Unified Hallucination Detection for Multimodal Large Language Models

Xiang Chen[♣♡], **Chenxi Wang**[♠♡], **Yida Xue**[♠♡], **Ningyu Zhang**[♠♡*], **Xiaoyan Yang**[♢]
Qiang Li[♢], **Yue Shen**[♢], **Lei Liang**[♢], **Jinjie Gu**[♢], **Huajun Chen**[♣♡*]
[♣]College of Computer Science and Technology, Zhejiang University
[♠]School of Software Technology, Zhejiang University
[♡]Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph [♢]Ant Group
{xiang_chen,zhangningyu}@zju.edu.cn

Abstract

Despite significant strides in multimodal tasks, Multimodal Large Language Models (MLLMs) are plagued by the critical issue of hallucination. The reliable detection of such hallucinations in MLLMs has, therefore, become a vital aspect of model evaluation and the safeguarding of practical application deployment. Prior research in this domain has been constrained by a narrow focus on singular tasks, an inadequate range of hallucination categories addressed, and a lack of detailed granularity. In response to these challenges, our work expands the investigative horizons of hallucination detection. We present a novel meta-evaluation benchmark, **MHaluBench**, meticulously crafted to facilitate the evaluation of advancements in hallucination detection methods. Additionally, we unveil a novel unified multimodal hallucination detection framework, **UNIHD**, which leverages a suite of auxiliary tools to validate the occurrence of hallucinations robustly. We demonstrate the effectiveness of **UNIHD** through meticulous evaluation and comprehensive analysis. We also provide strategic insights on the application of specific tools for addressing various categories of hallucinations[1].

1 Introduction

The recent emergence of MLLMs (Ho et al., 2020; OpenAI, 2023; Durante et al., 2024) that more closely mirror human cognition and learning has unleashed unprecedented possibilities for the future of artificial general intelligence (AGI). Despite MLLMs' impressive abilities, they are susceptible to generating seemingly credible content that contradicts input data or established world knowledge, a phenomenon termed "hallucination"(Liu et al., 2024; Wang et al., 2023a; Huang et al., 2023c; Tonmoy et al., 2024; Zhang et al., 2023a). These hallucinations hinder the practical deployment of MLLMs and contribute to the dissemination of misinformation. Consequently, detectors that could detect multimodal hallucinations (Yang et al., 2023) within responses from MLLMs are urgently needed to alert users to potential risks and drive the development of more reliable MLLMs.

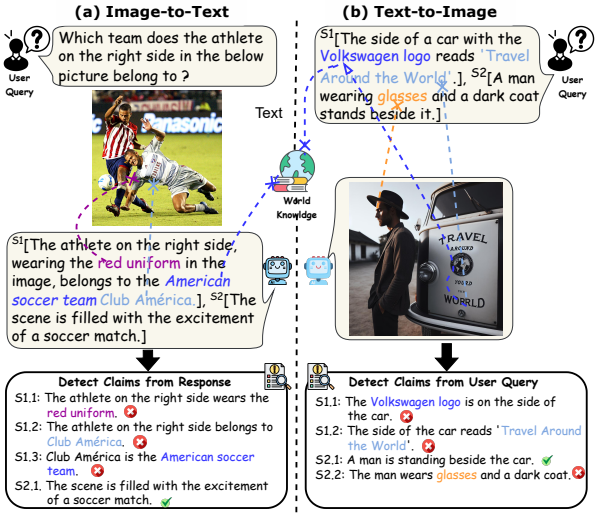

Figure 1: Unified multimodal hallucination detection aims to identify and detect modality-conflicting hallucinations at various levels such as object, attribute, and scene-text, as well as fact-conflicting hallucinations in both image-to-text and text-to-image generation. Our benchmark emphasizes fine-grained detection, with "S1" representing the segment and "S1.1" and "S1.2" denoting its corresponding claims.

Although several works have been conducted to detect hallucinations from MLLMs(Zhou et al., 2023; Zhai et al., 2023; Li et al., 2023b; Wang et al., 2023c) or alleviate hallucinations(Xing et al., 2024; Wu et al., 2024), these efforts operate in isolation and have certain limitations when compared with the aspects illustrated in Figure 1: *(1) Task Singularity:* Current research has primarily concentrated on specific tasks, such as image captioning while neglecting that text-to-image generation, an important component of AGI, also suffers from hal-

*Corresponding author.

[1]The code can be accessed via https://github.com/zjunlp/EasyDetect, and the demonstration is available at http://easydetect.openkg.cn.

lucinations induced by MLLMs. *(2) Limited Hallucination Categories:* Prior studies have focused on identifying hallucinations at the object level, yet they fail to consider the prevalence of scene-text or factual inconsistencies that also frequently occur in MLLMs. *(3) Incomplete Granularity:* It would be more valuable to assess hallucinations at a fine-grained level, examining individual claims within a response, rather than evaluating the entire response holistically. Considering these constraints hinder rapid progress in practical hallucination detection, it raises the question: *Can we develop a unified perspective for detecting hallucinations from MLLMs?*

To further investigate this problem, we have broadened the concept of multimodal hallucination within MLLMs to a holistic framework, integrating both image-to-text generation such as Image Captioning (**IC**) and Visual Question Answering (**VQA**), as well as text-to-image-synthesis (**T2I**) – to align with MLLMs' capabilities of performing varied multimodal tasks. We are committed to exploring a broad spectrum of hallucinatory categories and the intricate nuances of claim-level hallucination through a lens that integrates both modality-conflicting and fact-conflicting hallucinations. Based on the outlined perspectives, We have developed the **M**ulti**M**odal **Hallu**cination Detection **Bench**mark (**MHaluBench**) to assess the progress of unified multimodal hallucination detectors for MLLMs and embodied the data framework depicted in Figure 1.

At its core, leveraging MLLMs' inherent self-detection mechanisms to pinpoint diverse hallucinations encounters significant hurdles. We further develop a tool-augmented framework for unified hallucination detection, named **UNIHD**, which integrates evidence from multiple auxiliary tools through the following procedure: (1) *Essential Claim Extraction* involves extracting the core claims within the generated response for image-to-text generation or user queries in text-to-image generation; (2) *Autonomous Tool Selection via Query Formulation* prompts MLLMs (GPT-4/Gemini) to autonomously generate pertinent questions for each claim. These questions are crafted to determine the specific type of tool required for each claim and to establish the input for the tool's operation; (3) *Parallel Tool Execution* deploys a suite of specialized tools to operate concurrently, providing evidence from their outputs to reliably validate potential hallucinations; (4) *Hallucination Verification with Rationales* aggregates the collected evidence to instruct the underlying MLLM to judge whether the claim hallucinatory with rationals for explanation.

We have conducted a thorough evaluation of the **UNIHD** framework, utilizing the underlying MLLM against the MHaluBench benchmark. Our findings underscore the effectiveness of our approach and confirm that multimodal hallucination detection remains a formidable challenge. In a nutshell, We conclude our contributions as:

- We propose a more unified problem setting for hallucination detection in MLLMs, encompassing a broad spectrum of multimodal tasks and hallucination categories, thus enriching the unified understanding of hallucination in MLLMs.
- We unveil **MHaluBench**, a meta-evaluation benchmark that encompasses various hallucination categories and multimodal tasks. This benchmark is equipped with fine-grained analytical features, gauging the progress of hallucination detectors.
- We introduce **UNIHD**, a task-agnostic, tool-enhanced framework for the detection of hallucinations in content produced by MLLMs. Our extensive experiments demonstrate the efficacy of this method, underscoring that MHaluBench continues to be a challenging yet vital task.

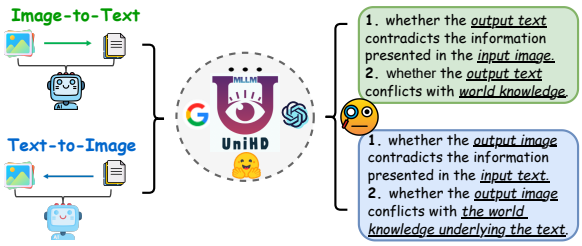

Figure 2: Unified multimodal hallucination detection.

2 Preliminaries

We explore a unified perspective on hallucination in MLLMs (illustrated in Figure 2) with the aspiration of developing a unified detection framework.

Unified View of Multimodal Hallucination Taxonomy. A prerequisite for unified detection is the coherent categorization of the principal categories of hallucinations within MLLMs. Our paper superficially examines the following Hallucination Taxonomy from a unified perspective:

- **Modality-Conflicting Hallucination.** MLLMs sometimes generate outputs that conflict with inputs from other modalities, leading to issues

Datasets	Response Generated by	Purpose	Granularity	Hallucination Types				Modality	Scenario Task
				Object	Attribute	Scene Text	Fact		
FactCC (Kryscinski et al., 2020)	Synthetic	Check.	Sentence				✔	Text	Text2Text
QAGS (Wang et al., 2020)	Model	Check.	Summary				✔	Text	Text2Text
HaluEval (Li et al., 2023a)	ChatGPT	Det.	Response				✔	Text	Text2Text
POPE (Li et al., 2023b)	-	Eval.	Response	✔				Multi.	Image2Text
HaELM (Wang et al., 2023c)	-	Det.	Response					Multi.	Image2Text
AMBER (Wang et al., 2023b)	-	Eval.	Response	✔	✔			Multi.	Image2Text
MHaluBench (Ours)	MMLMs	Det.	Res.,Seg.,Claim	✔	✔	✔	✔	Multi.	Image2Text/Text2Image

Table 1: A comparison of benchmarks w.r.t existing fact-checking or hallucination evaluation. "Check." indicates verifying factual consistency, "Eval." denotes evaluating hallucinations generated by different LLMs, and its response is based on different LLMs under test, while "Det." embodies the evaluation of a detector's capability in identifying hallucinations.

such as incorrect objects, attributes, or scene text. An example in Figure 1 (a) includes an MLLM inaccurately describing an athlete's uniform color, showcasing an attribute-level conflict due to MLLMs' limited ability to achieve fine-grained text-image alignment.

- **Fact-Conflicting Hallucination.** Outputs from MLLMs may contradict established factual knowledge. Image-to-text models can generate narratives that stray from the actual content by incorporating irrelevant facts, while text-to-image models may produce visuals that fail to reflect the factual knowledge contained in text prompts. These discrepancies underline the struggle of MLLMs to maintain factual consistency, representing a significant challenge in the domain.

Unified Detection Problem Formulation. Unified detection of multimodal hallucination necessitates the check of each image-text pair $a = \{v, x\}$, wherein v denotes either the visual input provided to an MLLM, or the visual output synthetic by it. Correspondingly, x signifies the MLLM's generated textual response based on the v or the textual user query for synthesizing v. Within this task, each x may contain multiple claims, denoted as $\{c_i\}_{i=1\cdots n}$. The objective for hallucination detectors is to assess each claim from a to determine whether it is "hallucinatory" or "non-hallucinatory", providing a rationale for their judgments based on the provided definition of hallucination. Text hallucination detection from LLMs denotes a sub-case in this setting, where v is null.

3 Construction of MHaluBench

To facilitate research in this area, we introduce the meta-evaluation benchmark MHaluBench, which encompasses the content from image-to-text and text-to-image generation, aiming to rigorously assess the advancements in multimodal hallucination detectors. Our benchmark has been meticulously curated to include a balanced distribution of instances across three pivotal tasks, which encompasses 200 exemplars for the task of IC 200 for VQA, and an additional 220 dedicated to Text-to-Image Generation. The comparison of MHaluBench with other benchmarks is detailed in Table 1 and the statistical details are provided in Figure 3 and Figure 4.

3.1 Hallucinatory Example Collection

Image-to-Text Generation. We focus on IC and VQA tasks, drawing samples from the MS-COCO 2014 validation set (Lin et al., 2014) and the TextVQA test set (Singh et al., 2019). We compile generative outputs from mPLUG (Ye et al., 2023), LLaVA (Liu et al., 2023c), and MiniGPT-4 (Zhu et al., 2023) to form the core dataset for MHaluBench. These models are representative of current leading MLLMs, characterized by their diverse content generation capabilities and a notable presence of hallucinations, as depicted in Figure 8.

Text-to-Image Generation. We source initial captions from DrawBench (Saharia et al., 2022) and T2I-CompBench (Huang et al., 2023a). These captions are augmented through ChatGPT to include more specific information such as objects, attributes, and factual details, among others. The refined caption guides the DALL-E 2 (Ramesh et al., 2022) and DALL-E 3 model (Betker et al., 2023) in producing visually detailed images.

3.2 Segment and Claim Extraction

Beyond evaluating overall responses, we introduce segmentation at both the segment and claim levels for a multi-granular assessment of hallucinations, enabling more precise feedback to improve model performance (Lightman et al., 2023). We leverage ChatGPT's advanced instruction-following ability to extract detailed segments and related claims. For image-to-text tasks, we split and extract the model's textual output into segments and claims; for text-to-image cases, we break down user queries

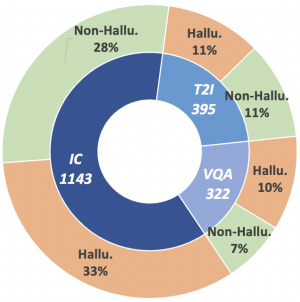

Figure 3: Claim-Level data statistics of MHaluBench. The claims are fine-grained atoms extracted from the complete "Query-Response" pairs.

into fundamental intent concepts, which are subsequently regarded as claims.

3.3 Human Annotation and Agreement.

Our annotation criteria evaluate whether image-to-text output conflicts with the input image or world knowledge and whether text-to-image visuals conflict with claims or world knowledge. Extracted claims are labeled as hallucinatory or non-hallucinatory, with a segment deemed hallucinatory if it contains any such claim; otherwise, it is labeled non-hallucinatory. An entire response is labeled hallucinatory if it includes even one hallucinatory segment. We allocate the dataset uniformly across three annotators with graduate-level qualifications for independent categorization. Decisions in uncertain cases were initially held by individual annotators and later resolved by majority rule. Inter-annotator reliability, measured by Fleiss's Kappa (κ), shows significant agreement ($\kappa = 0.822$) over the full annotated dataset, indicating a high level of concordance within the range $0.80 \leq \kappa \leq 1.00$.

4 UniHD: Unified Hallucination Detection Framework for MLLMs

We present **UniHD** in Figure 5 and follow. The specific prompts are listed in Appendix A

4.1 Essential Claim Extraction

To identify fine-grained hallucinations within the response, claim extraction is a prerequisite. Following the procedure in §3.2, we employ the advanced instruction-following abilities of MLLMs for efficient claim extraction. Specifically, GPT-4V/Gemini is adopted as the base LLM to efficiently derive verifiable claims from the outputs of image-to-text models (extracting each response into individual claims) and text-to-image models

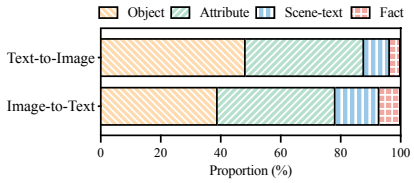

Figure 4: Distribution of hallucination categories within hallucination-labeled claims of MHaluBench.

(deconstructing user queries into distinct claims) [2].

4.2 Autonomous Tool Selection Via Query Formulation

After extracting essential claims from the input image-text pair $a = \{v, x\}$, the challenge of hallucination detection is to aptly match each claim with appropriate aspect-oriented tools. We approach this issue by assessing whether the underlying MLLMs can generate pertinent queries for a given set of claims $\{c_i\}_{i=1\cdots n}$ to provide relevant input to the specific aspect-oriented tool. To facilitate this, we prompt underlying MLLMs like GPT-4V/Gemini to autonomously formulate meaningful queries. Demonstrated in Figure 5, this module yields custom queries for each claim, or "`none`" when a tool is unnecessary. For example, the framework determines that claim1 calls for the attribute-oriented question "`What color is the uniform of the athlete on the right side?`" and the object-oriented inquiry "['`athlete`', '`uniform`']", bypassing the need for scene-text and fact-oriented tools.

4.3 Parallel Tool Execution

Leveraging queries autonomously generated from various perspectives, we simultaneously deploy these tools in response to the queries, gathering a comprehensive array of insights to underpin the verification of hallucinations. The specific tools employed in our framework are detailed below, selected for their ability to effectively address a wide range of multimodal hallucination scenarios:

- *Object-oriented tool:* We employ the open-set object detection model Grounding DINO (Liu et al., 2023d) for capturing visual object information, crucial for detecting object-level hallucinations. For instance, inputting "['`athlete`', '`uniform`']" prompts the model to return two

[2] In subsequent experiments, our framework builds upon the pre-annotated claims available in MHaluBench, and the claim extraction is only necessary in the open-domain setting.

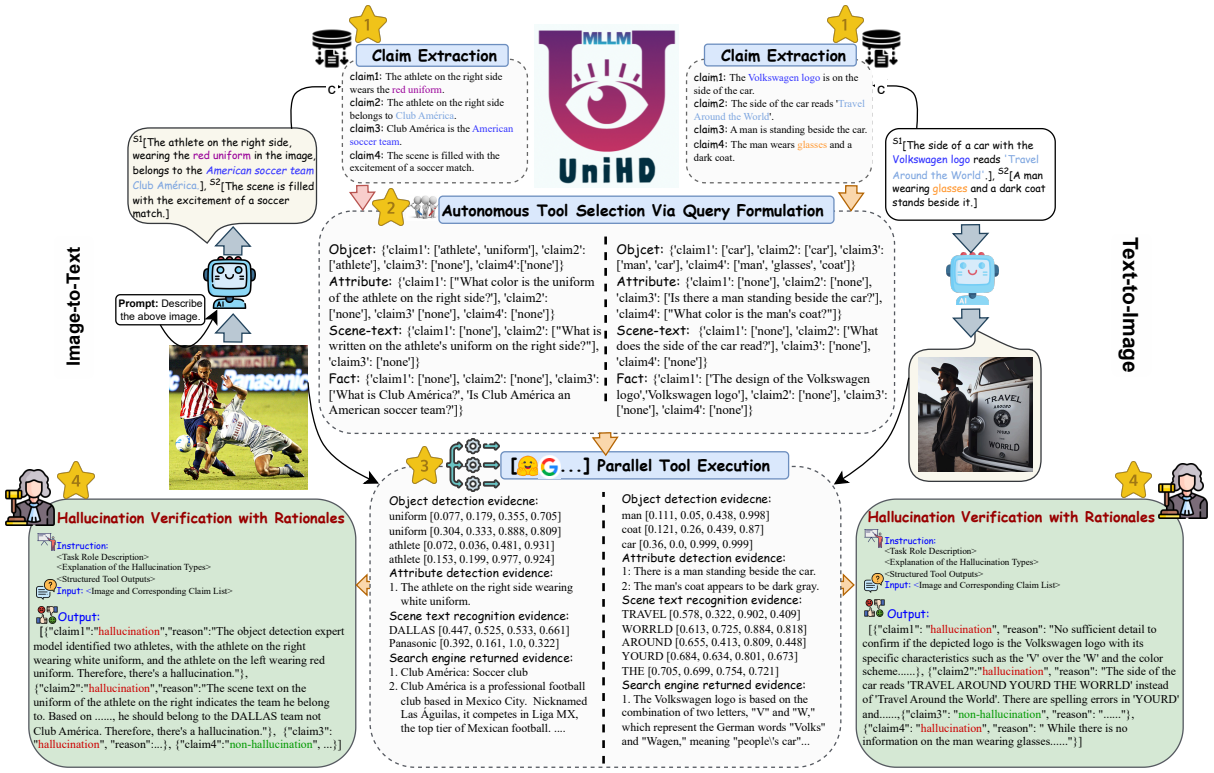

Figure 5: The specific illustration of UNIHD for unified multimodal hallucination detection.

uniform objects and two athlete objects, along with their normalized location coordinates.

- *Attribute-Oriented Tool:* Dealing with attributes such as positions, colors, and actions, we harness underlying MLLMs (such as GPT-4V and Gemini) to answer the specific attribute-level questions. These responses are leveraged for hallucination verification within the same MLLMs, mirroring a self-reflect akin to (Shinn et al., 2023).
- *Scene-Text-Oriented Tool:* Should the generated questions for scene text not be exclusively "none", we then invoke MAERec (Jiang et al., 2023) as our scene-text detection tool, which is capable of identifying scene text within images along with their corresponding normalized four-dimensional coordinates.
- *Fact-Oriented Tool:* To validate conflicting factual hallucinations, we harness the Serper Google Search API to perform web searches using specific fact-based questions. By extracting and scrutinizing the top results, we obtain a range of snippets from the API's responses for analysis.

Moreover, UNIHD is tool-agnostic, facilitating the seamless integration of emerging tools and detection strategies to amass tool knowledge, thereby bolstering the process of hallucination verification.

4.4 Hallucination Verification with Rationales

In the concluding phase of our process, we subject each claim, denoted as c_i, to a binary prediction to ascertain its hallucinatory status. Claims are categorized as either HALLUCINATORY or NON-HALLUCINATORY based on the level of evidence support. To accomplish this, we aggregate the collected evidence from tools with the original image and its corresponding claim list[3] into a comprehensive prompt. Subsequently, we instruct our chosen MLLM (GPT-4V or Gemini) to assess each claim's hallucinatory potential. In doing so, the MLLM also generates insightful explanations to elucidate the rationale behind its judgment.

5 Experiment

5.1 Experimental Settings

Baselines. We compare UNIHD on MHaluBench[4] with two baselines, Self-Check (2-shot)[5] and Self-Check (0-shot) based on

[3]Note that the set $a = \{v, x\}$, corresponding to the list of claims, is input into the detectors in a single batch. This operation allows the detectors to capture contextual information while also enhancing efficiency.

[4]In this paper, we conducted experiments using the evaluation benchmark from our published V0.1 version.

[5]Self-Check (2-shot) utilize two complete demonstrations based on $a = \{v, x\}$ rather than only two claims.

Tasks	LLMs	Methods	Levels	Hallucinatory			Non-Hallucinatory			Average			
				P	R	F1	P	R	F1	Acc.	P	R	Mac.F1
Image-to-Text	Gemini	Self-Check (0-shot)	Claim	83.17	42.15	55.95	55.64	89.48	68.61	63.34	69.41	65.82	62.28
			Segment	89.30	47.71	62.19	43.76	87.68	58.38	60.38	66.53	67.69	60.29
		Self-Check (2-shot)	Claim	84.24	66.75	74.48	67.35	84.60	75.00	74.74	75.80	75.68	74.74
			Segment	90.44	71.08	79.60	57.35	83.80	68.10	75.11	73.89	77.44	73.85
		UNIHD	Claim	84.44	72.44	77.98	71.08	83.54	76.80	77.41	77.76	77.99	77.39
			Segment	88.77	78.76	83.46	63.17	78.52	70.02	78.68	75.97	78.64	76.74
	GPT-4v	Self-Check (0-shot)	Claim	79.37	74.17	76.68	70.52	76.22	73.26	75.09	74.94	75.19	74.97
			Segment	84.78	80.07	82.35	61.64	69.01	65.12	76.56	73.21	74.54	73.73
		Self-Check (2-shot)	Claim	82.00	79.98	80.98	76.04	78.35	77.18	79.25	79.02	79.16	79.08
			Segment	86.54	85.13	85.83	69.05	71.48	70.24	80.80	77.80	78.30	78.04
		UNIHD	Claim	82.54	85.29	83.89	81.08	77.74	79.38	81.91	81.81	81.52	81.63
			Segment	87.03	91.01	88.98	78.52	70.77	74.44	84.60	82.77	80.89	81.71
Text-to-Image	Gemini	Self-Check (0-shot)	Claim	73.85	24.62	36.92	55.45	91.50	69.06	58.48	64.65	58.06	52.99
			Segment	87.27	30.00	44.65	32.53	88.52	47.58	46.15	59.90	59.26	46.11
		Self-Check (2-shot)	Claim	85.37	53.85	66.04	66.91	91.00	77.12	72.66	76.14	72.42	71.58
			Segment	91.67	61.88	73.88	46.02	85.25	59.77	68.33	68.84	73.56	66.83
		UNIHD	Claim	85.71	61.54	71.64	70.59	90.00	79.12	75.95	78.15	75.77	75.38
			Segment	93.28	69.37	79.57	51.96	86.89	65.03	74.21	72.62	78.13	72.30
	GPT-4v	Self-Check (0-shot)	Claim	88.55	59.49	71.17	70.08	92.50	79.74	76.20	79.31	75.99	75.45
			Segment	93.69	65.00	76.75	49.09	88.52	63.16	71.49	71.39	76.76	69.96
		Self-Check (2-shot)	Claim	84.39	74.87	79.35	77.93	86.50	81.99	80.76	81.16	80.69	80.67
			Segment	89.63	75.62	82.03	54.65	77.05	63.95	76.02	72.14	76.34	72.99
		UNIHD	Claim	84.92	86.67	85.79	86.73	85.00	85.86	85.82	85.83	85.83	85.82
			Segment	91.25	91.25	91.25	77.05	77.05	77.05	87.33	84.15	84.15	84.15

Table 2: Experimental results of UNIHD powered by Gemini and GPT-4V on Image-to-Text and Text-to-Image Generation. The default F1 score is Micro-F1, whereas Mac.F1 represents the Macro-F1 score.

CoT (Wei et al., 2022), which assess the capability of the underlying MLLM to identify hallucinations without external knowledge and have shown effectiveness across other various tasks (Chern et al., 2023; Xie et al., 2023). We prompt GPT-4V (`gpt-4-vision-preview`) and Gemini (Pro Vision) to recognize fine-grained hallucinations and explain the reasoning behind this determination.

Evaluation Perspective. We compute the recall, precision, and Micro-F1 metrics individually for both hallucinatory and non-hallucinatory categories. Additionally, we assess the overall performance by measuring the average Macro-F1 scores at the claim and segment levels. We categorize a segment as non-hallucinatory only if all associated claims are classified as non-hallucinatory; it is deemed hallucinatory if any associated claims do not meet this criterion.

5.2 Evaluation Results

MHaluBench poses a challenging benchmark for multimodal hallucination detection. The segment-level and response-level outcomes are presented in Table 2. Even though all hallucinatory instances in MHaluBench are obtained from open-source MLLMs' outputs rather than being generated by GPT-4V/Gemini itself, it is noteworthy that the majority of detectors achieve an overall Macro-F1 score ranging between 70%-80%, exhibiting subpar performance on MHaluBench.

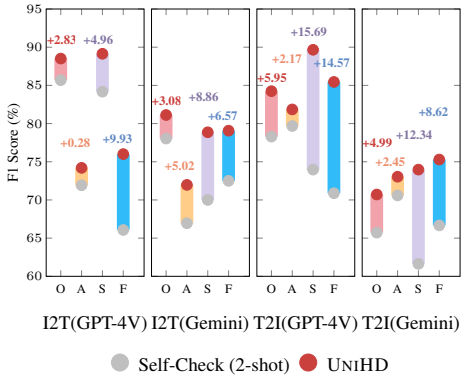

Figure 6: Comparative analysis to examine the augmentation in detection capabilities provided by specific tools on samples labeled with hallucinations. The x-axis labels "O", "A", "S" and "F" refer to object, attribute, scene-text, and fact, respectively.

GPT-4V surpasses Gemini as the detector base. GPT-4V-powered detectors consistently outperform Gemini counterparts, achieving higher Macro-F1 scores, especially in the text-to-image generation. For instance, Self-Check (0-shot) using GPT-4V achieves a claim-level Macro-F1 of 72.82, significantly surpassing Gemini's Macro-F1 score of 52.98. However, Gemini-powered detectors exhibit better performance in non-hallucinatory categories for image-to-text tasks, indicating a potential bias towards reduced sensitivity to hallucinations.

UNIHD Empowered by GPT-4V: Superior Detection Across the Board. Table 2 demonstrates that UNIHD, leveraging GPT-4V, consistently out-

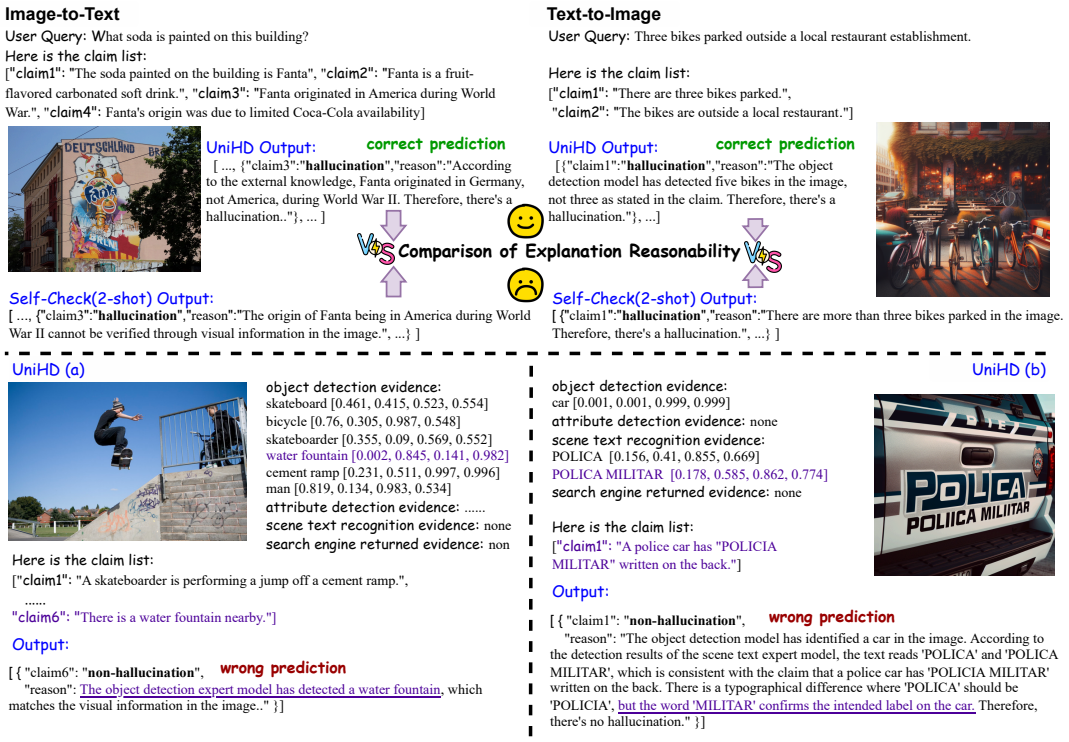

Figure 7: **Case Study.** The upper section depicts two exemplary cases where both UNIHD and Self-Check (2-shot) arrive at correct judgments, with a comparative demonstration of UNIHD providing explanations of superior reasonability. UNIHD (a) reveals a failure case where the tool presents erroneous evidence, leading to an incorrect verification outcome. Conversely, UNIHD (b) highlights a scenario where, despite the tool offering valid and correct evidence, GPT-4V persists in its original stance, resulting in a flawed verification.

performs other baseline detectors in image-to-text and text-to-image tasks. Despite the Self-Check (2-shot) showcasing GPT-4V and Gemini's robust in-context learning, UNIHD markedly exceeds its performance, emphasizing the benefits of integrating external tools for more robust evidence verification and reliable hallucination detection.

5.3 Analysis

Which Type of Hallucination Can Benefit the Most from Tool Enhancement? Figure 6 shows that UNIHD enhances the detection of **scene text** and **factual** hallucinations over Self-Check (2-shot), suggesting that GPT-4V or Gemini's inherent limitations make the evidence provided by the tool especially valuable. However, UNIHD exhibits minimal improvement in identifying **attribute-level** hallucinations, potentially attributed to a lack of specialized tools for direct attribute detection, with self-reflection methods based on GPT-4V/Gemini proving to be relatively weak.

Explanation Reasonability of UNIHD. As shown in the upper portion of Figure 7, both the fact-level hallucination "Fanta originated in America during World War." and the object-level hallucination "There are three bikes parked." are accurately identified by Self-Check (2-shot) and UNIHD. Comparative analysis reveals that UNIHD excels in synthesizing evidence to provide a more credible and compelling rationale.

Failure Analysis of UNIHD. As shown in the lower part of Figure 7, we present two instances where UNIHD exhibits limitations. The left case demonstrates situations where the tool either generates incorrect evidence or fails to provide useful information, leading to erroneous judgments by the MLLM. On the right, we observe cases where the MLLM maintains its initial bias despite receiving accurate evidence, resulting in incorrect decisions. These scenarios highlight areas for further research to enhance tool accuracy and to develop MLLMs dedicated to better hallucination detection.

Text-to-Image Hallucination vs. Image-to-Text Hallucination: Which is Easier to Detect? Both baselines and the GPT-4V-enhanced UNIHD show significantly improved performance in identifying hallucinations in text-to-image content over

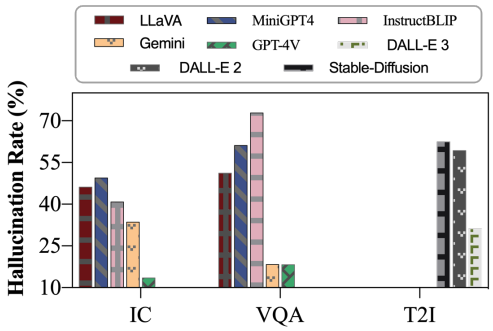

Figure 8: Comparison of claim-level hallucination ratios across MLLMs. We randomly select a set of 20 prompts from MHaluBench for each of the IC, VQA, and T2I. Responses for these prompts are generated by each of the evaluated MLLMs.

image-to-text content. This can be traced back to the structured nature of manually written user queries for text-to-image tasks, which yield more uniform images. while image-to-text confronts the complexity of natural images with background noise and content generated by MLLMs, characterized by greater diversity and fewer constraints. Consequently, it is intuitively easier to detect discrepancies between text and corresponding images in text-to-image tasks.

Explore UNIHD to Evaluate Hallucination of Modern MLLMs. We designate UNIHD powered by GPT-4V as the golden detector to assess the frequency of hallucinations in MLLMs, including GPT-4V, and Gemini, among others. The findings illustrated in Figure 8 indicate that (1) GPT-4V exhibits the lowest claim-level hallucination ratio across most tested conditions, and (2) the hallucination-based ranking of these MLLMs is generally in agreement with established leaderboards and human evaluation, demonstrating the potential of UNIHD for evaluating hallucinations.

6 Related Work

6.1 Hallucinations in MLLM

The advent of MLLMs (OpenAI, 2023; Liu et al., 2023c; Ye et al., 2023; Zhu et al., 2023) has highlighted the issue of hallucination (Hu et al., 2024; Zhang et al., 2023b; Huang et al., 2023b; Rawte et al., 2023; Ji et al., 2023), a crucial concern impacting their dependability. Previous research has primarily focused on three areas: evaluating (Li et al., 2023b; Liu et al., 2023a; Jing et al., 2023), detecting (Wang et al., 2023c; Yang et al., 2023; Yin et al., 2023), and mitigating hallucinations (Wan et al., 2024; Liu et al., 2023b; Huang et al., 2023c; Semnani et al., 2023; Zhao et al., 2024; Leng et al., 2023; Wang et al., 2024; Deng et al., 2024). In a complementary effort, HaELM (Wang et al., 2023c) scrutinizes the challenges associated with POPE (Li et al., 2023b) and suggests training a model based on simulated hallucination samples for detecting multimodal hallucinations. Diverging from prior efforts, this paper addresses a broader problem scope for hallucination detection, introducing a unified multimodal hallucination detection framework, UNIHD, along with meta-evaluation benchmarks, MHaluBench.

6.2 Harnessing Tool Resources for LLMs

Addressing the limitations of LLMs (Chen, 2023; Kang et al., 2024) due to their pre-training confinement, researchers have explored augmenting them with resources like knowledge bases, search engines, and external models, to expand their functionality. Notably, Schick et al. (2023); Hao et al. (2023); Qiao et al. (2023) have developed models that leverage external tools to improve performance in downstream tasks. More recently, Shen et al. (2023); Liang et al. (2023) has unveiled frameworks integrating LLMs with diverse AI models to tackle complex challenges. Building on this, researchers (Peng et al., 2023; Chen et al., 2023) have examined the utilization of external knowledge to mitigate or evaluate hallucinations in LLMs. Adapting these enhancements for MLLMs introduces unique challenges, necessitating the selection of appropriate tools for effective oversight. Our research focuses on automating the selection of functionally diverse tools to enhance multimodal hallucination detection.

7 Conclusion

We introduce a unified problem formulation for multimodal hallucination detection that encompasses a diverse range of multimodal tasks and hallucination types. A fine-grained benchmark dataset, MHaluBench, is also proposed to promote this challenging direction. Alongside this, we present the unified hallucination detection framework, UNIHD, capable of autonomously selecting external tools with capturing pertinent knowledge to support hallucination verification with rationales. Our experimental results indicate that UNIHD achieves better performance across both image-to-text and text-to-image generation tasks, confirming its universality and efficacy.

Limitations

This paper focuses on constructing a unified hallucination detection framework for MLLMs, dubbed UNIHD. Despite the best efforts, our paper still have some limitations.

The Scope of Multimodal Tasks. This paper primarily addresses the detection of multimodal hallucinations from a unified perspective, with a focus on image-to-text tasks (such as Image Captioning and VQA) and text-to-image generation tasks. Nonetheless, it is important to recognize that our framework does not yet encompass other multimodal tasks, such as video captioning, which are also susceptible to hallucinations. Moving forward, we aim to explore the possibilities of incorporating these additional domains into our UNIHD.

Limitations of Closed-Source MLLM Pricing and Inference Speed. Our UNIHD is primarily built upon powerful closed-source models as the foundation. However, closed-source models (Liu et al., 2023c; Zhu et al., 2023; Ye et al., 2023; Bai et al., 2023) often come with a cost, which introduces operational expenses. Additionally, our UNIHD relies on several external tools to provide evidence for enhanced illusion verification, resulting in additional inference time. In the future, we will further explore training open-source dedicated illusion detection models with the tool to further improve effectiveness and reduce costs.

The Scope of Hallucination Categories. In our commitment to developing a comprehensive hallucination detection framework, referred to as UNIHD, for MLLMs, we have made efforts to incorporate various prevalent hallucination categories within MHaluBench and UNIHD, including object, attribute, scene-text, and factual aspects, among others. However, it is important to acknowledge that there are additional categories of hallucinations that have not been covered in our framework, as discussed in the existing literature (Zhang et al., 2023b; Wang et al., 2023a; Mishra et al., 2024; Huang et al., 2023b; Rawte et al., 2023) . Moving forward, our research will expand its scope to adopt a unified approach towards a wider range of hallucination categories, to strengthen the robustness of our detection mechanisms.

Preliminary Attempts at Tool Utilization. In our early endeavors, we have configured a dedicated tool for detecting a specific type of hallucination, exemplified by the assignment of the Grounded DINO model as the object detection tool of choice. However, it should be acknowledged that the current selection of tools may not represent the optimum choice. It remains imperative to rigorously explore which SOTA object detection models are best suited for the task of multimodal hallucination detection. This necessitates an extensive evaluation of available models to pinpoint the most effective tool that aligns with the nuances and complexities of our multimodal detection objectives.

Acknowledgement

We are grateful for the API services provided by OpenAI and Google, which enabled us to process data and conduct some of our experiments. Part implementation of this work are assisted and inspired by the related hallucination toolkits including FactTool (Chern et al., 2023), Woodpecker (Yin et al., 2023), and others. We follow the same license for open-sourcing and thank them for their contributions to the community. This work also benefits from the public project of mPLUG-Owl[6], MiniGPT-4[7], LLaVA[8], GroundingDINO[9], and MAERec[10]. This work was supported by the National Natural Science Foundation of China (No. 62206246), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), CCF-Tencent Rhino-Bird Open Research Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work is supported by Ant Group.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

[6] https://github.com/X-PLUG/mPLUG-Owl

[7] https://github.com/Vision-CAIR/MiniGPT-4

[8] https://github.com/haotian-liu/LLaVA

[9] https://github.com/IDEA-Research/GroundingDINO

[10] https://github.com/Mountchicken/Union14M

James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Huajun Chen. 2023. Large knowledge model: Perspectives and challenges. *CoRR*, abs/2312.02706.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023. Factchd: Benchmarking fact-conflicting hallucination detection. *CoRR*, abs/2310.12086.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *CoRR*, abs/2402.15300.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent ai: Surveying the horizons of multimodal interaction.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *NeurIPS 2023*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Do large language models know about facts? *ICLR 2024*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023a. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *CoRR*, abs/2307.06350.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023c. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. 2023. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. FAITHSCORE: evaluating hallucinations in large vision-language models. *CoRR*, abs/2311.01477.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. 2024. C-RAG: certified generation risks for retrieval-augmented language models. *CoRR*, abs/2402.03181.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *CoRR*, abs/2311.16922.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

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

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *CoRR*, abs/2303.16434.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John

Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. *CoRR*, abs/2310.14566.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *CoRR*, abs/2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023d. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

OpenAI. 2023. Gpt-4 technical report. *OpenAI*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Shuofei Qiao, Honghao Gui, Huajun Chen, and Ningyu Zhang. 2023. Making language models better tool learners with execution feedback. *CoRR*, abs/2305.13068.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Vipula Rawte, Amit P. Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *CoRR*, abs/2309.05922.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *NeurIPS 2023*.

Sina J. Semnani, Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *NeurIPS 2023*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Mitigating hallucinations of large language models via knowledge consistent alignment. *CoRR*, abs/2401.10768.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023c. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *CoRR*, abs/2403.18715.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *CoRR*, abs/2402.11622.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023. Ask again, then fail: Large language models' vacillations in judgement. *CoRR*, abs/2310.02174.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. EFUF: efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *CoRR*, abs/2402.09801.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda R. Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of llms-generated content. *CoRR*, abs/2310.15654.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

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

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

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023a. Alleviating hallucinations of large language models through induced hallucinations. *CoRR*, abs/2312.15710.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *CoRR*, abs/2402.08680.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Prompt Templates

Within this section, we outline the prompt templates designed to guide the foundational MLLM for the autonomous query formulation (illustrated in Table 3-6) and verification of any hallucinated content (shown in Table 7-8).

SYSTEM:
You are a brilliant object extractor.

USER:
Given a list of claim, extract the objects from each claim for me.
Extract the common objects and summarize them as general categories without repetition, merge essentially similar objects.
Avoid extracting hypernyms, keep hyponyms!
Avoid extracting abstract or non-specific objects.
Extract object in the singular form.
Output all the extracted types of items separate each object type with a period.
If there is nothing to output, then output a single "none".
YOU MUST TO DISREGARD OBJECT WORDS THAT ARE NOT NATURAL OBJECTS, SUCH AS SCENES, AREA, SKY, GROUND, WORDS, ATMOSPHERES, COUNTRIES, NAMES, AND PLACES.IF THERE ARE NO NATURAL objects IN THE SENTENCE, RETURN 'none'.
YOU MUST RETURN THE RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS.
You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE.
response format: {{"claim1":"object1.object2.object3","claim2":"none","claim3":"object1.object2", ...}}

Here are three examples:
claim list:
claim1: The image depicts a man laying on the ground.
claim2: The man is next to a motorcycle.
claim3: The sun is shining upon the ground.
claim4: The light is very bright.
output:
{{"claim1":"man","claim2":"man.motorcycle","claim3":"none", "claim4":"none"}}

claim list:
claim1: The image shows a device.
claim2: The device has the words S̈amsung.̈
claim3: Samsung is a Korean company.
output:
{{"claim1":"device","claim2":"device", "claim3":"none"}}

claim list:
claim1: A man wears a green shirt.
claim2: The man's face is beaming with a smile.
claim3: The image shows the man in high spirits.
output:
{{"claim1":"man.shirt","claim2":"man","claim3":"man"}}

Now complete your output with following the above rules.
claim list:
{claims}
output:

Table 3: Prompt template of query formulation (object-level) for image-to-text generation.

SYSTEM:
You are a brilliant question generator.

USER:
Given a list of claim and some objects(each object is connected by a period), you're required to generate questions about attributes of the given objects.
The generated questions may involve basic attributes such as colors, actions and position mentioned in the claim.
Do not ask questions involving object counts or the existence of object. Do not ask questions involving scene text.
When asking questions about attributes, try to ask simple questions that only involve one object.
Ask questions that can be easily decided visually. Do not ask questions that require complex reasoning.
Do not ask semantically similar questions. Do not ask questions only about scenes or places.
Do not ask questions about uncertain or conjecture parts of the claim, for example, the parts described with "maybe" or "likely", etc.
It is no need to cover all the specified objects. If there is no question to ask, simply output 'none'.
YOU MUST RETURN THE RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS.
You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE.

response format: {{"claim1":["question1", "question2"],"claim2":["none"],"claim3":["question1", "question2"], ...}}

Here are three examples:
objects:
dog.cat
claim list:
claim1: There is one black dog on the left in the image.
claim2: There are two white cats on the right in the image.
output:
{{"claim1":["What color is the dog?", "Is there a dog on the left in the image?"],"claim2":["What color are the cat?", "Are there two cats on the right in the image?"]}}

objects:
man.baseball cap.wall
claim list:
claim1: The man is wearing a baseball cap.
claim2: The man appears to be smoking.
claim3: 'hello world' is written on the white wall.
output:
{{"claim1":["What is the man wearing?"], "claim2":["Does the man appear to be smoking?"], "claim3":[What color is the wall?]}}

objects:
kitchen.man.apron
claim list:
claim1: The image depicts a kitchen.
claim2: There is a man in a white apron.
claim3: The man is standing in the middle of the kitchen.
claim4: The overall atmosphere is very pleasant.
output:
"claim1":["none"], "claim2":["What does the man wear?", "What color is the apron?"], "claim3":["Is the man standing in the middle of the kitchen?"], "claim4": ["none"]

Now complete the following with following the above rules. DO NOT RESPOND WITH ANYTHING ELSE.
objects:
{objects}
claim list:
{claims}
output:

Table 4: Prompt template of query formulation (attribute-level) for image-to-text generation.

SYSTEM:
You are a brilliant question generator.

USER:
Given a list of claim, you're required to generate questions about scene text to assist users in verifying the accuracy of the claim.
If the information mentioned in this claim pertains to scene text, you'll need to generate question about the scene text.
If the claim is unrelated to the scene text information in the image, such as: objects, colors, actions, position etc, simply return 'none'.
YOU MUST RETURN THE RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS.
You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE.
response format: {{"claim1":["question1", "question2"],"claim2":["none"],"claim3":["question1", "question2"], ...}}

Here are three examples:
claim list:
claim1: There is a black device in the image.
claim2: The device is a brand of smartphones produced by Samsung Electronics.
output: {{"claim1":["none"],"claim2":["What is the brand of the device in the image?"]}}

claim list:
claim1: A stop sign is on the left.
claim2: The stop sign says stop eating animals.
output: {{"claim1":["none"],"claim2":["What does the stop sign say in the image?"]}}

claim list:
claim1: The words 'Hello World' are written on the car.
claim2: A man is standing beside the car.
output: {{"claim1":["What are written on the car?"],"claim2":["none"]}}

Now complete the following with following the above rules. DO NOT RESPOND WITH ANYTHING ELSE.
claim list:
{claims}
output:

Table 5: Prompt template of query formulation (scene-text-level) for image-to-text generation.

SYSTEM:
You are a brilliant question generator.

USER:
Given a list of claim, you're required to generate questions about related to factual visual information.
For a claim based on factual knowledge, Your primary task is to generate a Python list of two effective and skeptical search engine questions.
These questions should assist users in critically evaluating the factuality of a provided claim using search engines.
If a claim is not based on factual knowledge, simply return 'none'.
YOU MUST RETURN THE RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS.
You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE.
response format: {{"claim1":["question1", "question2"],"claim2":["none"],"claim3":["question1", "question2"], ...}}

Here are three examples:
claim list:
claim1: The image shows a black phone.
claim2: This black phone is manufactured by Huawei.
claim3: Huawei is a company located in Shenzhen, China.
output:
{{"claim1":["none"],"claim2":["none"],"claim3":["Where is Huawei headquartered?", "Huawei company"]}}

claim list:
claim1: The image shows an app of twitter.
claim2: The CEO of twitter is Bill Gates.
output: {{"claim1":["none"],"claim2":["Who is the CEO of twitter?", "CEO Twitter"]}}

claim list:
claim1: The man is playing baseball.
claim2: The man is wearing a colorful shirt.
output: {{"claim1":["none"],"claim2":["none"]}}

Now complete the following with following the above rules. DO NOT RESPOND WITH ANYTHING ELSE.
claim list:
{claims}
output:

Table 6: Prompt template of query formulation (fact-level) for image-to-text generation.

SYSTEM:
You are a brilliant hallucination judger.

USER:
Given a list of claims from Multimodal Large Language Models and an image, you are required to judge whether each claim in the list by the Multimodal Large Language Model model conflicts with the image, following these rules:
1. You must carefully judge from four aspects, including the object, attributes, scene text and fact. Here are specific descriptions of the four aspects for you to review:
"Object" specifically refers to whether the objects in the image exist and if the quantity of objects conflicts with the object information in the claims;
"Attributes" specifically refer to whether the color, position, action of objects in the image conflict with the attribute information in the claims;
"Scene Text" specifically refers to whether the textual information in the scene of the image conflicts with the required textual information in the claims.
"Fact" specifically refers to relevant factual knowledge obtained by querying a search engine. You can verify the factual accuracy of the claims based on the provided external knowledge.
2. You'll also receive detection results from the expert model. The object detection expert model will provide detected entity names along with their bounding box information in the image. When deriving position relationships between entity instances, try to also use the bounding boxes information, which are represented as [x1, y1, x2, y2] with floating numbers ranging from 0 to 1. These values correspond to the top left x1, top left y1, bottom right x2, and bottom right y2. The scene text expert model will provide detected specific text along with their bounding box information in the image. As long as there is a conflict between a single letter in the scene text and the text information required in the claim, it's considered a hallucination.
3. You must carefully judge whether the visual information in the image conflicts with each claim. If there is a conflict, the result for that statement is labeled as 'hallucination'; otherwise, it is labeled as 'non-hallucination'."
4. Finally, YOU MUST RETURN THE JUDGMENT RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS. You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE. response format: ["claim1":"hallucination", "reason":"The reason for your judgment.","claim2":"non-hallucination", "reason":"The reason for your judgment.","claim3":"hallucination", "reason":"The reason for your judgment.", ...]

[Begin of Example] (Image Entered)
Here is the object detection expert model's result:
people [0.345, 0.424, 0.408, 0.509]; people [0.197, 0.44, 0.28, 0.514]
people [0.517, 0.315, 0.561, 0.401]; people [0.441, 0.356, 0.47, 0.405]
chair [0.398, 0.595, 0.637, 0.901]; chair [0.621, 0.592, 0.789, 0.889]
umbrella [0.501, 0.334, 0.968, 0.88]

Here is the attribute detection expert model's result: none information

Here is the scene text recognition expert model's result: none information

Here is the external knowledge: none information

Here is the claim list:
claim1: The picture shows five people swimming.
claim2: On the beach, there is a chair, a umbrella, and a surfboard.
claim3: The green umbrella is on the right side of the chair.

Output: ["claim1":"hallucination","reason":"The object detection expert model identified four people, not five people. Based on the image information, they might be swimming. Therefore, there's a hallucination.", "claim2":"hallucination","reason":"According to the results of the object detection expert model and my judgment, there are two chairs and an umbrella in the picture, but there is no surfboard. Therefore, there's a hallucination.", "claim3":"non-hallucination","reason":"Based on the positional information of the bounding boxes and my judgment, the umbrella is to the right of the chairs. The umbrella is green. Therefore, there's no hallucination."]
.....
[End of Example]

<Input>:
<Output>:

Table 7: Prompt template of hallucination verification for image-to-text generation.

SYSTEM:
You are a brilliant hallucination judger.

USER:
Given a list of claims from human prompts, an image generated by the text-to-image model, you are required to judge whether the image conflicts with human-provided prompts, following these rules:
1. You must carefully judge from four aspects, including the object, attributes, scene text and fact. Here are specific descriptions of the four aspects for you to review:
"Object" specifically refers to whether the objects in the image exist and if the quantity of objects conflicts with the object information in the claims;
"Attributes" specifically refer to whether the color, position, action of objects in the image conflict with the attribute information in the claims;
"Scene Text" specifically refers to whether the textual information in the scene of the image conflicts with the required textual information in the claims.
"Fact" specifically refers to relevant factual knowledge obtained by querying a search engine. You can verify the factual accuracy of the claims based on the provided external knowledge.
2. You'll also receive detection results from the expert model. The object detection expert model will provide detected entity names along with their bounding box information in the image. When deriving position relationships between entity instances, try to also use the bounding boxes information, which are represented as [x1, y1, x2, y2] with floating numbers ranging from 0 to 1. These values correspond to the top left x1, top left y1, bottom right x2, and bottom right y2. The scene text expert model will provide detected specific text along with their bounding box information in the image. As long as there is a conflict between a single letter in the scene text and the text information required in the claim, it's considered a hallucination.
3. You must carefully judge whether the visual information in the image conflicts with each claim. If there is a conflict, the result for that statement is labeled as 'hallucination'; otherwise, it is labeled as 'non-hallucination'."
4. Finally, YOU MUST RETURN THE JUDGMENT RESULTS IN A DICTIONARY ACCORDING TO THE GIVEN ORDER OF THE LIST OF CLAIMS. You MUST only respond in the format as described below. DO NOT RESPOND WITH ANYTHING ELSE. response format: ["claim1":"hallucination", "reason":"The reason for your judgment.","claim2":"non-hallucination", "reason":"The reason for your judgment.","claim3":"hallucination", "reason":"The reason for your judgment.", ...]

[Begin of Example] (Image Entered)
Here is the object detection expert model's result:
basketball [0.741, 0.179, 0.848, 0.285]
boy [0.773, 0.299, 0.98, 0.828]
car [0.001, 0.304, 0.992, 0.854]

Here is the attribute detection expert model's result: none information

Here is the scene text recognition expert model's result:
worlld [0.405, 0.504, 0.726, 0.7]

Here is the external knowledge: none information

Here is the claim list:
claim1: The side of the car reads 'Hello World'
claim2: A boy is playing a yellow basketball beside a plant.

Output: ["claim1":"hallucination", "reason":"The object detection model has identified a car in the image. However, based on the detection results of the scene text expert model and my judgment, the text in the image is 'hello worlld' not 'hello world'. Therefore, there's a hallucination.","claim2":"hallucination", "reason":"The object detection model has identified a boy and a basketball in the image. And the boy is visible in the image playing with a yellow basketball. But according to the detection results of the object detection expert model and my judgment, there's no plant. Therefore, there's a hallucination."]
.....
[End of Example]

<Input>:
<Output>:

Table 8: Prompt template of hallucination verification for text-to-image generation.