arXiv:2509.19070v1 [cs.CV] 23 Sep 2025

COLORBLINDNESSEVAL: CAN VISION-LANGUAGE MODELS PASS COLOR BLINDNESS TESTS?

Zijian Ling*, **Han Zhang, Yazhuo Zhou, Jiahao Cui**
Apply U
United Kingdom
zijian.ling@applyu.ai

ABSTRACT

This paper presents ColorBlindnessEval, a novel benchmark designed to evaluate the robustness of Vision-Language Models (VLMs) in visually adversarial scenarios inspired by the Ishihara color blindness test. Our dataset comprises 500 Ishihara-like images featuring numbers from 0 to 99 with varying color combinations, challenging VLMs to accurately recognize numerical information embedded in complex visual patterns. We assess 9 VLMs using Yes/No and open-ended prompts and compare their performance with human participants. Our experiments reveal limitations in the models' ability to interpret numbers in adversarial contexts, highlighting prevalent hallucination issues. These findings underscore the need to improve the robustness of VLMs in complex visual environments. ColorBlindnessEval serves as a valuable tool for benchmarking and improving the reliability of VLMs in real-world applications where accuracy is critical. The code and dataset are available at https://github.com/ApplyU-ai/ColorBlindnessEval.

1 INTRODUCTION

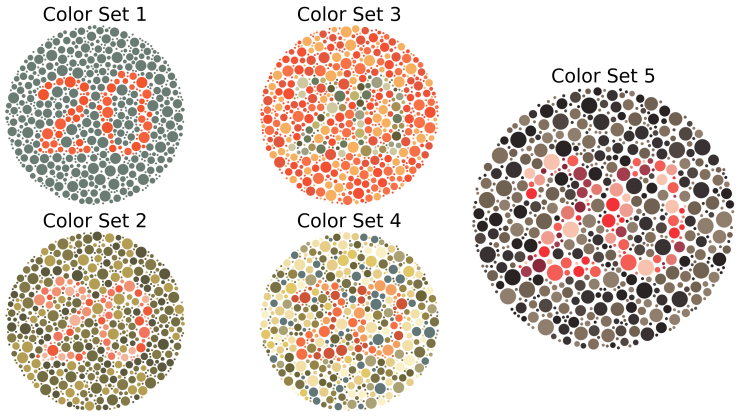

Figure 1: Example Ishihara-like images generated in the benchmark from 5 Color Sets. The number in each image is 20.

Vision-Language Models (VLMs), such as GPT-4o OpenAI (2024), Claude 3 Anthropic (2024), Qwen2-VL Wang et al. (2024), and Llama 3 Vision Dubey et al. (2024), represent a transformative advancement in artificial intelligence by integrating large language models (LLMs) with visual encoders to process multimodal information. These models have demonstrated remarkable abilities in tasks such as visual question answering Antol et al. (2015); Goyal et al. (2017); Hudson & Manning

*Corresponding Author

(2019), image captioning Dong et al. (2024), and multimodal reasoning Li et al. (2024b); Yue et al. (2024), showcasing their potential to understand and interpret complex visual and textual inputs. Nonetheless, an enduring challenge persists: *hallucination* Wang et al. (2023); Liu et al. (2024a). This phenomenon occurs when models produce outputs that are factually incorrect or fail to align with the provided visual or contextual information, raising serious concerns in any setting where reliability is paramount Ao (2023); Zhang et al. (2024a).

While the primary real-world use of the Ishihara test is diagnosing color blindness (e.g., in driving or aviation), the color-based adversarial complexity it introduces can generalize to other high-stakes domains. For instance, medical imaging often relies on color-coded pathology slides for tumor detection Litjens et al. (2017). Moreover, color-based adversarial examples illuminate broader vulnerabilities Zhao et al. (2024), including security contexts where camouflage or intricate patterns can undermine recognition Chen et al. (2024). By examining how VLMs respond to these Ishihara-like stimuli, we gain deeper insight into their robustness and uncover failure modes that may affect a range of visually demanding tasks.

However, recent benchmarks typically emphasize object-level or attribute-level recognition in relatively direct visual settings Liu et al. (2024a); Li et al. (2024a), overlooking whether VLMs can handle visually adversarial scenarios with fine-grained color manipulations. To address this gap, we introduce **ColorBlindnessEval**, a novel benchmark specifically designed to evaluate visual recognition hallucinations in VLMs. Our contributions can be described as (1) we introduce a novel benchmark that adds adversarial visual complexity to assess VLM robustness in numerical recognition tasks; (2) we evaluate state-of-the-art VLMs on ColorBlindnessEval, identifying their strengths and weaknesses in handling challenging visual inputs; and (3) we analyze VLM hallucination behaviors, providing insights into mitigating these issues and enhancing model reliability and safety in real-world applications.

2 COLORBLINDNESSEVAL

2.1 DATASET GENERATION

Our dataset is inspired by the Ishihara Test, a widely used tool for diagnosing red-green color blindness Clark (1924). Each image in the test consists of solid-colored dots arranged in a random pattern of varying color and size. A subset of these dots forms a number or shape that is easily visible to individuals with normal color vision but becomes difficult or impossible to distinguish for those with red-green color deficiencies. A discussion of related work is provided in Appendix C.

To address the insufficiency of the original 38-image Ishihara Test as a benchmark for VLMs, we propose a scalable pipeline to generate 500 Ishihara-like image pairs. Our data generation pipeline comprises three distinct stages. The first stage generated reference images containing numbers ranging from 0 to 99, where the numbers are rendered in black against a white background. Examples are provided in Appendix E. In the second stage, we modified and employed a Monte Carlo-based method to generate plates consisting of circles without any assigned color [1]. Finally, in the third stage, colors were assigned to the circles based on their location in the reference image. Circles within the number region are assigned foreground color(s), while those in the background are given contrasting color(s).

The latter two stages are detailed in Algorithm 1 in Appendix E. where we present our modifications to the original Monte Carlo-based method, which enhance its scalability.

For the foreground colors C_f and background colors C_b, we sampled five distinct sets derived from the images in the Ishihara Test, as detailed in Appendix D.

The final dataset generated using the five distinct color sets[2]. For each color set, we created 100 image pairs, where one image represents a standard Ishihara-like plate and the other represents a foreground-only Ishihara-like plate. This approach ensures a balanced representation of all five color combinations within the dataset.

[1] https://github.com/icfaust/IshiharaMC

[2] The dataset employs the Arial font

2.2 EVALUATION

The evaluation process can be described as follows:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\delta\left(f(\text{Image}_i, \text{Prompt}_i), \text{GroundTruth}_i\right), \tag{1}$$

where $f(\text{Image}_i, \text{Prompt}_i)$ represents the output answer from the Vision-Language Model (VLM) after receiving an input image and a prompt. GroundTruth_i is the true expected answer for the i-th input. $\delta(a, b)$ is an indicator function defined as:

$$\delta(a, b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases}$$

and N is the total number of input samples.

Models	$Y^*/N\uparrow$	$Y/N^*\uparrow$	**Open↑**	**Open-clear↑**
GPT-4o	0.936	0.554	0.308	0.962
GPT-4o-mini	0.780	0.780	**0.344**	**0.966**
Claude3-Haiku	0.008	0.998	0.008	0.124
Claude3.5-Sonnet	0.670	0.328	0.018	0.808
Qwen2-VL-Instruct-2B	0.060	0.998	0.098	0.958
Qwen2-VL-Instruct-7B	0.548	0.540	0.076	0.972
Qwen2-VL-Instruct-72B	0.966	0.076	0.036	0.916
Llama3.2-Vision-Instruct-11B	0.576	0.668	0.030	0.314
Llama3.2-Vision-Instruct-90B	0.108	0.928	0.032	0.510

Table 1: Evaluation results (Accuracy) for VLMs on Y/N Prompt Questions and Open Prompt Questions

3 EXPERIMENTS

3.1 SETUP

We evaluated nine VLMs, including both proprietary and open-source models. The proprietary models included OpenAI's GPT-4o and GPT-4o-mini OpenAI (2024), and Anthropic's Claude-3-Haiku and Claude-3.5-Sonnet Anthropic (2024). The open-source models were Qwen2-VL-Instruct with 2B, 7B, and 72B parameters Wang et al. (2024), and Llama3.2-Vision-Instruct with 11B and 90B parameters Dubey et al. (2024). All models utilized default hyperparameters as specified in their official codebases or API documentation. Proprietary models were accessed via their official APIs, while open-source models were deployed on four H100 GPUs.

We designed Yes / No (Y / N) prompts to evaluate the models under varying conditions, which were applied to ask if a specific number is present in the image, framed as a Yes/No question. To further analyze model behavior, we introduced:

- **Correct Number Prompt** (Y^*/N): The query includes the correct number.
- **Incorrect Number Prompt** (Y/N^*): The query includes an incorrect number.

Open-Ended (Open) Prompts were designed to require the models to identify the number present in the image, framed as an open-ended question. For comparison, we also included an *Open-clear* prompt, presenting the models with a foreground-only image to facilitate number recognition. For detailed evaluation prompts, please refer to Appendix F.

3.2 HUMAN EVALUATION

We conducted an online quiz using a representative sample of images. Two standard images and their corresponding clear images (foreground-only) were randomly selected from each of the five color sets, totaling 20 images.

Twenty participants who do not have color blindness completed the quiz, which comprised 10 Open questions (*Open-C*) based on standard Ishihara-like images and 10 Open-clear questions (*OpenClear-C*) featuring clear images. This design aimed to investigate differences in human accuracy when interpreting standard versus foreground-only images under identical conditions. See Appendix A for details.

4 DISCUSSION

4.1 VLMS OVERALL PERFORMANCE

Table 1 presents the evaluation results of nine state-of-the-art VLMs on our benchmark. Assessing a VLM's ability to recognize numbers requires considering both the Y^*/N and Y/N^* criteria. A high Y^*/N score indicates that the model often agrees with prompts containing correct numbers, while a high Y/N^* score means the model often disagrees with prompts containing incorrect numbers. Some models, such as Claude3-Haiku, Qwen2-VL-Instruct-2B, and Llama3.2-Vision-Instruct-90B, tended to respond "No" regardless of the prompt. We posit that balanced performance across both Y^*/N and Y/N^* indicates better overall capability.

Under the *Open-clear* criterion, most models performed exceptionally well. Notably, GPT-4o, GPT-4o-mini, and Qwen-2-VL-Instruct-72B achieved scores exceeding 96%, demonstrating strong ability to recognize numbers without adversarial backgrounds. However, under the *Open* criterion, performance dropped significantly for most models; for example, Claude3-Haiku provided few correct answers. Among all models, GPT-4o-mini emerged as the most accurate across both Open and Open-clear criteria. Additional experiments and discussions are provided in Appendix B.

4.2 HUMANS VS. VLMS

We compared human and VLMs performance on a calibration subset of the benchmark (Table 2). VLMs performed comparably to humans on the *OpenClear-C* condition, indicating their ability to recognize numbers in clear images composed of colorful dots. However, in the *Open-C* condition, including adversarial backgrounds, humans showed a slight performance drop of approximately 8%, while VLM performance decreased substantially. This highlights that, although humans maintain high proficiency under adversarial conditions, VLMs lag considerably and exhibit tendencies toward hallucination.

4.3 EFFECT OF VLMS' SCALE

The results in Table 1 show no strong correlation between the scale of VLMs within the same model family and performance variance. This suggests that factors other than model size—such as architecture design, fine-tuning strategies, or data quality—may play a more significant role in influencing performance differences.

4.4 DO VLMS HAVE PREFERRED COLOR SETS?

We analyzed the percentage of correct predictions for each color set and their contributions to the overall accuracy of each VLM, as shown in Figure 2. The analysis reveals that VLMs tend to perform better on *ColorSet1* and worst on *ColorSet3*. This suggests that the models are more effective at handling samples with large foreground-background contrasts.

5 CONCLUSION

ColorBlindnessEval provides insights into VLM performance under visually challenging conditions, inspired by Ishihara color blindness tests. The findings highlight significant gaps in the models' ability to handle complex patterns, leading to errors such as hallucinations. This underscores the need for improved training methods and fine-tuned data to enhance VLM reliability, particularly in accuracy-critical applications. By revealing areas of strength and weakness, this work lays the foundation for future innovations aimed at bridging these gaps and building more trustworthy AI systems for real-world use.

REFERENCES

Anthropic. Claude3. https://www.anthropic.com/news/claude-3-family, 2024.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuang Ao. Building safe and reliable ai systems for safety critical tasks with vision-language processing. In *European Conference on Information Retrieval*, pp. 423–428. Springer, 2023.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

JH Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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*, pp. 14375–14385, 2024.

Jeffery K Hovis. Diagnosis of defective colour vision, 2002.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024a.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2042–2051, 2021.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. Multimodal causal reasoning benchmark: Challenging vision large language models to infer causal links between siamese images. *arXiv preprint arXiv:2408.08105*, 2024b.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, dec 2017. doi: 10.1016/j.media.2017.07.005.

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. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Maureen Neitz and Jay Neitz. Molecular genetics of color vision and color vision defects. *Archives of ophthalmology*, 118(5):691–700, 2000.

OpenAI. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

Ahnaf Mozib Samin, M Firoz Ahmed, and Md Mushtaq Shahriyar Rafee. Colorfoil: Investigating color blindness in large vision and language models. *arXiv preprint arXiv:2405.11685*, 2024.

George Wald and Paul K Brown. Human color vision and color blindness. In *Cold Spring Harbor symposia on quantitative biology*, volume 30, pp. 345–361. Cold Spring Harbor Laboratory Press, 1965.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Gang Zhang, Xiaowei Fan, Jingquan Fang, Yanna Sun, Xiayang Shi, and Chunyang Lu. Unveiling vulnerabilities in large vision-language models: The savj jailbreak approach. In *International Conference on Artificial Neural Networks*, pp. 417–434. Springer, 2024a.

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*, 2024b.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

A HUMAN EVALUATION RESULTS

Models	**Open-C↑**	**OpenClear-C↑**
GPT-4o	0.20	0.90
GPT-4o-mini	0.30	0.90
Claude3-Haiku	0.00	0.10
Claude3.5-Sonnet	0.00	1.00
Qwen2-VL-Instruct-2B	0.00	1.00
Qwen2-VL-Instruct-7B	0.10	0.90
Qwen2-VL-Instruct-72B	0.00	1.00
Llama3.2-Vision-Instruct-11B	0.00	0.50
Llama3.2-Vision-Instruct-90B	0.10	0.60
Human	**0.89**	**0.97**

Table 2: The performance of Visual Language Models (VLMs) and humans was assessed using a calibration set comprising 10 selected items for both Open-C and OpenClear-C. For VLMs, performance is reported in terms of accuracy, while for human participants, performance is evaluated by calculating the mean accuracy across the entire participant group.

B ADDITIONAL EXPERIMENTS

This work select Qwen models (Qwen2-VL-Instruct-2B, Qwen2-VL-Instruct-7B and Qwen2-VL-Instruct-72B) in this section to explore additional experiments on performance on clear images, few-shot learning, and varying font styles.

B.1 PERFORMANCE EVALUATION ON THE CLEAR DATASET

Models	$Y^*/N\uparrow$	$Y/N^*\uparrow$
Qwen2-VL-Instruct-2B (Clear)	0.020	1.000
Qwen2-VL-Instruct-7B (Clear)	0.450	0.976
Qwen2-VL-Instruct-72B (Clear)	0.980	0.960

Table 3: Yes/No Performance on the Clear Dataset

By comparing Table 3 with Table 1, we observe that smaller models (e.g., 2B) tend to perform better on "No prompt" questions, whereas larger models (e.g., 72B) achieve higher accuracy on both "Yes prompt" and "No prompt" questions. Notably, Qwen2-VL-Instruct-72B demonstrates excellent performance on Yes-or-No questions. We argue that strong performance on both "Yes" and "No" responses indicates a model's ability to accurately interpret numerical content in the image. In contrast, extremely low performance on either type of question suggests a lack of true understanding or recognition of the visual content.

B.2 CAN FEW-SHOT LEARNING TO IMPROVE MODEL OUTCOMES?

Models	$Y^*/N\uparrow$	$Y/N^*\uparrow$	**Open↑**
Qwen2-VL-Instruct-2B	0.014	0.994	0.076
Qwen2-VL-Instruct-7B	0.006	0.998	0.038
Qwen2-VL-Instruct-72B	0.004	0.998	0.032

Table 4: Few-Shot Learning Performance

In the few-shot learning experiments, for Yes-or-No prompting, we randomly selected two examples under the same color configuration, one corresponding to a "Yes" response and one to a "No" response. For open prompting, a single image was randomly selected under the same color configuration as example. As shown in Table 4, few-shot learning did not lead to performance improvements;

in fact, it had a negative impact across various models and prompting methods compared to the zero-shot setting. These results suggest that the zero-shot approach may be more effective for this task. Nonetheless, there remains significant potential for further exploration, including investigations into sample selection strategies and alternative in-context learning paradigms, which we leave for future research.

B.3 Evaluating Model Performance Across Varying Font Styles

Models	$Y^*/N\uparrow$	$Y/N^*\uparrow$	**Open**$\uparrow$
Qwen2-VL-Instruct-2B	0.004	0.998	0.092
Qwen2-VL-Instruct-2B (Clear)	0.030	1.000	0.952
Qwen2-VL-Instruct-7B	0.012	1.000	0.032
Qwen2-VL-Instruct-7B (Clear)	0.386	0.984	0.704
Qwen2-VL-Instruct-72B	0.602	0.440	0.030
Qwen2-VL-Instruct-72B (Clear)	0.759	0.947	0.876

Table 5: Evaluating Model Performance on DejaVuSans Font

We constructed an additional dataset using the DejaVuSans font, while keeping all other conditions consistent with the experiments presented in Table 1. We observed that the performance trends between the two fonts on both standard Ishihara images and clear images remained consistent: models generally performed better on clear images than on standard ones, and larger models consistently outperformed smaller ones. However, we also noted notable performance differences between fonts. For example, Qwen2-VL-Instruct-7B achieved 70.4% accuracy on DejaVu Sans clear images, compared to 97.2% on Arial clear images.

These results suggest that font choice can have a non-negligible impact on model performance. One possible explanation is that differences in font style may subtly alter the spatial arrangement of colors, thereby affecting model predictions. Future research could explore font-specific visual biases and develop strategies to enhance font-agnostic robustness in vision-language models.

C Background

C.1 Color Blindness Tests

Color blindness tests like the Ishihara test assess color discrimination by presenting patterns of colored dots forming numbers or shapes discernible only to those with normal color vision Clark (1924). Inspired by these tests, recent works like ColorFoil and BlindTest have adapted similar challenges to evaluate VLMs' ability to perceive and interpret color patterns Rahmanzadehgervi et al. (2024); Samin et al. (2024). These studies reveal that VLMs often struggle to distinguish subtle color differences or interpret manipulated visual content, highlighting gaps in robustness and alignment with human perception.

C.2 Vision-Language Models

VLMs such as CLIP Radford et al. (2021), LLaVA Liu et al. (2024b), GPT-4o OpenAI (2024), Claude Anthropic (2024), and Qwen2 Wang et al. (2024) have significantly advanced multi-modal understanding by integrating vision and language processing. These models excel in common tasks like image captioning Dong et al. (2024), visual question answering Antol et al. (2015); Goyal et al. (2017); Hudson & Manning (2019), and image-text retrieval Zhang et al. (2024b). Typically based on transformer architectures, they leverage techniques like contrastive learning and multi-modal attention to align and process image-text data effectively. While these functionalities enable sophisticated tasks such as image-grounded dialogues and high-precision understanding, they also introduce vulnerabilities, as models may rely on learned correlations rather than genuine multi-modal comprehension, raising concerns about robustness Zhang et al. (2024a).

C.3 HALLUCINATIONS BENCHMARKS

Despite advancements, VLMs remain vulnerable to hallucinations—outputs deviating from input data—posing safety risks in critical applications Liu et al. (2024a). They struggle with geometric reasoning, object counting, and complex image interpretation due to deficiencies in low-level vision and over-reliance on parametric memory Rahmanzadehgervi et al. (2024); Guan et al. (2024). This degrades reliability, especially in autonomous systems and medical diagnostics Wang et al. (2023); Li et al. (2023), undermining trust and leading to unsafe decisions, particularly under adversarial attacks Chen et al. (2024); Li et al. (2021); Zhao et al. (2024). Robust evaluation frameworks are needed to mitigate these weaknesses.

Some existed benchmarks address hallucinations by studying biases in training data and object hallucination Guan et al. (2024); Li et al. (2023). While they quantify vulnerabilities Liu et al. (2024a), they focus on object recognition in natural images and may not fully assess model robustness in visually adversarial scenarios inspired by human visual deficiencies.

C.4 THEORETICAL FRAMEWORK

The theoretical framework of our benchmark, ColorBlindnessEval, is centered on emulating the diagnostic specificity inherent in clinical assessments of human color vision deficiencies (CVDs), exemplified by tools like the Ishihara plates. Human color vision, theoretically understood through frameworks like trichromacy (Young-Helmholtz) and opponent processing (Hering), relies on specific photoreceptor sensitivities (L, M, S cones) and subsequent neural pathways that give rise to perceptual phenomena like color confusion lines in individuals with CVDs Wald & Brown (1965). These tests are rigorously designed based on the physiological principles of human trichromatic vision and its most common hereditary forms Neitz & Neitz (2000), they leverage the principle of pseudo-isochromatic stimuli: color combinations are meticulously selected according to precise colorimetric principles to fall along established color confusion lines for specific CVD types—primarily protan (L-cone related) and deutan (M-cone related) deficiencies affecting red-green perception—rendering the embedded figures difficult or impossible to distinguish for affected individuals Hovis (2002).

By adopting color schemes derived from this established and theoretically grounded methodology, ColorBlindnessEval moves beyond presenting arbitrarily difficult colors. Instead, it specifically challenges VLMs along these well-defined critical axes of human color perception failure. This design enables a more targeted evaluation: examining VLM failure patterns on these theoretically motivated stimuli can reveal potential analogous vulnerabilities or biases within their internal color representation and processing mechanisms. Consequently, this approach offers deeper, more structured insights into VLM robustness compared to evaluations using generic visual challenges. It allows us to pinpoint where the highly optimized, statistically driven pattern recognition of AI diverges from the evolved, neurobiologically constrained perception of humans. This comparative understanding is crucial for the responsible development and deployment of AI, particularly in domains requiring nuanced visual interpretation, such as clinical settings. Knowing how and why AI perception differs from human perception is essential for defining the operational boundaries of AI systems and ensuring they augment, rather than compromise, human expertise.

D DETAILS ON DIFFERENT COLOR SETS

D.1 COLORSET DEFINITION

Based on the Ishihara Test Clark (1924), we sampled colors directly from its numeral plates to ensure consistency with the visual principles of the original test. The following five background-foreground color sets were selected for use in the benchmark, the colors are represented as RGB values.

COLOR SET 1

- Background: (106, 124, 115)
- Foreground: (245, 97, 60), (242, 85, 45)

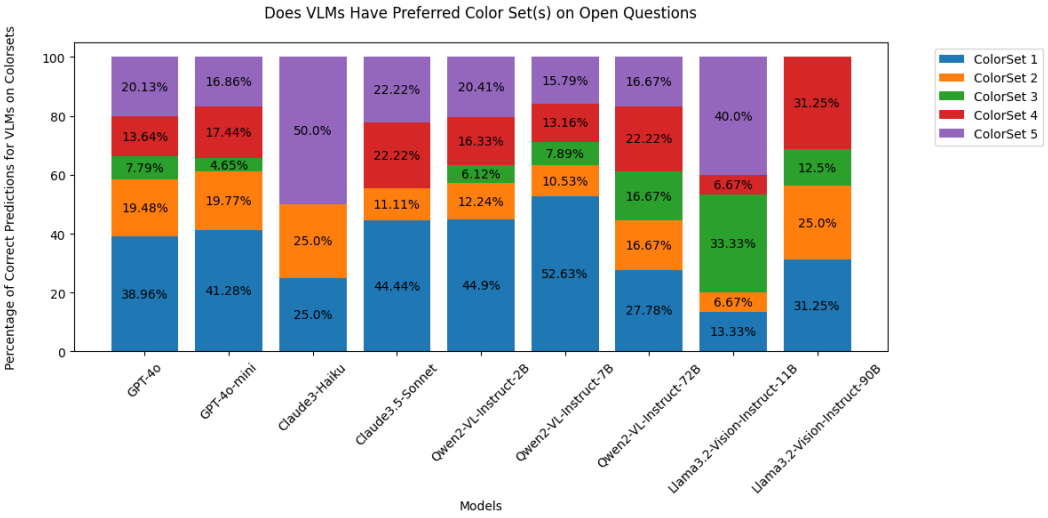

Figure 2: VLM's Performance on Different Color Sets

COLOR SET 2

- Background: (180, 158, 83), (91, 88, 62), (132, 123, 73), (115, 109, 66)
- Foreground: (238, 91, 59), (242, 180, 154), (240, 146, 114), (242, 118, 94)

COLOR SET 3

- Background: (248, 175, 96), (249, 113, 71), (244, 80, 51), (228, 87, 62)
- Foreground: (192, 179, 108), (107, 122, 91), (207, 201, 161), (99, 93, 56), (167, 144, 84), (158, 159, 131)

COLOR SET 4

- Background: (226, 199, 102), (108, 101, 56), (250, 241, 199), (122, 114, 70), (148, 132, 69), (170, 161, 117), (242, 224, 167), (230, 205, 136), (98, 119, 120)
- Foreground: (244, 160, 96), (245, 112, 66), (206, 84, 55)

COLOR SET 5

- Background: (130, 112, 94), (57, 50, 51), (80, 70, 66), (41, 35, 35), (113, 98, 82), (144, 127, 110)
- Foreground: (244, 94, 86), (243, 50, 55), (137, 41, 60), (163, 62, 78), (228, 123, 113), (239, 157, 144), (248, 195, 175)

These color sets were chosen to provide a diverse range of contrasts and challenges for evaluating Visual Language Models (VLMs) and human participants, replicating the nuanced visual conditions found in the Ishihara Test.

E DATA GENERATION

E.1 EXAMPLE REFERENCE IMAGES

We present example reference images in Figure 3.

Figure 3: Example reference images used for guiding the following color assign to foreground and background

Algorithm 1 Generation of Ishihara-Plate-like images

Require: Reference image I, foreground colors C_f, background colors C_b
Ensure: Ishihara plate P
  Initialize empty plate $P \leftarrow \emptyset$
  // Generate circles in multiple layers
  Max layer number L_{max}
  for $i = 1$ to L_{max} **do**
    Set number of circles n_i for layer i
    Set size range $[s_{min}, s_{max}]_i$ for layer i
    for $j = 1$ to n_i **do**
      valid $\leftarrow$ false
      while not valid and attempts $<$ max_attempts **do**
        Generate random (x, y) within unit circle
        Generate random radius $r \in [s_{min}, s_{max}]_i$
        if (x, y, r) does not overlap existing circles **then**
          valid $\leftarrow$ true
          $P \leftarrow P \cup \{(x, y, r)\}$
        end if
      end while
    end for
  end for
  // Assign colors based on reference image
  for each circle (x, y, r) in P **do**
    Map (x, y) to image coordinates (px, py)
    if $I(px, py) < threshold$ **then**
      Assign random color from C_f to circle
    else
      Assign random color from C_b to circle
    end if
  end for
  return P

E.2 Data Generation Algorithm

We present data generation algorithm in Algorithm 1.

F EVALUATION PROMPTS

For **Yes/No (Y/N) Prompts**, we use the following prompt:

```
This is an Ishihara plate designed for color blindness testing. The
    number is made up of dots of different colors. What number do you
    see in the image? Output the number you see only.
```

For **Open-Ended (Open) Prompts**, we use the following prompt:

```
This is an Ishihara plate designed for color blindness testing. Is the
     number in the image is <NUM>? Answer 'yes' or 'no'.
```

We replace `<NUM>` with the correct number when using the Correct Number Prompt, and with a random incorrect number between 0 and 99 when using the Incorrect Number Prompt.