

Evaluating Model Perception of Color Illusions in Photorealistic Scenes

Supplementary Material

A. Data

A.1. Prompts for Question Generation

Figures 1, 2, and 3 contain the prompts for generating questions for our three different illusion types.

The Prompt for Generating Questions For Contrast Illusion with GPT-4o

```
messages = [{"role": "system", "content": "" You will be given some images, and your task is to generate a question for each image that prompt participants to compare the colors of two distinct areas marked as A and B in the image.

The presented areas are usually rectangular and symmetrical on both the left and right sides, and their colors are quite similar, so make sure you describe the correct areas. Please note that participants cannot see the labels A and B, so you should guide them to observe the differences without explicitly using specific labels such as 'A' and 'B'. Ensure that the description of the areas in the questions is accurate and unambiguous to facilitate participant orientation, and that the generated questions have practical significance. When describing the areas, avoid ambiguity, be mindful of the area boundaries, and avoid interference from scenes outside the area. You are recommended to use directional terms such as 'upper left corner' and color terms like 'the two green areas on the left and right' to describe the regions. Make sure the listener can easily understand the two regions you are describing. However, avoid mentioning 'lighter' or 'darker' to prevent influencing the listener's judgment."}]

for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample[image]})
    messages.append({"role": "assistant", "content": sample[question]})
```

Figure 1. Prompt design for generating color comparison questions on contrast illusions using GPT-4o.

The Prompt for Generating Questions For Stripe Illusion with GPT-4o

```
messages = [{"role": "system", "content": "" You will be given several images, and your task is to generate a question for each image that prompts participants to compare the colors of two striped regions on the left and right sides of the image. These regions are symmetrical and alternate with black stripes, with colors that are very similar. Please ensure that your description of the regions is accurate and unambiguous to avoid any confusion.

When describing the regions, avoid vague expressions, be mindful of the boundaries, and exclude interference from areas outside the target regions. Use directional terms like "upper left corner" or "lower right side" and descriptive color terms such as "the two green areas on the left and right" to specify the regions clearly. The goal is to make it easy for participants to understand which two areas are being referred to. However, refrain from using terms like "brighter" or "darker" to prevent influencing their judgments."}]

for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample[image]})
    messages.append({"role": "assistant", "content": sample[question]})
```

Figure 2. Prompt design for generating color comparison questions on stripe illusions using GPT-4o.

The Prompt for Generating Questions For Filter Illusion with GPT-4o

```
messages = [{"role": "system", "content": "" You will be given several images, and your task is to generate a question for each image that prompts participants to identify the color of a specific object. First, identify the object in the image that is colored (blue, yellow, red, etc.), and then ask the participant about its color. The question should be direct and clear, but make sure not to reveal the object's color to avoid influencing the participant's judgment."}]

for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample[image]})
    messages.append({"role": "assistant", "content": sample[question]})
```

Figure 3. Prompt design for generating color recognition questions on filter illusions using GPT-4o.

A.2. Human Data Collection

We use Prolific to acquire illusion judgments. Our annotation interface is shown in Figure 4. Before annotators begin the task, we provide recommendations for display screen settings (e.g. scaling size, brightness, resolution) and encourage participants to answer in a relatively dark environment to minimize the impact of external lighting. For contrast illusions, participants can toggle between labeled (i.e., with target regions identified with 'A' and 'B') and unlabeled versions of the image by clicking a button. After every 50 questions completed, the system will enforce a half-minute break to prevent visual fatigue. After human validation and selection, approximately 86% of non-illusion images were retained, while 8% were relabeled as illusion images. For illusion images, about 56% were retained, while 24% were relabeled as non-illusion images.

B. Experiments

B.1. Fine-tuning Details

We conduct all fine-tuning and training on an RTX 6000 using a total of three GPUs. During the pre-training phase, which aims to enhance the model's ability to recognize regions and compare colors, we use 6,000 non-illusion simple images along with our generated realistic non-illusion training set (2,000 for contrast, 3,000 for stripe, and 500 for filter), training for 5 epochs. In subsequent experiments investigating the impact of illusion images during instruction tuning, we add 2,000 contrast illusions, 1,500 stripe illusions, and 500 filter illusions to the training set.

B.2. Impact of Model Size

We additionally provide an example from Unified-IO to demonstrate how model size impacts VLM responses to



Q1. Looking at the two natural rock formations creating arches against the sky, one on the left side with an archway and the other being a standalone pillar on the right side, can you describe the differences in color and shade between these two silhouettes?

They are exact same left is lighter right is lighter

Figure 4. Website interface designed for collecting human responses to color illusion perception tasks.

contrast illusions, shown in Figure 5. As the model size increases, the proportion of human-like responses also grows.

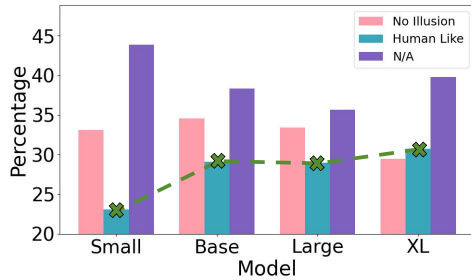


Figure 5. Proportions of ‘No Illusion,’ ‘Human Like,’ and ‘N/A’ responses for Unified-IO models of different sizes on contrast illusion images.

B.3. Performance of Purely Visual Models on Realistic Illusion Images

We also use generated realistic illusion images to test purely vision model for a better comparison with VLMs. We find while their accuracy on non-illusion images is low, their consistency with pixel values still drops on illusion images, shown in Figure 6.

B.4. Impact of Prior Knowledge

Figure 8 illustrates how GPT-4o generates different responses for the same gray color block contrast illusion depending on whether the images are sourced from websites or generated. For website-sourced images, GPT-4o often appears to rely on memorized answers rather than deriving responses based on its own observations. To further

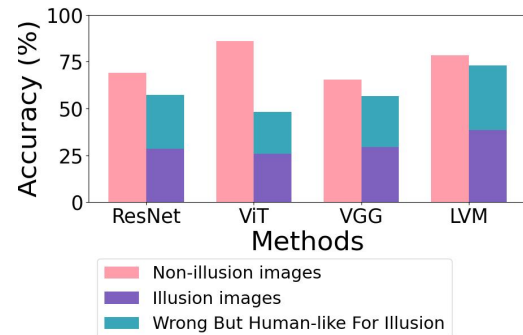


Figure 6. Performance of purely vision models on realistic illusion images.

investigate this behavior, we conduct a statistical analysis of existing illusion datasets, such as IllusionVQA [2]. For each image, we ask GPT-4o to provide an explanation. In 59.3% of IllusionVQA examples, GPT-4o’s responses include phrases like “well-known” or “famous.” For simple rectangle-based illusion images, 84.2% of GPT-4o’s responses use words like “typical” or “classic,” suggesting that these images were likely part of its training data.

B.5. VLM Performance on Illusion Images Without Fine-tuning for Color QA

We evaluate the performance of closed-source models (GPT-4o, Gemini) and the base open-source model (LLaVA-7B) on color illusion images and their corresponding control groups. In contrast to our main paper results, none of these models are fine-tuned on the task of answering questions about color comparisons within an image. The results indicate that, without such fine-tuning, VLMs perform

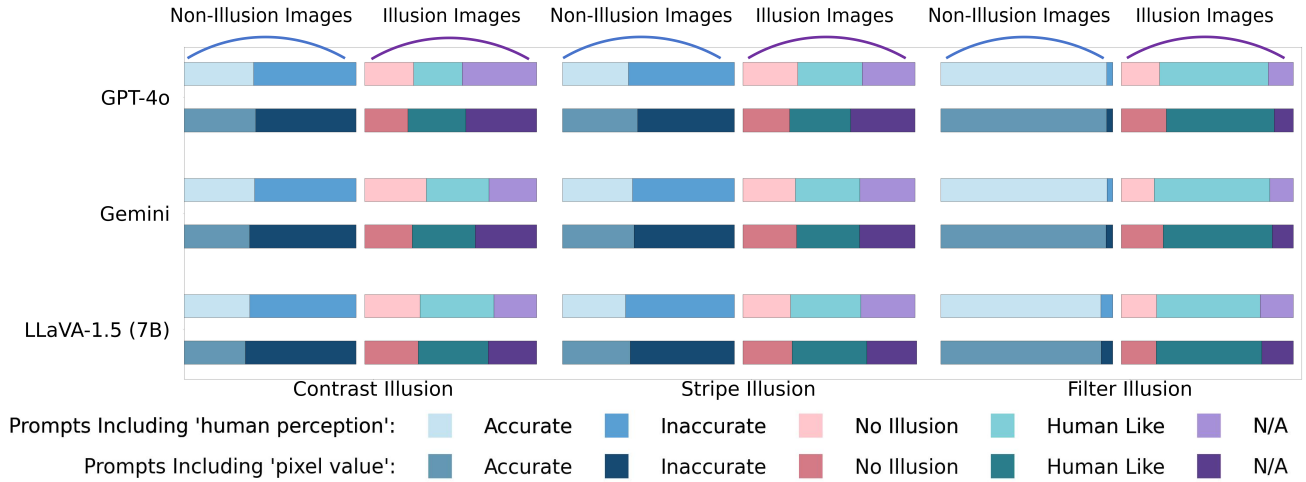
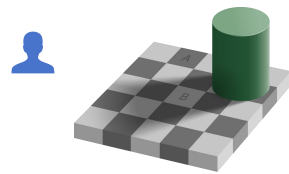


Figure 7. This figure illustrates the proportion of responses from different models across three types of illusions on our development set. The results indicate that VLMs generally perform poorly on tasks involving color comparison (contrast illusion and stripe illusion), while they excel in color recognition tasks (filter illusion). Overall, the responses of VLMs are significantly influenced by color illusions.



Based on the pixel values of the image, are the color and brightness of Square A and Square B the same?



The image you shared is a well-known optical illusion called the 'Checker Shadow Illusion.' In this illusion, Square A and Square B appear to have different shades due to the shadow cast by the cylinder. However, the pixel values of Square A and Square B are **actually the same**.

(a) Ask GPT-4 using a website-sourced image



Based on the pixel values of the image, how do the color and brightness of the left and right ground shadows, A and B, compare?



In the image you provided, ground shadow on the skier's left is **light gray**, while the one on the right appears to be a **slightly darker shade**. If you need a more detailed pixel analysis or have any further questions, feel free to let me know!

(b) Ask GPT-4 using a generated image

Figure 8. (a) When presented with a well-known color illusion scraped from the web, GPT-4o can recognize the illusion and identify that the two squares are indeed the same color gray. (b) However, when presenting GPT-4o with an image generated with the same exact shade of gray, it does identify the image as an illusion and is “deceived”.

poorly on color depth comparison tasks (contrast illusions and stripe illusions), even on non-illusion images, with accuracy only slightly above random guessing (33.3%). In contrast, most VLMs perform well on object color recognition tasks (filter illusions). We hypothesize that this may be due to the lack of color comparison data in the training datasets of current VLMs [1]. For all three types of color illusions, the proportion of VLM responses consistent with pixel values decreases when tested on illusion images, reflecting a certain degree of visual bias similar to human perception, shown in Figure 7.

B.6. Examples of Color Perception Task For Purely Vision Models

Figure 9 shows several examples of the images used in the vision-only Color Perception Task described in Section ??.



Figure 9. Examples of image color perception tasks.