UNIFIED HALLUCINATION DETECTION FOR MULTIMODAL LARGE LANGUAGE MODELS

Xiang Chen[♣♡*], **Chenxi Wang**[♣♡*], **Ningyu Zhang**[♣♡†], **Yida Xue**[♣♡], **Xiaoyan Yang**[♢], **Qiang Li**[♢], **Yue Shen**[♢], **Jinjie Gu**[♢], **Huajun Chen**[♣♡†]
♣ Zhejiang University ♢ Ant Group
♡ Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph
{xiang_chen,zhangningyu}@zju.edu.cn

https://openkg-org.github.io/easydetect/

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; Yin et al., 2023). 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.

Although several works Zhou et al. (2023); Zhai et al. (2023); Li et al. (2023); Wang et al. (2023b) have been conducted to evaluate or detect hallucinations from MLLMs, 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 hallucinations 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

[*] Equal contribution.
[†] Corresponding author.
[1]The code can be accessed via https://github.com/OpenKG-ORG/EasyDetect, and the demonstration is available at http://easydetect.openkg.cn.

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**ultiModal **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.

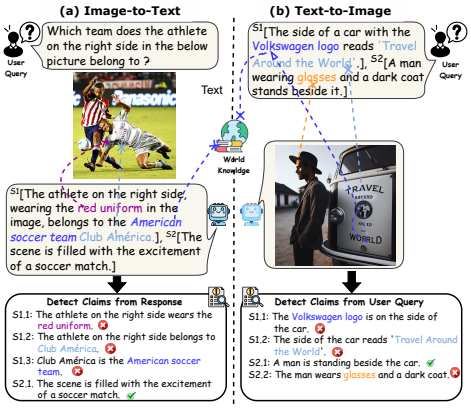

Figure 1: Unified multimodal hallucination detection aims to identify and detect modality-conflicting hallucinations at various levels such as object, attribute, and scene-text, as well as fact-conflicting hallucinations in both image-to-text and text-to-image generation. Our benchmark emphasizes fine-grained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

At its core, leveraging MLLMs' inherent self-detection mechanisms to pinpoint diverse hallucinations encounters significant hurdles. Inspire by Chern et al. (2023), 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.

2 UNIHD: UNIFIED HALLUCINATION DETECTION FRAMEWORK FOR MLLMS

Addressing the key challenges in hallucination detection, we introduce a tool-enhanced framework that systematically tackles hallucination identification in multi-modal large language models (MLLMs) for both image-to-text and text-to-image tasks. Our framework capitalizes on the domain-specific strengths of various tools to efficiently gather multi-modal evidence for confirming hallucinations. The components of the framework, UNIHD, are detailed in Fig. 2. Due to the page limit, we provide background and details of dataset construction in Appendix A and B.

2.1 ESSENTIAL CLAIM EXTRACTION

To identify fine-grained hallucinations within the response, claim extraction is a prerequisite. Following the procedure of "claim collection" in Appendix B.2, we employ the advanced instruction-following abilities of MLLMs for efficient claim extraction, bypassing the extensive resources typically needed for model training. Specifically, GPT-4V/Gemini is adopted as the base LLM

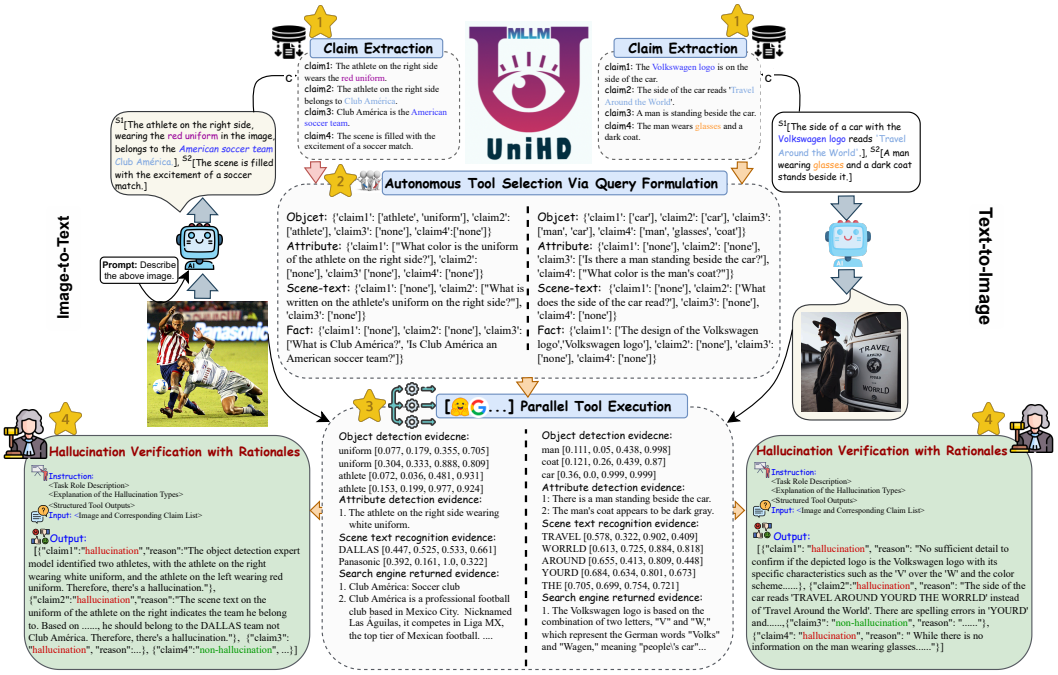

Figure 2: The specific illustration of UNIHD for unified multimodal hallucination detection.

to efficiently derive verifiable claims from the outputs of image-to-text models (extracting each response into individual claims) and text-to-image models (deconstructing user instructions into distinct claims). These claims are represented as $\{c_i\}_{i=1\cdots n}$ [2].

2.2 AUTONOMOUS TOOL SELECTION FOR CLAIM

After extracting essential claims in 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 questions 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 leverage underlying MLLMs like GPT-4V/Gemini, providing them with contextual examples to autonomously generate meaningful questions. Demonstrated in Figure 1, our autonomous tool selection yields custom queries for each claim, or "`none`" when a tool is unnecessary. For example, as seen in Figure 1, 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.

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

[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.

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 1: Experimental results of UNIHD powered by Gemini and GPT-4v on Image-to-Text Generation and Text-to-Image Generation.

- *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.

2.4 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 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.

3 EXPERIMENT

Due to the page limit, we further analyze "**Which Type of Hallucination Detection Can Be More Effectively Enhanced by Tools?**", "**Explanation Reasonability of UNIHD.**", "**Failure Analysis of UNIHD.**", "**Text-to-Image Hallucinations vs. Image-to-Text Hallucinations: Which is Easier to Detect?**", "**Explore UNIHD to Evaluate Hallucination of Modern MLLMs.**" in Appendix D.2.

[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.

3.1 EVALUATION RESULTS

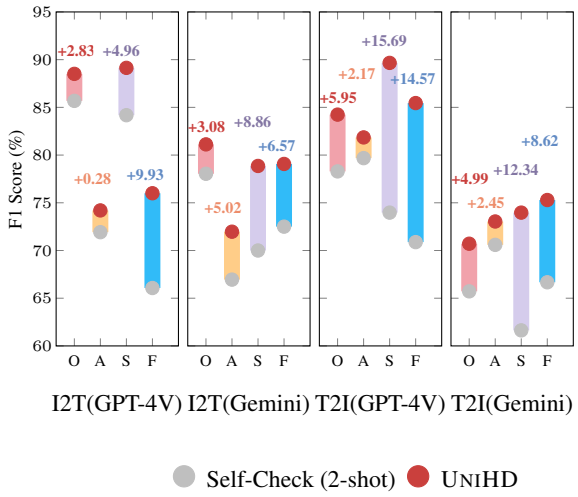

Figure 3: The statistical analysis was conducted on samples with hallucinatory labels. In this analysis, the x-axis labels "O", "A", "S" and "F" refer to object, attribute, scene-text, and fact, respectively.

MHaluBench poses a challenging benchmark for multimodal hallucination detection. The segment-level and response-level outcomes are presented in Table 1. Even though all hallucinatory instances in MHaluBench were 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.

GPT-4V surpasses Gemini as the detector base. GPT-4V-powered detectors consistently outperform Gemini counterparts, achieving higher Macro-F1 scores, especially in 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 1 demonstrates that UNIHD, leveraging GPT-4V, consistently outperforms 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.

4 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. Looking ahead, we aim to expand UNIHD into an API service that can assess the prevalence of hallucinations in MLLMs and serve as a foundation for model editing.

REFERENCES

James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, Yunxin-Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Huajun Chen. Large knowledge model: Perspectives and challenges. *CoRR*, abs/2312.02706, 2023. doi: 10.48550/ARXIV.2312.02706. URL https://doi.org/10.48550/arXiv.2312.02706.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. Factchd: Benchmarking fact-conflicting hallucination detection, 2024.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528, 2023. doi: 10.48550/ARXIV.2307.13528. URL https://doi.org/10.48550/arXiv.2307.13528.

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. Agent ai: Surveying the horizons of multimodal interaction, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *CoRR*, abs/2307.06350, 2023a. doi: 10.48550/ARXIV.2307.06350. URL https://doi.org/10.48550/arXiv.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. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023b. doi: 10.48550/ARXIV.2311.05232. URL https://doi.org/10.48550/arXiv.2311.05232.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911, 2023c. doi: 10.48550/ARXIV.2311.17911. URL https://doi.org/10.48550/arXiv.2311.17911.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *EMNLP*, 2023. URL https://doi.org/10.48550/arXiv.2305.10355.

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. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *CoRR*, abs/2303.16434, 2023. doi: 10.48550/ARXIV.2303.16434. URL https://doi.org/10.48550/arXiv.2303.16434.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

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

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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, 2023a. doi: 10.48550/ARXIV.2310.14566. URL https://doi.org/10.48550/arXiv.2310.14566.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565, 2023b. doi: 10.48550/ARXIV.2306.14565. URL https://doi.org/10.48550/arXiv.2306.14565.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023c. doi: 10.48550/ARXIV.2304.08485. URL https://doi.org/10.48550/arXiv.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. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023d. doi: 10.48550/ARXIV.2303.05499. URL https://doi.org/10.48550/arXiv.2303.05499.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

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

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813, 2023. doi: 10.48550/ARXIV.2302.12813. URL https://doi.org/10.48550/arXiv.2302.12813.

Shuofei Qiao, Honghao Gui, Huajun Chen, and Ningyu Zhang. Making language models better tool learners with execution feedback. *CoRR*, abs/2305.13068, 2023. doi: 10.48550/ARXIV.2305.13068. URL https://doi.org/10.48550/arXiv.2305.13068.

Vipula Rawte, Amit P. Sheth, and Amitava Das. A survey of hallucination in large foundation models. *CoRR*, abs/2309.05922, 2023. doi: 10.48550/ARXIV.2309.05922. URL https://doi.org/10.48550/arXiv.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. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *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*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

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

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

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

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 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*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

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

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. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521, 2023a. doi: 10.48550/ARXIV.2310.07521. URL https://doi.org/10.48550/arXiv.2310.07521.

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. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126, 2023b. doi: 10.48550/ARXIV.2308.15126. URL https://doi.org/10.48550/arXiv.2308.15126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models' vacillations in judgement. *CoRR*, abs/2310.02174, 2023. doi: 10.48550/ARXIV.2310.02174. URL https://doi.org/10.48550/arXiv.2310.02174.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda R. Petzold, William Yang Wang, and Wei Cheng. A survey on detection of llms-generated content. *CoRR*, abs/2310.15654, 2023. doi: 10.48550/ARXIV.2310.15654. URL https://doi.org/10.48550/arXiv.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. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023. doi: 10.48550/ARXIV.2304.14178. URL https://doi.org/10.48550/arXiv.2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045, 2023. doi: 10.48550/ARXIV.2310.16045. URL https://doi.org/10.48550/arXiv.2310.16045.

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

doi: 10.48550/ARXIV.2310.01779. URL https://doi.org/10.48550/arXiv.2310.01779.

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. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. doi: 10.48550/arXiv.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *CoRR*, abs/2310.00754, 2023. doi: 10.48550/ARXIV.2310.00754. URL https://doi.org/10.48550/arXiv.2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023. doi: 10.48550/ARXIV.2304.10592. URL https://doi.org/10.48550/arXiv.2304.10592.

A PRELIMINARIES

Figure 4 illustrates our extension of hallucination detection for MLLMs to cover both image-to-text and text-to-image generation. Further, we explore a unified perspective on hallucination in MLLMs with the aspiration of developing a unified framework for hallucination detection.

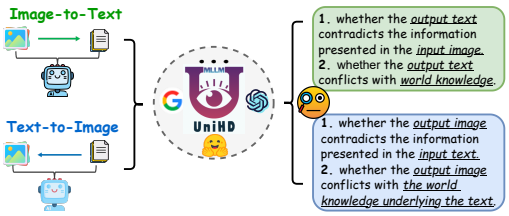

Figure 4: Unified View ( ) of multimodal hallucination detection.

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 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.

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

Table 2: 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.

B 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. Statistical details about MHaluBench are provided in Figure 5 and Figure 6.

B.1 HALLUCINATORY EXAMPLE COLLECTION

Image-to-Text Generation. We concentrate on IC and VQA tasks, sampling examples from validation of MS-COCO 2014 Lin et al. (2014) and testing of TextVQA Singh et al. (2019). Based on these samples, we aggregate generative outputs from mplug Ye et al. (2023), LLaVA Liu et al. (2023c), and MiniGPT-4 Zhu et al. (2023) as foundational data for MHaluBench.

Text-to-Image Generation. To curate a dataset of text-prompted images exhibiting typical hallucinatory features, 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 3 model Betker et al. (2023) in producing visually detailed images.

B.2 FINE-GRAINED HUMAN ANNOTATION

Beyond response evaluation, we implement claim-level fine-grained annotation to pinpoint hallucinations, facilitating targeted feedback for enhancing model capability Lightman et al. (2023).

Segment and Claim Collection for Detection. We propose utilizing ChatGPT's advanced instruction-following prowess to extract fine-grained segments and associated claims. For image-to-text tasks, we capture the model's textual output, while in text-to-image scenarios, we distill user queries into constituent intent concepts, which are then treated as claims.

Human Annotation Principles. 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.

Annotator Collaboration and Agreement. 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 ($\kappa = 0.855$) in a random subset of 100 annotations, indicating a high level of concordance within the range $0.80 \leq \kappa \leq 1.00$.

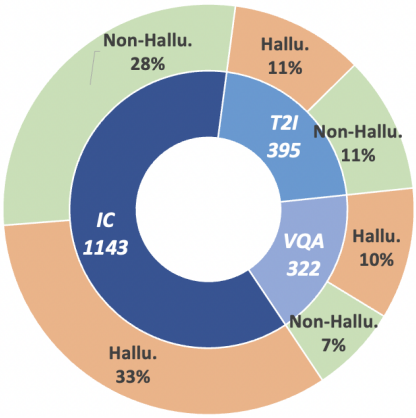

Figure 5: Claim-Level data statistics of MHaluBench. "IC" signifies Image Captioning and "T2I" indicates Text-to-Image synthesis, respectively.

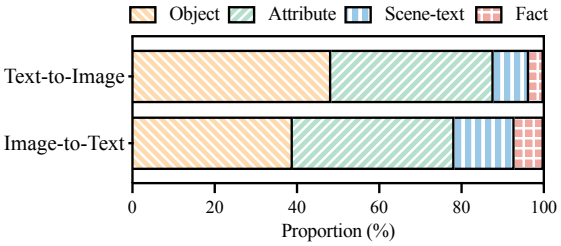

Figure 6: Distribution of hallucination categories within hallucination-labeled claims of MHaluBench.

C Related Work

C.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 Zhang et al. (2023); Huang et al. (2023b); Rawte et al. (2023); Ji et al. (2023); Yin et al. (2023), a crucial concern impacting their dependability. Previous research has primarily focused on three areas: evaluating Li et al. (2023); Liu et al. (2023a), detecting Wang et al. (2023b); Yang et al. (2023), and mitigating hallucinations Liu et al. (2023b); Huang et al. (2023c); Semnani et al. (2023). In a complementary effort, HaELM Wang et al. (2023b) scrutinizes the challenges associated with POPE Li et al. (2023) 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.

C.2 Harnessing Tool Resources for LLMs

Addressing the limitations of LLMs Chen (2023) 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. (2024) 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.

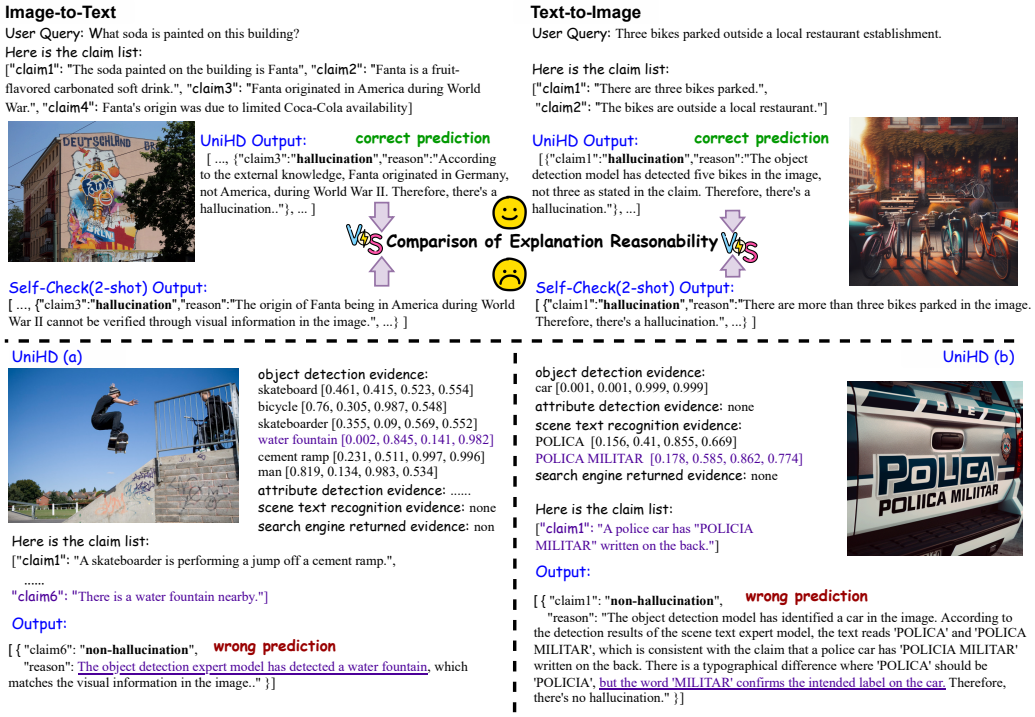

Figure 7: **Case Study.** The upper section depicts two exemplary cases where both UNIHD and Self-Check (2-shot) arrive at correct judgments, with a comparative demonstration of UNIHD providing explanations of superior reasonability. UNIHD (a) reveals a failure case where the tool presents erroneous 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.

D EXPERIMENT

D.1 EXPERIMENTAL SETTINGS

Baselines. We compare UNIHD with two baselines, Self-Check (2-shot)[4] and Self-Check (0-shot) based on 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 Madaan et al. (2023); Chern et al. (2023); Xie et al. (2023). In practice, we prompt GPT-4V (`gpt-4-vision-preview`) and Gemini[5] to recognize fine-grained hallucination 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.

D.2 EXPERIMENTAL ANALYSIS

Which Type of Hallucination Detection Can Be More Effectively Enhanced by Tools? Figure 3 shows that UNIHD significantly 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

[4]Self-Check (2-shot) utilize two complete demonstrations based on $a = \{v, x\}$ rather than only two claims.
[5]`https://deepmind.google/technologies/gemini/`

provided by the tool especially valuable. However, UNIHD exhibits minimal improvement in identifying attribute-level hallucinations, likely due to a dearth of tools tailored for direct attribute detection. The gains achieved through self-reflection attribute detection based on GPT-4V/Gemini, are relatively limited.

Explanation Reasonability of UNIHD. As shown in the upper portion of Figure 7, both the fact-level hallucinatory claim "Fanta originated in America during World War." and the object-level hallucinatory claim "There are three bikes parked." are correctly judged by both Self-Check (2-shot) and UNIHD. Upon further comparison of the reasons provided by these detectors, it is apparent that UNIHD is adept at combining the evidence returned by the tool to furnish a more reliable and persuasive explanation.

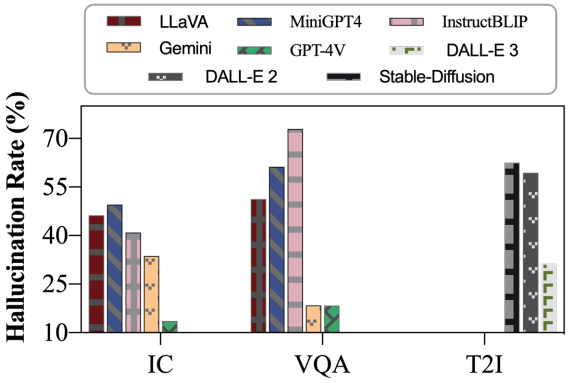

Figure 8: Comparison of claim-level hallucination ratios across MLLMs. We randomly select a set of 20 prompts from MHaluBench for each of the IC, VQA, and T2I. Responses for these prompts are generated by each of the evaluated MLLMs.

Failure Analysis of UNIHD. As depicted in the lower section of Figure 7, we present two failure modes of UNIHD. The left case illustrates an instance where, despite obtaining accurate evidence, the MLLM persists in its initial bias, leading to a wrong decision. On the right, we see cases where the tool produces incorrect evidence or provides no helpful information, causing the MLLM to make erroneous judgments. These scenarios highlight areas for further research to enhance tool accuracy and to develop MLLMs dedicated to better hallucination detection.

Text-to-Image Hallucinations vs. Image-to-Text Hallucinations: 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 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 goledn 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, demonstrating the potential of UNIHD for evaluating hallucinations.