Color Perception in Vision-Language Models: From Feature Extraction to Instruction Tuning Methods

Aya Nuseir[1][0000−0001−7905−4199] and Marc Ebner[1][0000−0003−2725−2454]

University of Greifswald, Institute of Mathematics and Computer Science, Germany
{s-aynuse,marc.ebner}@uni-greifswald.de

Abstract. This study evaluates the color perception capabilities in vision language models (VLM) by studying the impact of adopting three different visual feature extraction methods and the Visual Instruction Tuning method on color understanding. We created a customized Visual Question Answering (VQA) dataset of simple geometric shapes in various colors against a colored background, paired with diverse question types ranging from basic color identification to relational reasoning. Twelve VLMs are used in our evaluation and divided into two groups: the first group contains VLP models that employ three distinct visual feature extraction approaches: region-based (VisualBERT-FRCNN), grid-based (VisualBERT-ResNet), and patch-based (ViLT, BLIP). The second group includes recent state-of-the-art Multi-modal Large Language Models (MLLMs) that adopt visual instruction tuning. The performance of the two groups varies across different color perception tasks, where the first group failed to identify the color categorization task and showed an inability to identify the cyan and magenta colors. The second group, including models like BLIP-2, InstructBLIP, LLaVA-1.5, Gemma 3, Qwen2.5-VL, GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano, show varying performances across the color perception tasks. BLIP-2, InstructBLIP, and LLaVA-1.5 were unable to identify the cyan and magenta colors, but can identify the color category. The performance of GPT-4.1 outperforms all the adopted models, including the Qwen2.5-VL models, which shows competitive performance despite being open source.

Keywords: Vision-Language Models · Visual Question Answering · color identification · visual feature extraction · Visual Instruction Tuning.

1 Introduction

Over the last decade, Vision-Language Models (VLMs) have captured researchers' attention and witnessed rapid advances in their ability to understand and reason about visual content, making them powerful tools for multimodal applications ranging from image captioning to visual reasoning [1,2]. These models were trained on massive and extensive image-text datasets to learn the representations of both visual and textual data[3]. Since the VLMs are becoming growingly integrated into real-world applications, their ability to perceive and reason about fundamental visual elements becomes important [4,5]. One of these elements is color perception[6], which involves using top-down knowledge from experience or trained data for bottom-up processing of raw visual data [7]. However, the perception and reasoning about colors represent a particular challenge in VLMs[13]. While humans naturally and easily process color information like categorising colors, understanding relationships between colors, and applying color-based reasoning, current vision-language models face several obstacles in reaching similar abilities [31]. Although

color perception is important, most current VLM benchmarks combine it with other visual tasks. That makes it hard to focus on and test how well models understand color on their own. Visual Question Answering (VQA) is considered a VLM downstream task that combines computer vision (CV) and natural language processing (NLP) fields to answer questions about images [14,15]. While great efforts are made to develop robust VQA models that can answer complex questions, there is a gap in the fine-grained understanding of visual elements, such as color perception and understanding. This research fills that gap by creating an evaluation framework to check how well VQA handles color perception. We created a custom dataset consisting of three simple shapes (a triangle, a circle, and a square) in nine different colors, displayed on various colored backgrounds [1]. By making sure the color of the shapes is distinct from the background, we created an environment that focuses only on testing color perception without including other visual reasoning challenges. Our question set is diverse, covering four different features of color understanding, including color identification, which focuses on nine specific colors. The second feature is the color categorization, where nine colors are categorized into three categories: primary, secondary, and achromatic. The third and fourth features cover the relationship between colors, where the third feature focuses on whether the color is warm or cool. The final feature is complementary colors, which are pairs of colors that, when combined, produce a grayscale color (such as white or black) [8], Figure 1 shows the complementary color pairs. Notably, we adopted the RGB additive color model when developing our dataset. We use the RGB color model because it directly relates to how the human visual system perceives color and is the fundamental color space for electronic devices that display images [9,10].

Fig. 1. Complementary colors in the RGB additive color, where the first pair is red-cyan (the upper pair), the second pair is green–magenta (the middle pair), and finally, the third pair is blue–yellow.

Through this research, we classify our experiments into two groups; the first group analyzes the performance of several important Vision-Language Models that adopt different visual feature extraction approaches: ViLT [17], VisualBERT [16] (with both FRCNN and ResNet backbones), and BLIP [18]. These models adopt three different types of visual feature extraction: (1) Region Feature-based models like VisualBERT-FRCNN, which extract features from main object regions; (2) Patch Feature-based models such as ViLT and BLIP, which process images as sequences of equal patches; and (3) Grid Feature-based models like VisualBERT-ResNet, which extract features from uniform grid cells across the entire image. By adopting these models with different visual feature extraction methods, we investigate their impact on color perception. Figure 2 illustrates

[1] https://github.com/ASNuseir/Color-Perception-in-VL/tree/main

the visual feature extraction methods by example, utilizing an input image that contains one colored shape. The second group examines the performance of the recent state-of-the-art Multi-modal Large Language Models (MLLMs) that adopt visual instruction tuning. The process begins with data preparation, where the input data is formatted as {Instruction, Input, Output}. Instruction represents a textual description of the task. Input consists of $\{< image >, < text >\}$, and Output is the response following the given instruction. The general architecture for these models comprises three elements: a visual encoder for processing images, an LLM for understanding and generating text, and a projector that aligns embeddings between the two modalities [12]. Through the work in this group, we evaluated state-of-the-art MLLMs, including BLIP-2 [19], InstructBLIP [20], LLaVA-1.5 [21], Gemma 3 [22], Qwen2.5-VL [23], GPT-4.1, GPT-4.1 mini and GPT-4.1 nano [24] (the last three models are considered closed-source commercial models).

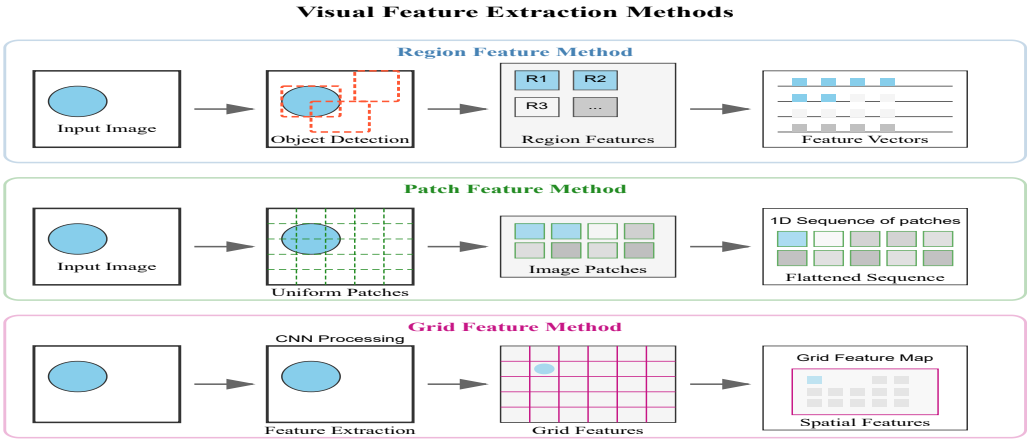

Fig. 2. The main three visual feature extraction methods used in VLMs.

The contributions of this paper are : (1) we present a constructed dataset for evaluating color perception in VLMs that separates the color perception from other visual reasoning tasks; (2) We created a structure of color perception questions that covers identification, categorization, and relational reasoning; (3) We provide an evaluation of four VLMs that employ three different visual feature extraction methods; and (4) we evaluate the recent state-of-the-art MLLMs that adopt visual instruction tuning method. In this work, we use 'color perception' to cover all color understanding capabilities, from basic identification to more complex relational tasks.

This paper is organized as follows. Section 2 provides an overview of VLP models that address the topic of color perception. Section 3 describes the features and characteristics of the developed datasets. Section 4 reviews the models that we have used in evaluating the created datasets. Section 5 presents the evaluation of the experiments, and the paper concludes in Section 6.

2 Related Work

Recently, there has been an increasing interest in studying the capability of VLMs to understand visual attributes, including color perception. Liang et al. [25] introduced ColorBench, a comprehensive benchmark specifically designed to assess the color understanding capabilities of VLMs across perception, reasoning, and robustness. The evaluation of 32 VLMs with different architectures indicated that while scaling laws generally hold, language models play a more crucial role than vision encoders in color understanding tasks. Hyeon-Woo et al. [26] investigate how a VLM perceives images by an eye examination process in the context of color, shape, and semantics. For this purpose, they introduce Learning ElemeNt for visual Sensory (LENS), a synthetic dataset categorized into basic visual elements such as color, shape, and semantics. They found that VLMs have varying sensitivity to different colors while consistently showing insensitivity to green across different VLMs. While ColorBench assesses whether VLMs can use color information in practical applications, our dataset identifies specific color perception mechanisms that are fit or not. Both approaches are scientifically necessary: ColorBench shows practical limitations, while our dataset defines their causes. For example, a model failing our cyan identification task will likely struggle with ColorBench's cyan-related questions. However, ColorBench's complex scenes cannot isolate whether failures are from color perception deficits, object recognition errors, or reasoning limitations. Our dataset provides this diagnostic specificity.

Akbarinia [30] analyzed categorical color perception in artificial neural networks (ANNs). The study evaluates how unimodal vision models (ImageNet-trained networks) and multimodal vision-language models (CLIP) represent color categories. Results show that vision networks alone explain approximately 80% of human-like color categorization. At the same time, language-modulated models account for the remainder, suggesting that color categories are a language-independent representation, although linguistic color terms partly shape it during their development. A second set of experiments using Taskonomy networks demonstrates that human-like color categories mostly emerge in networks trained on high-level semantic and 3D tasks rather than low-level 2D tasks, indicating that these categories serve functional roles in vision. Arias et al. [27] explored how color is learned in CLIP. Through their study, they investigated if color is encoded as an object attribute by asking questions on a basic color/object dataset. They examined the capability of CLIP to recognize and read color with a Stroop test dataset. Although existing work is considered foundational for evaluating the VLP model and has begun investigating specific perceptual capabilities, there is a significant gap in the evaluation of color perception using custom VQA benchmarks. Most available benchmarks focus on general multimodal understanding or provide limited tasks related to color. This demands the development of custom VQA datasets that are designed to evaluate color perception across various visual contexts and represent a significant advancement in VLM evaluation methodology.

3 Dataset

In this paper, we present a custom dataset designed to evaluate the understanding of color features in state-of-the-art vision-language models. Our dataset consists of images of single geometric shapes (circle, square, triangle) with one of nine colors against colored backgrounds, where the background and the geometric shape should not have the same color. The list of colors we adopted is: red, green, blue, cyan, magenta, yellow, white, gray, and black. All images feature a 3x3 matrix of patches, and each one of the shapes' sizes is 32x32 pixels, resulting in images with 96x96 pixels. The 3x3 matrix offers us nine positions for the geometric shape, and since we adopt nine possible colors for the shape, eight colors for the background, and nine positions to place the shape, we obtain 648 images for each of the three geometric shapes. As a result, the dataset has 1944 images. As we mentioned, we avoid identical colors; some combinations may appear visually similar. Our color selection ensures perceptual separation with 60° hue differences between any chromatic color in HSV space. For example, red (0°) and magenta (300°) are separated by 60° and the same perceptual distance as red-yellow, yellow-green, green-cyan, cyan-blue, blue-magenta, and magenta-red. Through this uniform separation, we ensure that real color recognition is required rather than depending on contrast differences.

3.1 Color Space and Geometric Shape Selection

Our color selection contains both primary colors (red, green, blue), secondary colors (cyan, magenta, yellow), and achromatic colors (white, gray, black), and the purpose of this selection is to provide comprehensive coverage of fundamental color categories. We ensure equal representation across the RGB color space (primary, secondary, and achromatic) to avoid bias toward specific color categories. This design enables us to evaluate VLMs' understanding of different color classifications and perceptual categories. Keeping the shape and background colors separate ensures that all evaluation methods require actual color differentiation rather than relying on illumination or intensity differences. Our purpose in selecting the three basic geometric shapes (circle, square, triangle) is to offer variety in shapes while keeping simplicity, which helps ensure that color perception is the primary focus of the evaluation. The 3x3 grid positioning system allows us to assess whether spatial location affects the accuracy of color perception. Figure 3 shows samples from the dataset images.

Fig. 3. samples from the dataset images, where three geometric shapes were adopted during the creation of our dataset

3.2 Question Design

Each image is paired with a set of questions focusing on different features of color perception, including basic color identification, color categorization (primary/secondary or achromatic), and color relationships (complementary colors, warm/cool). Our question set is diverse to cover different features of color understanding, like:

- Basic recognition: ' What is the color of the circle in the image?'
- Color categorization: 'Is the background color a primary, secondary, or achromatic color?'
- Temperature perception: 'Is the circle colored with a warm color?'
- Relationships between colors: 'What is the complementary color of the green square?' or 'Are the square color and the background color complementary colors?'

The total number of questions generated is 22,032, distributed as shown in Figure 4. For the color identification question type, each color appears 432 times, representing 17.5% of the dataset questions, and each category of three-color categorization occurs 1296 times, representing 17.5% of the dataset questions. Two forms of questions are used for the complementary color question: a binary answer and a color name answer. The questions with color names are answered equally, and each of the six complementary colors is answered 216 times. This type of question represents 12% of the dataset questions. For the warm/cool questions, we use a binary question form, which represents 12% of the dataset questions. Finally, the remaining dataset questions are in binary form and represent 41% of the dataset questions; these questions concern whether the color of the shape or background is visible in the image. Our four question types (identification, categorization, temperature, complementary) range from basic recognition to a higher level of color reasoning, enabling diagnostic evaluation of where specific models fail in color processing.

3.3 Dataset Characteristics and Scope

Color perception in vision-language models requires a human-like understanding of visual content and runs through complex interactions between visual feature extraction, semantic understanding, and linguistic representation [11]. We developed our synthetic dataset to isolate the specific contribution of color processing from other factors. Real-world images introduce multiple issues: texture variations, lighting conditions, object semantics, and scene complexity interact with color perception. For example, a "green banana" requires the model to employ object recognition (banana), color identification (green), and reasoning (bananas are generally yellow/green). Failure on such tasks could be attributed to object recognition errors, color perception deficits, or reasoning failures, and our geometric shapes eliminate these obstacles, creating a clean basis for evaluating color. The synthetic images enable us to evaluate how color perception works.

4 VLP Models and visual feature extraction methods

Our evaluation of the adopted models is divided into two groups of experiments: the first group focuses on selecting VLP models using three different visual feature extraction

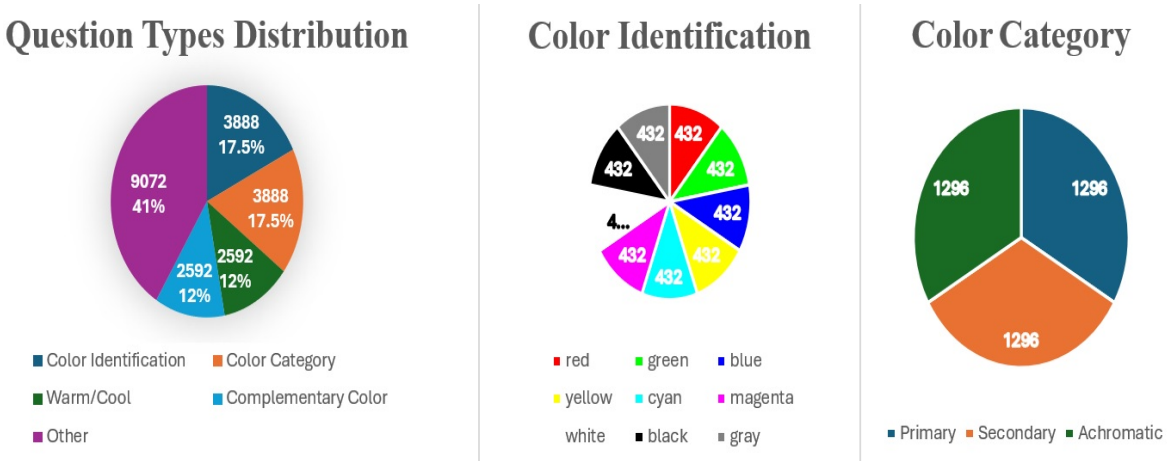

Fig. 4. Distribution of question types and related answers in the dataset. The left chart shows the overall distribution of question types, with "Other" (9,072 questions), followed by "Color Identification" and "Color Category" (each with 3,888 questions), and "Warm/Cool" and "Complementary Color" (each with 2,592 questions). The middle chart displays the distribution of answers for Color Identification questions, showing equal representation across nine color categories with 432 instances each. The chart to the right presents the distribution of Color Category question answers, with equal representation across three categories.

methods. The second group is based on the evaluation of the Visual Instruction tuning models. For the first group of experiments, we adopt four Vision-Language Models that employ different approaches to visual feature extraction. By comparing models across these visual feature extraction methods, we aim to understand how architectural choices in visual feature extraction affect a model's ability to perceive and reason about color features. The second group analyzes recent models whose architecture involves instruction tuning within Multi-modal Large Language Models (MLLMs). In this section, we provide an overview of the adopted models, classifying them into two groups: models with three visual feature extraction methods and Visual Instruction tuning models. All models are implemented in PyTorch, and the Hugging Face transformers library for the open-source models is employed to ensure reproducibility and standardization across experiments.

4.1 Models with three Visual Feature Extraction methods

Region-based Feature Method: For this method, we adopted VisualBERT [16] with Faster-RCNN (FRCNN) [28] backbone. FRCNN identifies regions of interest (ROIs) within the image, then extracts the image features from these regions to align with text features (that are extracted using BERT transformer [29]) through multiple transformer layers. This object-based approach focuses on detected objects rather than processing the entire image uniformly.

Grid-based Feature Method: The other version of VisualBERT model is the one that uses a ResNet backbone to extract grid features from the image, by producing a uniform

representation of the visual raw input without detecting objects. These grid features are then processed together with text through a transformer architecture. Unlike FRCNN, this approach captures information from the whole image in a regular grid frame.

Patch-based Feature Method: ViLT [17] patch projection embedding Vision Transformers (ViT) [32], dividing input images into fixed-size patches and linearly projecting these patches into embeddings. These patch embeddings are then combined with text embedding and fed into a unified transformer architecture to learn the interaction between images and text. BLIP [18] uses a ViT backbone [32] to extract features from images, which are then processed through a Multimodal mixture of Encoder-Decoder (MED) architecture. The MED operates in three modes: unimodal encoder, image-grounded text encoder, and image-grounded text decoder. The model is jointly pre-trained using image-text contrastive learning, image-text matching, and language modeling. BLIP introduces a novel bootstrapping approach that generates synthetic captions to improve training efficiency.

4.2 Visual Instruction Tuning Models

We also evaluate the following recent state-of-the-art MLLMs that utilize visual instruction tuning, starting with BLIP-2 [19], InstructBLIP [20], LLaVA-1.5 [21], Gemma 3 [22], Qwen2.5-VL [23], GPT-4.1, GPT-4.1 mini and GPT-4.1 nano [24]. Table 1 illustrates the number of parameters of each model used in this group and the datasets used in the training. In addition to the datasets that were used in pretraining InstructBLIP, LLaVA-1.5 uses a mixture of datasets that cover both academic task-oriented benchmarks(such as TextVQA[39] and ScienceQA [40]) and vision-language instruction tuning datasets (like LLaVA-Instruct-150K [21], ShareGPT[41]). The Gemma 3 model used a set of multilingual datasets with doubled multilingual data coverage, multimodal image-text pairs, 2T-14T tokens depending on model size. Qwen2.5-VL expanded the volume of their pre-training data to 4.1T multimodal tokens, including interleaved image-text data, grounding data with absolute coordinates, document omni-parsing (HTML format), video data with dynamic frames per second (FPS), and agent interaction data from UI screenshots.

Table 1. The number of model parameters and the adopted pre-trained datasets.

Model	Parameters number	Pre-training dataset
BLIP-2	2.7B	COCO [34], Visual Genome [35], SBU [36], LAION[38],CC3M[37], CC12M
InstructBLIP FlanT5XL	4B	26 datasets cover 11 task
LLaVA-1.5	7B	A mixture of datasets
Gemma 3	4B	A custom corpus (2T-14T tokens), multimodal data
Qwen2.5-VL	7B	4.1T multimodal tokens
GPT-4.1	Not specified	Not specified
GPT-4.1 nano	Not specified	Not specified
GPT-4.1 mini	Not specified	Not specified

5 Discussion

Our evaluation of the first group shows notable differences in how these vision-language models process and reason about color perception. From Table 2, BLIP shows the best performance accuracy (57.87%), followed by ViLT (40.93%) and VisualBERT variants (33.26 % with FRCNN and 32.67 % with ResNet50). However, the performance varied across different question types and color categories, suggesting that each model architecture encodes color information differently. All models show a higher performance on certain questions (yes/no answers) compared to the remaining types of questions. The results show that all models in this group failed to return the color categories, while identifying the color names is easier for these models than determining the complementary color names. BLIP again outperforms the other models with (8.82%) accuracy on questions about warm vs. cool color questions, and the other models in this group show moderate performance close to BLIP, which reveals that BLIP has made a deeper understanding of color properties beyond just naming colors. Notably, BLIP performs worse on complementary color tasks than the other models in the group.

Table 2. Overall Performance Comparison of Vision-Language Models

Model	**Overall Acc**	**Micro F1**	**CI**	**CC**	**CoC**	**WC**	**Other**
Models with three Visual Feature Extraction methods							
VisualBERT (FRCNN backbone)	33.26	0.333	01.67	00.00	05.72	06.00	19.85
VisualBERT (ResNet50 backbone)	32.67	0.327	01.96	00.00	04.94	05.91	19.85
ViLT	40.93	0.409	04.67	00.00	04.28	05.79	26.18
BLIP	57.87	0.579	12.20	00.00	01.37	08.82	35.46
Visual Instruction Tuning Models							
BLIP-2	43.14	0.431	05.40	05.51	01.67	05.89	24.65
InstructBLIP	55.61	0.556	09.69	05.88	04.28	07.82	27.90
LLaVA-1.5	66.18	0.662	12.65	06.20	03.49	08.70	35.12
Gemma 3	58.73	0.587	14.73	10.89	06.84	06.42	19.82
Qwen2.5-VL	86.32	0.863	16.62	13.58	08.38	10.12	37.63
GPT-4.1	**91.69**	**0.917**	**17.39**	**15.45**	**10.30**	**10.81**	**37.72**
GPT-4.1 nano	65.17	0.652	11.96	05.50	06.40	08.04	33.25
GPT-4.1 mini	53.30	0.533	12.11	09.59	10.20	05.91	15.47

CI: Color Identification, CC: Color Categorization, WC: Warm/Cool, CoC: Complementary Color

To provide more detailed insights into this group, we analyzed model performance across the color identification and complementary tasks, as shown in Table 3 and Figure 5, where we calculated the F1-score for each color. Per-color F1-scores reveal dramatic variations in model performance across the nine different colors. Unusually, cyan and magenta are challenging for all models in Group 1, with all achieving F1-scores of 0.000 for both colors, indicating complete failure in identification. BLIP achieved the highest F1-scores for yellow (0.578) and black (0.834) among Group 1 models. The second group significantly improved across the different color perception tasks, especially GPT-4.1 and Qwen2.5-VL models. However, BLIP-2 shows poor performance similar to the models in the first group but shows the ability to recognize the color categories with accuracy

(5.51%). From Table 3, BLIP-2, InstructBLIP, and LLaVA-1.5 are unable to correctly identify cyan and magenta. For instance, instead of magenta, they refer to this color as pink, LLaVA-1.5 refers to cyan as blue, InstructBLIP as white, and Blip-2 refers to it in some cases as blue, in others as green, as demonstrated in Figure 6, which shows the confusion matrices for the InstructBLIP and LLaVA-1.5 models. Although Gemma 3 shows the ability to identify the cyan and magenta colors, its overall performance (58.73%) compared to LLaVA-1.5 (66.18 %) is lower and slightly improved compared to InstructBLIP (55.61%). The inability of the three models (BLIP-2, InstructBLIP, and LLaVA-1.5) to identify the cyan and magenta colors affects their performance in identifying the complementary pairs of color task, where these two colors are part of this task. The rest of the models in this group can identify these two colors (cyan and magenta), which improves their performance in the complementary color task. The two commercial models, GPT-4.1 mini and GPT-4.1 nano, overall performance (respectively 65.17 % and 53.30%) are similar to or less than the LLaVA-1.5(66.18%). However, these two models can identify the cyan and magenta colors, which directly affects the ability to identify the complementary color pairs task. Also, these two models can categorize colors into three categories.

The GPT-4.1 and Qwen2.5-VL models show impressive performances in all color perception tasks, and their performances on warm/cool and the other tasks (binary questions about the color presence in the image) are closely similar; the differences can be seen in the complementary color task, where the GPT-4.1 outperforms the Qwen2.5-VL in recognizing colors like cyan and magenta. However, GPT-4.1 is a commercial model, while Qwen2.5-VL is an open-source model that can be used freely. The slight difference between these two models could be attributed to the quality and quantity of datasets used in their pre-training. However, the Qwen2.5-VL model is a good choice for people looking for open-source and competitive performance similar to or better than commercial models like GPT-4.1 mini and GPT-4.1 nano. To study the performance of the second group on the color categorization task, we compute the model accuracy of each one of the three categories (primary, secondary, and achromatic), where these three categories are distributed equally through the color categorization task(each one of the categories represents 33.33% of this task). Table 4 presents the performance of this group on the color categorization task. The first three models in this group failed to identify the secondary category, which contains yellow, magenta, and cyan colors. InstructBLIP shows a bias toward the primary category, where all of its responses for this task are in the primary category. LLaVA-1.5 shows a slight improvement compared to InstructBLIP, which shows the ability to identify the achromatic category besides the primary category.When compared with the two commercial models, GPT-4.1 mini and GPT-4.1 nano, Gemma 3 outperforms them. Although these two models exceed Gemma -3 in correctly identifying the primary category, Gemma 3 outperforms them within the other two categories. Qwen2.5-VL outperforms GPT-4.1 in identifying the primary category with an accuracy equal to 25.56% (this means that around 76 % of the primary category answers are correctly recognized). In contrast, GPT-4.1 in the secondary and achromatic categories, with 31.14 % and 32.71 %, respectively(that means GPT-4.1 answers the secondary category correctly with an accuracy of around 93% and with an accuracy of around 98% for the achromatic category), is better than the

Qwen2.5-VL model. We measured latency and throughput for GPT-4.1 (API-hosted) and Qwen (locally deployed). Qwen showed lower average latency (0.52 s vs 1.20 s for GPT-4.1) and higher throughput (41.3 vs 18.3 tokens/s), as shown in Table 5. These results indicate that Qwen, which is locally running, can generate faster responses on appropriate hardware. However, the comparison should consider the following points:

– GPT-4.1 results include API overhead and network latency, which increase latency and reduce throughput compared to the raw inference speed of the model.
– Our Qwen results are related to the specific GPU hardware and software environment we used; different setups may yield slower or faster performance.

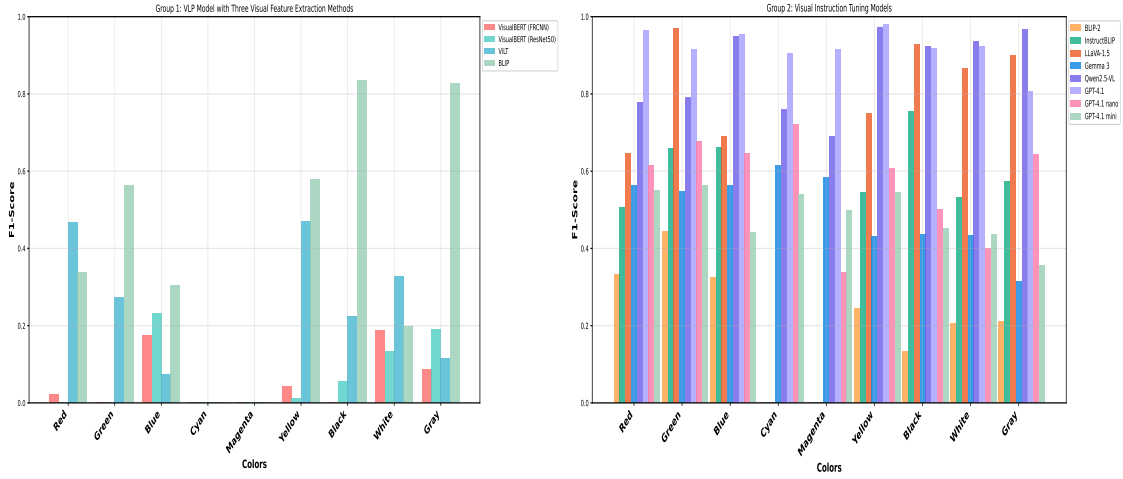

Fig. 5. Per-color F1-score performance for color identification tasks. (A) Group 1 shows poor performance for cyan and magenta across all models. (B) Group 2 exhibits significant improvement, with GPT-4.1 achieving strong performance across all colors.

6 Conclusions and Future Research Directions

Our evaluation of color perception in Vision-Language Models reveals a deep overview of how architectural design choices and the quantity of the datasets utilized in pre-training affect a model's ability to understand and reason about colors. In the first group, BLIP outperforms other models across all color types and question categories, suggesting its patch-based visual feature extraction method may better encode color information. However, all tested models struggled with the categorization task and could not identify the two colors, magenta and cyan, showing an important gap between human and machine color perception capabilities. In the second group, GPT-4.1 outperforms all other models with an overall accuracy of around 91.7%, exceeding the open-source Qwen2.5-VL model, which obtains an overall accuracy of around 86.3%. These results impact future VLM development, we recommend that researchers consider visual feature extraction

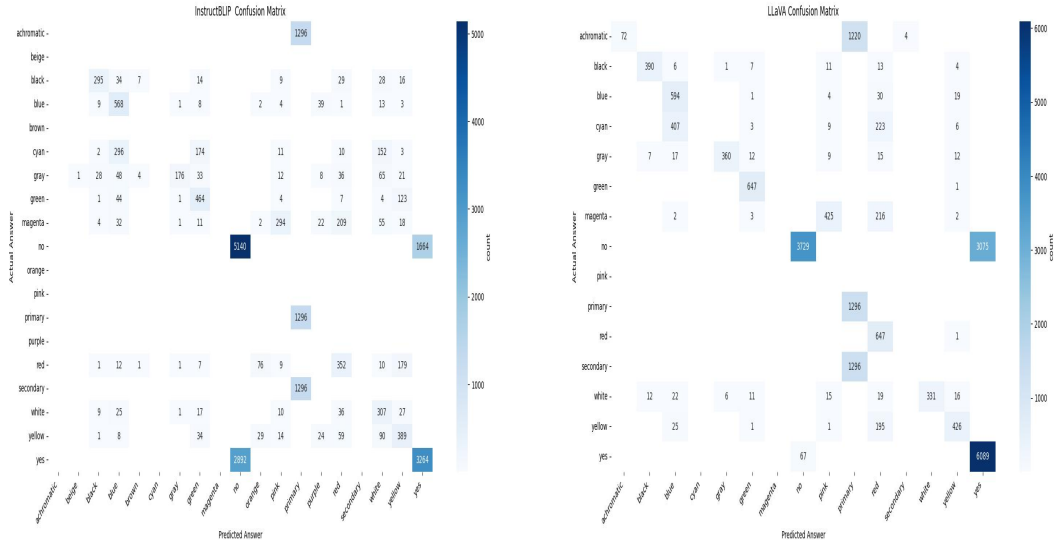

Fig. 6. The confusion matrices of the InstructBLIP and LLaVA-1.5 models show the failure in cyan and magenta colors. InstructBLIP identifies magenta as pink or red and cyan as blue, green, or white. Similarly, LLaVA-1.5 identifies magenta as pink or red, while cyan is blue or red.

Table 3. Per-Color F1-Score for Color Identification and Complementary Color Tasks

Model	Red	Green	Blue	Cyan	Magenta	Yellow	Black	White	Gray
Models with three Visual Feature Extraction methods									
VisualBERT (FRCNN)	0.022	0.000	0.174	0.000	0.000	0.044	0.000	0.187	0.088
VisualBERT (ResNet50)	0.000	0.000	0.233	0.000	0.000	0.011	0.055	0.133	0.190
ViLT	0.468	0.272	0.074	0.000	0.000	0.469	0.224	0.327	0.115
BLIP	0.338	0.564	0.305	0.000	0.000	0.578	0.834	0.199	0.827
Visual Instruction Tuning Models									
BLIP-2	0.333	0.444	0.325	0.000	0.000	0.244	0.134	0.205	0.211
InstructBLIP	0.507	0.658	0.662	0.000	0.000	0.545	0.754	0.531	0.574
LLaVA-1.5	0.645	0.970	0.690	0.000	0.000	0.750	0.927	0.867	0.901
Gemma 3	0.562	0.548	0.564	0.614	0.583	0.430	0.435	0.434	0.314
Qwen2.5-VL	0.781	0.793	0.946	0.761	0.694	0.969	**0.924**	**0.934**	**0.970**
GPT-4.1	**0.969**	**0.924**	**0.960**	**0.906**	**0.908**	**0.977**	0.907	0.922	0.803
GPT-4.1 nano	0.616	0.678	0.646	0.722	0.338	0.608	0.500	0.399	0.644
GPT-4.1 mini	0.550	0.563	0.442	0.541	0.498	0.546	0.451	0.435	0.356

Table 4. Performance of the second group models(visual instruction tuning models) on the color categorization task by category type.

Model	Primary(%)	Secondary (%)	Achromatic (%)
BLIP-2	08.69	00.00	22.55
InstructBLIP	33.33	00.00	00.00
LLaVA-1.5	33.33	00.00	01.85
Gemma 3	18.41	18.87	24.45
Qwen2.5-VL	**25.56**	21.75	29.65
GPT-4.1	23.68	**31.14**	**32.71**
GPT-4.1 nano	19.16	01.95	10.05
GPT-4.1 mini	20.16	16.53	17.66

Table 5. Comparison of Latency (s) and throughput (tokens/s) for GPT-4.1 (API) and Qwen (local). Note: GPT-4.1 includes network transit times and the actual time it takes to come up with the answer; Qwen depends on GPU hardware.

Model	Avg. Latency (s)	Throughput (tokens/s)
Qwen2.5-VL	0.5227	41.32
GPT-4.1	1.2030	18.35

methods when designing models for tasks requiring fine-grained visual understanding. Further, it shows that open-source models, like the Qwen2.5-VL model, can be a good choice for such tasks, whereas the Qwen2.5-VL model shows a competitive performance compared with commercial models like GPT-4.1. In addition, the customized dataset and evaluation framework developed in our study provide valuable tools for evaluating and improving color perception in multimodal AI models. Future work should focus on enhancing models' capabilities for color perception and studying whether similar patterns occur for other main visual elements, rather than color. Future work should vary in complexity from our geometric shapes to natural objects on colored backgrounds and then to complex real-world scenes. This would help us define where the color perception capabilities break down. Although our RGB-based evaluation provides precise control, extending the framework to Lab or HSV color spaces would test whether observed limitations reflect RGB-specific encoding or general color perception deficiencies. Also, future research could examine attention patterns and intermediate representations to understand why specific architectures, like patch-based or region-based architectures, show differential color sensitivity.

References

1. Ho, Huu-Tuong, Luong Vuong Nguyen, Minh-Tien Pham, Quang-Huy Pham, Quang-Duong Tran, Duong Nguyen Minh Huy, and Tri-Hai Nguyen. "A review on vision-language-based approaches: Challenges and applications." *Computers, Materials & Continua*, vol. 82, no. 2, 2025.
2. Huang, Kung-Hsiang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. "Why vision language models struggle with visual arithmetic? Towards enhanced chart and geometry understanding." *arXiv preprint*, arXiv:2502.11492, 2025.
3. Zhou, Mingyang, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. "Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment." In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16485–16494, 2022.
4. Hu, Zhe, Yixiao Ren, Jing Li, and Yu Yin. "VIVA: A benchmark for vision-grounded decision-making with human values." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2294–2311, 2024.
5. Hu, Zhe, Jing Li, and Yu Yin. "When words outperform vision: VLMs can self-improve via text-only training for human-centered decision making." *arXiv preprint*, arXiv:2503.16965, 2025.
6. Xia, Weihao, Raoul De Charette, Cengiz Oztireli, and Jing-Hao Xue. "DREAM: Visual decoding from reversing human visual system." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8226–8235, 2024.
7. Schüz, Simeon, and Sina Zarrieß. "Knowledge supports visual language grounding: A case study on colour terms." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6536–6542, 2020.
8. Wikipedia contributors. "Complementary Colors." *Wikipedia*, The Free Encyclopedia. `https://en.wikipedia.org/wiki/Complementary_colors`, accessed June 12, 2025.
9. Wu, Yijia, Yanjing Mao, Kaiqiang Feng, Donglai Wei, and Liang Song. "Decoding of the Neural Representation of the Visual RGB Color Model." *PeerJ Computer Science*, vol. 9, p. e1376, 2023.
10. Loesdau, Martin, Sébastien Chabrier, and Alban Gabillon. "Hue and Saturation in the RGB Color Space." In *International Conference on Image and Signal Processing*, pp. 203–212, 2014. Springer.
11. Samin, Ahnaf Mozib, M. Firoz Ahmed, and Md Mushtaq Shahriyar Rafee. "ColorFoil: Investigating Color Blindness in Large Vision and Language Models." *arXiv preprint*, 2024.
12. Wadekar, Shakti N., Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. "The Evolution of Multimodal Model Architectures." *arXiv preprint*, 2024.
13. Sun, Haoran, Suyang Yu, Yijiang Li, Qingying Gao, Haiyun Lyu, Hokin Deng, and Dezhi Luo. "Probing perceptual constancy in large vision language models." *arXiv preprint*, 2025.
14. Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–9, 2015.
15. Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "VQA: Visual question answering." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.
16. Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "VisualBERT: A simple and performant baseline for vision and language." *arXiv preprint*, arXiv:1908.03557, 2019.
17. Kim, Wonjae, Bokyung Son, and Ildoo Kim. "ViLT: Vision-and-language transformer without convolution or region supervision." In *International Conference on Machine Learning*, pp. 5583–5594, 2021. PMLR.
18. Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In *International Conference on Machine Learning*, pp. 12888–12900, 2022. PMLR.
19. Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In *International Conference on Machine Learning*, pp. 19730–19742, 2023. PMLR.
20. Dai, Wenliang, Li, Junnan, Li, Dongxu, Tiong, Anthony Meng Huat, Zhao, Junqi, Wang, Weisheng, Li, Boyang, Fung, Pascale, and Hoi, Steven. "InstructBLIP: Towards general-purpose vision-language models with instruction tuning." In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.
21. Liu, Haotian, Li, Chunyuan, Li, Yuheng, and Lee, Yong Jae. "Improved baselines with visual instruction tuning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.
22. Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. "Gemma 3 Technical Report." *arXiv preprint*, 2025.
23. Bai, Shuai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. "Qwen2.5-VL Technical Report." *arXiv preprint3*, 2025.

24. OpenAI. "GPT-4.1 Technical Report." *OpenAI*, 2025. https://openai.com/index/gpt-4-1/
25. Liang, Yijun, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. "ColorBench: Can VLMs see and understand the colorful world? A comprehensive benchmark for color perception, reasoning, and robustness." *arXiv preprint*, 2025.
26. Nam, Hyeon-Woo, Ye-Bin Moon, Wonseok Choi, Lee Hyun, and Tae-Hyun Oh. "VLM's Eye Examination: Instruct and Inspect Visual Competency of Vision Language Models." *arXiv preprint*, 2024.
27. Arias, Guillem, Ramon Baldrich, and Maria Vanrell. "Color in Visual-Language Models: CLIP Deficiencies." *arXiv preprint* , 2025.
28. Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
29. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 Conference of the NAACL-HLT*, 2019.
30. Akbarinia, Arash. "Exploring the categorical nature of colour perception: Insights from artificial networks." *Neural Networks*, vol. 181, p. 106758, 2025. Elsevier.
31. Yang, Jiachen, Chenguang Wang, Bin Jiang, Houbing Song, and Qinggang Meng. "Visual perception enabled industry intelligence: State of the art, challenges and prospects." *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2204–2219, 2020. IEEE.
32. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16×16 words: Transformers for image recognition at scale." *arXiv preprint*, 2020.
33. Yang, Jinyu, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. "Vision-language pre-training with triple contrastive learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.
34. Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. "Microsoft COCO: Common objects in context." In *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, pp. 740–755, 2014. Springer.
35. Krishna, Ranjay, Zhu, Yuke, Groth, Oliver, Johnson, Justin, Hata, Kenji, Kravitz, Joshua, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2017. Springer.
36. Ordonez, Vicente, Kulkarni, Girish, and Berg, Tamara. "Im2text: Describing images using 1 million captioned photographs." *Advances in Neural Information Processing Systems*, vol. 24, 2011.
37. Sharma, Piyush, Ding, Nan, Goodman, Sebastian, and Soricut, Radu. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.
38. Schuhmann, Christoph, Vencu, Richard, Beaumont, Romain, Kaczmarczyk, Robert, Mullis, Clayton, Katta, Aarush, Coombes, Theo, Jitsev, Jenia, and Komatsuzaki, Aran. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." *arXiv preprint*, 2021.
39. Singh, Amanpreet, Natarjan, Vivek, Shah, Meet, Jiang, Yu, Chen, Xinlei, Parikh, Devi, and Rohrbach, Marcus. "Towards VQA models that can read." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.
40. Lu, Pan, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering." *Advances in Neural Information Processing Systems*, 2022.
41. ShareGPT. "ShareGPT." *ShareGPT*, 2023. https://sharegpt.com/

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.