IllusionBench: A Large-scale and Comprehensive Benchmark for Visual Illusion Understanding in Vision-Language Models

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, Xiongkuo Min*
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China
{ming_zhang_sjtu, zzc1998, moj-will, xiaohongliu, zhaiguangtao, minxiongkuo}@sjtu.edu.cn

***Abstract*—Current Visual Language Models (VLMs) show impressive image understanding but struggle with visual illusions, especially in real-world scenarios. Existing benchmarks focus on classical cognitive illusions, which have been learned by state-of-the-art (SOTA) VLMs, revealing issues such as hallucinations and limited perceptual abilities. To address this gap, we introduce IllusionBench, a comprehensive visual illusion dataset that encompasses not only classic cognitive illusions but also real-world scene illusions. This dataset features 1,051 images, 5,548 question-answer pairs, and 1,051 golden text descriptions that address the presence, causes, and content of the illusions. We evaluate ten SOTA VLMs on this dataset using true-or-false, multiple-choice, and open-ended tasks. In addition to real-world illusions, we design trap illusions that resemble classical patterns but differ in reality, highlighting hallucination issues in SOTA models. The top-performing model, GPT-4o, achieves 80.59% accuracy on true-or-false tasks and 76.75% on multiple-choice questions, but still lags behind human performance. In the semantic description task, GPT-4o's hallucinations on classical illusions result in low scores for trap illusions, even falling behind some open-source models. IllusionBench is, to the best of our knowledge, the largest and most comprehensive benchmark for visual illusions in VLMs to date.**

***Index Terms*—Benchmark, VLM, Visual Illusion**

I. Introduction

Visual illusions are perceptual anomalies caused by the visual system, characterized by a discrepancy between visual perception and reality [1]. However, Richard Gregory's classification [2], [3] provides a framework by dividing visual illusions into three main categories: physical illusions, physiological illusions, and cognitive illusions. Among these, cognitive visual illusions are the result of unconscious inferences and are perhaps the most widely recognized.

These classic cognitive illusion images share a common feature: they are all artificially synthesized and inherently ambiguous.

In addition to artificially synthesized images, a small proportion of images captured in real-world scenes also exhibit visual illusions. The fundamental cause of this phenomenon is the inverse projection problem, where information is irreversibly lost during the projection from the three-dimensional world to two-dimensional images [4]. This results in difficulties such as information loss, ambiguity, and multiple possible interpretations when attempting to infer three-dimensional objects and scenes from two-dimensional images (light and shadow projections) [5].

While the human brain compensates for the missing depth information in two-dimensional images (retinal projections) through binocular disparity and motion parallax [6], [7], this issue remains unresolved in two-dimensional images captured by cameras. Consequently, both humans and vision models may experience visual illusions, leading to difficulties or errors in interpreting these images [8].

Fig. 1: Error cases from IllusionBench.

To address this challenge, the human visual system leverages contextual cues for cognitive reasoning and utilizes monocular cues, such as perspective, occlusion relationships, shadows, and lighting, to alleviate the difficulties in information interpretation [9]. However, the extent to which current VLMs can recognize and interpret these visual illusions in real-world scenes remains an open question, as shown in Fig. 1.

Recent advancements in VLMs, like GPT-4o and Gemini-pro-1.5, have greatly improved visual question answering (VQA) [10]–[12]. The improvements highlight their growing ability to bridge the gap between visual and textual information, enabling them to understand visual illusions.

Previous research has used artificially synthesized classic

*Corresponding authors
This work was supported in part by the National Natural Science Foundation of China under Grant 62271312 and Grant 62132006, and in part by STCSM under Grant 22DZ2229005.

TABLE I: Comparison of IllusionBench with other illusion datasets

Dataset	Base Image	Question Type	Number of Instance	Text Description?	Image Type
GVIL	16	Binary	1600	×	Color & Size illusions and variant
HallusionBench	72	Binary	1129*	×	Color & Size illusions and variant
IllusionVQA	374	Multiple-choice	1435	×	12 types, mainly classical synthetic cognitive illusions
Ours	1051	Binary, Multiple-choice, Open-ended description	6599	✓	Classic illusions, real scene illusions, trap illusions, no illusion, and Ishihara images

* Note: The instances in HallusionBench include more than just visual illusions.

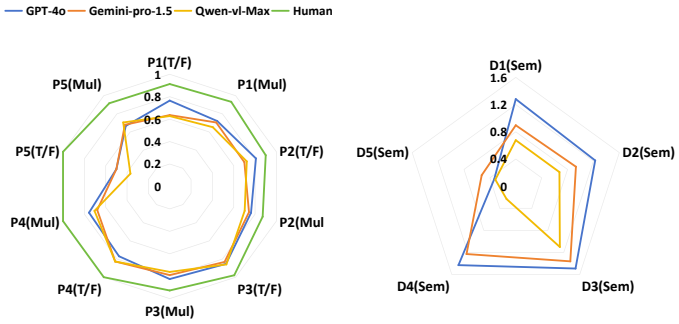

Fig. 2: Performance of advanced VLMs and human evaluators on IllusionBench perception tasks (left) and description tasks (right). The left image shows P1-P5 representing perception tasks on the subsets of **Classic Cognitive Illusion, Real Scene Illusion, No Illusion, Ishihara Image, and Trap Illusion**, respectively. Similarly, the right image shows D1-D5 representing description tasks on these subsets. "T/F", "Mul", "Sem", and "Illu" respectively represent true-or-false, multiple-choice and semantic descriptions.

cognitive visual illusion images as benchmarks for VLMs to explore the similarities between artificial intelligence and human visual cognition and to evaluate VLMs' understanding of visual illusions [13]–[15]. Unlike previous studies, our work includes not only classic cognitive illusion images, which lack real-world context, but also a large collection of real-scene visual illusions. These real-world images better represent practical applications and assess VLMs' ability to use contextual cues, similar to human perception. Additionally, because SOTA VLMs may have already learned classical illusions, these images may no longer be sufficient to test the visual perception ability of VLMs. To address the issue of potential overfitting to classic cognitive illusions, we introduce Ishihara color blindness detection images and trap illusion images. These images are accompanied by carefully crafted, manually annotated question-answer pairs, as well as image descriptions that cover image semantics, the presence of visual illusions, and their underlying causes.

Using our testing framework, we comprehensively evaluate the latest SOTA VLMs, such as GPT-4o, Gemini-pro-1.5, and several open-source models. Specifically, our framework includes true-or-false, multiple-choice, open-ended questions. Additionally, each image in our dataset is assigned a manually annotated cognitive difficulty level, and we conduct human testing to provide a multidimensional, fine-grained comparison between human performance and VLMs' performance on visual illusion cognition tasks. Fig. 2 shows the performances of SOTA VLMs on these tasks. Our contributions can be summarized as follows:

- **IllusionBench Dataset:** We build a large-scale dataset that includes both classic and real-world visual illusions, color blindness test images, and trap illusions, supplemented with question-answer pairs and detailed annotations on image semantics, the presence of illusions, and their causes. To the best of our knowledge, IlluisionBench is the **largest** and most comprehensive visual illusion benchmark for VLMs to date.
- **Comprehensive Testing Framework:** We apply a rigorous framework to evaluate SOTA VLMs, such as GPT-4o and Gemini-pro-1.5, using a range of question types including true-or-false, multiple-choice, open-ended, ensuring a thorough evaluation of the models' capabilities in understanding visual illusions.

II. Related Work

Existing research has demonstrated that visual illusions for humans can induce equivalent illusions in models [16], [17]. However, these studies primarily focus on specific types of illusions, including motion [18], brightness and color [19], and completion [20].

Recent studies have explored VLMs' ability to perceive visual illusions through natural language. A pioneering work [13] tested this by using a dataset of 1,600 variants from 16 root images, focusing on color and geometric distortions. The study aims to evaluate if SOTA VLMs align with human perception in visual illusions. Results show that while larger models perform better in localization tasks, VLMs generally struggle to interpret visual illusions as humans do.

Another study [14] introduced a benchmark to evaluate VLMs' handling of visual illusions and language hallucinations using a dataset of 346 images, including 72 focused on illusions, paired with question-answer tasks. The models, including GPT-4V, struggle with these illusions and hallucinations, achieving only 31.42% accuracy. This highlights a misalignment with human perception and suggests that SOTA

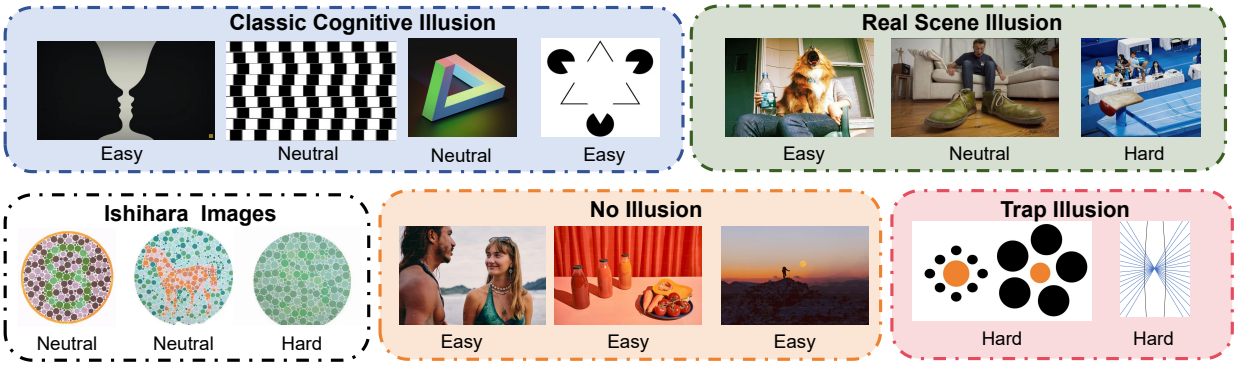

Fig. 3: **Categories in IllusionBench.** The annotations under each image represent the human cognitive difficulty score.

VLMs may overfit classic illusions, making them less effective for testing complex visual understanding.

Additionally, [15] introduced a dataset of 374 classic cognitive illusion images, generating 439 question-answer pairs to test VLMs' understanding and localization of challenging visual content. The study finds that advanced VLMs like GPT-4V and Gemini Pro perform poorly on visual illusions, with accuracy below that of human evaluators, highlighting current limitations in interpreting complex visual scenes.

Previous studies mainly focus on synthetic cognitive illusions, but our study expands this by including real-world scenes with visual illusions to better assess VLMs' use of contextual cues. Additionally, we introduce Ishihara and trap illusions to evaluate potential overfitting, ensuring a more precise alignment with human visual perception. To the best of our knowledge, IllusionBench is the largest and most comprehensive visual illusion benchmark for VLMs to date, as shown in Table I.

III. IllusionBench

To evaluate VLMs' understanding of visual illusions, we create IllusionBench with more than 1K images equipped with 5K QA pairs and manually annotated golden descriptions, as shown in Fig. 4. IllusionBench includes five image types: classical illusions, real scene illusions, no illusions, Ishihara images, and trap illusions (Fig. 3). Testing tasks involve judgment, multiple-choice, and descriptive questions focused on illusion existence, causes, and content. This section details the dataset composition, question generation methods, and tasks.

A. Collection and Composition of Images

We collect 1K+ images from various online repositories. After manual selection, 780 images are confirmed to contain visual illusions, 26 are Ishihara color blindness detection images, and 245 images have no illusions, as shown Appendix Fig. 1. The details are:

- **Classic Cognitive Illusion Images** These include blur, distortion, paradox, and fictitious illusions—key examples of traditional synthetic illusions. Designed by psychologists, these ambiguous images test VLMs' alignment with human perception. However, their classic nature and limited number may reduce their effectiveness, as they could be part of advanced VLMs' training datasets.
- **Trap Illusion** Trap illusions are edited versions of classic visual illusions, resembling them in appearance but differing in physical properties. These images test whether VLMs overfit classic cognitive illusions, potentially causing hallucinations.
- **Real Scene Illusion Images** IllusionBench includes 597 real-scene images with visual illusions. These images depict real-world objects and scenes, with unique and definite semantic descriptions. The illusions arise from the inverse projection problem, where information is lost in the transition from 3D to 2D. Understanding these images requires monocular cues like perspective, occlusion, shadows, and lighting, as well as contextual reasoning.
- **Ishihara Color Blindness Detection Images** IllusionBench includes 26 Ishihara images, verified by vision-healthy individuals, where the patterns convey unique and definite semantics. These images test whether VLMs' visual cognition aligns with human perception, specifically regarding Gestalt principles such as grouping, similarity, and proximity.
- **No Illusion Images** IllusionBench contains 245 images with no illusions, depicting diverse subjects such as people, landscapes, and objects. These images provide a baseline for evaluating VLMs' visual understanding and the impact of illusions and evaluate the models' yes-bias when addressing questions about illusion presence.

B. Benchamrk on Illusion Perception Ability

1) Question Types and QA Pairs Generation

The question-answer pairs in IllusionBench include both binary (true-or-false) and multiple-choice questions. Each image is accompanied by at least two binary questions and three multiple-choice questions, all manually annotated by humans. Each image also has a manually assigned cognitive difficulty rating, categorized as Easy, Neutral, or Hard, with all questions related to a given image sharing the same difficulty level.

- **True-or-false Question:** IllusionBench includes over 2,200 binary questions focused on semantic content and the presence of illusions, with 57% of correct answers

Check if the following description is correct
Q: The man is drinking from the can. A: False
Q: There is a visual illusion in the image. A: True

What is the man wearing?
i. A hat ii. A t-shirt iii. A jacket iv. A dress
What causes the visual illusion?
i. The can is closer to the camera ii. The man is a giant
iii. The can is digitally altered iv. The sky is artificially colored

In this picture, a man stands with his mouth open in front of a beer can held in another person's hand. There is a visual illusion in the image. The illusion is created by the clever use of forced perspective, camera angle and positioning, making it appear as though the man is about to drink from an enormous beer can, even though the can is actually much closer to the camera than the man.

Fig. 4: Example of real scene illusion in IllusionBench. Each image in IllusionBench is equipped with at least two true-or-false questions, three multiple-choice questions, and a description that summarizes the semantic content of the image, the existence of visual illusions, and their causes.

TABLE II: Performance of VLMs across different image categories and difficulty levels on IllusionBench true-or-false task. The best performance is marked in **bold**. "Human" refers to the average performance of two human evaluators.

Sub-category	Image Category					Difficulty Rating			All
VLMs	Classical (P1)	Real Scene (P2)	No Illusion (P3)	Ishihara (P4)	Trap (P5)	Easy	Neutral	Hard	
Closed-Source VLMs									
GPT-4o	**0.7653**	**0.8082**	0.8532	0.7692	0.5000	0.8526	0.8040	0.7397	**0.8059**
Gemini-pro-1.5	0.6363	0.6998	0.8319	0.8269	0.5000	0.7907	0.6943	0.6591	0.7183
Qwen-vl-Max	0.6276	0.7223	**0.8589**	0.8269	0.3684	0.8295	0.6913	0.6777	0.7338
Qwen-vl-plus	0.5522	0.6479	0.8250	**0.9800**	0.3055	0.7592	0.6447	0.6058	0.6742
Opened-Source VLMs									
CogVLM-17B (Vicuna-v1.5-7B)	0.4028	0.4291	0.4431	0.4808	**0.5263**	0.4263	0.4352	0.4286	0.4308
DeepSeek-VL-7B-chat	0.3994	0.4705	0.4812	0.4694	**0.5263**	0.4682	0.4649	0.4478	0.4626
InternLM-XComposer2-VL-7B (InternLM2)	0.5552	0.6456	0.8033	0.7500	0.3158	0.7436	0.6379	0.5914	0.6625
LLaVA-v1.5 (Vicuna-v1.5-7B)	0.4128	0.4192	0.4741	0.4808	**0.5263**	0.4437	0.4319	0.4204	0.4333
LLaVA-v1.5 (Vicuna-v1.5-13B)	0.5145	0.7092	0.7847	0.7692	0.2632	0.7693	0.6719	0.6061	0.6895
LLaVA-NeXT (Llama3-8B)	0.6221	0.6735	0.8302	0.7885	**0.5263**	0.7897	0.6586	0.6489	0.6995
mPLUG-Owl2 (LLaMA-7B)	0.5843	0.6154	0.7557	0.6731	0.3421	0.7001	0.6026	0.616	0.6375
Qwen-VL-Chat	0.3866	0.4449	0.4534	0.4423	**0.5263**	0.4464	0.4416	0.4230	0.4391
Human	0.9130	0.9000	0.9787	1.0000	1.0000	0.9394	0.9170	0.9142	0.9234

marked as False to counteract yes-bias in some VLMs. Semantic statements are intentionally misleading by human visual standards as shown in Fig. 4.

- **Multiple-choice Questions:** IllusionBench also features over 3,300 multiple-choice questions targeting fine-grained perception of image content and illusion causes. Each question offers four options, with one correct answer. Options are shuffled during evaluation.

2) LLM-assisted Evaluation for VQA

Our observations reveal that some VLMs do not output answers in the specified format. So we employ a *LLM-Assisted Evaluation* method, which involves inputting the questions, correct answers, and VLM responses into a large language model (LLM) to evaluate the accuracy of the responses. Qwen-plus assisted in the evaluation of all models for 5 rounds.

While *LLM-Assisted Evaluation* is efficient, it can sometimes err when the model's output significantly deviates from the standard answer format. To address this, we manually review and correct all cases marked incorrect by the LLM. Thus, our evaluation combines manual and LLM-assisted methods for accuracy. Further details are in the Appendix. B.

C. Benchmarks on Illusion Description Ability

1) Golden Description Definition and Question Type

In addition to multiple question-answer pairs, each image is also accompanied by a manually crafted *golden description*, covering the main content of the image, the existence of any visual illusion, and the causes of the illusion. The average length of each description is 53.21 words. All descriptions follow the format:

In this picture, [image semantics content]. There [is/is no] visual illusion in the image. The reason for visual illusions is [illusion causes].

Supported by the golden descriptions, we conduct open-ended question-answer testing VLMs' semantic describing ability. To evaluate whether VLMs can accurately describe the semantic content of the image with illusions, the prompt is:

 # user: Please provide a description of the content in this image.

2) LLM-assisted Evaluation for Description

This work examines how VLMs understand visual illusions, which often lead to challenges and inaccuracies in image interpretation. We evaluate VLM performance by assessing the accuracy of their descriptions, specifically their alignment with physical reality and human perception.

Previous studies have shown that single-modal language models are effective for evaluating language tasks [21]. After collecting open-ended responses from the VLMs, we use advanced LLMs to quantitatively evaluate multimodal description tasks. Specifically, both the model's output and the golden description are input into the LLM, which compares the two to identify significant conflicts. Preciseness is scored on a scale of [0, 1, 2]. The evaluations of all models are assisted by Qwen-

TABLE III: Performance of VLMs across different image categories and difficulty ratings on IllusionBench multiple-choice task. The best performance is marked in **bold**. "Human" refers to the average performance of two human evaluators.

Sub-category	Image Category					Difficulty Rating			All
VLMs	Classical (P1)	Real Scene (P2)	No Illusion (P3)	Ishihara (P4)	Trap (P5)	Easy	Neutral	Hard	
Closed-Source VLMs									
GPT-4o	**0.7206**	**0.7620**	**0.8255**	**0.7564**	0.6667	**0.8163**	**0.7558**	**0.7172**	**0.7675**
Gemini-pro-1.5	0.7050	0.7432	0.7901	0.6795	0.6795	0.7998	0.7335	0.6818	0.7444
Qwen-vl-Max	0.6531	0.7026	0.7620	0.7051	**0.7051**	0.7608	0.6981	0.6392	0.7064
Qwen-vl-plus	0.5038	0.5715	0.7020	0.6410	0.5556	0.6720	0.5563	0.5339	0.5903
Opened-Source VLMs									
CogVLM-17B (Vicuna-v1.5-7B)	0.4624	0.4979	0.5943	0.5256	0.5256	0.5904	0.4980	0.4147	0.5112
DeepSeek-VL-7B-chat	0.3158	0.3603	0.3623	0.1538	0.1538	0.3843	0.3388	0.3078	0.3473
InternLM-XComposer2-VL-7B (InternLM2)	0.5188	0.6138	0.6961	0.4103	0.4103	0.6801	0.5863	0.5404	0.6077
LLaVA-v1.5 (Vicuna-v1.5-7B)	0.3195	0.3345	0.3835	0.2692	0.2692	0.3848	0.3134	0.3250	0.3395
LLaVA-v1.5 (Vicuna-v1.5-13B)	0.5075	0.5594	0.6274	0.4872	0.4872	0.6295	0.5322	0.5173	0.5612
LLaVA-NeXT (Llama3-8B)	0.5094	0.6042	0.6771	0.6410	0.6410	0.6515	0.6013	0.5404	0.6050
mPLUG-Owl2 (LLaMA-7B)	0.4530	0.5137	0.5794	0.3333	0.3333	0.5621	0.4997	0.4540	0.5107
Qwen-VL-Chat	0.3158	0.3614	0.3609	0.1538	0.1538	0.3834	0.3390	0.3106	0.3477
Human	0.9327	0.8712	0.9275	1.0000	0.9167	0.9170	0.8889	0.8889	0.8975

plus for 5 rounds. Our human study shows that Spearman's rank correlation coefficient (SRCC) between LLM and human evaluation results exceeds 0.9. Details regarding prompts and other specifics can be found in the Appendix. C.

IV. Experiment Setup

A. Vision Language Models

We test four SOTA closed-source models and eight open-source models. The closed-source models include GPT-4o (version 2024-05-13) [22], Gemini-pro-1.5 (latest update in May 2024) [23], Qwen-VL-Plus, and Qwen-VL-Max [24]. We use the latest versions available at the time of writing, with their default API parameters. The open-source models include CogVLM-17B (Vicuna-v1.5-7B) [25], DeepSeek-VL-7B-chat [26], InternLM-XComposer2-VL-7B (InternLM2) [27], LLaVA-v1.5 (Vicuna-v1.5-7B), LLaVA-v1.5 (Vicuna-v1.5-13B), LLaVA-NeXT (Llama3-8B) [28], mPLUG-Owl2 (LLaMA-7B) [29], and Qwen-VL-Chat [24]. These models span different architectures and parameter scales, are trained on a wide range of vision-language tasks, and exhibit strong visual understanding capabilities.

B. Human vs VLMs

To evaluate the alignment of the perception of visual illusion between VLMs and human, we utilize a subset of IllusionBench to evaluate human visual illusion perception. We recruited two human evaluators and provided them with a subset of 200 sampled images from the dataset, proportionally sampled according to image categories. The human evaluators completed all multiple-choice and judgment questions within this subset. We then quantify human cognitive abilities using the same LLM-assisted method described earlier.

V. Result on IllusionBench

A. Result on Illusion Perception

The existence of visual illusions significantly affects the visual perception of VLMs. We evaluate VLMs' ability to perceive visual illusions using true-or-false and multiple-choice tasks, with results in Table II and Table III, revealing several key insights:

1) GPT-4o performs best in both tasks, with a true-or-false accuracy of 0.8059 and multiple-choice accuracy of 0.7675, but still lags behind human performance, indicating room for improvement in handling illusions.

2) Performances of all VLMs vary across image categories, with higher accuracy for no-illusion images and real-scene illusions compared to classical cognition illusions. GPT-4o excels in classic illusions but underperforms in trap illusions, likely due to hallucinations when encountering patterns similar to classic ones, suggesting that testing VLMs with only classic illusions is insufficient.

3) We also use Ishihara color blindness test images to examine if VLMs' perception aligns with Gestalt principles. Qwen-vl-plus shows the highest judgment accuracy (0.98), nearing the human level, but the multiple-choice performance is weaker, highlighting gaps in fine-grained perception and specific knowledge of the Ishihara test. Other VLMs all have gaps with humans in both tasks.

B. Result on Illusion Description

TABLE IV: Performance of VLMs on IllusionBench description task. The best performance is marked in **bold**. The blue part represents the standard deviation between samples.

VLMs	Semantic Content Description
Closed-Source VLMs	
GPT-4o	**1.2872** ± 0.9315
Gemini-pro-1.5	1.0257 ± 0.9789
Qwen-vl-Max	0.7571 ± 0.9492
Qwen-vl-plus	0.7924 ± 0.9490
Opened-Source VLMs	
CogVLM-17B (Vicuna-v1.5-7B)	0.9001 ± 0.9703
DeepSeek-VL-7B-chat	0.7550 ± 0.9518
InternLM-XComposer2-VL-7B (InternLM2)	0.7431 ± 0.9313
LLaVA-v1.5 (Vicuna-v1.5-7B)	0.4814 ± 0.8260
LLaVA-v1.5 (Vicuna-v1.5-13B)	0.5290 ± 0.8540
LLaVA-NeXT (Llama3-8B)	0.7364 ± 0.9431
mPLUG-Owl2 (LLaMA-7B)	0.5975 ± 0.8916
Qwen-VL-Chat	0.7336 ± 0.9408

The performance results for VLMs on the open-ended

description task are shown in Table IV and Appendix Table II , revealing several key insights:

As shown in Table IV, GPT-4o achieves the highest overall performance in the description task. The open-source model CogVLM-17B performs comparably to the closed-source Qwen-vl series. However, as shown in Fig. 2, GPT-4o performs poorly in the trap illusion subset, even worse than some open-source models. Detailed test results can be found in Appendix Table II. This is because GPT-4o exhibits significant hallucinations regarding classical cognitive illusions, which affects its perceptual ability with trap illusion images. This also indicates that classical cognitive illusions have already been learned by some SOTA models, making them insufficient for testing the perceptual abilities of these models. The real-world scene illusions in this study benefit from their diverse sources and larger quantity, which not only enhance illusion-related research but also compensate for the limitations of classical cognitive illusions.

VI. Conclusion

In this study, we introduce IllusionBench, the most extensive and comprehensive benchmark for evaluating VLMs on visual illusions. Our findings demonstrate that while SOTA VLMs, like GPT-4o, perform well in various tasks, they still struggle to interpret visual illusions accurately, highlighting a significant gap between model performance and human perception. However, because of the significant hallucinations regarding classical cognitive illusions, GPT-4o performs poorly in the trap illusion subset. The persistent challenges indicate that there is still much room for improvement in aligning VLMs with human visual cognition. IllusionBench can bring VLMs closer to human-like understanding and interpretation of complex visual scenes.

References

[1] Dejan Todorović, "What are visual illusions?," *Perception*, vol. 49, no. 11, pp. 1128–1199, 2020.

[2] Richard L Gregory, "Putting illusions in their place," *Perception*, vol. 20, no. 1, pp. 1–4, 1991.

[3] Richard L Gregory, "Visual illusions classified," *Trends in cognitive sciences*, vol. 1, no. 5, pp. 190–194, 1997.

[4] E. Bruce Goldstein and Laura Cacciamani, *Sensation and Perception*, Cengage Learning, Boston, MA, 11th edition, 2022.

[5] Stephen E Palmer, *Vision Science: Photons to Phenomenology*, MIT Press, 1999.

[6] Ian P Howard and Brian J Rogers, *Binocular Vision and Stereopsis*, Oxford University Press, 1995.

[7] Ken Nakayama, Shinsuke Shimojo, and Vilayanur S Ramachandran, "Structure from motion," *The Cognitive Neurosciences*, vol. 2, pp. 425–437, 1994.

[8] Antonio Torralba and Alexei A Efros, "Unbiased look at dataset bias," *CVPR 2011*, pp. 1521–1528, 2011.

[9] Richard L Gregory, "Eye and brain: The psychology of seeing," *Princeton University Press*, 1998.

[10] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[11] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan, "Seed-bench: Benchmarking multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13299–13308.

[12] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[13] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai, "Grounding visual illusions in language: Do vision-language models perceive illusions like humans?," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5718–5728.

[14] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al., "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14375–14385.

[15] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar, "Illusionvqa: A challenging optical illusion dataset for vision language models," *arXiv preprint arXiv:2403.15952*, 2024.

[16] Alex Gomez-Villa, Adrián Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo, "On the synthesis of visual illusions using deep generative models," *Journal of Vision*, vol. 22, no. 8, pp. 2–2, 2022.

[17] Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo, "Color illusions also deceive cnns for low-level vision tasks: Analysis and implications," *Vision Research*, vol. 176, pp. 156–174, 2020.

[18] Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka, "Illusory motion reproduced by deep neural networks trained for prediction," *Frontiers in psychology*, vol. 9, pp. 345, 2018.

[19] Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, and Marcelo Bertalmio, "Convolutional neural networks can be deceived by visual illusions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Been Kim, Emily Reif, Martin Wattenberg, and Samy Bengio, "Do neural networks show gestalt phenomena? an exploration of the law of closure," *arXiv preprint arXiv:1903.01069*, vol. 2, no. 8, 2019.

[21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al., "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[23] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[24] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[25] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al., "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[26] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al., "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[27] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al., "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," 2024.

[29] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 13040–13051.