between the graphs and generated images to identify hallucinations. This structured approach establishes a quantitative, interpretable framework for evaluating image-text alignment that closely correlates with human assessment standards.

Davidsonian Scene Graph (DSG) (Cho et al., 2024) employs a graph-based framework with VQA to detect hallucinations in generated images. Unlike GraphQA’s approach (Qin et al., 2024), this methodology integrates dependency graphs to ensure comprehensive semantic coverage while preventing contradictory or redundant question generation. Drawing on linguistic formalization theories, the construction process recursively decomposes sentences into atomic propositions implemented as directed acyclic graphs. During VQA validation, negative responses to parent questions automatically terminate dependent sub-question evaluation. This linguistically-grounded framework resolves critical reliability challenges by guaranteeing question uniqueness, eliminating answer inconsistencies, and providing adaptable semantic coverage for robust image-text faithfulness evaluation.

VQAScore (Lin et al., 2024) leverages VQA architectures to quantify fine-grained text-visual alignment. The framework reformulates prompt-image consistency as a binary classification task, evaluating through VQA models with queries: “Does this figure show {Text}?” Alignment scores derive from the probability of affirmative responses. To enhance detection capabilities, the method incorporates CLIP-FlanT5, a bidirectional image-question encoder fine-tuned specifically for hallucination identification. This approach establishes a state-of-the-art compositional evaluation metric that overcomes lexical ambiguity limitations in conventional methods, enabling reliable assessment of complex prompts across eight benchmarks.

Caption-based Detection. LLMScore (Lu et al., 2023) integrates object detection architectures with LLMs to identify hallucinations across object-level and scene-level granularity. Object detector localizes the visual entities, which is followed by region-to-text transformers and image captioning systems that generate localized and holistic descriptions. These descriptions are evaluated against the source text prompts via LLM-based alignment assessment to detect hallucinations. Results show that the framework demonstrates stronger correlation with human judgment than conventional metrics CLIP and BLIP.

Summary of T2I Hallucination Detection. Hallucination detection methods for T2I models mainly evaluate the consistency between generated images and textual prompts using external or fine-tuned models. Early works leverage specialized tools such as object detector (Gokhale et al., 2022). These methods are limited in coverage and mainly focus on coarse-granularity hallucination, such as object and spatial relationship. Another common approach is to use LLMs to decompose the prompt into relevant questions, and then employ VLMs to answer these questions based on the generated image (Qin et al., 2024; Cho et al., 2024), enabling the detection of various types of hallucinations. In addition, methods concerning factuality hallucinations are relatively few. Therefore, exploring comprehensive detection methods remains an important direction for future research.

4.3 Unified Hallucination Detection Methods

Unified hallucination detection for both I2T and T2I models is emerging as a promising direction due to its potential to bridge cross-modal evaluation. Some studies have explored hallucination detection from a unified perspective that integrates both I2T and T2I tasks through image-text alignment (Chen et al., 2024; Fei et al., 2024). Compared to methods typically designed for I2T or T2I models, unified approaches capture both visual and textual content, requiring the integration of more external models. Recent unified methods are primarily black-box approaches. Current black-box unified hallucination detection methods

rely on external models to analyze the image-text pairs and identify cross-modal inconsistencies.

VQA-based Detection. UNIHD (Chen et al., 2024) introduces a unified multi-modal framework for detecting hallucinations in T2I and I2T tasks by assessing image-text consistency across four categories: object existence, attributes, scenes, and knowledge-based factuality. The methodology decomposes text prompts into atomic facts using GPT, and then generates verification questions from multiple perspectives. For visual validation, specialized tools, including object detectors, attribute classifiers (GPT-based), and external knowledge retrievers (e.g., Google Search), operate in parallel to gather evidentiary support. These outputs, combined with task metadata, are integrated into a VLM for hallucination assessment. This comprehensive approach establishes a novel paradigm for multi-category hallucination detection. It overcomes the limitations of task-specific methods through tool-assisted evidence validation and enables granular evaluation via the MHaluBench benchmark.

CSN (Cross-graph Siamese Network) (Fei et al., 2024) establishes a unified hallucination detection framework for I2T and T2I tasks using structured scene graph representations. This approach detects object-existence, attribute, and relationship hallucinations through comparative analysis of dual scene graphs. Initially, a scene graph parser constructs a Visual Scene Graph (VSG) from generated images and a Textual Scene Graph (TSG) from input prompts. Subsequently, a pretrained Cross Graph Siamese Network with cross-attention mechanisms identifies semantic discrepancies between VSG and TSG representations. Ultimately, a large language model adjudicates hallucination occurrences based on detected inconsistencies. This structured framework pioneers fine-grained multi-modal hallucination detection by resolving semantic gaps in comprehension and generation processes, outperforming baselines on MHaluBench.

Summary of Unified Hallucination Detection. Unified hallucination detection methods aim to build a comprehensive cross-modal framework for identifying hallucinations in both I2T and T2I tasks. By leveraging cross-modal

consistency, question decomposition, and graph-based reasoning, these methods enable a unified approach to hallucination detection, often assisted by external models such as VQA models or LLMs. However, challenges remain, including the reliance on external models, varying hallucination patterns across modalities, and limited detection efficiency. Developing more robust, efficient, and generalizable hallucination detection frameworks is an important future research direction.

4.4 Relationship between Hallucination Detection in I2T and T2I Models

As shown in Tab. 3 and analysis above, hallucination detection in I2T and T2I models shares the common goal of identifying semantic inconsistencies between the input and output modalities (Chen et al., 2024; Fei et al., 2024). Both approaches commonly depend on external models such as VQA models (e.g., CutPaste & Find, TIFA) (Hu et al., 2023; Nguyen et al., 2025) or large language models (e.g., GAVIE, LLM-Score) (Liu et al., 2023; Lu et al., 2023) to verify cross-modal consistency.

In addition to these shared characteristics, a difference lies in white-box methods. I2T models take advantage of techniques such as attention analysis (e.g., DHCP) (Zhang et al., 2024b) and feature (e.g., HallE-Control) (Zhai et al., 2023), which enable direct inspection of the generation process and help trace the origins of hallucinated content. T2I hallucination detection primarily relies on external tools, including object detectors and VQA models, and typically lacks direct access to internal generation signals. Future work could consider incorporating white-box strategies into T2I models to enhance detection effectiveness.

Current unified hallucination detection methods typically check the alignment between image and text content by combining multi-modal features (Chen et al., 2024; Fei et al., 2024). These methods combine outputs from multiple external pretrained models, allowing for fine-grained identification of hallucinations. Future research could explore new mechanisms, such as fine-grained uncertainty estimation, building effective unified

frameworks that better capture hallucinations in both T2I and I2T tasks.

5 Challenges and Future Directions

Despite growing efforts and progress in MLLM hallucination evaluation, several critical challenges remain unresolved and underexplored. Below, we summarize these key issues to offer guidance and suggestions for future research.

5.1 Scalable and In-depth Hallucination Evaluation

With the advancement of reasoning capabilities in MLLMs, the complexity and diversity of their generated outputs have increased significantly (Mu et al., 2023; Shao et al., 2024; Zhao et al., 2025). This evolution underscores the urgent need for more sophisticated automatic evaluation methods that transcend simple binary or surface-level assessments of hallucination (Jiang et al., 2024; Jing et al., 2023; Rawte et al., 2025; Zhou et al., 2024). Future research should prioritize the development of comprehensive evaluation frameworks capable of automatically capturing fine-grained characteristics of hallucinations. Given the high cost of human annotation, robust and efficient evaluation tools are essential for scalable model development and deployment. Moreover, rather than merely categorizing outputs as hallucinated or non-hallucinated, evaluation frameworks should quantify the degree, type, and impact of hallucinations. This includes distinguishing minor inconsistencies from critical factual errors and identifying sources such as semantic drift, object misrecognition, or commonsense violations.

5.2 Rigorous and Comprehensive Hallucination Evaluation

As MLLMs continue to advance in their reasoning capabilities and generate increasingly complex multi-modal outputs, ensuring the robustness and reliability of hallucination evaluation becomes ever more critical (Zhang et al., 2024; Guo et al., 2024). As noted in Sec. 3.1 and Sec. 3.2, existing evaluation methods often rely on limited datasets, heuristic metrics, or binary labels, which fail to capture the nuanced, diverse,

and context-dependent nature of hallucinations. Moreover, issues such as data leakage, insufficient coverage of hallucination types, and a lack of standardized evaluation protocols undermine the fairness and credibility of current assessments (Guan et al., 2023; Chen et al., 2024). Additionally, cross-modal and cross-task consistency checks, the use of external knowledge, and the adoption of standardized reporting protocols will be essential to achieve more trustworthy and reproducible evaluations (Ganesh et al., 2025; Wu et al., 2025). Ultimately, rigorous and comprehensive evaluation practices will lay a more solid foundation for understanding, comparing, and mitigating hallucinations in next-generation MLLMs.

5.3 Explainable and Reliable Hallucination Evaluation

Hallucinations in VLMs often result from subtle misalignments between visual and textual modalities, involving complex, non-linear, and context-dependent interactions (Rudin, 2019; Yang et al., 2023). Developing explainability methods that can disentangle and trace these multi-modal reasoning processes remains a significant challenge (Yan et al., 2024; Ben Melech Stan et al., 2024). As mentioned in Sec. 3.1 and Sec. 3.2, existing hallucination evaluation benchmarks predominantly rely on response-level assessments, which oversimplify the problem by ignoring the nuanced spectrum of hallucination types and severities. This lack of interpretability limits insight into the origins and locations of hallucinations, impeding effective diagnosis and mitigation. Advancing explainable and reliable hallucination evaluation is essential to enable root cause analysis, improve model design, and support trustworthy deployment, ultimately fostering greater user confidence through transparent evidence of model output validity.

5.4 Unified Detection for I2T and T2I Models

Current research on I2T and T2I hallucination detection faces notable gaps, including limited investigation and fragmented evaluation. Future efforts could explore unified modular paradigms for granular hallucination analysis across both modalities. Approaches like UNIHD (Chen et al.,

2024) illustrate the feasibility of consistency-based paradigms for detecting hallucinations in both I2T and T2I models, reducing reliance on modality-specific pipelines. Modular architectures (e.g., VP-Eval (Cho et al., 2023), UNIHD (Chen et al., 2024)) offer precise hallucination localization through specialized components but often require integration with external tools. Establishing standardized, modular frameworks would not only enhance comparability across studies but also improve reproducibility in future research.

5.5 Domain-specific Factuality Hallucination Evaluation

Evaluating factuality hallucinations in MLLMs poses a particular challenge when applied to specialized domains such as medicine (Ayaz et al., 2024), agriculture (Arshad et al., 2025), embodied intelligence (Wu et al., 2025; Sarch et al., 2024), and autonomous driving (You et al., 2024). Unlike general-domain scenarios, domain-specific contexts require precise and accurate knowledge grounded in complex, technical, or regulated information. However, efforts to systematically evaluate hallucinations in these domain-specific contexts remain limited. As Tab. 2 shows, benchmarks like Med-HallMark (Chen et al., 2024) assess medical VLMs using radiology images, yet comparable frameworks for other domains are scarce. Future research could establish specialized benchmarks to quantify factual inconsistencies in domain-adaptive scenarios, particularly for safety-sensitive applications.

6 Conclusion

Recent MLLMs have demonstrated strong capabilities in multi-modal tasks, but still suffer from the hallucination problem. This survey provides a comprehensive review of hallucination issues observed in existing MLLMs, with a particular focus on both I2T and T2I paradigms. We systematically categorize hallucinations into faithfulness and factuality types and analyze current evaluation methodologies. We highlight several key trends in the field, including the emergence of automated benchmark construction, the movement towards fine-grained evaluation, and the shift towards a unified perspective that bridges

evaluation approaches between I2T and T2I tasks. Furthermore, our review of hallucination detection methods highlights the growing reliance on external tools to enable black-box detection strategies. Based on these insights, we outline several promising future directions for advancing hallucination evaluation in MLLMs. We hope this survey could provide researchers in this field with a clearer understanding of the current landscape and offer valuable guidance for future exploration.

References

- Open AI: Hello GPT-4o. <https://openai.com/zh-Hans-CN/index/hello-gpt-4o/>. Accessed: 2025-07-13 (2024)
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. *CoRR* (2023) [2312.11805](#)
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR* (2024) [2409.12191](#)
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. *CVPR*, 10684–10695 (2022)
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. *ICML*, 8821–8831 (2021). Pmlr
- Zhang, H., Shao, W., Liu, H., Ma, Y., Luo, P., Qiao, Y., Zheng, N., Zhang, K.: B-avibench: Towards evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *TIFS* (2024)
- Guo, Q., Pang, S., Jia, X., Liu, Y., Guo, Q.: Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *TIFS* (2024)
- Aafaq, N., Akhtar, N., Liu, W., Shah, M., Mian, A.: Language model agnostic gray-box adversarial attack on image captioning. *TIFS* **18**, 626–638 (2022)
- Liu, H., Wang, W., Sun, H., Rocha, A., Li, H.: Robust domain misinformation detection via multi-modal feature alignment. *TIFS* **19**, 793–806 (2023)
- Vice, J., Akhtar, N., Hartley, R., Mian, A.: Bagm: A backdoor attack for manipulating text-to-image generative models. *TIFS* **19**, 4865–4880 (2024)
- Zhang, H., Tang, H., Sun, Y., He, S., Li, Z.: Modality-specific interactive attack for vision-language pre-training models. *TIFS* (2025)
- Fan, W., Li, H., Jiang, W., Hao, M., Yu, S., Zhang, X.: Stealthy targeted backdoor attacks against image captioning. *TIFS* **19**, 5655–5667 (2024)
- Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS* **36**, 78723–78747 (2023)
- Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. *CoRR* (2024) [2404.18930](#)
- Lan, W., Chen, W., Chen, Q., Pan, S., Zhou, H., Pan, Y.: A survey of hallucination in large visual language models. *CoRR* (2024) [2410.15359](#)
- Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. *CoRR* (2024) [2402.00253](#)
- Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. *EMNLP*, 4035–4045 (2018)
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.-r.: Evaluating object hallucination in large vision-language models. *EMNLP* (2023)
- Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H.: Unified hallucination detection for multimodal large language models. *ACL* (2024)
- Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *ICCV*, 20349–20360 (2023)
- Wu, M.-K., Ji, J., Huang, O., Li, J., Wu, Y., Sun, X., Ji, R.: Evaluating and analyzing relationship hallucinations in large vision-language models. *ICML* (2024)
- Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. *CoRR* (2023) [2306.13394](#)
- Seth, A., Manocha, D., Agarwal, C.: Hallucinogen: A benchmark for evaluating object hallucination in large visual-language models. *CoRR* (2024) [2412.20622](#)
- Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W.X., Chen, X., Lin, Y., Wen, J.-R., Han, J.: Don't make your llm an evaluation benchmark cheater. *CoRR* (2023) [2311.01964](#)
- Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. *ICLR* (2023)
- Gokhale, T., Palangi, H., Nushi, B., Vineet, V.,

- Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. *CoRR* (2022) [2212.10015](#)
- Huang, Z., He, W., Long, Q., Wang, Y., Li, H., Yu, Z., Shu, F., Chan, L., Jiang, H., Gan, L., et al.: T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *CoRR* (2024) [2412.04300](#)
- Hu, H., Zhang, J., Zhao, M., Sun, Z.: Ciem: Contrastive instruction evaluation method for better instruction tuning. *NeurIPS Workshop* (2023)
- Chen, Z., Zhu, Y., Zhan, Y., Li, Z., Zhao, C., Wang, J., Tang, M.: Mitigating hallucination in visual language models with visual supervision. *CoRR* (2023) [2311.16479](#)
- Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *CVPR*, 14375–14385 (2023)
- Wang, L., He, J., Li, S., Liu, N., Lim, E.-P.: Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *MMM* (2023)
- Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D., Chai, J.: Multi-object hallucination in vision language models. *NeurIPS* **37**, 44393–44418 (2024)
- Qiu, H., Huang, J., Gao, P., Qi, Q., Zhang, X., Shao, L., Lu, S.: Longhalqa: Long-context hallucination evaluation for multimodal large language models. *CoRR* (2024) [2410.09962](#)
- Liu, J., Fu, Y., Xie, R., Xie, R., Sun, X., Lian, F., Kang, Z., Li, X.: Phd: A chatgpt-prompted visual hallucination evaluation dataset. *CVPR*, 19857–19866 (2025)
- Cao, Q., Cheng, J., Liang, X., Lin, L.: Visdiahalbench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. *ACL*, 12161–12176 (2024)
- Wu, X., Guan, T., Li, D., Huang, S., Liu, X., Wang, X., Xian, R., Shrivastava, A., Huang, F., Boyd-Graber, J., et al.: Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *Findings of the EMNLP*, 8395–8419 (2024)
- Ben-Kish, A., Yanuka, M., Alper, M., Giryes, R., Averbuch-Elor, H.: Mitigating open-vocabulary caption hallucinations. *EMNLP*, 22680–22698 (2024)
- Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., Li, C., Li, M.: Halle-control: controlling object hallucination in large multimodal models. *CoRR* (2023) [2310.01779](#)
- Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y.,
- Gan, C., Gui, L., Wang, Y.-X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. *Findings of the ACL*, 13088–13110 (2024)
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. *ICLR* (2023)
- Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., Fung, P.: Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ALVR Workshop*, 37–58 (2024)
- Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al.: Evaluation and analysis of hallucination in large vision-language models. *CoRR* (2023) [2308.15126](#)
- Rani, A., Rawte, V., Sharma, H., Anand, N., Rajbangshi, K., Sheth, A., Das, A.: Visual hallucination: Definition, quantification, and prescriptive remediations. *CoRR* (2024) [2403.17306](#)
- Villa, A., Léon, J., Soto, A., Ghanem, B.: Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *CVPR*, 492–502 (2025)
- Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR* (2023) [2311.07397](#)
- Huang, W., Liu, H., Guo, M., Gong, N.: Visual hallucinations of multi-modal large language models. *Findings of the ACL*, 9614–9631 (2024)
- Jiang, C., Ye, W., Dong, M., Jia, H., Xu, H., Yan, M., Zhang, J., Zhang, S.: Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *ACM MM* (2024)
- Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. *CoRR* (2024) [2406.10185](#)
- Tu, Y., Hu, R., Sang, J.: Ode: Open-set evaluation of hallucinations in multimodal large language models. *CVPR*, 19836–19845 (2025)
- Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. *ICCV*, 3043–3054 (2023)
- Bakr, E.M., Sun, P., Shen, X., Khan, F.F., Li, L.E., Elhoseiny, M.: Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *ICCV*, 20041–20053 (2023)
- Lim, Y., Choi, H., Shim, H.: Evaluating image hallucination in text-to-image generation with question-answering. *AAAI* **39**(25), 26290–26298 (2025)
- Niu, Y., Ning, M., Zheng, M., Lin, B., Jin, P., Liao, J., Ning, K., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *CoRR* (2025) [2503.07265](#)

- Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: Evaluating and improving compositional text-to-visual generation. *CVPR*, 5290–5301 (2024)
- Jing, L., Li, R., Chen, Y., Du, X.: Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *EMNLP* (2023)
- Kaul, P., Li, Z., Yang, H., Dukler, Y., Swaminathan, A., Taylor, C., Soatto, S.: Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *CVPR*, 27228–27238 (2024)
- Nguyen, C.-D., Wu, X., Vu, D.A., Zhao, S., Nguyen, T., Luu, A.T.: Cutpaste&find: Efficient multimodal hallucination detector with visual-aid knowledge base. *CoRR* (2025) [2502.12591](#)
- Huang, J.-H., Zhu, H., Shen, Y., Rudinac, S., Kanoulas, E.: Image2text2image: A novel framework for label-free evaluation of image-to-text generation with text-to-image diffusion models. *MMM*, 413–427 (2025). Springer
- Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. *AAAI* (2023)
- Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *AAAI* (2025)
- Park, E., Kim, M., Kim, G.: Halloc: Token-level localization of hallucinations for vision language models. *CVPR*, 29893–29903 (2025)
- Heiman, A., Zhang, X., Chen, E., Kim, S.E., Rajpurkar, P.: Factcheckr: Mitigating measurement hallucinations in chest x-ray report generation models. *CVPR*, 30787–30796 (2025)
- Zhang, R., Zhang, H., Zheng, Z.: Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *CoRR* (2024) [2411.11919](#)
- Zhang, Y., Xie, R., Chen, J., Sun, X., Wang, Y., et al.: Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. *CoRR* (2024) [2411.18659](#)
- Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CVPR*, 13418–13427 (2024)
- Jiang, N., Kachinthaya, A., Petryk, S., Gandelsman, Y.: Interpreting and editing vision-language representations to mitigate hallucinations. *ICLR* (2025)
- Phukan, A., Divyansh, D., Morj, H.K., Vaishnavi, V., Saxena, A., Goswami, K.: Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. *NAACL*, 9661–9675 (2025)
- Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. *ICCV*, 3020–3031 (2022)
- Cho, J., Zala, A., Bansal, M.: Visual programming for step-by-step text-to-image generation and evaluation. *NeurIPS* **36**, 6048–6069 (2023)
- Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y.: Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *NeurIPS* **36**, 23075–23093 (2023)
- Qin, Z., Cheng, D., Wang, H., Yi, H., Shao, Y., Fan, Z., Li, K., Lao, Q.: Evaluating hallucination in text-to-image diffusion models with scene-graph based question-answering agent. *CoRR* (2024) [2412.05722](#)
- Cho, J., Hu, Y., Baldridge, J.M., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *ICLR* (2024)
- Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. *ECCV*, 366–384 (2024). Springer
- Fei, H., Luo, M., Xu, J., Wu, S., Ji, W., Lee, M.-L., Hsu, W.: Fine-grained structural hallucination detection for unified visual comprehension and generation in multimodal llm. *ACM Multimedia Workshop*, 13–22 (2024)
- Bentall, R.P.: The illusion of reality: a review and integration of psychological research on hallucinations. *Psychological bulletin* **107**(1), 82 (1990)
- Slade, P.D., Bentall, R.P.: Sensory Deception: A Scientific Analysis of Hallucination. Johns Hopkins series in contemporary medicine and public health. Johns Hopkins University Press, London (1988)
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**(12), 1–38 (2023)
- Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. *CVPR*, 5505–5514 (2018)
- Baker, S., Kanade, T.: Hallucinating faces. *IEEE FG*, 83–88 (2000). IEEE
- Liu, C., Shum, H.-Y., Freeman, W.T.: Face hallucination: Theory and practice. *IJCV* **75**(1), 115–134 (2007)
- Wang, N., Tao, D., Gao, X., Li, X., Li, J.: A comprehensive survey to face hallucination. *IJCV* **106**(1), 9–30 (2014)
- Huang, H., He, R., Sun, Z., Tan, T.: Wavelet domain generative adversarial network for multi-scale face hallucination. *IJCV* **127**(6), 763–784 (2019)

- Quan, W., Chen, J., Liu, Y., Yan, D.-M., Wonka, P.: Deep learning-based image and video inpainting: A survey. *IJCV* **132**(7), 2367–2400 (2024)
- Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. *ICCV*, 3018–3027 (2017)
- Yang, M., Wang, Z.: Image synthesis under limited data: A survey and taxonomy. *IJCV* **133**(6), 3689–3726 (2025)
- Kayhan, O.S., Vredebregt, B., Van Gemert, J.C.: Hallucination in object detection—a study in visual part verification. *ICIP*, 2234–2238 (2021). IEEE
- Piasco, N., Sidibé, D., Gouet-Brunet, V., Demonceaux, C.: Improving image description with auxiliary modality for visual localization in challenging conditions. *IJCV* **129**(1), 185–202 (2021)
- You, J., Shi, H., Jiang, Z., Huang, Z., Gan, R., Wu, K., Cheng, X., Li, X., Ran, B.: V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *CoRR* (2024) [2408.09251](#)
- Ayaz, M., Khan, M., Saqib, M., Khelifi, A., Sajjad, M., Elsaddik, A.: Medvlm: Medical vision-language model for consumer devices. *IEEE Consumer Electronics Magazine* (2024)
- Aithal, S.K., Maini, P., Lipton, Z., Kolter, J.Z.: Understanding hallucinations in diffusion models through mode interpolation. *NeurIPS* **37**, 134614–134644 (2024)
- Kim, S., Jin, C., Diethe, T., Figini, M., Tregidgo, H.F., Mollokandov, A., Teare, P., Alexander, D.C.: Tackling structural hallucination in image translation with local diffusion. *ECCV*, 87–103 (2024). Springer
- Wang, H., Cao, J., Liu, J., Zhou, X., Huang, H., He, R.: Hallo3d: Multi-modal hallucination detection and mitigation for consistent 3d content generation. *NeurIPS* **37**, 118883–118906 (2024)
- Saleh, M., Tabatabaei, A.: Building trustworthy multimodal ai: A review of fairness, transparency, and ethics in vision-language tasks. *IJWR* **8**(2), 11–24 (2025)
- Agnese, J., Herrera, J., Tao, H., Zhu, X.: A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(4), 1345 (2020)
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM TOIS* **43**(2), 1–55 (2025)
- Dai, W., Liu, Z., Ji, Z., Su, D., Fung, P.: Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *EACL*, 2136–2148 (2023)
- Nesteruk, S., Svetlana, I., Andrey, S.: Image dataset augmentation a survey and taxonomy. *Measurements and Instrumentation for Machine Vision*, 110–136 (2024)
- Arshad, M.A., Jubery, T.Z., Roy, T., Nassiri, R., Singh, A.K., Singh, A., Hegde, C., Ganapathysubramanian, B., Balu, A., Krishnamurthy, A., et al.: Leveraging vision language models for specialized agricultural tasks. *WACV*, 6320–6329 (2025). IEEE
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photo-realistic text-to-image diffusion models with deep language understanding. *NeurIPS* **35**, 36479–36494 (2022)
- Meng, F., Shao, W., Luo, L., Wang, Y., Chen, Y., Lu, Q., Yang, Y., Yang, T., Zhang, K., Qiao, Y., et al.: Phybench: A physical commonsense benchmark for evaluating text-to-image models. *CoRR* (2024) [2406.11802](#)
- Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence* **7**, 1430984 (2024)
- Ye, J., Tang, H.: Multimodal large language models for medicine: A comprehensive survey. *CoRR* (2025) [2504.21051](#)
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: *ICLR* (2020)
- Uzunova, H., Ehrhardt, J., Jacob, F., Frydrychowicz, A., Handels, H.: Multi-scale gans for memory-efficient generation of high resolution medical images. *MICCAI*, 112–120 (2019). Springer
- Pinaya, W.H., Tudosi, P.-D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. *MICCAI Workshop*, 117–126 (2022). Springer
- Polamreddy, L.R., Roy, K., Yueh, S.-H., Mahato, D., Kuppili, S., Li, J., Zhang, Y.: Leapfrog latent consistency model (llcm) for medical images generation. *CoRR* (2024) [2411.15084](#)
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS* **30** (2017)
- Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 7514–7528 (2021)
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. *ECCV*, 740–755 (2014)

- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. *CVPR*, 633–641 (2017)
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* **123**, 32–73 (2016)
- Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. *ICCV*, 8430–8439 (2019)
- Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 6700–6709 (2019)
- Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. *ICCV*, 1965–1973 (2017)
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallochi, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* **128**(7), 1956–1981 (2020)
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CVPR*, 2097–2106 (2017)
- Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ISBI*, 1650–1654 (2021). IEEE
- Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. *CoRR* (2024) [2404.18930](#)
- Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. *ICCV*, 8948–8957 (2019)
- Abaid, A., Farooq, M.A., Hynes, N., Corcoran, P., Ullah, I.: Synthesizing cta image data for type-b aortic dissection using stable diffusion models. *EMBC*, 1–5 (2024). IEEE
- Chakraborty, N., Ornik, M., Driggs-Campbell, K.: Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Comput. Surv.* (2025)
- Yan, B., Zhang, J., Yuan, Z., Shan, S., Chen, X.: Evaluating the quality of hallucination benchmarks for large vision-language models. *CoRR* (2024) [2406.17115](#)
- Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. *NeurIPS* **36**, 25081–25094 (2023)
- Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: *NeurIPS* (2024)
- Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *CVPR*, 1702–1713 (2025)
- Rawte, V., Mishra, A., Sheth, A., Das, A.: Defining and quantifying visual hallucinations in vision-language models. *TrustNLP Workshop*, 501–510 (2025)
- Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., Yao, H.: Calibrated self-rewarding vision language models. *NeurIPS* **37**, 51503–51531 (2024)
- Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? *NeurIPS* **37**, 27056–27087 (2024)
- Ganesh, P., Shokri, R., Farnadi, G.: Rethinking hallucinations: Correctness, consistency, and prompt multiplicity. *ICLR Workshop* (2025)
- Wu, Y., Zhang, L., Yao, H., Du, J., Yan, K., Ding, S., Wu, Y., Li, X.: Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception. *CVPR*, 14646–14656 (2025)
- Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**(5), 206–215 (2019)
- Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *CVPR*, 19187–19197 (2023)
- Ben Melech Stan, G., Aflalo, E., Rohekar, R.Y., Bhawandiwala, A., Tseng, S.-Y., Olson, M.L., Gurwicz, Y., Wu, C., Duan, N., Lal, V.: Lvlm-intrepert: An interpretability tool for large vision-language models. *CVPR*, 8182–8187 (2024)
- Wu, K., Xu, S., Chen, H., Wang, C., Li, Z., Wang, Y., Zhong, F.: Vlm can be a good assistant: Enhancing embodied visual tracking with self-improving visual-language models. *CoRR* (2025) [2505.20718](#)
- Sarch, G., Jang, L., Tarr, M., Cohen, W.W., Marino, K., Fragkiadaki, K.: Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *NeurIPS* **37**, 75942–75985 (2024)