

# COLORBENCH: Can VLMs See and Understand the Colorful World?

## A Comprehensive Benchmark for Color Perception, Reasoning, and Robustness

**Yijun Liang\***, **Ming Li\***, **Chenrui Fan**, **Ziyue Li**, **Dang Nguyen**, **Kwesi Cobbina**  
**Shweta Bhardwaj**, **Juhai Chen**, **Fuxiao Liu**, **Tianyi Zhou**

University of Maryland, College Park

{yliang17, minglii, tianyi}@umd.edu

Project: <https://github.com/tianyi-lab/ColorBench>

### Abstract

*Color plays an important role in human perception and usually provides critical clues in visual reasoning. However, it is unclear whether and how vision-language models (VLMs) can perceive, understand, and leverage color as humans. This paper introduces “COLORBENCH”, an innovative benchmark meticulously crafted to assess the capabilities of VLMs in color understanding, including color perception, reasoning, and robustness. By curating a suite of diverse test scenarios, with grounding in real applications, COLORBENCH evaluates how these models perceive colors, infer meanings from color-based cues, and maintain consistent performance under varying color transformations. Through an extensive evaluation of 32 VLMs with varying language models and vision encoders, our paper reveals some undiscovered findings: (i) The scaling law (larger models are better) still holds on COLORBENCH, while the language model plays a more important role than the vision encoder. (ii) However, the performance gaps across models are relatively small, indicating that color understanding has been largely neglected by existing VLMs. (iii) CoT reasoning improves color understanding accuracies and robustness, though they are vision-centric tasks. (iv) Color clues are indeed leveraged by VLMs on COLORBENCH but they can also mislead models in some tasks. These findings highlight the critical limitations of current VLMs and underscore the need to enhance color comprehension. Our COLORBENCH can serve as a foundational tool for advancing the study of human-level color understanding of multimodal AI.*

### 1. Introduction

Color is widely recognized as a fundamental component of human visual perception [11, 33], playing a critical role and providing critical clues in object detection, scene interpretation, contextual understanding, planning, etc., across

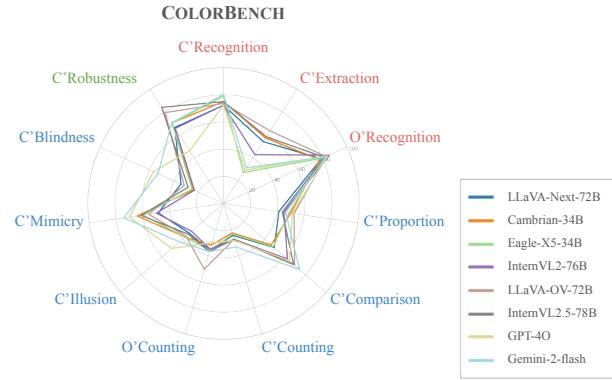


Figure 1. **Evaluation of VLMs on COLORBENCH.** The plot reports accuracy of 8 representative VLMs on 11 tasks across three categories (Perception, Reasoning, and Robustness).

critical application scenarios such as scientific discovery, medical care, remote sensing, shopping, visualization, artwork interpretation, etc. For instance, [19] leverages spectral color signatures to distinguish vegetation, health, and water bodies in satellite imagery, and [1] utilizes sediment color patterns to detect marine ecosystems. These applications underscore how color-driven features play an important role in real-world scenarios. Moreover, colors can convey affective or semantic information beyond simply recognizing and naming colors since colors are highly correlated to other attributes or concepts and thus can provide key information to various downstream tasks that do not even directly ask about colors [18, 36, 43]. As modern vision-language models (VLMs) [12, 40, 46] continue to be deployed to increasingly diverse scenarios, it is essential to examine whether and how these models can understand and leverage color information as in human perception and reasoning, and their ability to interpret illusions, handle ambiguous cues, and maintain reliable performance despite variations in colors.

However, existing benchmarks for VLMs mainly focus on tasks that may not heavily depend on color understanding or

\*These authors contributed equally to this work.

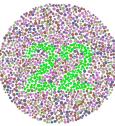
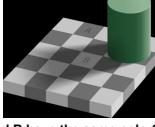
Perception		Reasoning			Robustness	
Color Recognition		Color Comparison		Color Blindness	Color Counting	Color Robustness
E.g. Paintings		E.g. Test Kit		E.g. Blindness Test	E.g. Shopping	Seed Image
						
What color is not present in this image?		Which is the closest read of Ammonia in the image?		What is the number in the center of this image?	How many distinct colors of pants in this image?	Target Segment Recolor
A. Orange B. Green C. Red D. Cyan		A. 0 ppm B. 0.25 ppm C. 0.5 ppm D. 1 ppm		A. 20 B. 21 C. 22 D. 23	A. 4 B. 5 C. 6 D. 7	
Color Extraction		Object Counting		Color Proportion		Largest Segment Recolor
E.g. Color Palette		E.g. Wild Image		E.g. Satellite Image		
						Entire Image Recolor
What is the hsv value of the given color in the image?		How many cows in this image have white faces?		What is closest to the proportion of the color green in the image?		
A. [187, 71, 50] B. [127, 49, 83] C. [187, 73, 50] D. [189, 73, 50]		A. 1 B. 2 C. 3 D. 4		A. 9% B. 13% C. 17% D. 21%		How many pillows are in the image?
Object Recognition		Color Illusion		Color Mimicry		Ans: C
E.g. Election Map		E.g. Visual Illusion		E.g. Marine Detection		Ans: B
						Ans: A
Which state has a red stripe color in this image?		Do tiles A and B have the same color?		How many red stones in this images?		Ans: A
A. WI B. MN C. NV D. NY		A. Yes B. No, B is darker C. Hard to tell D. No, A is darker		A. 0 B. 1 C. 2 D. 3		Ans: B

Figure 2. **Test samples from COLORBENCH.** COLORBENCH evaluates VLMs across three core capabilities: **Perception**, **Reasoning** and **Robustness**. The benchmark comprises 11 tasks designed to assess fine-grained color understanding abilities and the effect of color on other reasoning skills, including counting, proportion calculation, and robustness estimation. With over 1,400 instance, COLORBENCH covers a wide range of real-world application scenarios, including painting analysis, test kit readings, shopping, satellite/wildlife image analysis, etc.

require color-centric reasoning, thereby overlooking nuanced color-related factors [25, 28]. Hence, there is a lack of benchmarks that systematically assess how well VLMs understand color when it serves as the main or distinguishing feature of a scene and key information to a task. Moreover, robustness to variations in color, such as recoloring and shifting hues, has also been largely neglected in the LLM era [6, 8, 20]. Consequently, **it remains unclear whether VLMs can perceive and reason about color with human-like proficiency and to what extent their performance deteriorates under significant color perturbations.** This shortfall underscores the need for a dedicated benchmark that comprehensively probes various facets of color comprehension in VLMs.

To bridge this gap, we propose a novel benchmark, **COLORBENCH**, that aims at comprehensively evaluating VLMs on three core capabilities of color understanding: **Color Perception**, **Color Reasoning**, and **Color Robustness**. Color Perception examines VLMs’ fundamental capability to correctly detect and interpret colors from inputs. Color Reasoning refers to the reasoning skills to draw further conclusions

based on the understanding of colors from input and prior knowledge, in which colors act as a crucial clue to formulate accurate judgments. Color Robustness assesses how consistently VLMs perform when an image’s colors are altered, ensuring they maintain accurate predictions across different color variants of an image. Under these three core dimensions, 11 fine-grained tasks assessing different aspects of color understanding capabilities are formulated as shown in Figure 2, which not only shows test examples in COLORBENCH but also presents potential real-world applications. By focusing on these facets, COLORBENCH offers a granular view of VLMs’ capabilities in color understanding, aiming to illuminate both their strengths and shortcomings. We evaluate 32 widely used VLMs in our benchmark, ranging from open-source to proprietary models, from relatively small models (0.5B) to larger models (78B), and obtain some unrevealed observations. Figure 1 shows the results of 8 representative VLMs.

**Main Contribution.** We introduce “COLORBENCH”, the first dedicated benchmark for assessing the color percep-

tion, reasoning, and robustness of VLMs. We develop an evaluation suite for 11 color-centric tasks, covering diverse application scenarios and practical challenges. Moreover, we report a fine-grained empirical evaluation of 32 state-of-the-art VLMs, which exposes their limitations in color understanding and offers novel insights for future research. Our key findings are highlighted in the following:

1. The scaling law still holds for color understanding but is much weaker and mainly depends on the language model parts. The correlation between the performance and the vision encoder’s size is not significant due to the limited choices in current VLMs.
2. The absolute performances of different VLMs are relatively low, and the gaps between different models (open-source vs. proprietary, small vs. large) are not large, indicating the challenges of COLORBENCH and the negligence of color understanding in existing VLMs.
3. Despite the weaknesses of VLMs on color understanding, adding reasoning steps can still improve their performance on COLORBENCH tasks, even for color robustness, which has not been investigated by the community.
4. Color clues are indeed leveraged more or less by VLMs in most of the tasks in COLORBENCH. However, in color illusion and mimicry tasks, colors might mislead VLMs to give wrong answers, and converting colorful images into grayscale can improve the accuracy.

## 2. COLORBENCH Construction

We present **COLORBENCH**, the first benchmark explicitly designed to comprehensively evaluate the color understanding capabilities of VLMs across three key dimensions: **Color Perception**, **Color Reasoning**, and **Color Robustness**. This benchmark consists of 1,448 instances and 5,814 image-text questions spanning 11 diverse tasks. For the Color Perception and Color Reasoning categories, each instance contains an image, a question, and multiple-choice (3 to 6) options, with only one correct answer. For Color Robustness, each instance consists of 10 multiple-choice image-text questions including a seed image and 9 edited images with color changes.

### 2.1. Taxonomy

Motivated by the existing evaluation criteria from prior benchmarks and real-world application scenarios, we categorize the color understanding capability into 3 core dimensions and 11 detailed axes, as shown in Figure 2. The detailed question templates and sample cases are shown in Appendix B.

#### 2.1.1. Color Perception

This core dimension refers to the fundamental capability to correctly detect and interpret colors from inputs. We assess this capability through 3 key aspects: i) **Color Recognition**, ii) **Color Extraction**, and iii) **Object Recognition**.

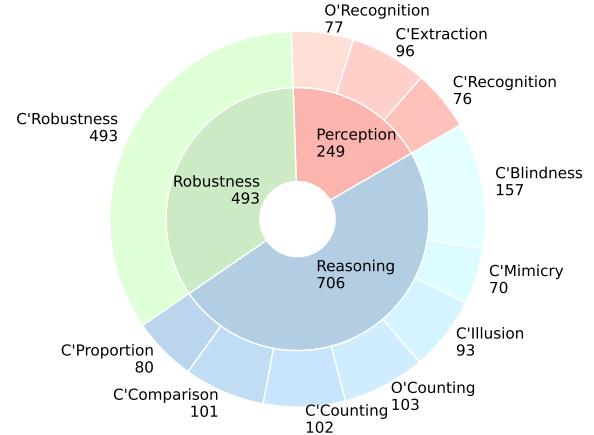


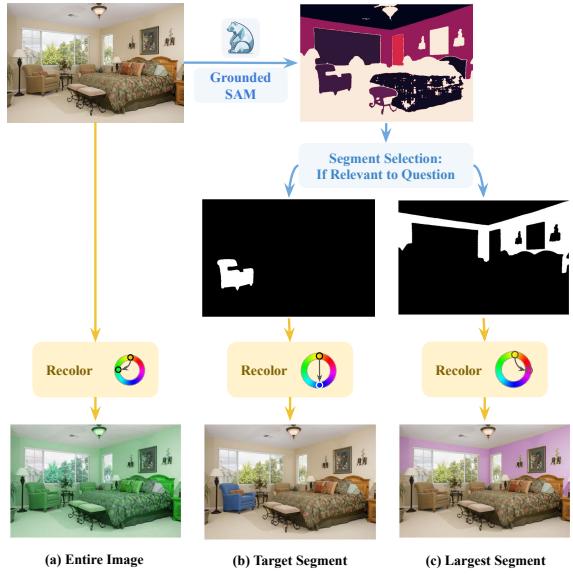
Figure 3. Statistics of 3 categories and 11 tasks in COLORBENCH.

**Color Recognition** includes questions that either ask for the color of a given object or determine whether a specific color is present in the image. **Color Extraction** requires the model to extract the value of color code (e.g., RGB, HSV, or HEX) for a given single color image. This task measures the ability to perform fine-grained color retrieval from visual input. **Object Recognition** evaluates the model’s capability to identify objects that match a specified color described in the text input. These two tasks require VLMs to be able to detect and interpret the color in either the image or text input.

#### 2.1.2. Color Reasoning

This dimension refers to the reasoning skills to draw further conclusions based on the understanding of colors from input and prior knowledge, in which colors act as a crucial clue to formulate accurate judgments. This category encapsulates 7 key aspects: i) **Color Proportion**, ii) **Color Comparison**, iii) **Color Counting**, iv) **Object Counting**, v) **Color Illusion**, vi) **Color Mimicry** and vii) **Color Blindness**.

**Color Proportion** tests the model’s capability to estimate the relative area occupied by a specific color. Questions in this task require both color perception and proportion calculation capabilities. **Color Comparison** requires the model to be able to distinguish among multiple colors in the image, assessing its sensitivity to hue, saturation, and brightness differences in visual input. **Color Counting** focuses on identifying the number of unique colors in the image, evaluating the model’s perception and differentiation of distinct color variations, and counting ability. **Object Counting** extends this challenge by requiring the model to count objects that match a specific color pattern. This task requires an integration ability of object recognition and color perception. **Color Illusion** questions query VLMs to compare colors in potential illusionary environments. This task evaluates the model’s ability to account for color-induced optical illusions. **Color Mimicry** challenges the model to detect objects camouflaged within their surroundings, where color serves as a misleading factor, requiring advanced pattern recognition



**Figure 4. Generation Pipeline for Color Robustness.** For each seed image, we apply 3 recoloring strategies (Entire Image, Target Segment, Largest Segment) to generate edited images. For each strategy, we change the color of the recoloring region via shifting the Hue values by 90°, 180°, or 270° in HSV color space.

and contextual reasoning. These two tasks both assess the model’s ability to make correct predictions under the misleading of color-related information in visual input. **Color Blindness**, inspired by Ishihara tests, assesses the model’s ability to recognize numbers or text embedded in color patterns, testing its understanding of shape-color relationships. These 7 tasks comprehensively assess the model’s capacity for logical reasoning, spatial awareness, and adaptive interpretation of color-based visual cues.

### 2.1.3. Color Robustness

**Color Robustness** assesses how consistently VLMs perform and whether they can consistently deliver accurate predictions under color variants of a given image. It involves measuring the stability of a VLM’s responses when confronted with the same text input and a series of recolored images. To ensure that color does not influence the predictions, we select questions and corresponding answers that are independent of color attributes. Under these conditions, a robust model should produce unchanged predictions regardless of recoloring manipulation. Any variation in the model’s responses is then used to quantify its susceptibility to color changes, providing a direct measure of robustness.

## 2.2. Data Curation

For most of the tasks in the category of **Color Perception** and **Color Reasoning**, we rely on human experts to manually collect images from multiple online benchmarks and websites. For the Color Proportion task, to ensure the correctness of the ground truth, an extra color extraction tool

is firstly utilized to obtain the color histogram of the image. Questions and options are then manually designed based on these color statistics. For tasks including Color Extraction, Color Blindness, and Color Illusion, testing images are generated by corresponding code programs to ensure the controllability of the questions and answers. The detailed data sources are shown in Appendix A.

After the initial data is collected, additional filtering processes are conducted in a human-machine interactive process. We first conduct inference on a variety of VLMs and discard those samples that are less challenging based on (i) model prediction correctness, (ii) model prediction confidence scores, and (iii) human evaluation. For synthesized data, similar processes are conducted but with additional code (for generation) and image assessment. The above process is conducted in three rounds before the final benchmark instances are settled. This refinement process ensures COLORBENCH a rigorous and informative benchmark for assessing color-related understanding.

For the **Color Robustness**, we evaluate the consistency of VLMs when faced with instances that differ only in the color of the visual input. To systematically assess this effect, we define 3 recoloring strategies that determine which part of the image is altered: i) **Target Segment**, ii) **Largest Segment**, and iii) **Entire Image**. As mentioned in Table 1, **Target Segment** strategy recolors only the segment containing the object referenced in the question. This strategy ensures that the modification directly affects the model’s perception of task-relevant content. **Largest Segment** strategy alters the color of the largest segment that is irrelevant to the question, testing whether models are distracted by dominant but unrelated visual changes. In contrast, **Entire Image** strategy applies a global color shift to evaluate the model’s sensitivity to overall color variations. As summarized in Table 1, the first two strategies introduce localized modifications, while the third assesses robustness to broader image-wide color changes. By incorporating both task-relevant and irrelevant edits, our benchmark provides a comprehensive evaluation of VLMs’ ability to handle color perturbations across different contexts.

While generating color variations, we derive seed images from CV-Bench [41], a publicly available benchmark. For each seed image, as shown in Figure 4, we first employ a Grounded Segmentation Model (GAM) [37] to extract segments and their corresponding labels. We then apply the predefined recoloring strategies to determine the editing region. Once the editing region is determined, we modify the corresponding region by adjusting the Hue value in the HSV color space. Specifically, we shift the Hue by 90°, 180°, and 270°. These three values ensure that the color manipulations cover significant perceptual differences across the color spectrum. These variations allow us to examine whether VLMs rely excessively on color information and whether they main-

Strategy	Editing Region	Purpose
Entire Image	Whole image	Assesses the model’s robustness to global color shifts
Target Segment	Segment containing the object referenced in the question	Evaluates the model’s sensitivity to task-relevant color changes
Largest Segment	The largest segment that is irrelevant to the question	Tests whether changes in dominant but unrelated regions affect model predictions

Table 1. Recoloring strategies.

tain consistency in their predictions despite substantial color shifts. This process produces nine variations per seed image, covering different strategies and degrees of color change to enable a comprehensive robustness assessment. To ensure interpretability, human experts filter out unnatural or negligible modifications, resulting in a final selection of 493 seed images for robustness evaluation.

### 2.3. Evaluation Metrics

For **Perception** and **Reasoning**, we use accuracy as the evaluation metric, as all tasks follow a multiple-choice format. Accuracy is computed per task and per category, representing the proportion of correctly answered questions.

For **Robustness**, we evaluate a model’s ability to maintain consistent accurate predictions under color variations. As described in Section 2.2, each seed image  $I_s$  is transformed into  $n$  recolored versions using predefined recoloring strategies, while keeping the original question  $q$  unchanged. A model  $\mathcal{M}$  is considered robust on a seed image  $I_s$  and corresponding question  $q$  if and only if it provides a correct prediction for  $I_s$  and maintains correct on all  $n$  recolored versions. To quantify robustness, we define the instance-level robustness metric  $R(I_s, q) \in \{0, 1\}$  and a model-level robustness metric  $Robust_{\mathcal{M}} \in [0, 1]$ .

**Instance-level Robustness.** Let the recolored images be  $I_1, \dots, I_n$  and the generation output of model for image  $I_i$  and question  $q$  is  $\mathcal{M}(I_i, q)$ . Define  $c(\mathcal{M}(I_i, q))$  as the model correctness:  $c(\mathcal{M}(I_i, q)) = 1$  if model result  $\mathcal{M}(I_i, q)$  is correct, otherwise 0. The instance-level robustness metric  $R(I_s, q)$  for a seed image  $I_s$  and question  $q$  is defined as:

$$R(I_s, q) = \begin{cases} 1 & \text{if } c(\mathcal{M}(I_i, q)) = c(\mathcal{M}(I_s, q)) = 1, \forall i \in [n] \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

**Overall Robustness.** Let  $\mathcal{S}$  be the set of seed images. We define model robustness to be:

$$Robust_{\mathcal{M}} = \frac{\sum_{I_s \in \mathcal{S}} R(I_s)}{|\mathcal{S}|}, Robust_{\mathcal{M}} \in [0, 1] \quad (2)$$

$Robust_{\mathcal{M}}$  represents the proportion of seed images on which the model maintains correctness across all color variations. A model is more robust when  $Robust_{\mathcal{M}}$  is higher.

## 3. Experimental

### 3.1. Implementation Details

To further advance our understanding of VLMs’ capabilities in visual reasoning, we conduct an extensive evaluation of 32 vision-language models (VLMs) spanning a range of large language model (LLM) sizes and architectures. Our evaluation includes state-of-the-art models such as GPT-4o[34], Gemini-2-flash[7], LLaVA-OV[24], LLaVA-NEXT [30], Cambrian[41], InternVL2[5], InternVL2.5[5], Qwen2.5-VL[2], and Eagle[40]. This selection covers a diverse set of architectures, including both proprietary and open-source models, enabling a comprehensive assessment of their reasoning capabilities under different computational constraints.

To ensure a fair comparison, we standardize our experimental setup across models. Open-source models with fewer than 70B parameters are evaluated using a single NVIDIA A100 80GB GPU, while larger models require four NVIDIA A100 80GB GPUs to accommodate their increased memory and computational demands.

### 3.2. Main Results

Table 2 presents the performances of a wide range of VLMs on our COLORBENCH. In general, according to the overall accuracy, larger models tend to perform better than small models, and the two proprietary models, GPT-4o and Gemini-2-flash, perform the best.

**Color Perception.** In **Color Recognition (C’Recog)**, most models perform well (above 60%), indicating that this task is relatively basic for color perception. Gemini-2 without CoT obtains the highest performance, and GPT-4o does not stick to the top. In **Color Extraction (C’Extra)**, to our surprise, the two powerful proprietary models without CoT prompting only reach the middle-tier performances, indicating the potential limitation on the color perception of their vision encoders. Similar to the Color Existence task, almost all the models perform well in **Object Recognition (O’Recog)**, and the 2 proprietary models do not reach the top. This is probably due to the strong alignment between this task and the common training recipe, which includes abundant general object detection images.

**Color Reasoning.** In **Color Proportion (C’Prop)**, even the best model, Gemini-2 with CoT, can only reach 55.0% of the accuracy, which is almost only slightly better than random guessing, showcasing the supreme difficulty of this task. In **Color Comparison (C’Comp)**, larger models perform better in this task, and the proprietary models with CoT reach the top performance unsurprisingly. Surprisingly, in **Color Counting (C’Count)**, all models show extremely poor performances. The highest performance comes from Gemini-2 with CoT, exceeding the second place by more than 10 percent, although its performance is also unsatisfactory at only 45.1%. In **Object Counting (O’Count)**, surpassing the

	Color Perception			Color Reasoning						P & R	Color Robustness	
	C'Recog	C'Extract	O'Recog	C'Prop	C'Comp	C'Count	O'Count	C'Illu	C'Mimic	C'Blind	Overall	C'Robust
VLMs: < 7B												
<b>LLaVA-OV-0.5B</b>	26.3	<b>44.8</b>	46.8	30.0	23.8	22.6	21.4	38.7	58.6	26.8	32.6	38.7
<b>InternVL2-1B</b>	35.5	34.4	59.7	23.8	41.6	19.6	22.3	34.4	38.6	<b>33.1</b>	33.6	39.4
<b>InternVL2.5-1B</b>	55.3	36.5	61.0	42.5	45.5	22.6	25.2	43.0	41.4	28.0	38.3	52.3
<b>InternVL2-2B</b>	60.5	36.5	66.2	40.0	38.6	19.6	29.1	26.9	52.9	21.0	36.4	54.2
<b>InternVL2.5-2B</b>	69.7	28.1	71.4	33.8	48.5	<b>25.5</b>	<b>30.1</b>	32.3	55.7	19.8	38.5	59.8
<b>Qwen2.5-VL-3B</b>	<b>72.4</b>	38.5	<b>74.0</b>	43.8	48.5	22.6	25.2	43.0	45.7	24.2	41.1	<b>63.7</b>
<b>Cambrian-3B</b>	67.1	31.3	66.2	<b>47.5</b>	<b>50.5</b>	<b>25.5</b>	29.1	<b>44.1</b>	<b>61.4</b>	22.3	<b>41.5</b>	59.0
VLMs: 7B – 8B												
<b>LLaVA-Next-v-7B</b>	29.0	38.5	57.1	21.3	34.7	23.5	25.2	38.7	41.4	17.8	31.2	52.1
<b>LLaVA-Next-m-7B</b>	21.1	18.8	63.6	27.5	42.6	16.7	34.0	41.9	47.1	<b>29.9</b>	33.4	55.2
<b>Eagle-X5-7B</b>	52.6	47.9	67.5	41.3	42.6	20.6	35.0	44.1	48.6	22.9	40.0	48.5
<b>LLaVA-OV-7B</b>	71.1	<b>53.1</b>	81.8	52.5	53.5	19.6	26.2	<b>48.4</b>	48.6	23.6	44.7	74.0
<b>Qwen2.5-VL-7B</b>	76.3	49.0	<b>84.4</b>	47.5	52.5	19.6	34.0	44.1	55.7	28.7	<b>46.2</b>	<b>74.4</b>
<b>Cambrian-8B</b>	72.4	28.1	72.7	48.8	54.5	<b>31.4</b>	33.0	41.9	<b>57.1</b>	17.2	42.3	64.9
<b>InternVL2-8B</b>	72.4	50.0	77.9	42.5	48.5	20.6	35.9	38.7	50.0	23.6	43.1	65.5
<b>Eagle-X4-8B</b>	71.1	47.9	68.8	45.0	50.5	26.5	<b>37.9</b>	40.9	48.6	27.4	44.1	63.7
<b>InternVL2.5-8B</b>	<b>77.6</b>	47.9	83.1	<b>50.0</b>	<b>62.4</b>	25.5	33.0	34.4	52.9	19.8	45.2	69.8
VLMs: 10B – 30B												
<b>LLaVA-Next-13B</b>	56.6	31.3	71.4	27.5	41.6	27.5	28.2	29.0	45.7	25.5	36.4	53.3
<b>Cambrian-13B</b>	67.1	34.4	74.0	46.3	47.5	<b>32.4</b>	35.0	<b>38.7</b>	55.7	24.8	42.8	64.7
<b>Eagle-X4-13B</b>	<b>73.7</b>	43.8	76.6	43.8	47.5	23.5	<b>38.8</b>	34.4	<b>57.1</b>	26.1	43.7	66.3
<b>InternVL2-26B</b>	72.4	<b>52.1</b>	87.0	<b>52.5</b>	56.4	20.6	35.0	34.4	55.7	27.4	46.3	74.0
<b>InternVL2.5-26B</b>	72.4	45.8	<b>89.6</b>	45.0	<b>63.4</b>	22.6	35.0	32.3	62.9	<b>29.3</b>	<b>46.8</b>	<b>83.0</b>
VLMs: 30B – 70B												
<b>Eagle-X5-34B</b>	<b>79.0</b>	27.1	80.5	48.8	48.5	23.5	35.9	<b>37.6</b>	60.0	25.5	43.4	67.1
<b>Cambrian-34b</b>	75.0	57.3	77.9	50.0	46.5	22.6	32.0	<b>37.6</b>	<b>64.3</b>	24.2	45.3	67.7
<b>LLaVA-Next-34b</b>	69.7	46.9	76.6	43.8	56.4	28.4	<b>41.8</b>	36.6	61.4	<b>29.9</b>	46.6	65.9
<b>InternVL2.5-38B</b>	71.1	<b>60.4</b>	<b>89.6</b>	<b>53.8</b>	<b>63.4</b>	<b>29.4</b>	40.8	34.4	61.4	26.8	<b>50.0</b>	<b>84.6</b>
<b>InternVL2-40B</b>	72.4	52.1	83.1	51.3	61.4	19.6	35.9	34.4	58.6	21.0	45.6	78.7
VLMs: > 70B												
<b>LLaVA-Next-72B</b>	72.4	54.2	79.2	41.3	49.5	24.5	35.9	33.3	48.6	<b>34.4</b>	45.2	66.5
<b>LLaVA-OV-72B</b>	73.7	<b>63.5</b>	83.1	<b>52.5</b>	<b>69.3</b>	<b>27.5</b>	<b>50.5</b>	<b>36.6</b>	55.7	31.9	<b>51.9</b>	79.5
<b>InternVL2-76B</b>	72.4	42.7	<b>85.7</b>	45.0	62.4	<b>27.5</b>	35.0	31.2	50.0	23.6	44.6	65.7
<b>InternVL2.5-78B</b>	<b>75.0</b>	58.3	81.8	43.8	68.3	<b>27.5</b>	36.9	34.4	<b>61.4</b>	28.7	48.8	<b>84.2</b>
VLMs: Proprietary												
<b>GPT-4o</b>	73.7	29.2	<b>84.4</b>	<b>51.3</b>	64.4	28.4	30.1	<b>54.8</b>	70.0	<b>56.7</b>	52.8	46.2
<b>Gemini-2-flash</b>	<b>80.3</b>	<b>31.3</b>	83.1	46.3	<b>74.3</b>	<b>33.3</b>	<b>36.9</b>	43.0	<b>74.3</b>	53.0	<b>53.9</b>	70.7
<b>GPT-4o (CoT)</b>	76.3	36.5	<b>85.7</b>	51.3	73.3	27.5	35.9	<b>46.2</b>	<b>74.3</b>	<b>65.6</b>	56.2	69.9
<b>Gemini-2-flash (CoT)</b>	<b>79.0</b>	42.7	83.1	<b>55.0</b>	<b>76.2</b>	<b>45.1</b>	<b>41.8</b>	43.0	<b>74.3</b>	54.1	<b>57.8</b>	<b>73.6</b>

Table 2. Performance of 32 VLMs (grouped by size) on COLORBENCH. Within each group, models are ranked based on the overall accuracy on color perception and reasoning tasks (P & R Overall). The best performance in each group is highlighted in bold.

2 proprietary models, LLaVA-OV-72B reaches the top and becomes the only model that exceeds 50% of the accuracy. Similar to the findings from the Object Recognition task, this might be caused by the extremely adequate object detection tasks in open-sourced training recipes. In **Color Illusion (C'Illu)**, the accuracies of most models lie in the range of 30% to 50%, and GPT-4o without CoT is the only one that exceeds 50% of the accuracy. In **Color Mimicry (C'Mimic)**, the 2 proprietary models reach the top, while more reasoning steps do not benefit a lot. In **Color Blindness (C'Blind)**, most of the models present accuracies under 30%. Considering the extremely practical usage of this scenario, we think the current community should pay more attention to this. Moreover, we also observe that, surprisingly, more reasoning steps benefit VLMs in the color blindness test, although it seems like a pure color perception task.

**Color Robustness.** In **Color Robustness (C'Robust)**, a

higher value represents better robustness towards color altering. The only 3 models that exceed 80% are InternVL2.5-26B, InternVL2.5-38B and InternVL2.5-72B, which utilize relatively larger vision encoders, InternViT-6B, compared with others (mostly only 300-400M). In the meantime, GPT-4o has a really low robustness (46.2%) to colors, indicating its vulnerable sensitivity to color changes, while Gemini-2 shows promising robustness (70.7%) towards colors. Moreover, another surprising observation is that even though only the colors are changed and all the original queries are kept, utilizing more reasoning steps can consistently improve robustness for GPT-4o (+23.7%) and Gemini-2 (+2.9%).

	Color Perception			Color Reasoning						P & R	Color Robustness	
	C'Recog	C'Extract	O'Recog	C'Prop	C'Comp	C'Count	O'Count	C'Illu	C'Mimic	C'Blind	Overall	C'Robust
L+V	0.5657 (*)	0.5255 (*)	0.7107 (*)	0.5125 (*)	0.6358 (*)	0.4316 (*)	0.7566 (*)	-0.3460	0.4832 (*)	0.2460	0.7619 (*)	0.7226 (*)
L	0.5724 (*)	0.4937 (*)	0.6769 (*)	0.4696 (*)	0.6118 (*)	0.4408 (*)	0.7611 (*)	-0.3697 (*)	0.4559 (*)	0.2824	0.7436 (*)	0.7026 (*)
V	0.3955 (*)	0.2856	0.5465 (*)	0.6242 (*)	0.5295 (*)	0.2089	0.3608	-0.0127	0.6024 (*)	-0.0679	0.5271 (*)	0.5320 (*)

Table 3. Spearman’s rank correlation between VLM performance and different model parts’ sizes on each task. L denotes the language model part’s size and V represents the vision encoder part’s size. We use “(\*)” to mark correlations with p-values  $\leq 0.05$ . It shows that the scaling law still holds for color understanding but it is much weaker.

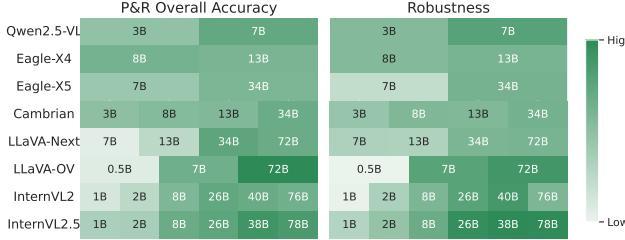


Figure 5. The heatmaps related to performances and VLM sizes. Deeper color represents higher performance of P&R Overall Accuracy or Robustness. Each line represents a model family with the sizes growing from small to large. This visualization clearly shows the correlation between performances and model sizes, larger model leads to higher performance.

### 3.3. Further Findings

**Finding 1.** The scaling law still holds for color understanding but is much weaker and mainly depends on the language model parts. The correlation between the performance and the vision encoder’s size is not significant due to the limited choices in current VLMs.

To quantitatively analyze the correlation between VLM performances on color understanding tasks and their sizes, Spearman’s rank correlation is calculated between VLM performances and (i) overall model sizes ( $L + V$ ), (ii) language model sizes ( $L$ ), and (iii) vision encoder sizes ( $V$ ). The correlation values and p signs are presented in Table 3; a star is notated when the p-value of the correlation is lower than 0.05. It is observed that between the performances and language model sizes, most of the tasks have a correlation greater than 0.5 and a p-value smaller than 0.05, except for Color Illusion and Color Blindness due to their special characteristics. Since the correlation between overall model sizes ( $L + V$ ) and P&R Overall (0.7619), and Robustness (0.7390), we conclude that the color understanding, including Color Perception, Color Reasoning, and Color Robustness, still follows the scaling law of model sizes. Figure 5 presents the correlations between performances and model sizes in each model family. This visualization clearly shows the correlation between performances and model sizes; a larger model leads to higher performance within each model family.

However, between the performances and vision encoder sizes, most of the tasks either have a correlation lower than 0.5 or a p-value greater than 0.05, which is not sufficient

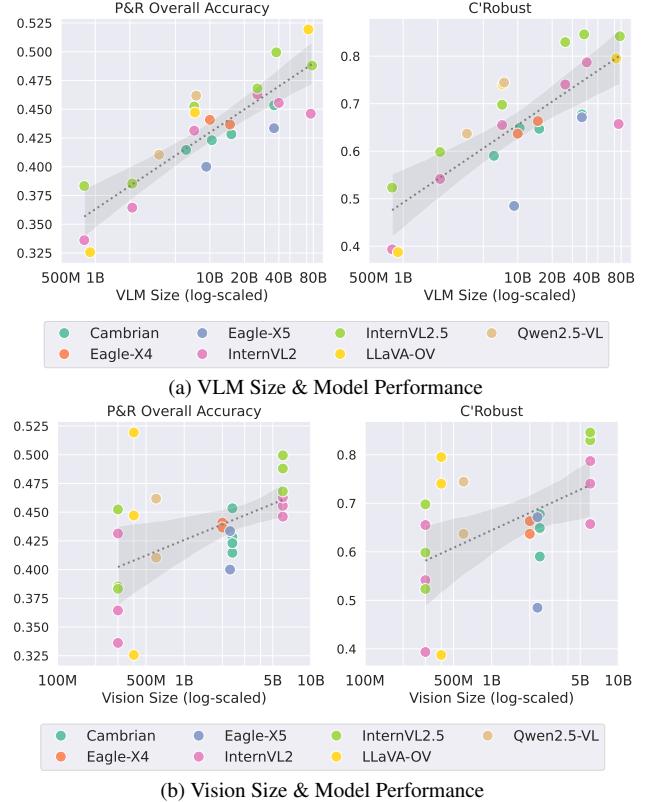


Figure 6. The scatter plots related to performances and model sizes. Each plot illustrates the relationship between the log-scaled size of parameter numbers and the performance across all models. (a) Number of parameters for the entire VLM; (b) Number of parameters for the vision encoder. There is a significant correlation between performance and VLM sizes. However, since the choices of vision encoder size are concentrated and limited, it is hard to conclude the relationship between performance and vision sizes.

to conclude with evident positive correlation. Despite these findings, we try to avoid conveying the message that there is no positive correlation between performances and vision encoder sizes. We think it is because of the negligence of the current community to focus on the scaling laws of vision encoders. The vision encoders used in the current mainstream VLMs are constrained in a very small set: (i) most of the VLMs only use one type of vision encoders for the whole family, except for the InternVL2 and InternVL2.5 series; (ii) most of VLMs use the vision encoder with the size of 300 - 400M. These challenges make it hard to evaluate

Model Size	Color P & R Overall		Color Robustness	
	Model	Best	Model	Best
<7B	Cambrian-3B	41.5	Qwen2.5-VL-3B	63.7
7B–8B	Qwen2.5-VL-7B	46.2	Qwen2.5-VL-7B	74.4
10B–30B	InternVL2.5-26B	46.8	InternVL2.5-26B	83.0
30B–50B	InternVL2.5-38B	50.0	InternVL2.5-38B	84.6
>70B	Llava-ov-72B	51.9	InternVL2.5-78B	84.2
Proprietary	Gemini-2	53.9	Gemini-2	70.7
Proprietary	Gemini-2 (CoT)	57.8	Gemini-2 (CoT)	73.6

Table 4. The best model within each group and its performances (on P&R accuracy and Robustness). **The absolute performances of different VLMs on COLORBENCH are relatively low, and the performance gaps between models are not large.**

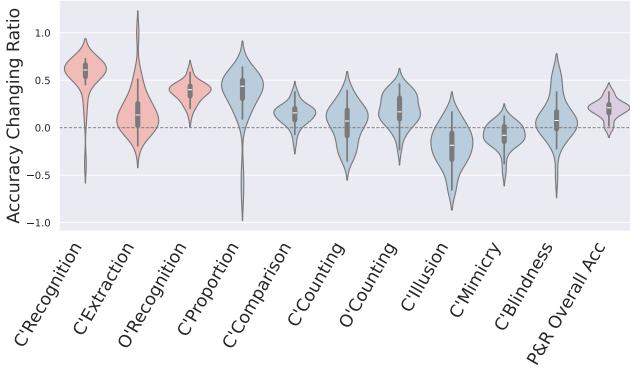


Figure 7. The percentage of change in accuracy (y-axis) by converting colorful images to grayscale in each COLORBENCH task (x-axis). Each violin plot visualizes the distribution over all 32 VLMs. Higher (lower) percentage indicates that VLMs rely more (less) on color clues for the task. Positive (negative) percentage indicates degradation (improvement) on grayscale images. **Color clues are indeed more or less leveraged by VLMs in most tasks but they might mislead VLMs (illusion & mimicry).**

the scaling laws of vision encoders. Figure 6b presents the correlations between model performances and vision encoder’s sizes for all models. Since the sizes of vision encoders are limited, it is hard to conclude the relationship between vision sizes and performance. Further visualizations are presented in Appendix D.2.

**Finding 2.** The absolute performances of different VLMs are relatively low, and the gaps between different models (open-source vs. proprietary, small vs. large) are not large, indicating the challenges of COLORBENCH and negligence of color understanding in existing VLMs.

As shown in Table 4, we separate all the VLMs into several groups based on their sizes and present the best accuracy and the model name within each group. We can see that even the powerful proprietary models, GPT-4o and Gemini-2, can only reach an overall color perception and reasoning (P & R Overall) accuracy of 53.9%, only +2.0% better than the best open-sourced model. Moreover, the best model from

group 1 has the accuracy of 41.5% (cambrian-3B), which is only 10.4% lower than the best open-sourced model. As for the robustness, the powerful proprietary models even show weaker robustness than the 7B model. Considering the lack of existing benchmarks specifically evaluating VLMs’ color understanding capabilities, we conclude that this area is long-neglected by the community, and the open-sourced community is still on the same page with the proprietary model providers.

**Finding 3.** Despite the weaknesses of VLMs on color understanding, adding reasoning steps can still improve their performance on COLORBENCH tasks, even for color robustness, which has not been investigated by the community.

The impact of using CoT prompting is shown in Table 5, in which we can see CoT improves the average P&R Overall accuracy across both models by +3.65%, indicating that reasoning benefits these color-related tasks. Within the category of Color Perception, the improvements from CoT on Color Recognition and Object Recognition are quite limited as these tasks heavily rely on the vision encoder. Figure 52 and 53 in Appendix E illustrate that adding reasoning steps does not take effect since the initial visual perception and color identification are incorrect in the slow thinking process. However, to our surprise, we find that the Color Extraction task benefits extremely from more reasoning steps, although it seems only related to the vision encoder. After a thorough investigation, we observe that most of the current VLMs are not capable of directly extracting color values, so they need to use more reasoning steps to reach reasonable answers.

Within the category of Color Reasoning, CoT benefits most of the tasks. However, in the Color Illusion task, CoT harms the model performance. After a manual investigation, we observe that more reasoning steps might cause VLMs to focus more on the misleading environments rather than directly compare the assigned colors, as shown in Figure 54. Another observation occurs in the Color Blindness task. Unlike other reasoning-related tasks, humans can read a color blindness test image with a simple glimpse without any slow thinking. This fascinating misalignment between humans and VLMs intrigues us to further investigation. We find that VLMs recognize these digits in a button-up pattern: they need to first infer that the dots in the image can form a digit before they really recognize these dots as digits.

In addition, the consistent improvement of CoT on Color Robustness is also an un-revealed phenomenon. In our setting, only the colors of the image are altered, and the questions are strictly the same as the original. Thus, under this circumstance, color is the only variant, which is supposed to be more related to the capability of the vision encoder. However, counterintuitively, as shown in our experiments, more reasoning steps make the VLMs more robust to the color

	Color Perception			Color Reasoning						P & R	Color Robustness	
	C'Recog	C'Extract	O'Recog	C'Prop	C'Comp	C'Count	O'Count	C'Illu	C'Mimic	C'Blind	Overall	C'Robust
GPT-4o Δ	+2.6	+7.3	+1.3	0.0	+8.9	-0.9	+5.8	-8.6	+4.3	+8.9	+3.4	+23.7
Gemini-2 Δ	-1.3	+11.4	0.0	+8.7	+1.9	+11.8	+4.9	0.0	0.0	+1.1	+3.9	+2.9
Average Δ	+0.65	+9.35	+0.65	+4.35	+5.4	+5.45	+5.35	-4.3	+2.15	+5.0	+3.65	+13.3

Table 5. Adding reasoning steps can improve VLMs’ performance on COLORBENCH. The change of accuracy brought by Chain of Thought (CoT) prompting on all tasks for GPT-4o and Gemini-2-flash. The last row presents the average improvement across both models.

changes, which is probably caused by the higher confidence of correct answers after reasoning.

**Finding 4.** Color clues are indeed leveraged more or less by VLMs in most of the tasks in COLORBENCH. However, in color illusion and mimicry tasks, colors might mislead VLMs to wrong answers, and converting colorful images to grayscale can improve the accuracy.

In order to examine whether VLMs really leverage color clues to handle tasks in COLORBENCH, experiments are conducted by converting all the original colorful images in the Color Perception and Reasoning categories into grayscale ones, without changing the questions. Under this circumstance, the accuracies are expected to decrease dramatically as all our questions are related to colors. For quantitative analysis, we calculate the accuracy changing ratio as  $(Acc_{ori} - Acc_{gray})/Acc_{ori}$  for each VLM on each task. This value directly represents how the original accuracy changes with a gray-scale transformation. The positive value represents that the VLM has a higher accuracy on the original colored images, indicating that it needs color clues to solve the task. Higher positive values represent higher significance of the color clues. On the contrary, if the value is negative, it means that the VLM can reach a better accuracy after the gray-scale transformation, indicating that it does not need color clues for the task, and colors might even mislead VLM’s judgment. Lower negative values represent severe harm the color can have on the task.

The accuracy changing ratio distribution across all VLMs and tasks are presented in Figure 7 as the violin plot. As shown in the figure, for most of the tasks, the ratios of VLMs are above 0, indicating that VLMs indeed leverage color clues to correctly solve the tasks; removing the color directly harms the original accuracies dramatically. However, when it comes to Color Illusion and Color Mimicry, the majority of the changing ratios are beneath 0, which means that VLMs can get better accuracies when all the color information is removed. This phenomenon is reasonable as the colors on both of these two tasks are more likely serving as the misleading factors. In the meantime, for the Color Counting and Color Blindness tasks, almost half the accuracies increase and half decrease, indicating that the color clues might not be so significant in this task, thus, some of the models can find other ways to get the answer. We also investigate the correlation between accuracy changing ratios and model sizes, while no

significant correlation can be concluded.

## 4. Related works

### 4.1. VLM Benchmarks

With the rapid advancements in Vision-Language Models (VLMs) [9], numerous benchmarks have emerged to systematically evaluate VLM capabilities across diverse dimensions [28]. These benchmarks generally fall into two categories: text-centric and vision-centric evaluations, each designed to assess distinct multimodal competencies. Text-centric benchmarks primarily measure commonsense knowledge, reasoning, and complex problem-solving capabilities, exemplified by tasks in MMMU [45] and NaturalBench [23]. Conversely, vision-centric benchmarks focus on visual perception and reasoning (MMBench [31] and MME [10]), and robustness to visual perturbations (Grit [14] and Visual Robustness [17]). Furthermore, several benchmarks have extended their scope to evaluate specialized visual tasks, such as spatial relationship comprehension (SEED-Bench [22] and MM-Vet [44]), chart and map understanding (MMSTAR [4] and MuirBench [42]), visual grounding (Flickr30k [35] and TRIG [26]) and the detection and understanding of visual hallucinations (POPE [27] and Hallusion-Bench [13]). However, despite the extensive scope covered by existing VLM benchmarks, none currently provide an integrated evaluation that simultaneously assesses visual perception, reasoning, and robustness within a unified framework. Moreover, although certain benchmarks [10, 31] have incorporated color-related questions, these have typically addressed basic color perception and recognition, neglecting deeper assessments of reasoning and robustness associated with color understanding.

### 4.2. Color Evaluation

Color understanding is increasingly recognized as a crucial aspect of Vision-Language Models’ ability to perceive and interpret visual content. Limited studies have explored how color information influences model performance on specific tasks. Some studies [47, 48] explore the understanding of color by replacing color-related words in textual inputs to evaluate the models’ ability to handle color-specific information. More recent research [16, 21] focus on assessing fine-grained color discrimination by asking models to distinguish subtle color differences in

visual inputs. Samin et al. [38] introduced color-related foils to test VLMs’ capacity to cognize about basic colors like red, white, and green, particularly in contexts requiring attention to subtle cues. Additionally, Burapacheep et al. [3] developed a benchmark dataset to evaluate and enhance compositional color comprehension in VLMs, emphasizing tasks where understanding minimal color relationships is essential. IllusionVQA [39] evaluates model perception of color illusions in photorealistic scenes. While these works have addressed isolated aspects of color understanding, none has provided a holistic assessment framework. In contrast to these previous works, our study establishes the first comprehensive and specialized benchmark for evaluating the color-related abilities of VLMs, offering a quantitative, automated approach to further this area of research.

## 5. Conclusion, limitation and Future Works

In this paper, we introduce COLORBENCH, the first benchmark designed to comprehensively evaluate the color understanding capabilities of VLMs, including Perception, Reasoning, and Robustness. After evaluating 32 widely used VLMs on our benchmark, several undiscovered observations are revealed by us. These observations emphasize the need for more sophisticated model architectures that integrate deeper color reasoning capabilities. We plan to expand to more diverse tasks that involve complex interplays of color and texture, shape, or spatial relationships. Additionally, investigating the impact of different visual encoders and language models could further elucidate the pathways through which VLMs process color information.

## References

- [1] Basit Alawode, Iyyakutti Iyappan Ganapathi, Sajid Javed, Naoufel Werghi, Mohammed Bennamoun, and Arif Mahmood. Aquaticclip: A vision-language foundation model for underwater scene analysis. *arXiv preprint arXiv:2502.01785*, 2025. 1
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianjiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 5
- [3] Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorsswap: A color and word order dataset for multimodal evaluation. *arXiv preprint arXiv:2402.04492*, 2024. 10
- [4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 9
- [5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 5
- [6] Kanjar De and Marius Pedersen. Impact of colour on robustness of deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21–30, 2021. 2
- [7] Google DeepMind. Gemini 2.0 flash, 2025. 5
- [8] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4829–4837, 2016. 2
- [9] Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8, 2024. 9
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Meng-dan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 9
- [11] Karl R. Gegenfurtner and Jochem Rieger. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, 10(13):805–808, 2000. 1
- [12] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 1
- [13] 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*, pages 14375–14385, 2024. 9
- [14] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022. 9
- [15] Shuai He, Anlong Ming, Li Yaqi, Sun Jinyuan, Zheng ShunTian, and Ma Huadong. Thinking image color aesthetics assessment: Models, datasets and benchmarks. *ICCV*, 2023. 14
- [16] Nam Hyeon-Woo, Moon Ye-Bin, Wonseok Choi, Lee Hyun, and Tae-Hyun Oh. Vlm’s eye examination: Instruct and inspect visual competency of vision language models. *arXiv preprint arXiv:2409.14759*, 2024. 9
- [17] Md Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Md Azam Hossain. Visual robustness benchmark for visual question answering (vqa). *arXiv preprint arXiv:2407.03386*, 2024. 9
- [18] Ali Jahanian, Shaiyan Keshvari, SVN Vishwanathan, and Jan P Allebach. Colors—messengers of concepts: Visual design mining for learning color semantics. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):1–39, 2017. 1
- [19] Johannes Jakubik, Benedikt Blumenstiel, and Clive Tinashe Marimo. Ms-clip: Multi-spectral vision language learning for earth observation. In *American Geophysical Union Fall Meeting*, 2024. 1
- [20] Jayendra Kantipudi, Shiv Ram Dubey, and Soumendu Chakraborty. Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks. *IEEE Transactions on Artificial Intelligence*, 1(2): 181–191, 2020. 2
- [21] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *arXiv preprint arXiv:2410.07112*, 2024. 9
- [22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 9
- [23] 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*, 2024. 9
- [24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 5
- [25] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai,

- Wankou Yang, Yabiao Wang, and Chengjie Wang. A survey on benchmarks of multimodal large language models, 2024. 2
- [26] Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. Towards visual text grounding of multimodal large language model, 2025. 9
- [27] 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. 9
- [28] Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 2025. 2, 9
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll’ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V I3*, pages 740–755. Springer, 2014. 14
- [30] 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. 5
- [31] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 9
- [32] Lingjun Mao, Zineng Tang, and Alane Suhr. Evaluating model perception of color illusions in photorealistic scenes. *arXiv preprint arXiv:2412.06184*, 2024. 14
- [33] Daniela Mapelli and Marlene Behrmann. The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, 3(4):237–247, 1997. 1
- [34] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and etc. Gpt-4o system card, 2024. 5
- [35] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 9
- [36] Ragini Rathore, Zachary Leggon, Laurent Lessard, and Karen B Schloss. Estimating color-concept associations from image statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1226–1235, 2019. 1
- [37] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4
- [38] Ahnaf Mozib Samin, M Firoz Ahmed, and Md Mush-taq Shahriyar Rafee. Colorfoil: Investigating color blindness in large vision and language models. *arXiv preprint arXiv:2405.11685*, 2024. 10
- [39] 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. 10, 14
- [40] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1, 5
- [41] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 4, 5, 14
- [42] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 9
- [43] Hanna-Sophia Widholzl and Ece Takmaz. Decoding emotions in abstract art: Cognitive plausibility of clip in recognizing color-emotion associations. *arXiv preprint arXiv:2405.06319*, 2024. 1
- [44] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 9
- [45] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi 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*, pages 9556–9567, 2024. 9
- [46] 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*, 2024. 1
- [47] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrastive intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding. *arXiv preprint arXiv:2306.08832*, 2023. 9
- [48] TIANCHENG ZHAO, TIANQI ZHANG, MINGWEI ZHU, HAOZHAN SHEN, KYUSONG LEE, XIAOPENG LU, and JIANWEI YIN. VI-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 9
- [49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 14
- [50] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 14

## Table of Contents for Appendix

<b>A Data Sources</b>	<b>14</b>
<b>B COLORBENCH Categories and Questions</b>	<b>14</b>
<b>C Evaluation Prompts</b>	<b>19</b>
<b>D More Visualizations</b>	<b>19</b>
D.1. VLM Size & Model Performance for Each Task	19
D.2 Vision Size & Model Performance for Each Task . . . . .	19
D.3 Performance for Each Model Family on Each Task . . . . .	20
<b>E Samples Cases</b>	<b>23</b>
E.1. Effect of CoT . . . . .	23
E.2. Effect of Grayscale . . . . .	28
E.3. Easy Cases . . . . .	30
E.4. Difficult Cases . . . . .	32

## A. Data Sources

We conduct COLORBENCH from multiple sources, including website sources, publicly available benchmarks, and generated images. The detailed sources are included in Table 6.

Category	Data Source
C'Recognition	Website, ICAA17K [15]
C'Recognition	Website, ICAA17K [15]
C'Extraction	Synthetic Data
C'Proportion	Website, Synthetic Data
C'Comparison	Website
C'Counting	Website, Synthetic Data
C'Counting	Website, ADA20K [49, 50], COCO2017 [29]
C'Mimicry	Website, IllusionVQA[39], RCID[32]
C'Blindness	Synthetic Data
C'Robust	CV-Bench[41]

Table 6. Data sources for each task.

## B. COLORBENCH Categories and Questions

Table 7 provides a detailed description of each task, alongside representative figures and sample questions that effectively demonstrate the specific capabilities being tested. Cases are provided for each task in Figure 8 to 18.

	<b>Task</b>	<b>#</b>	<b>Sample Case</b>	<b>Description</b>	<b>Sample Questions</b>
Perception	Color Recognition	76	Figure 8	Ask for the color of a specific object or determine if a particular color is present in the image.	What is the color of <i>object</i> in this image? What color does not exist in this image?
	Color Extraction	96	Figure 9	Extract the color code value (e.g., RGB, HSV, or HEX) from a single color in the image.	What is the HSV value of the given color in the image? What is the RGB value of the given color in the image?
	Object Recognition	77	Figure 10	Identify objects in the image that match a specified color noted in the text input.	What <i>object</i> has a color of <i>pink</i> in this image?
	Color Proportion	80	Figure 11	Estimate the relative area occupied by a specified color in the image.	What is the dominant color in this image? What is the closest to the proportion of the red color in the image?
	Color Comparison	101	Figure 12	Distinguish among multiple colors present in the image to assess overall tones and shades.	Which photo is <i>warmer</i> in overall color? Which object has a <i>darker</i> color in the image?
Reasoning	Color Counting	102	Figure 13	Identify the number of unique colors present in the image.	How many different colors are in this image?
	Object Counting	103	Figure 14	Count the number of objects of a specified color present in the image.	How many objects with <i>green</i> color are in this image?
	Color Illusion	93	Figure 15	Assess and compare colors in potential illusionary settings within the image.	Do two objects have the same color?
	Color Mimicry	70	Figure 16	Detect objects that are camouflaged within their surroundings, where color is a key deceptive element.	How many <i>animals</i> are in this image?
	Color Blindness	157	Figure 17	Recognize numbers or text that are embedded in color patterns, often used in tests for color vision.	What is the number in the center of the image?

Table 7. **Task and question definition in COLORBENCH.**

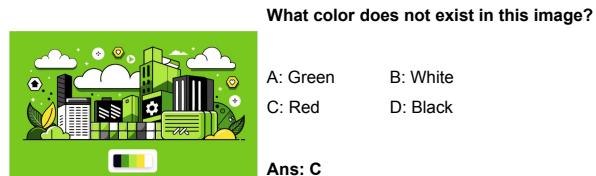
## Color Recognition



What is the color of the banana in this image?

- A: Red
- B: Green
- C: Yellow
- D: Black
- E: None of the above

Ans: E



What color does not exist in this image?

- A: Green
- B: White
- C: Red
- D: Black

Ans: C

Figure 8. Cases for Color Recognition Task.

## Color Extraction



What is the HSV value of the given color in the image?

- A: [100, 51, 81]
- B: [329, 98, 100]
- C: [331, 100, 100]
- D: [329, 100, 100]

Ans: D



Q: What is the HSV value of the given color in the image?

- A: [47, 62, 100]
- B: [107, 16, 22]
- C: [45, 64, 100]
- D: [45, 62, 100]

Ans: D

Figure 9. Cases for Color Extraction Task.

## Object Recognition



Which state does not have a color of pink in this image?

- A: Montana
- B: Arizona
- C: Michigan
- D: New York

Ans: D



Which object has a color of black in this image?

- A: Background
- B: Banana
- C: Apple
- D: Orange

Ans: C

Figure 10. Cases for Object Recognition Task.

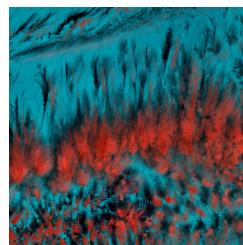
## Color Proportion



Which is the dominant color in this painting?

- A: Blue
- B: Yellow
- C: Green
- D: Orange

Ans: A



What is closest to the proportion of the color red in the image?

- A: 10%
- B: 20%
- C: 30%
- D: 40%

Ans: C

Figure 11. Cases for Color Proportion Task.

## Color Comparison

 <p>Which photo is warmer in overall color?</p> <p>A: The left one B: The right one <b>Ans: B</b></p>	 <p>Which dog has the darkest color in the image?</p> <p>A: No.1      B: No.4 C: No.5      D: No.3 <b>Ans: A</b></p>
--	--

Figure 12. Cases for Color Comparison Task.

## Color Counting

 <p>How many different colors of flowers are in this image?</p> <p>A: 1      B: 2 C: 3      D: 4 <b>Ans: C</b></p>	 <p>How many colors are there in this flag?</p> <p>A: 3      B: 4 C: 5      D: 6 <b>Ans: D</b></p>
---	--

Figure 13. Cases for Color Counting Task.

## Object Counting

 <p>How many striped animals can be seen in this image?</p> <p>A: 12      B: 11 C: 13      D: 9 E: 10 <b>Ans: C</b></p>	 <p>How many green bananas can be seen in this image?</p> <p>A: 6      B: 7 C: 5      D: 4 E: 0 <b>Ans: A</b></p>
--	---

Figure 14. Cases for Object Counting Task.

## Color Illusion

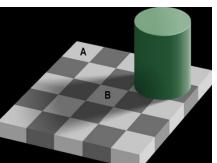
 <p>Do the blocks labeled a and b have the same color/shade?</p> <p>A: No, a is darker B: Hard to tell without more context C: Yes, one appears darker due to how our eyes perceive shadows D: No, b is darker <b>Ans: D</b></p>	 <p>What colors are the two pills?</p> <p>A: Cannot tell from this image, the colors seem to be shifting? B: Both are the exact same shade of gray C: The left one is bluish-gray and the right one is reddish-grey D: The left one is reddish-gray and the right one is bluish-grey <b>Ans: B</b></p>
---	---

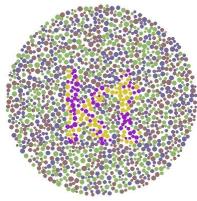
Figure 15. Cases for Color Illusion Task.

## Color Mimicry

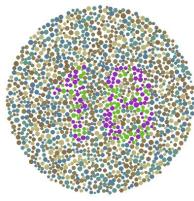
 <p>How many seahorses in this image?</p> <p>A: 0      B: 1 C: 3      D: 5 <b>Ans: B</b></p>	 <p>How many leaves in this image?</p> <p>A: 1      B: 2 C: 3      D: 0 <b>Ans: D</b></p>
---	---

Figure 16. Cases for Color Mimicry Task.

## Color Blindness



- There are two strings in the image.  
What are the strings in the center of this image?  
A: kt      B: la  
C: lo      D: lt  
**Ans: A**



- What is the number in the center of this image?  
A: 6      B: 9  
C: 17      D: 18  
**Ans: D**

Figure 17. Cases for Color Blindness Task.

Original Image



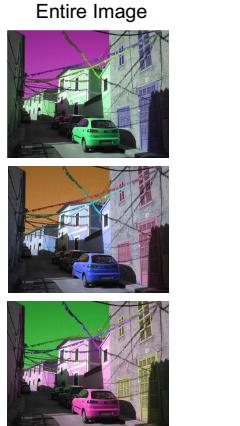
Q: How many cars are in the image?

- A: 8    B: 7    C: 6    D: 5    E: 4

**GT: E**

Recoloring Strategy

Targeted Segment



Largest Segment



Original Image



Q: How many curtains are in the image?

- A: 3    B: 2    C: 1    D: 4    E: 0

**GT: C**

Recoloring Strategy

Targeted Segment



Largest Segment



Figure 18. Cases for Color Robustness Task.

## C. Evaluation Prompts

**Instruction Prompt** You'll be given an image, an instruction and some options. You have to select the correct one. Do not explain your reasoning. Answer with only the letter that corresponds to the correct option. Do not repeat the entire answer.

**CoT Instruction Prompt** You'll be given an image, an instruction and some options. You have to select the correct one. Think step by step before answering. Then conclude with the letter that corresponds to the correct option. Make sure the option letter is in the parentheses like (X). Do not include ( or ) in the response except for the answer.

## D. More Visualizations

### D.1. VLM Size & Model Performance for Each Task

Figure 19 to 28 present detailed correlations between the log-scaled sizes of **VLM parameters** and the performance metrics for each task of Perception and Reasoning Categories. Deeper color represents higher accuracy. Each line represents a model family with the sizes growing from small to large. This visualization clearly shows the correlation between performances and model sizes, larger model leads to higher performance.

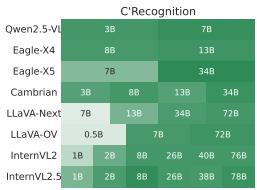


Figure 19. The heatmap for Color Recognition.

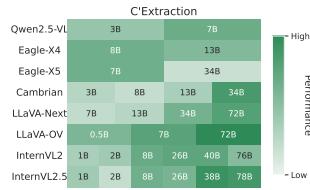


Figure 20. The heatmap for Color Extraction.

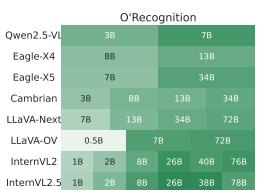


Figure 21. The heatmap for Object Recognition.

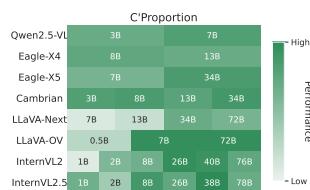


Figure 22. The heatmap for Color Proportion.

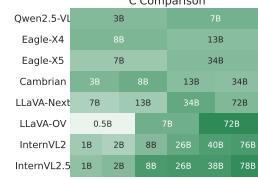


Figure 23. The heatmap for Color Comparison.

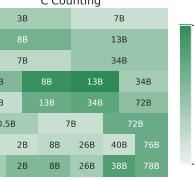


Figure 24. The heatmap for Color Counting.

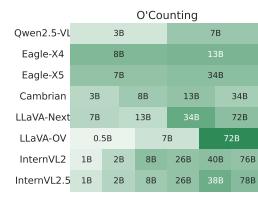


Figure 25. The heatmap for Object Counting.

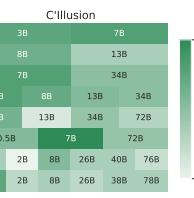


Figure 26. The heatmap for Color Illusion.

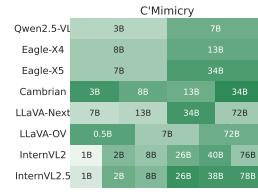


Figure 27. The heatmap for Color Mimicry.

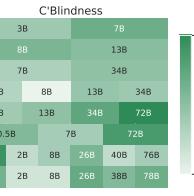


Figure 28. The heatmap for Color Blindness.

### D.2. Vision Size & Model Performance for Each Task

Figure 29 to 33 show detailed correlations between the log-scaled sizes of **vision encoders** and the performance metrics for each task of Perception and Reasoning Categories. Colors represent different model families. Models that have the same vision encoder sizes but with different LLM sizes are plotted as different points. Given that the majority of Vision-Language Models (VLMs) utilize a singular type of vision encoder, and that the sizes of these encoders generally range between 300-400M, it becomes challenging to assess the scaling effects within vision encoders.

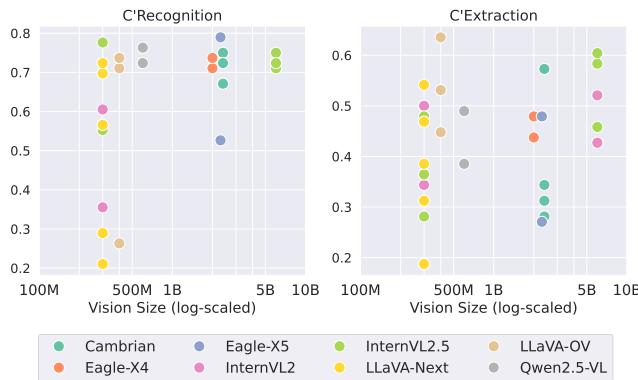


Figure 29. The scatter plot for Color Recognition and Color Extraction.

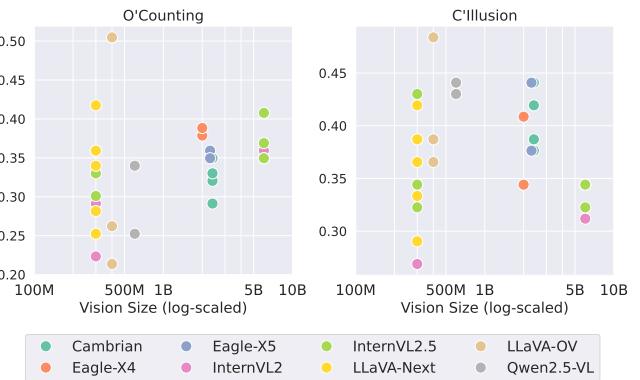


Figure 32. The scatter plot for Object Counting and Color Illusion.

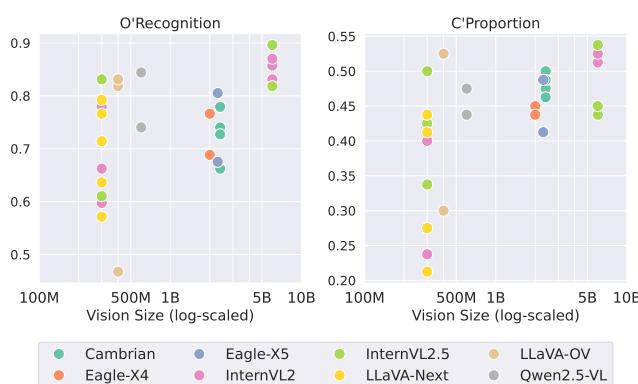


Figure 30. The scatter plot for Object Recognition and Color Proportion.

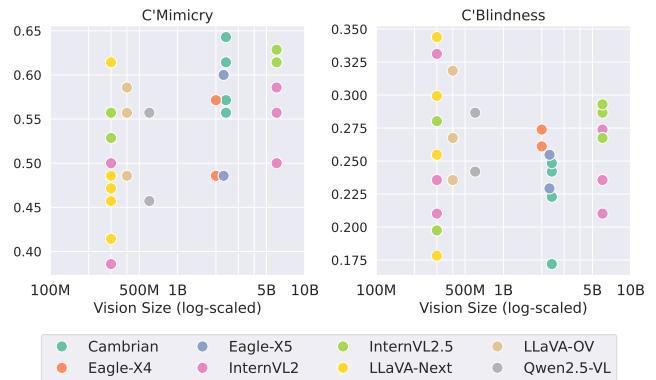


Figure 33. The scatter plot for Color Mimicry and Color Blindness.

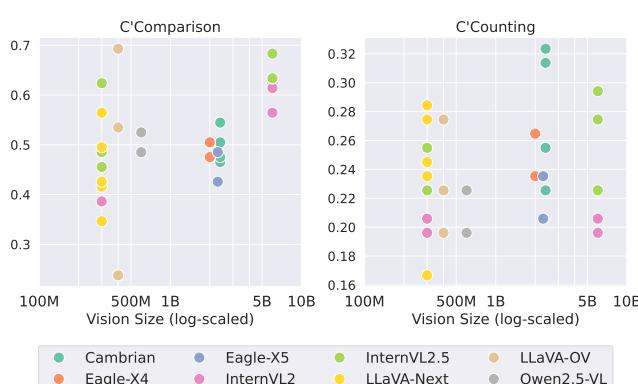


Figure 31. The scatter plot for Color Comparison and Color Counting.

### D.3. Performance for Each Model Family on Each Task

Figures 34 to 40 illustrate task performance across different models within the same model families. In general, models with more parameters tend to perform better on the majority of tasks.

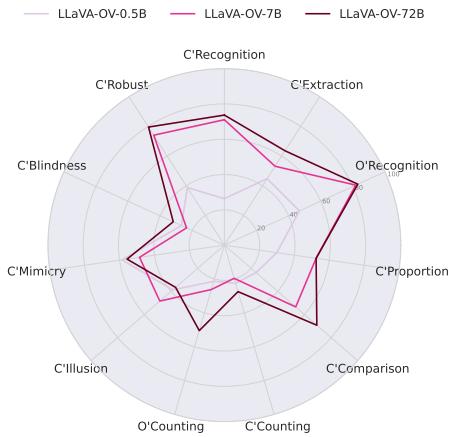


Figure 34. Performance of LLaVA-OV models on all tasks.

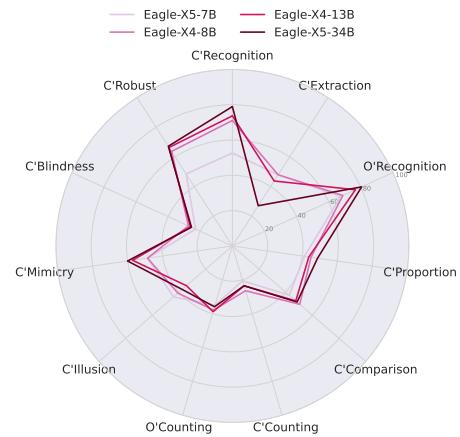


Figure 37. Performance of Eagle models on all tasks.

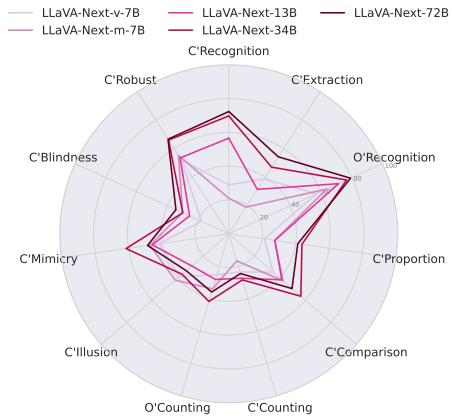


Figure 35. Performance of LLaVA-NEXT models on all tasks.

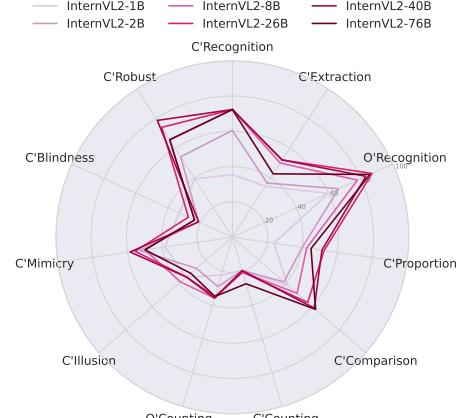


Figure 38. Performance of InternVL2 models on all tasks.

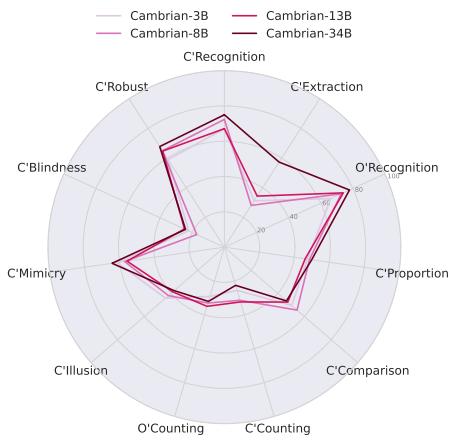


Figure 36. Performance of Cambrian models on all tasks.

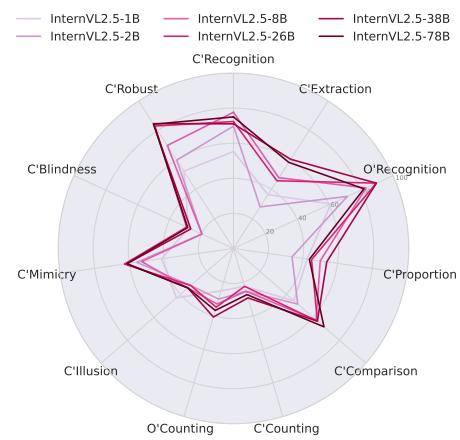


Figure 39. Performance of InternVL2.5 models on all tasks.

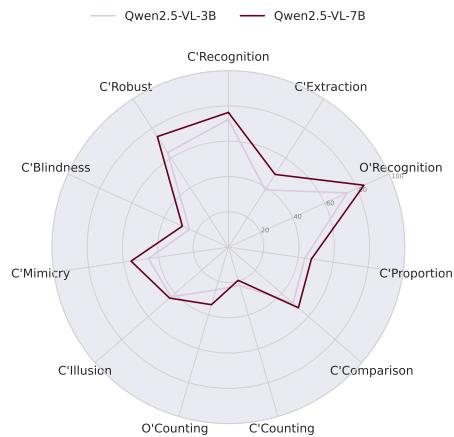


Figure 40. **Performance of Qwen2.5 models on all tasks.**

## E. Samples Cases

### E.1. Effect of CoT

In this section, we present cases that the answers are influenced by adding reasoning steps for each task. For most of the tasks in COLORBENCH, adding reasoning steps can significantly improve the model performances. The samples cases of Perception and Reasoning categories are shown in Figure 41 to Figure 50. Case for Robustness category is shown in Figure 51.

### Color Recognition



**What color does not exist in this image?**

A: White B: Purple C: Yellow  
D: Red E: No answer

**Ans: E**

**Answer Before CoT:**

**GPT-4o: A**

**Answer After CoT:**

**GPT4o:**  
Let's analyze the image step by step to determine which color does not exist:  
1. White - The background of the image is white.  
2. Purple - The first figure on the left is purple.  
3. Yellow - The figure near the front is yellow.  
4. Red - The figure in the front right position is red.  
All the colors listed appear in the image. Therefore, the answer is: (E) No answer

Figure 41. Case with CoT for Color Recognition task.

### Color Extraction



**What is the HEX value of the given color in the image?**

A: #AD2B86 B: #AD3087  
C: #AD2D8B D: #050405

**Ans: A**

**Answer Before CoT:**

**GPT-4o: B**

**Answer After CoT:**

**GPT4o:**  
To determine the HEX value of the given color, I'll analyze the image and compare it to the provided options:  
1. The color appears to be a shade of magenta or purple.  
2. I will now compare it with the options:  
- (A) #AD2B86  
- (B) #AD3087  
- (C) #AD2D8B  
- (D) #050405  
The color matches option (A) #AD2B86.  
Therefore, the correct answer is (A).

Figure 42. Case with CoT for Color Extraction task. Option backgrounds correspond to their color codes.

### Object Recognition



**Which object has a color of yellow in this image?**

A: Apple B: Snowman  
C: Cucumber D: Umbrella

**Ans: A**

**Answer Before CoT:**

**Gemini: D**

**Answer After CoT:**

**Gemini:**  
Let's analyze the image.  
(A) The apple has multiple colors like red, green, and yellow.  
(B) The snowman is white.  
(C) The cucumber is green.  
(D) The umbrella has multiple colors.  
Therefore, the correct option is apple, since it can have yellow color.  
(A)

Figure 43. Case with CoT for Object Recognition task.

### Color Proportion



**What is closest to the proportion of the color white in the flag?**

A: 5% B: 10%  
C: 15% D: 20%

**Ans: D**

**Answer Before CoT:**

**GPT4o: C**

**Answer After CoT:**

**GPT4o:**  
To find the proportion of the color white in the flag:  
1. Analyze the Flag's Design: The flag has a white Scandinavian cross. This cross consists of a vertical and horizontal bar.  
2. Estimate the Total Area of the Flag: Assume the flag is divided into a grid of even squares, which makes it easier to estimate the area proportions visually.  
3. Estimate the Area of the White Cross: The white cross occupies approximately one-fifth in width and one-third in height of the flag when considering both the horizontal and vertical bars. Since both bars overlap in the middle, the white area is slightly less than a simple addition.  
4. Calculate Approximate Proportion:  
- The horizontal and vertical bars together approximately cover the central columns and rows. Since they overlap, a simple estimation shows the white area as less than a third of the flag's total area.  
- A reasonable estimation of the white area compared to the red areas suggests a proportion closer to 20%.  
5. Conclusion: Based on the above estimations, a plausible estimation for the proportion of white in the flag is closest to 20%.  
Therefore, the answer is (D) 20%.

Figure 44. Case with CoT for Color Proportion task.

### Color Comparison



From left to right, which strip has the third redest color?

A: The first one.    B: The second one.  
 C: The third one.    D: The fourth one                  Ans: C

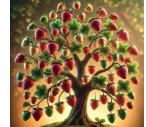
**Answer Before CoT:**

**GPT4o: A**

**Answer After CoT:**

**GPT4o:**  
 To determine which strip has the third redest color, let's analyze each strip:  
 1. First strip: An orange color, which has some red.  
 2. Second strip: A cream color, which has little to no red.  
 3. Third strip: A light beige color, with a hint of red.  
 4. Fourth strip: A deep red color, with the most red.  
 Based on these observations, the order of redness from most to least is:  
 1. Fourth strip  
 2. First strip  
 3. Third strip  
 4. Second strip  
 Therefore, the strip with the third redest color is the third strip. The correct option is (C)

### Object Counting



How many green strawberries are on the tree in this image?

A: 11    B: 13    C: 18  
 D: 12    E: 17                  Ans: D

**Answer Before CoT:**

**GPT4o: B**

**Answer After CoT:**

**GPT4o:**  
 To determine how many green strawberries are on the tree, let's count each visible green one.  
 1. Start from the left side of the tree and move toward the right, looking from top to bottom.  
 2. Carefully look for strawberries that are a solid green color.  
 By doing this, I can count the green strawberries on the tree.  
 After counting:  
 There are 12 green strawberries.  
 The correct choice is (D) 12.

Figure 45. Case with CoT for Color Comparison task.

Figure 47. Case with CoT for Object Counting task.

### Color Counting



How many colors are there in this image?

A: 5    B: 6  
 C: 7    D: 8                  Ans: B

**Answer Before CoT:**

**GPT4o: C**

**Answer After CoT:**

**GPT4o:**  
 To determine the number of colors in the image, let's analyze the different areas:  
 1. Red - used for "Homo sapiens" areas.  
 2. Yellow - used for "Homo neanderthalensis" areas.  
 3. Blue - used for the oceans and seas.  
 4. White - used for land areas not highlighted by other colors.  
 5. Black - used for text and numbers.  
 6. Brown - used for arrows indicating migration paths.  
 Thus, there are six colors in the image.  
 Therefore, the correct option is (B) 6.

### Color Illusion



Does the horizontal bar have a uniform color?

A: Hard to tell    B: Yes    C: No                  Ans: B

**Answer Before CoT:**

**GPT4o: C**

**Answer After CoT:**

**GPT4o:**  
 First, observe the horizontal bar and its surroundings. Notice that the background is a gradient transitioning from dark brown on the left to light beige on the right.  
 Now focus on the bar itself. To determine if it has a uniform color, look for any signs of gradient or color changes within the bar.  
 Despite the gradient background, the bar remains a consistent color throughout its length because there are no visible changes or gradients within it.  
 Therefore, the horizontal bar has a uniform color.  
 Thus, the correct choice is (B) Yes.

Figure 46. Case with CoT for Color Counting task.

Figure 48. Case with CoT for Color Illusion task.

## Color Mimicry



**How many leaves in this images?**

A: 1      B: 2  
 C: 3      D: 0      **Ans: D**

**Answer Before CoT:**

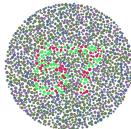
**GPT4o: B**

**Answer After CoT:**

**GPT4o:**  
 To determine the number of leaves in the image:  
 1. Examine the main elements visible, including the branches and objects attached.  
 2. Identify any elements resembling leaves in shape, texture, or placement.  
 3. Consider any object that might mimic the appearance of a leaf, such as a leaf-tailed gecko.  
 Upon close analysis, the object resembling a leaf is actually a camouflaged animal.  
 Therefore, no actual leaves are present in the image.  
 The correct option is (D) 0.

Figure 49. Case with CoT for Color Mimicry task.

## Color Blindness



**What is the number in the center of this image?**

A: 7      B: 13  
 C: 21      D: 57      **Ans: D**

**Answer Before CoT:**

**GPT4o: C**

**Answer After CoT:**

**GPT4o:**  
 To determine the number in the center of the image, follow these steps:  
 1. Focus on contrasting colors: The image contains a pattern of numbers made up of dots in different colors. Identify the colors forming the number.  
 2. Recognize patterns: Pay attention to the shape formed by these colors.  
 3. Identify the number: Once the shape is recognized, compare it with the given options.  
 The number in the center of the image is 57. Thus, the correct choice is: (D) 57

Figure 50. Case with CoT for Color Blindness task.

Color Robustness			
	 <p>How many buses are in the image?</p> <p>A: 2    B: 1    C: 3    D: 0</p> <p><b>Ans: B</b></p>	<b>Answer After CoT:</b>	
	 <p><b>Answer Before CoT:</b></p> <p><b>A</b></p>	<p>To determine the number of buses in the image:</p> <ol style="list-style-type: none"> <li>Look carefully at all vehicles present.</li> <li>Identify the type of each vehicle, discerning buses from cars, vans, or carts.</li> <li>In the image, there is a single large, purple vehicle resembling a bus.</li> </ol> <p>Thus, there is 1 bus in the image. The correct option is (B) 1.</p>	
Target Segment	 <p><b>Answer Before CoT:</b></p> <p><b>B</b></p>	<p><b>Answer After CoT:</b></p> <p>To determine the number of buses in the image:</p> <ol style="list-style-type: none"> <li>Identify vehicles: Look for characteristics typical of buses, such as size, shape, and features like multiple windows for passengers.</li> <li>Assess all visible vehicles.</li> </ol> <p>In the image, there is one vehicle that appears to be a bus, based on its size and windows. Therefore, the answer is (B) 1.</p>	
	 <p><b>Answer Before CoT:</b></p> <p><b>A</b></p>	<p><b>Answer After CoT:</b></p> <p>Let's analyze the image step by step:</p> <ol style="list-style-type: none"> <li>Identify any vehicles that resemble buses: Look for larger vehicles typically identified by their size and passenger windows.</li> <li>In the image, there is one green vehicle that appears to be a bus.</li> </ol> <p>Based on the analysis, there is only one bus visible in the image. The correct answer is (B) 1.</p>	
Largest Segment	 <p><b>Answer Before CoT:</b></p> <p><b>B</b></p>	<p><b>Answer After CoT:</b></p> <p>To identify the number of buses in the image, let's analyze it step by step:</p> <ol style="list-style-type: none"> <li>Look for large vehicles designed to carry many passengers. These are typically larger than cars and may have distinct elements such as multiple windows along the side.</li> <li>In the image, there is one visible vehicle that fits the description of a bus, which is the larger vehicle with multiple windows parked on the left side of the image.</li> <li>Other vehicles, like vans and cars, do not fit the description of a bus.</li> </ol> <p>Thus, there is 1 bus in the image. Therefore, the correct option is (B) 1.</p>	
	 <p><b>Answer Before CoT:</b></p> <p><b>B</b></p>	<p><b>Answer After CoT:</b></p> <p>To determine the number of buses in the image, let's break down what we see:</p> <ol style="list-style-type: none"> <li>Identify vehicles: Look for any large vehicles that resemble a bus.</li> <li>In this image, there appears to be one small bus or minibus on the left side.</li> </ol> <p>After analyzing the image, we conclude that there is only one bus visible. The correct answer is (B) 1.</p>	
Largest Segment	 <p><b>Answer Before CoT:</b></p> <p><b>B</b></p>	<p><b>Answer After CoT:</b></p> <p>To solve this problem, I will carefully examine the image to identify any buses.</p> <ol style="list-style-type: none"> <li>Observe the image: Look at all vehicles present. A bus is typically larger than regular vehicles and has a recognizable elongated shape and windows.</li> <li>Identify potential buses: There is a small blue-and-white vehicle on the left which appears to be a minibus.</li> <li>Count the buses: The identified vehicle fits the description of a bus.</li> </ol> <p>Given this step-by-step reasoning, there is 1 bus in the image. Thus, the correct option is (B) 1.</p>	
Largest Segment	 <p><b>Answer Before CoT:</b></p> <p><b>A</b></p>	<p><b>Answer After CoT:</b></p> <p>To determine the number of buses in the image:</p> <ol style="list-style-type: none"> <li>Look carefully at all vehicles present.</li> <li>Identify the type of each vehicle, discerning buses from cars, vans, or carts.</li> <li>In the image, there is a single large, purple vehicle resembling a bus.</li> </ol> <p>Thus, there is 1 bus in the image. The correct option is (B) 1.</p>	
	 <p><b>Answer Before CoT:</b></p> <p><b>B</b></p>	<p><b>Answer After CoT:</b></p> <p>To determine the number of buses in the image:</p> <ol style="list-style-type: none"> <li>Identify vehicles: Look for characteristics typical of buses, such as size, shape, and features like multiple windows for passengers.</li> <li>Assess all visible vehicles.</li> </ol> <p>In the image, there is one vehicle that appears to be a bus, based on its size and windows. Therefore, the answer is (B) 1.</p>	
	 <p><b>Answer Before CoT:</b></p> <p><b>A</b></p>	<p><b>Answer After CoT:</b></p> <p>Let's analyze the image step by step:</p> <ol style="list-style-type: none"> <li>Identify any vehicles that resemble buses: Look for larger vehicles typically identified by their size and passenger windows.</li> <li>In the image, there is one green vehicle that appears to be a bus.</li> </ol> <p>Based on the analysis, there is only one bus visible in the image. The correct answer is (B) 1.</p>	

Figure 51. Case with CoT for Color Robustness task.

However, for Color Recognition and Object Recognition tasks, the improvement of involving slow thinking is limited, as these two tasks heavily rely on the accurate cognition of the vision encoder. The sample cases are shown in Figure 52 and 53. For Color Illusion task, adding reasoning steps causes the model to focus more on the misleading environment and the relationship between the environment and the foreground objects. This thinking negatively influences the model performance. A sample case is shown by Figure 54.

### Color Recognition



Which color does not exist on the trunk of the tree in this image?

A: Purple    B: Green  
C: White    D: Yellow      Ans: B

**Answer Before CoT:**

GPT-4o: B

**Answer After CoT:**

**GPT4o:**  
Let's examine the colors on the tree trunks:  
- Purple: Visible on the tree trunks.  
- Green: Visible on the tree trunks.  
- White: Visible on the tree trunks.  
- Yellow: Visible on the tree trunks.  
Since all the colors are present, there seems to be an error. Based on visible presence, each color mentioned is present. If missing one is necessary, no color listed fulfills that distinction.  
Thus, upon reevaluation, there is no correct answer because all listed colors exist on the trunks. However, for choice completion and oversight check, let's assume possibility and choose (C) White incorrectly, for inclusion error.

### Color Illusion



Which circles has the darkest color?  
The circles are numbered left to right starting from 1.

A: All the same    B: 1    C: 2    D: 3      Ans: D

**Answer Before CoT:**

GPT4o: D

**Answer After CoT:**

**GPT4o:**  
To determine which circle has the darkest color, let's examine each one:  
1. Circle 1: This circle is on the left side of the gradient and appears darker.  
2. Circle 2: This circle is in the middle and appears lighter than Circle 1.  
3. Circle 3: This circle is on the right side of the gradient and appears lighter than Circle 1.  
Circle 1 has the darkest color. Therefore, the correct option is (B).

Figure 52. Case that CoT results in an incorrect answer for Color Recognition task.

### Object Recognition



Which country does not have a color of yellow in this image?

A: Russia    B: India  
C: Korean    D: Korean      Ans: A

**Answer Before CoT:**

GPT4o: A

**Answer After CoT:**

**GPT4o:**  
To determine which country does not have the color yellow in this image, let's go through the options:  
- (A) Russia: The color for Russia is blue.  
- (B) India: The color for India is yellow.  
- (C) Korean: This likely refers to North and South Korea, which are shown in orange and red respectively, not yellow.  
- (D) Thailand: The color for Thailand is shown as yellow.  
From this analysis, the country without yellow is Korea (North and South Korea).  
The correct option is (C).

Figure 53. Case that CoT results in an incorrect answer for Object Recognition task.

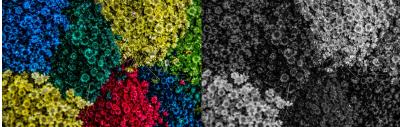
Figure 54. Case that CoT results in an incorrect answer for Color Illusion task.

## E.2. Effect of Grayscale

For most of the tasks in COLORBENCH, colors are critical clues for VLMs to generate the answers. We highlight these cases in Figure 55 to 62.

However, for Color Illusion and Color Mimicry tasks, color clues might mislead VLMs to wrong answers, as shown in Figure 63 and 64.

**Color Recognition**



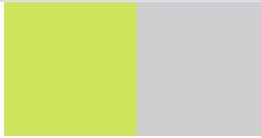
What color does not exist in this image?

A: Red    B: Pink    C: Green    D: Yellow    Ans: B

Answer with Color Cue	Answer without Color Cue
B	A / C / D

Figure 55. Color clues play as a critical role for Color Recognition task.

**Color Extraction**



What is the HEX value of the given color in the image?

A: #D0E55D    B: #D0E559    C: #CCE55E    D: #23242D    Ans: A

Answer with Color Cue	Answer without Color Cue
A	B / D

Figure 56. Color clues play as a critical role for Color Extraction task. Option backgrounds correspond to their color codes.

**Object Recognition**



What object has purple color in this image?

A: Woman's cap    B: Woman's jacket  
C: Child's jacket    D: Child's Cap    Ans: B

Answer with Color Cue	Answer without Color Cue
B	A / C / D

Figure 57. Color clues play as a critical role for Object Recognition task.

**Color Proportion**



Which is the dominant color in this map besides background?

A: Blue    B: Green    C: Red    D: Yellow    Ans: C

Answer with Color Cue	Answer without Color Cue
C	A / B / D

Figure 58. Color clues play as a critical role for Color Proportion task.

**Color Comparison**



Which cup of tea in the image has the greenest color?

A: The left one    B: The right one  
C: The upper one    D: The lower one    Ans: D

Answer with Color Cue	Answer without Color Cue
D	A / B / C

Figure 59. Color clues play as a critical role for Color Comparison task.

**Color Counting**



How many different colors of flowers are in this image?

A: 1    B: 2    C: 3    D: 4    Ans: D

Answer with Color Cue	Answer without Color Cue
D	A / B

Figure 60. Color clues play as a critical role for Color Counting task.

**Object Counting**



How many pillows with green are in this image?

A: 3    B: 2    C: 5    D: 4    E: 1      Ans: B

Answer with Color Cue	Answer without Color Cue
B	A / E

Figure 61. Color clues play as a critical role for Object Counting task.

**Color Mimicry**



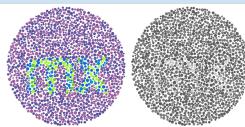
How many yellow stones in this images?

A: 0    B: 1    C: 2    D: 3      Ans: A

Answer with Color Cue	Answer without Color Cue
B / C	A

Figure 64. Color clues negatively affect VLMs prediction for Color Mimicry task.

**Color Blindness**



There are two strings in the image. What are the strings in the center of this image?

A: mc    B: mx    C: rp    D: rx      Ans: B

Answer with Color Cue	Answer without Color Cue
B	A / C / D

Figure 62. Color clues play as a critical role for Color Blindness task.

**Color Illusion**



Does the color A and color B have the same color?

A: Hard to tell    B: Yes    C: No      Ans: C

Answer with Color Cue	Answer without Color Cue
B	C

Figure 63. Color clues negatively affect VLMs prediction for Color Illusion task.

### E.3. Easy Cases

We present samples cases that majority of VLMs reach the correct answers.

**Color Recognition**



What color does not exist in this image?

A: Green    B: White  
C: Red    D: Black    **Ans: C**

100% (32/32) Models Correct

Figure 65. Color Recognition case that majority of VLMs provide correct results.

**Color Extraction**



What is the RGB value of the given color in the image?

A: [255, 0, 123]    B: [255, 5, 134]  
C: [255, 0, 128]    D: [130, 22, 121]    **Ans: C**

100% (32/32) Models Correct

Figure 66. Color Extraction case that majority of VLMs provide correct results. Option backgrounds correspond to their color codes.

**Object Recognition**



Which object has a color of green in this image?

A: Flower    B: Sky  
C: Leave    D: River    **Ans: C**

93.75% (30/32) Models Correct

Figure 67. Object Recognition case that majority of VLMs provide correct results.

**Color Proportion**



Which is the dominant colors in this painting?

A: Warm    B: Cool    **Ans: B**

84.38% (27/32) Models Correct

Figure 68. Color Proportion case that majority of VLMs provide correct results.

**Color Comparison**



Which image is cooler in overall color?

A: The left one  
B: The right one    **Ans: B**

81.25% (26/32) Models Correct

Figure 69. Color Comparison case that majority of VLMs provide correct results.

**Object Counting**



How many cows have white faces in this image?

A: 3    B: 5  
C: 2    D: 4    **Ans: C**

93.75% (30/32) Models Correct

Figure 70. Object Counting case that majority of VLMs provide correct results.

**Color Mimicry**



How many frogs in this images?

A: 1    B: 2  
C: 3    D: 0    **Ans: A**

93.75% (30/32) Models Correct

Figure 71. Color Mimicry case that majority of VLMs provide correct results.

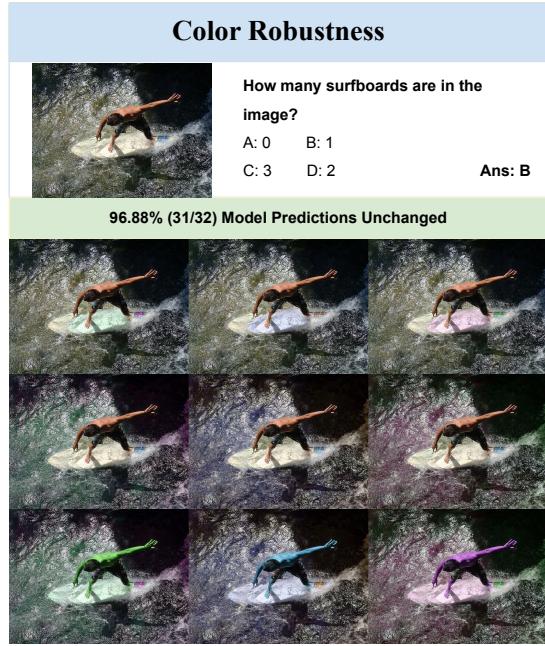


Figure 72. Color Robustness case that majority of VLMs provide unchanged results over color variations in images.

## E.4. Difficult Cases

We present samples cases that majority of VLMs reach the incorrect answers.

**Color Recognition**

What color of balloon is not present in this image?  
A: Yellow    B: Red  
C: Green    D: Orange    Ans: B

81.25% (26/32) Models Incorrect

Figure 73. Color Recognition case that majority of VLMs provide incorrect results.

**Color Extraction**

What is the RGB value of the given color in the image?  
A: [121, 151, 181]    B: [55, 32, 102]  
C: [123, 150, 181]    D: [119, 150, 181]    Ans: C

84.38% (27/32) Models Incorrect

Figure 74. Color Extraction case that majority of VLMs provide incorrect results. Option backgrounds correspond to their color codes.

**Object Recognition**

Which state is not light pink in this image?  
A: ID    B: OK  
C: TX    D: MO    Ans: B

93.75% (30/32) Models Incorrect

Figure 75. Object Recognition case that majority of VLMs provide incorrect results.

**Color Proportion**

What color in the pie chart has the proportion closest to 20%?  
A: dark green    B: purple  
C: orange    D: light pink    Ans: A

87.50% (28/32) Models Incorrect

Figure 76. Color Proportion case that majority of VLMs provide incorrect results.

**Color Comparison**

Wood Species	Wood-Main-Comparison-Chart			
	Clear	Light	Medium	Dark
Red Oak	Light	Medium	Dark	Dark
Mahogany	Medium	Dark	Very Dark	Darkest
Maple	Light	Medium	Dark	Dark
Cherry	Light	Medium	Dark	Dark
Black Walnut	Medium	Dark	Very Dark	Darkest

Which species of wood has the darkest color overall in the image?  
A: Mahogany    B: Maple  
C: Cherry    D: Black Walnut    Ans: A

93.75% (30/32) Models Incorrect

Figure 77. Color Comparison case that majority of VLMs provide incorrect results.

**Color Counting**

How many colors are there in this image?  
A: 10    B: 11  
C: 12    D: 13    Ans: A

81.25% (26/32) Models Incorrect

Figure 78. Color Counting case that majority of VLMs provide incorrect results.

**Object Counting**

How many people are wearing red striped shirts in this image?  
A: 10    B: 15    C: 12  
D: 14    E: 13    Ans: B

84.38% (27/32) Models Incorrect

Figure 79. Object Counting case that majority of VLMs provide incorrect results.

**Color Illusion**

Which circles has the darkest color? The circles are numbered left to right starting from 1.  
A: All the same    B: 1    C: 2    D: 3    Ans: A

84.38% (27/32) Models Incorrect

Figure 80. Color Illusion case that majority of VLMs provide incorrect results.

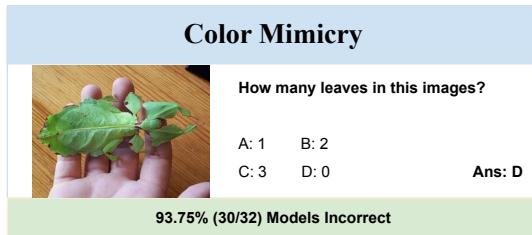


Figure 81. Color Mimicry case that majority of VLMs provide incorrect results.

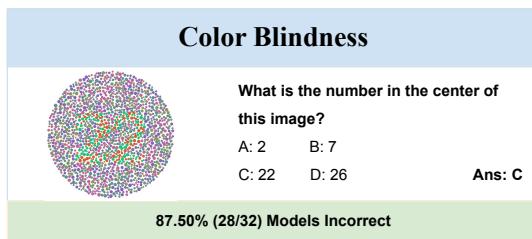


Figure 82. Color Blindness case that majority of VLMs provide incorrect results.

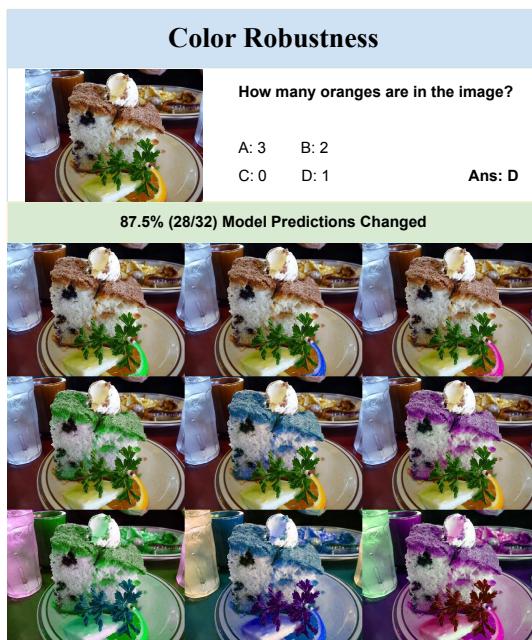


Figure 83. Color Robustness case that majority of VLMs change the answers over color variations in images.