This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Evaluating Model Perception of Color Illusions in Photorealistic Scenes

Lingjun Mao Zineng Tang Alane Suhr
University of California, Berkeley
{lingjun, terran, suhr}@berkeley.edu

Abstract

We study the perception of color illusions by vision-language models. Color illusion, where a person's visual system perceives color differently from actual color, is well-studied in human vision. However, it remains underexplored whether vision-language models (VLMs), trained on large-scale human data, exhibit similar perceptual biases when confronted with such color illusions. We propose an automated framework for generating color illusion images, resulting in RCID (Realistic Color Illusion Dataset), a dataset of 19,000 realistic illusion images. Our experiments show that all studied VLMs exhibit perceptual biases similar human vision. Finally, we train a model to distinguish both human perception and actual pixel differences.

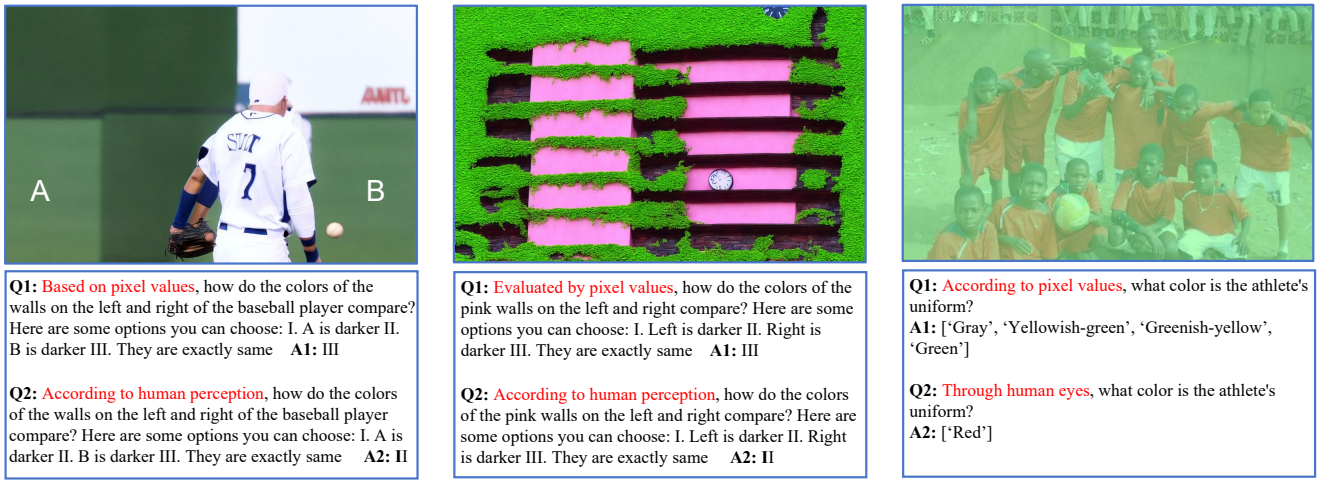
1. Introduction

Human visual perception does not always align perfectly with the real world [4, 12, 17]; we are easily influenced by contextual factors such as lighting conditions or surrounding objects and patterns. Figure 1 shows several examples of color illusions, where human judgment of object color and brightness does not align with actual pixel values. Currently, many large vision-language models (VLMs) [9, 22, 23, 39] exhibit behavioral similaritities to human visual perception in various aspects, including object recognition [41, 44], scene understanding [24, 36], and spatial reasoning [5]. But because the human visual system behaves idiosyncratically when encountering these illusions, we aim to study whether these models perceive such illusions in the same way humans do. Prior work [34, 43] has explored this question by sourcing illusion images from the internet. However, this approach has a drawback: the number of illusion images on the web is limited, and most images (e.g., 60% for the IllusionVQA dataset [34]) are well-known examples of these illusions; thus, VLMs have likely memorized humanlike behavioral responses to them. Additionally, the limited scale also restricts the depth and variety of analyses that can be conducted.

To address these limitations, we propose an automatic framework for generating realistic color illusion images, covering many scenarios with illusions that we might genuinely encounter in our real life. We cover three different types of color illusions: *contrast* (Figure 1a), *stripe* (Figure 1b), and *filter* (Figure 1c) illusions. For contrast and stripe illusions, we first generate simple arrangements of shapes exhibiting specific illusions, which are then processed through ControlNet [42] to generate more realistic images. For filter illusions, we directly apply filters to the original MS COCO 2017 images. In the end, we produce a dataset consisting of total 19,000 images, half of which include an illusion, and the other half being a control set without illusions. Each image is paired with a synthetically-generated natural language question, asking either about color differences between specific regions or directly inquiring about the color of a particular object in the image.

We test recent VLMs using our generated illusion images. We find that, after appropriate fine-tuning, they perform well on non-illusion images, but their accuracy drops significantly on illusion images. Some of their responses exhibit visual biases similar to those of humans, while some are entirely incorrect, failing to align with either pixel values or human perception. We also observe that external prompts, such as Chain of Thought (CoT) [40] and few-shot examples, can influence VLM performance to some extent. Furthermore, fine-tuning on illusion images enhances the models' ability to understand these color illusions. Based on this, we propose a foundational baseline method that enables models to capture both human perceptual biases and true pixel values. Finally, we investigate the underlying causes of VLM susceptibility to color illusions, including biases from both the visual system and prior knowledge.

In conclusion, our contributions are as follows: (1) We propose an automated framework for generating realistic illusion images and create a large, realistic dataset of color illusion images, named **RCID** (Realistic Color Illusion Dataset), to enhance the fairness and accuracy of model testing. (2) We investigate underlying mechanisms of color illusions in VLMs, highlighting the combined influence of the vision system and prior knowledge. We also explore how external prompts and instruction tuning



(a) Contrast Illusion (b) Stripe Illusion (c) Filter Illusion

Figure 1. **Examples of color illusions.** (a) Contrast illusion: The two green squares on the left and right are identical in color, but most people perceive the square on the right as darker due to the brighter background. (b) Stripe illusion: The left and right walls seem to be different colors, but this is an illusion caused by the dark stripes interfering with our perception. (c) Filter illusion: The player's uniform appears red, but in reality, there are no red pixels present.

impact the models' performance on these illusions. (3) We propose a simple training method that enables models to understand human perception while also recognizing the actual pixel values. Our code and data are released under an open-source license at the following URL: https://github.com/mao1207/RCID.

2. Related Work

Visual illusion is a phenomenon where human perception differs from the actual environment [4, 12, 17], often triggered by specific visual stimuli [7]. These illusions offer valuable insights into human perception [16, 31] and have been widely used in artworks. [13, 25]. Many studies [28, 29] have attempted to explore the causes of visual illusions, with the *bottom-up* and *top-down* theories being the most widely accepted explanations. Proponents [33] of the *bottom-up* theory argue that low-level visual signals, such as contrast, brightness, edges, and motion, are correlated with human perception of illusions. However, supporters of the *top-down* theory contend that prior knowledge is also a major factor in these illusions. Evidence shows that people in industrialized societies are more susceptible to the Müller-Lyer illusion due to being misled by implied arrow meanings, and individuals from language communities with fewer color terms are less affected by color illusions [1, 30].

Motivated by illusion phenomena in human visual perception, recent research has begun to explore whether models are also susceptible to such illusions. Early research [14, 15] suggests that traditional vision models, such as CNNs, are influenced by visual illusions in a way similar to humans. More recent studies [34, 43] have shown that VLMs also exhibit similar perceptual biases. However, images on which these models have been tested are either sourced from the internet or manual compositions of simple shapes (e.g., two squares of different colors), making it easy for the VLM to infer results based on their knowledge rather than actual observation. Additionally, the number of these images is quite limited. To address these limitations, we develop a large-scale dataset that embeds visual illusions into realistic scenes, enabling robust evaluation in complex, real-world contexts. Our study fills a gap by investigating the mechanisms behind VLMs' susceptibility to illusions.

3. Illusion Dataset Construction

The construction of our dataset involves three steps (Figure 2): (1) **Image Generation.** For contrast and stripe illusions, we use procedural code to generate simple illusion images, which are then processed by ControlNet to create realistic illusion images. For filter illusions, we directly apply contrasting color filters to the original images. Each type of illusion also includes a corresponding control group without any illusions for comparison. (2) **Question Generation.** We use GPT-4o to generate image-specific questions that are designed to evaluate the model's understanding of the illusion. (3) **Human Feedback.** We collect human participants' feedback on these images and adjust the original classification of "illusion" and "non-illusion" based on whether participants are deceived.

3.1. Contrast and Stripe Illusion Image Generation

ControlNet Training Stage: We begin by training a ControlNet to map from simple illusion images to photoreal-

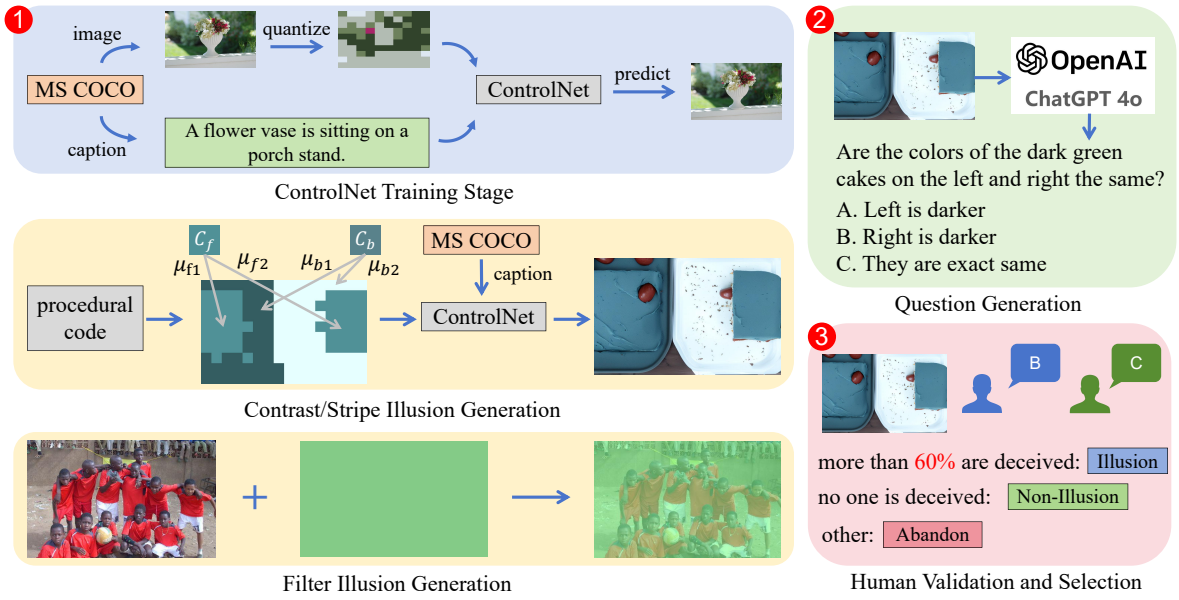


Figure 2. Process for generating our dataset.

istic images with illusions. We train using modified image-caption pairs (I, T) from the MS COCO 2017 dataset. First, we quantize each original image I into a 10×10 grid G, where each cell $G_{x,y}$ represents the average color of its corresponding region:

$$G_{x,y} = \frac{1}{|R_{x,y}|} \sum_{p \in R_{x,y}} I(p)$$

Here, $R_{x,y}$ denotes the set of pixels in the (x, y)-th grid cell, and $I(p)$ is the color value of pixel p in the original image.

During training, Gaussian noise is added to the quantized image $x_0 = G$, producing a noisy image x_t at time step t. The model, conditioned on the text prompt T and color distribution G, is trained to predict the added noise ϵ_θ by minimizing the expected squared error:

$$\mathcal{L} = \mathbb{E}_{x_0,t,C,\epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(x_t, t, T, G)\|_2^2 \right]$$

Through this loss, the model learns to progressively refine x_t into a realistic image during inference, guided by color distribution and the text prompt.

Procedural Generation Stage: We use a procedural function f to generate simple illusion images, denoted as S, which consist of basic shapes and colors, on which we will condition the trained ControlNet model.

For *Contrast Illusions*, we randomly select two base colors from RGB space: C_b for the background and C_f for the foreground (squares). For a color $C = (r, g, b)$, we use the function p to adjust its brightness by $p(C, \mu) = (r \cdot \mu, g \cdot \mu, b \cdot \mu)$, where μ scales each RGB component. The background is divided into a darker region $p(C_b, \mu_{b1})$ and a brighter region $p(C_b, \mu_{b2})$, where $\mu_{b1} < 1$ and $\mu_{b2} > 1$. Additionally, two foreground squares, colored $p(C_f, \mu_{f1})$ and $p(C_f, \mu_{f2})$, are symmetrically placed at random positions (x_1, y_1) and (x_2, y_2) on the background.

If human perception differs from the actual pixel values, the image is classified as an illusion; otherwise, it belongs to the control group (no illusion). We observe that a darker background enhances the perceived brightness of the foreground color, with this effect being strongest when C_b and C_f are similar and weakening as the contrast increases. Section 4.2.3 validates this conclusion. Based on this principle, we generate theoretical illusion and non-illusion images, which are later adjusted through human validation. To increase image diversity, we apply random noise η to introduce irregularities at the square edges or make the boundaries less smooth. The generated image S_i for the contrast illusion can be represented as:

$$S = f\big(p(C_b, \mu_{b1}), p(C_b, \mu_{b2}), p(C_f, \mu_{f1}), p(C_f, \mu_{f2}), (x_1, y_1), (x_2, y_2), \eta\big)$$

Similarly, for *Stripe Illusions*, we randomly generate the background color C_b and the stripe color C_s. The colored stripes alternate with black stripes, and their colors are adjusted to $p(C_s, \mu_{s1})$ and $p(C_s, \mu_{s2})$. Based on an initial estimate of human perception of μ_{s1} and μ_{s2}, the images are classified as either *illusion* or *non-illusion*. The generated image S for the stripe illusion is represented as:

$$S = f\left(C_b, p(C_s, \mu_{s1}), p(C_s, \mu_{s2}), \theta, N, \eta\right)$$

Here the stripe direction θ (horizontal, vertical, or diagonal) and the number of stripes N are randomly selected. Random noise η is applied to create irregularities, such as curved stripe arrangements or slight misalignments.

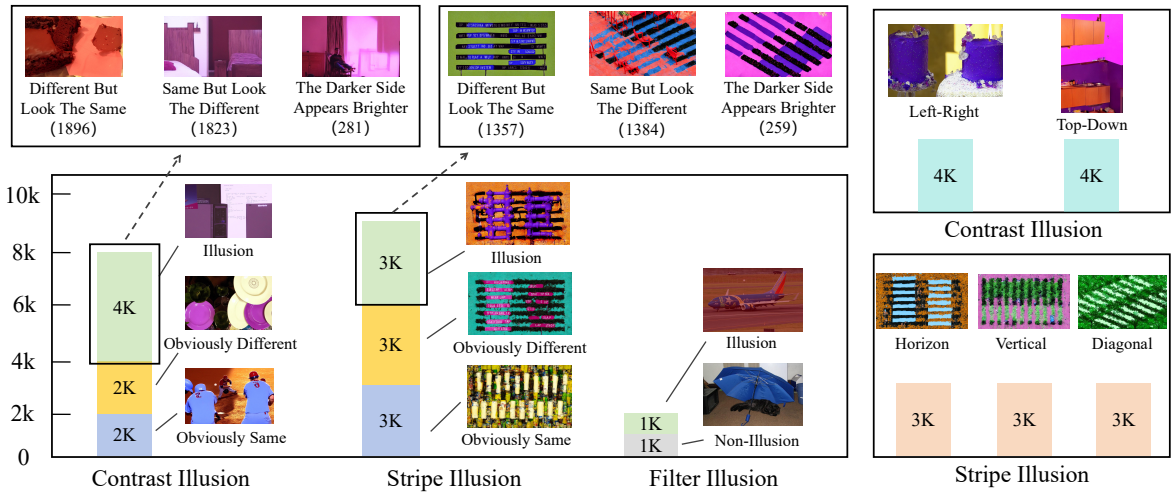


Figure 3. Data statistics of RCID (Realistic Color Illusion Dataset).

ControlNet Generating Stage: We pair each simple illusion image S with a randomly selected caption T from the MS COCO 2017 dataset, which together serve as input to our model. Starting from the quantized image $x_0 = S$, the model recursively denoises over multiple time steps following a diffusion-based approach, gradually refining the image to produce a realistic output I' that retains the intended illusion effects. The final image I' is obtained by summing the contributions from all time steps T:

$$I' = \sum_{t=0}^{T} (x_t + \epsilon_\theta(x_t, t, C, S))$$

3.2. Filter Illusion Image Generation

To generate filter illusion images, we first select images from MS COCO 2017 that predominantly contain a specific color C_d (e.g. red, blue, yellow). We then apply a contrasting color filter C_f which overlays the original image in the HSV (Hue, Saturation, Value) color space to suppress C_d by shifting the hue H of filter and ensuring no pixels fall within C_d's HSV range. Despite this, the filtered regions may still visually evoke the original color.

3.3. Question Generation

For each image I', we use GPT-4o to generate an image-grounded question $\mathcal{Q}$ that asks human respondents to compare colors between two regions or identify the color of a specific object. For example, in Figure 1a, the generated questions refer to *the walls on the left and right of the baseball player*. We then prefix the question with two distinct prompts: one focused on "human perception" (e.g., "Based on human perception", "Through human eye") and the other on "pixel values", each associated with a different answer depending on whether the image contains an illusion. For contrast and stripe illusions, answer options are appended to the question. Examples are shown in Figure 1.

3.4. Human Validation and Selection

To ensure that our generated images truly contain illusions, we conduct a survey on Prolific with a total of 241 participants. Each participant is shown 400 images, including both illusion and non-illusion examples. We use respondents' answers as the human perception ground truth. For every image in our dataset, we assign 5 participants. If more than 3 participants are deceived by the illusion, the image is classified as an *illusion* image. Conversely, if all 5 participants provide answers consistent with the pixel values, the image is classified as *non-illusion* and added to the control group. On average, we pay $21 USD per hour.

3.5. Dataset Details

Realistic Color Illusion Dataset (RCID, Figure 3) contains 19,000 images, comprising 8,000 contrast, 9,000 stripe, and 2,000 filter images. We divide our dataset into 9,500 images for training, 4,750 for development, and the remaining 4,750 for the test set. Each illusion type has distinct structures and subcategories.

4. Evaluating Vision-Language Models

4.1. Experimental Setup

In this experiment, we evaluate different models' performance on color illusions and explore the impact of external prompts and instruction-tuning on them. We also provide a foundational baseline model based on LLaVA-1.5 (7B) for understanding color illusions.

Vision-Language Models. We test recent models with our generated illusion dataset to examine whether VLMs exhibit perceptual responses similar to humans. To ensure

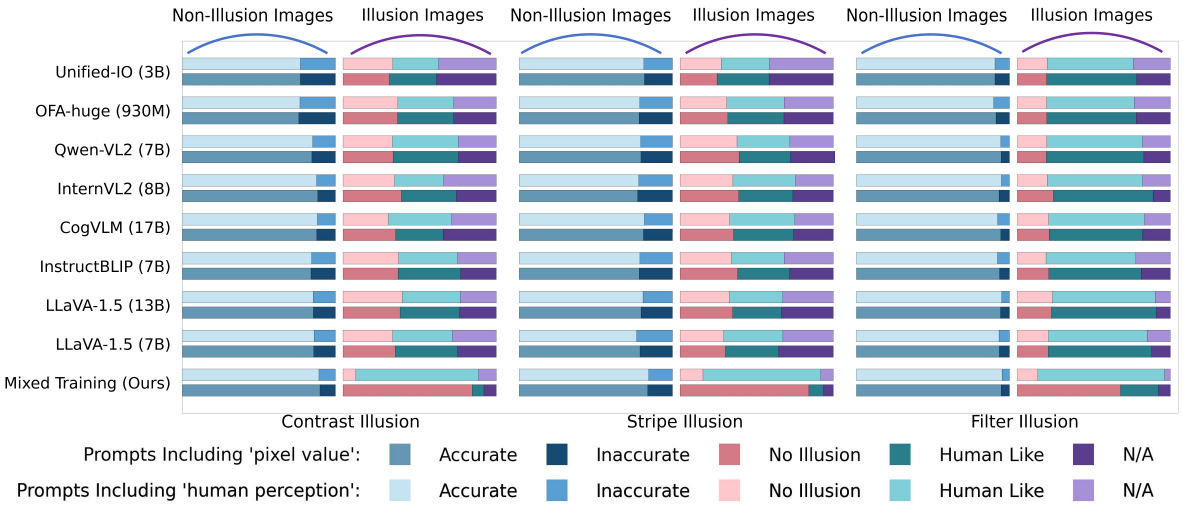


Figure 4. This figure shows the proportion of different model responses across three types of illusions (Contrast, Filter, and Stripe) on our development set. For non-illusion images, we report the proportions of "Accurate" and "Wrong" responses. For illusion images, we categorize responses as "No Illusion" (consistent with pixel values), "Human like," and "N/A." Each image is evaluated using two types of prompts: one based on pixel values and the other based on human perception.

that their performance on illusion images reflects actual susceptibility to illusions rather than challenges in identifying specified regions and comparing color brightness [11, 37], we focus on open-source models, and fine-tune them on our generated non-illusion simple images and a small set of realistic images to enhance their ability in answering questions about relative colors of different regions in an image. During training, we randomly apply prompts such as 'based on human perception,' 'based on pixel values,' or provide no specific guidance prompt at all. Our models tested include LLaVA [22], InstructBLIP [9], CogVLM [39], InternVL2 [6], Qwen-VL2 [2], OFA [38] and Unified-IO [26]. During inference for contrast and stripe illusions, questions are presented in a multiple-choice format, while filter illusion questions require open-ended responses about the color of specific objects in the image, as illustrated in Figure 1. We also include experiments with closed-source models like GPT-4o and Gemini in the appendix. While their overall color comparison ability is weak, there is still a significant drop in performance on illusion images.

Metrics. For images containing illusions, we use the same metrics as in GVIL [43]. We track the *Human-like* rate, which measures alignment between human perception and VLM responses by calculating the percentage of cases where the model's responses match human answers. When the prompt explicitly includes "Based on pixel values," this can also be referred to as *Deception Rate*. The *No-Illusion* rate indicates the proportion of examples where the model's responses align with actual pixel values. If the model's answer aligns with neither the human response nor the pixel values, we categorize those instances as *Not Applicable (N/A)*. For images without illusions, where human perception matches pixel values, we classify model responses as "accurate" or "inaccurate".

4.2. Results and Analysis

We evaluate a range of open-source vision-language models using generated illusion and non-illusion images in our development set, with questions explicitly asking for color judgments 'Based on pixel values' or 'Based on human perception'. The results show that, after fine-tuning on non-illusion images, these models achieve high accuracy (>75%) on non-illusion images, while their accuracy on illusion images is significantly lower. We find that explicitly querying for color judgments 'Based on pixel values' or 'Based on human perception' does not lead to significant changes in model performance. In addition, models are likely to produce responses that are completely inaccurate, matching neither pixel values nor human perception. This suggests that while the models are misled by color illusions to some extent, they still struggle to fully model human perception. Detailed results can be found in Figure 4.

Inter-Annotator Agreement. We refine our final set of illusion images through human validation, retaining only those that deceive at least 3 out of 5 human participants. Before this selection, the average Fleiss' kappa score among the five respondents on the illusion images is 0.648, which increases to 0.806 after filtering. 21.6% of the selected images deceive all five participants. We use this stricter subset of illusion images as a challenging test set for our models. The results show a slight increase (<5%) in the proportion of Human-Like responses by the VLMs across all three types of illusions, but overall, the difference remains

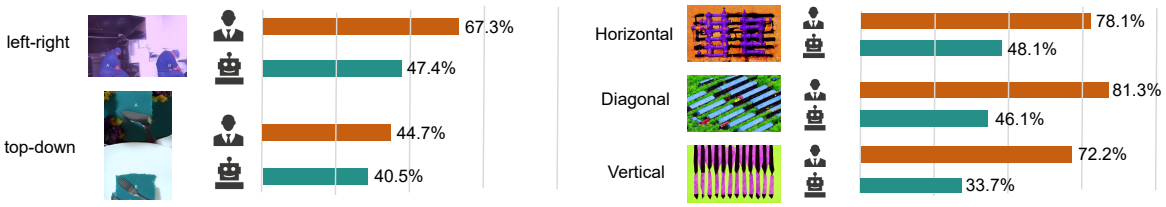


Figure 5. Deception rates of humans and VLMs across different structural patterns.

minimal, indicating that our filtering criterion (more than 3 humans are deceived) is robust.

Generated Image Realism. As some generated illusion images appear realistic while others still exhibit an "AI-looking" appearance, we assess LLaVA-7B and two human participants on three 500-image sets with varying realism levels. Our results show no significant correlation between realism and accuracy, with $r = 0.084$ for human participants and $r = 0.091$ for the model.

Model Size. To investigate the effect of model size, we evaluate the performance of two groups of models, OFA and Unified-IO. These groups consist of models of varying sizes but share the same overall architecture. We conduct the experiment on contrast illusions, and explicitly emphasize "Based on pixel values" in the questions. Our result shows, with the model size increases, the proportion of human-like responses also increases, while the proportion of responses consistent with actual pixel values decreases. We present OFA's performance in Figure 6 and provide Unified-IO's performance details in the appendix.

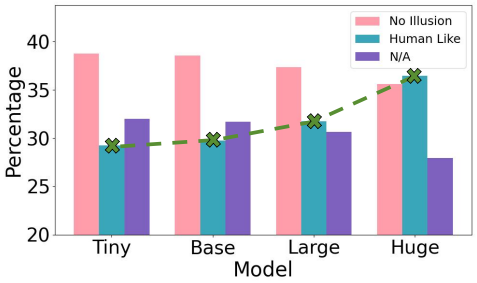


Figure 6. Proportions of 'No Illusion,' 'Human Like,' and 'N/A' responses for OFA models of different sizes on contrast illusions.

Illusion Design. We explore a range of visual factors that may influence the strength of color illusions and compare whether these factors affect human perception and VLMs in the same way. We focus on three potential influencing factors: the orientation of the illusion, the contrast between foreground and background colors (for contrast illusions only), and the number of stripes (for filter illusions only). Overall, our findings indicate that these factors significantly impact the strength of the illusion. For example, altering certain factors, such as increasing the color contrast between the foreground and background, can turn a non-illusion image into an illusion image. And these effects are consistent across both humans and VLMs. In these experiments, we use LLaVA-1.5 (7B) as our tested model.

Illusion orientation. We randomly generate a set of images with various structural arrangements to analyze how the location of the illusion affects humans and VLMs. For contrast illusions, the images feature left-right and up-down arrangements, while for stripe illusions, they include horizontal, vertical, and diagonal stripes (Figure 3). We evaluate the error rates of both humans and VLMs across these structures. As illustrated in Figure 5, the results show that for contrast illusions, both humans and VLMs are more frequently deceived by left-right arrangements, while for stripe illusions, they are more often misled by horizontal and diagonal stripes. We find that humans can quickly and accurately perceive color differences between left and right, but are slightly less sensitive to top-bottom differences.

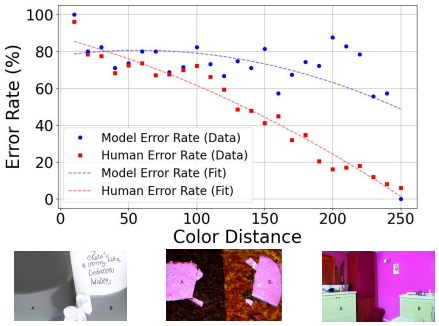
Color distance between the foreground and background. For contrast illusions, we calculate the correlation between the color distance in RGB space between the foreground and background, and the illusion's intensity. Figure 7a shows that as the foreground and background colors become more similar, the contrast illusion becomes most pronounced for both humans and VLMs. Conversely, when the foreground and background colors are distinctly different, humans and VLMs are better at distinguishing them.

Image spatial frequency. For stripe illusions, we examine the effect of image spatial frequency on the strength of the illusion. After converting the image to the frequency domain using the Fourier transform, we compute spatial frequency as the ratio of high-frequency energy to the total image energy [20]. Our findings suggest that the stripe illusion becomes stronger for both humans and models as image frequency increases, which typically corresponds to a higher number of stripes, as shown in Figure 7b.

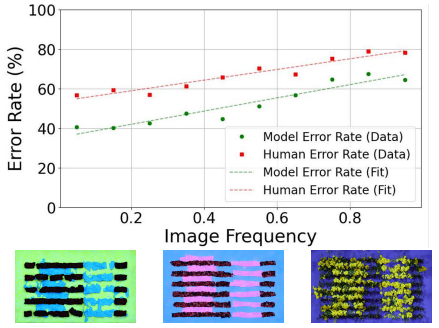
Prompting Methods. We explore whether external prompts could alter VLM responses. In contrast to our main experiments shown in Figure 4, here we aim to identify the model's bias towards predicting answers which are human-like or based on pixel values without specifying this in the question posed to the model. We consider several prompting methods: a simple prompt that only contains questions about color comparisons; chain of thought (CoT), where the model is prompted to first consider if any factors in the image might affect color perception; emphasizing the illusion

Illusion Category	Contrast Illusion		Filter Illusion		Stripe Illusion	
Prompts	**Pixel-Based**	**Human-Based**	**Pixel-Based**	**Human-Based**	**Pixel-Based**	**Human-Based**
LLaVA-1.5 (7B)	35.4 (41.2)	32.4 (37.4)	20.0 (67.8)	19.8 (64.6)	30.3 (35.2)	28.3 (40.1)
Mixed Training (Ours)	83.9 (6.9)	8.1 (79.0)	68.0 (24.3)	12.1 (82.3)	84.5 (9.8)	15.4 (76.7)

Table 1. Model's improved understanding of illusions with mixed training. Two prompts are used: "based on pixel value" and "based on human perception." The values represent the proportion of *no-illusion* responses, with parentheses indicating *Human-like* rates.



(a) Influence of background–foreground color difference on intensity of contrast illusion.



(b) Influence of image spatial frequency on intensity of stripe illusions.

Figure 7. Error rates of humans and LLaVA across different conditions in color illusions.

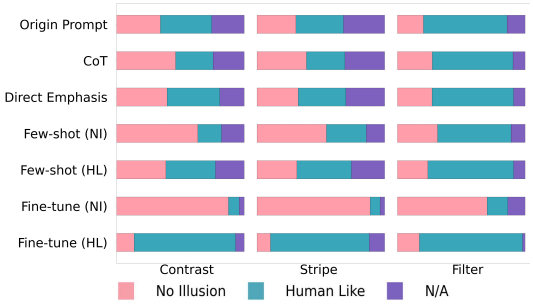


Figure 8. Effect of different prompts and fine-tuning on VLM response preferences across illusion Types. "NI" denotes datasets using no-illusion examples, while "HL" represents datasets with human-like examples.

by directly informing the model that the image contains an illusion without specifying its type; and providing few-shot examples, using no-illusion (NI) or human-like (HL) answers as 3-shot examples. Because we use the LLaVA-7B model here, which doesn't support multi-image input, we combine the few-shot examples and the query image into a single composite image before feeding it to the model [22].

Figure 8 show that external prompts have minimal effect on filter illusions. But for contrast and stripe illusions, the effect is more pronounced, particularly when no-illusion answers are used as few-shot examples, increasing the proportion of no-illusion responses by approximately 25%. However, using human-like responses as few-shot examples does not effectively enhance the model's reproduction of human illusion perception, the proportion of human-like responses even decreases slightly for contrast illusions. Additionally, both direct emphasis and CoT prompts partially increase the proportion of no-illusion responses, although the effect of direct emphasis is relatively weak.

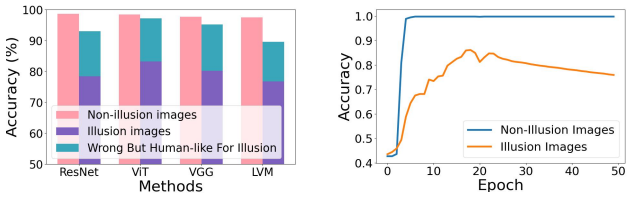
Fine-Tuning. First, we fine-tune two separate LLaVA-7B models on illusion images, with one model trained with responses according to pixel values (NI), and the other trained on human-like responses (HL). The bottom two rows of Figure 8 show a significant effect on the test set, particularly when fine-tuning on no-illusion responses. For both contrast and stripe illusions, fine-tuning on no-illusion responses increases the proportion of no-illusion answers to over 90%. In contrast, fine-tuning on human-like responses has a weaker effect, suggesting that even with human-like responses, the model struggles to fully grasp human perception in the context of illusion images.

Next, we experiment with a mixed training approach that combines both types of responses. We train on question-answer pairs for both types of responses, where questions asking "Based on pixel values" are paired with pixel-value responses, those asking "Based on human perception" are paired with a human-like response. Table 1 shows the influence of this mixed training on the test set, compared to the base model. These results confirm that training model can effectively distinguish between pixel-based and human perception cues, yielding responses aligned with user's query.

5. Why Color Illusions Affect VLMs

5.1. Perceptual Biases from the Visual System

To investigate whether VLMs' perceptual biases originate from their visual components, we test a range of purely visual models, including ResNet [18], ViT [10], VGG [35],



(a) Performance of purely vision models on non-illusion and illusion images.

(b) Accuracy of ViT on non-illusion and illusion images over each training epochs.

Figure 9. Perceptual biases in purely vision models.

MobileNet [19] and Large Vision Model (LVM) [3]. We evaluate these models on both simple, rectangle-based contrast illusion images and generated realistic illusion images. For simple illusion images, we generate backgrounds with varying brightness on both sides and place a rectangle on each side. The vision model is trained as a three-way classifier on 6,000 non-illusion images to predict which rectangle appears darker, or if their colors are identical, stopping when the loss shows no significant decrease. It is then tested on 1,000 non-illusion and 1,000 illusion images. The results are shown in Figure 9a. For realistic illusion images, we train the model on the training set of our generated dataset and evaluate it on the test set, shown in the appendix. While overall accuracy (the proportion of classifications aligning with actual pixel values) on realistic illusion images drops compared to simple rectangle-based images, in both cases, accuracy on illusion images remains significantly lower than on non-illusion images. We also observe that during training, accuracy on non-illusion images quickly improves, while accuracy on illusion images fluctuates significantly. We use ViT as an example in Figure 9b. These results suggests that even purely vision-based models can be influenced by color illusions.

5.2. Perceptual Biases from Prior Knowledge

Influence of Language. We test the multilingual vision-language model PALO [32] on contrast illusions, posing questions in English, Chinese, and Russian to analyze error rates on images containing target regions across several colors. We hypothesize that colors with more frequent textual modifiers (e.g., *dark blue*, *sky blue*) lead to responses aligned with pixel values rather than human perception. To validate this, we first analyze color descriptor diversity in the training data. For example, in English, 50.7% of *red* instances include modifiers like *dark red* or *bright red*, while 49.3% use *red* alone. Figure 10 shows the correlation between color descriptor diversity and pixel-value accuracy for English, Chinese, and Russian.[1] We find that richer descriptions improve model sensitivity to that color, reducing susceptibility to color illusions in that color range.

[1]In Russian, blue has 100% descriptor diversity due to distinct terms for *light blue* (*goluboj*) and *deep blue* (*sinij*).

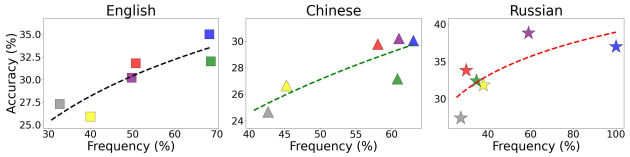


Figure 10. Comparison of color term accuracy and frequency across languages for the vision-language model. Each shape and color represents a specific color term (e.g., blue, red, gray).

Influence of Commonsense Knowledge. Humans are often deceived by visual illusions due to our prior knowledge. For instance, research suggests that in filter illusions, humans perceive colors that aren't actually present due to a phenomenon called *chromatic adaptation* [27]. This allows us to use surrounding context to infer the presence of a filter, mentally "removing" it to perceive the original color. One example is that we always recognize a piece of white paper as white, regardless of the lighting conditions.

To examine whether commonsense knowledge similarly influences VLMs' susceptibility to illusions, we categorized filter illusion images into three groups: single objects, multiple-object compositions, and complex real-world scenes, then analyzed the deception rates for both humans and VLMs across these groups. On non-illusion images, VLMs perform well across all three categories, suggesting that scene complexity does not affect their basic color recognition ability. However, for illusion images, our results in Figure 11 show that both humans and VLMs are less likely to be deceived by single-object images, while complex scenes lead to the highest deception rates, underscoring the role of prior knowledge in visual illusions.

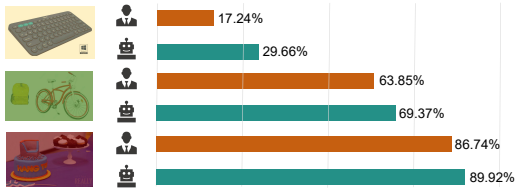


Figure 11. Deception rates of VLMs and human participants across different types of filter illusions.

6. Conclusion

Our experiments show that current VLMs, influenced by their visual systems and prior knowledge, can exhibit visual illusions similar to humans. This raises a key question: as models become more human-like, might they also inherit human perceptual biases? In fields like biomedicine [8, 21], such biases could result in flawed visual judgments, while in human-centric tasks like image generation, understanding human perception can benefit. Therefore, VLM behavior should be tailored to application needs, with careful consideration of human-like perceptual biases.

References

[1] Adam Alter. Drunk tank pink: And other unexpected forces that shape how we think, feel, and behave. 2014. 2

[2] 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. 5

[3] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *CVPR*, 2024. 8

[4] Claus-Christian Carbon. Understanding human perception by human-made illusions. *Frontiers in human neuroscience*, 2014. 1, 2

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *CVPR*, 2024. 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CVPR*, 2024. 5

[7] Fan L Cheng, Tomoyasu Horikawa, Kei Majima, Misato Tanaka, Mohamed Abdelhack, Shuntaro C Aoki, Jin Hirano, and Yukiyasu Kamitani. Reconstructing visual illusory experiences from human brain activity. *Science Advances*, 2023. 2

[8] Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. Biomedical visual instruction tuning with clinician preference alignment. *NeurIPS*, 2024. 8

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1, 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021. 7

[11] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *ECCV*, 2024. 5

[12] Maurizio Gentilucci, Sergio Chieffi, Elena Daprati, M Cristina Saetti, and Ivan Toni. Visual illusion and action. *Neuropsychologia*, 1996. 1, 2

[13] Ernst Hans Gombrich. Art and illusion: A study in the psychology of pictorial representation-millennium edition. 2023. 2

[14] Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, and Marcelo Bertalmío. Convolutional neural networks can be deceived by visual illusions. *CVPR*, 2019. 2

[15] 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*, 2020. 2

[16] Richard Langton Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1968. 2

[17] Richard L Gregory. Visual illusions. *Scientific American*, 1968. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 7

[19] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 8

[20] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. *CVPR*, 2006. 6

[21] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 2023. 8

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeuRIPS*, 2023. 1, 5, 7

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CVPR*, 2024. 1

[24] Sichao Liu, Jianjing Zhang, Robert X. Gao, Xi Vincent Wang, and Lihui Wang. Vision-language model-driven scene understanding and robotic object manipulation. *IEEE CASE*, 2024. 1

[25] Margaret S Livingstone. Art, illusion and the visual system. *Scientific American*, 1988. 2

[26] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *CVPR*, 2023. 5

[27] D. L. MacAdam. Chromatic adaptation∗. *J. Opt. Soc. Am.*, 1956. 8

[28] O Maddaluno, A Facchin, D Zavagno, N Bolognini, R Daini, et al. Top-down and bottom-up neuromodulation over two different visual illusions. *Atti del 40th European Conference on Visual Perception*, 2018. 2

[29] Glenn E Meyer and Susan Petry. Top-down and bottom-up: The illusory contour as a microcosm of issues in perception. 1987. 2

[30] William L Phillips. Cross-cultural differences in visual perception of color, illusions, depth, and pictures. *Cross-Cultural Psychology: Contemporary Themes and Perspectives*, 2019. 2

[31] Jiang Qiu, Hong Li, Qinglin Zhang, Qiang Liu, and Fenghua Zhang. The müller–lyer illusion seen by the brain: An event-related brain potentials study. *Biological psychology*, 2008. 2

[32] Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. Palo: A large multilingual multimodal language model. *WACV*, 2025. 8

[33] Alireza Saeedi, Kun Wang, Ghazaleh Nikpourian, Andreas Bartels, Nikos K Logothetis, Nelson K Totah, and Masataka Watanabe. Brightness illusions drive a neuronal response in the primary visual cortex under top-down modulation. *Nature Communications*, 2024. 2

[34] 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. 1, 2

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015. 7

[36] Zineng Tang, Lingjun Mao, and Alane Suhr. Grounding language in multi-perspective referential communication. *EMNLP*, 2024. 1

[37] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? exploring the visual shortcomings of multimodal llms. *CVPR*, 2024. 5

[38] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *PMLR*, 2022. 5

[39] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. *NeuRIPS*, 2024. 1, 5

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeuRIPS*, 2022. 1

[41] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2024. 1

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *CVPR*, 2023. 1

[43] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding Visual Illusions in Language: Do vision-language models perceive illusions like humans? *EMNLP*, 2023. 1, 2, 5

[44] Zeliang Zhang, Xin Liang, Mingqian Feng, Susan Liang, and Chenliang Xu. Will the inclusion of generated data amplify bias across generations in future image classification models? *arXiv preprint arXiv:2410.10160*, 2024. 1