

# A Benchmark for Systematic Testing of How Sensitive Visio-Linguistic Models are to Color Naming

**Marie Bexte**  
 CATALPA  
 FernUniversität in Hagen  
 Germany

**Andrea Horbach**  
 Hildesheim University  
 Germany

**Torsten Zesch**  
 CATALPA  
 FernUniversität in Hagen  
 Germany

## Abstract

With the recent emergence of powerful visio-linguistic models comes the question of how fine-grained their multi-modal understanding is. This has led to the release of several probing datasets. Results point towards models having trouble with prepositions and verbs, but being relatively robust when it comes to color. To gauge how deep this understanding goes, we compile a comprehensive probing dataset to systematically test multi-modal alignment around color. We demonstrate how human perception influences descriptions of color and pay special attention to the extent to which this is reflected within the predictions of a visio-linguistic model. Probing a set of models with diverse properties with our benchmark confirms the superiority of models that do not rely on pre-extracted image features, and demonstrates that augmentation with too much noisy pre-training data can produce an inferior model. While the benchmark remains challenging for all models we test, the overall result pattern suggests well-founded alignment of color terms with hues. Analyses do however reveal uncertainty regarding the boundaries between neighboring color terms.

## 1 Introduction

Visio-linguistic models, which jointly process image and text, have started to yield increasingly promising results on tasks such as Visual Question Answering (Antol et al., 2015), Image Captioning (Stefanini et al., 2023), and Image-Text Retrieval (Peng et al., 2018). With their success come efforts of probing the depth of alignment between the two modalities. A frequently used approach that is visualized in Figure 1 is to use two minimally different descriptions of the same image. While one description matches the image, the other does not. The task is for a model to compare the descriptions to the image, ideally accepting the matching description and rejecting the wrong one. Such

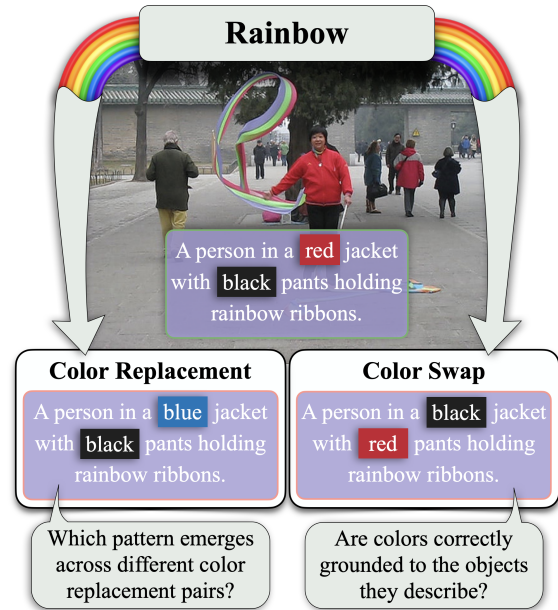


Figure 1: Overview of our color probing setup.

probing has revealed different abilities across linguistic categories. Results are more promising for categories that are more strongly associated with certain image areas, such as nouns, than for other aspects, such as prepositional information or verbs (Shekhar et al., 2017a; Parcalabescu et al., 2022).

In this paper, we focus on **color** as an important visual aspect that is central to image description and immediately associated with certain image areas. When it comes to visio-linguistic alignment around color, results are reasonably promising (Pan et al., 2019; Salin et al., 2022; Boukhers et al., 2022), although in probing datasets, color is often merged into a more general category of *attribute* understanding (Yuksekgonul et al., 2023; Wang et al., 2023). This prohibits straightforward performance evaluation on color probes alone.

To provide a benchmark that is tailored specifically towards color, we release 🌈Rainbow, which collects suitable image-text pairs from existing data

and enriches them by systematically varying the color words in the textual descriptions. The resulting probes are split into two categories, which are visualized in Figure 1. In the **color replacements**, color mentions are systematically replaced with alternative colors. This allows to assess whether a consistent performance pattern emerges, such as results for neighboring color pairs differing from results for complementary ones. To gauge whether models are not just checking for the overall presence of the colors mentioned in a description, we also include **color swaps**. We derive these from descriptions that contain exactly two color mentions: Swapping the two colors creates misaligned descriptions where models have to check whether the respective objects appear in the color they are described as.

We provide a language-only baseline for the probes in 🌈Rainbow🌈 and test a set of eight visio-linguistic models to get an impression of how well current models are able to solve them. In addition, we enrich a subset of 🌈Rainbow🌈 with the RGB values corresponding to the color names in the descriptions. This permits analysis of how human perception influences color naming and how the actual hue is related to classification decisions of models. We release 🌈Rainbow🌈 publicly and make all our code available.<sup>1</sup>

As this work focuses on color, it is our responsibility to nonetheless ensure accessibility to the extent possible: We include RGB values and names of colors instead of merely relying on the depiction of the hues we discuss. Where color is used to encode information, we use shades that are distinguishable for people with deuteranopia.

## 2 Probing Visio-Linguistic Models

With the growing success of visio-linguistic models comes interest in their inner workings and limitations. A promising approach to probing their understanding is to design contrast sets (Gardner et al., 2020). Their idea is to create minimally different examples that fall in different classes and thus test the decision boundary of a model. For probing visio-linguistic models, this usually means to alter a word in a description of an image, thereby deriving a mismatched description.

Such datasets have been designed to target nouns (Shekhar et al., 2017b), numbers (Parcalabescu et al., 2021), verbs (Jiang et al., 2022; Hendricks

and Nematzadeh, 2021; Nikolaus et al., 2022) and a number of other linguistic categories (Shekhar et al., 2017a; Wang et al., 2023; Parcalabescu et al., 2022; Zhao et al., 2022).

Some of these datasets include probes around attributes, among which are color probes (Yuksekonul et al., 2023; Wang et al., 2023). Since these are however often mixed in with probes targeting other attribute types, they do not permit a dedicated estimate of performance regarding color. VL-CheckList (Zhao et al., 2022) includes such a set of probes dedicated to color alone, and in their experiments model performance on these probes is higher than for the other attribute types they test, such as size or material.

This is in line with other previous work on visio-linguistic model performance around color: Tasking a model with editing an image so that the response to a color-focused question about this image changes (Boukhers et al., 2022) works better than asking for alterations targeting size (Pan et al., 2019). Similarly, classification models that predict a masked-out color word based on a visio-linguistic embedding yield promising results (Salin et al., 2022).

What these rather technical approaches do however not take into account is how the probed aspects are processed in human perception. In this vein, Kajić and Nematzadeh (2022) assess the extent to which human processing of number information is mirrored in visio-linguistic models.

In this work, we take a similarly human-informed approach. We aim to systematically assess color alignment in visio-linguistic models by creating an extensive collection of probing examples. During evaluation, we take into account the relations between colors and the influence of human perception on color naming.

## 3 Color Naming

Computers process color with the objectiveness of a machine: In the RGB system of the digital world, colors are combinations of intensities of red, green, and blue. The textual data processed by visio-linguistic models does however consist of descriptions of these hues in human language. Models therefore have to pick up on how color words align with certain shades. This is complicated by three factors rooted in the human perceptual system and the language used to describe color: context, coverage, and subjectivity.

<sup>1</sup><https://github.com/mariebexte/vl-probing>



Figure 2: Color illusion that demonstrates contrast effect (designed by Akiyoshi Kitaoka<sup>2</sup>). Both center squares are the same shade.

While we do not experience it as such, we do not perceive every shade as its exact value. A key driver in this is **context**: The perceptual system demonstrates a number of constancy phenomena. One of these is color constancy, which makes the same sheet of paper appear white both in direct sunlight and a dimly lit room (Walsh and Kulikowski, 1998). We are thus applying a sort of *color correction* to account for the overall setting. This helpful quirk does however make us susceptible to illusions (see Kitaoka (2010) for an overview). One such illusion is depicted in Figure 2: The surrounding colors make the center square appear more yellow on the left and more orange on the right, even though both are the same exact shade. This means that in visio-linguistic datasets the same hue will not necessarily always be described the same, because its context may cause it to be perceived differently. Another relevant dimension of context is the interplay of a color word and the object it refers to: White skin has a different shade of white than a white shirt, just like a red wine is a different red than red hair.

In general, the complementary RGB color pairs *blue-yellow* and *red-cyan* align well with how color is processed by the human perception system (Hurvich and Jameson, 1957; Pridmore, 2011), but the **coverage** of the RGB space with color terms varies. Berlin and Kay (1969) formulate a set of eleven basic colors, which they postulate to follow a certain order of emergence across languages. Some languages do not use all eleven colors (Bornstein, 2007) and others have separate terms for different shades of the same basic color (Thierry et al., 2009; Kim et al., 2019), but these eleven basic colors match the common color terms in the English language. These terms are not equidistant on the RGB color wheel (see Figure 3): Leaving *white*, *gray* and *black* to their own axis, six of the remaining eight colors are located in the upper half. This aligns well with communication around warm colors, i.e. the upper half of the color wheel, being more ef-

<sup>2</sup><http://www.psy.ritsumei.ac.jp/~akitaoka/color2e.html>

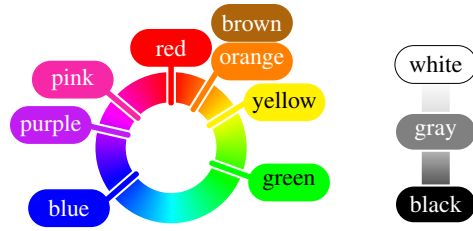





Figure 3: Overview of the colors covered by , which correspond to the eleven basic colors Berlin and Kay (1969) postulate.

fective than for cool colors (Gibson et al., 2017). Gibson et al. (2017) postulate that this emerged out of a necessity to discriminate objects from background, finding that objects tend to have warm and backgrounds cool colors. The uneven coverage of the RGB space with color names means that a visio-linguistic model has to notice how variations in the RGB values matter more in some areas than in others: While there is a rather restricted *orange* area, it is a much wider range of values that can be described as *blue*.

A third aspect that skews human color descriptions is that there is a certain level of **subjectiveness** to them: Some hues may permit multiple descriptions, or even cause disagreement regarding the appropriate color word to describe them. An example of this is the tank top of the woman in the exemplary Flickr30k image in Table 1: While the crowd annotator described it as *green*, it might just as well be described as *blue*. This would present as noise in a visio-linguistic dataset. What can also play into these differences is variation in the appearance of digitally displayed colors depending on the respective display.


In summary, learning the association between RGB values and color names is not a straightforward case of discretizing equal-sized subareas of the color space into a set of vocabulary. Color naming is influenced by the context in which a shade occurs, the language a person speaks and may at times be ambivalent. To assess these effects on crowdworker’s descriptions of images, we manually enrich one of the subsets of  with RGB values of the described elements.


#### 4

We now describe , our benchmark to systematically test the sensitivity of visio-linguistic models to color naming. We start by describing the underlying datasets. Then we outline how these

datasets were processed to collect descriptions that mention color, from which we derive two kinds of probes: color replacements and color swaps.

#### 4.1 Source Datasets

We first screened an extensive set of existing datasets of English image descriptions for descriptions that mention color. The aim was to find a diverse set of source datasets that all include a substantial amount of mentions of a shared set of colors. Table 1 gives an overview of the source datasets  builds on: **Flickr30k** (Young et al., 2014) and **MS COCO** (Lin et al., 2014) consist of high-quality crowdsourced descriptions of images that were collected from Flickr.<sup>3</sup> The **EQ-GEBC** subset of the EqBen (Wang et al., 2023) benchmark has descriptions of video frames. Since this dataset also includes negative descriptions of things that are *not in the scene*, we exclude any such cases. **ARO** (Yuksekgonul et al., 2023), **VL-CheckList** (Zhao et al., 2022) and the **EQ-Kubric** subset of EqBen contain descriptions that were generated, which means that descriptions from these datasets have a fixed structure. Both ARO and VL-CheckList derive descriptions off of the scene graph annotations in the Visual Genome (Krishna et al., 2017) dataset. From VL-CheckList, we collect all descriptions found in the *color attribute* and *object* subsets and split them into two categories: Those that are short and merely consist of a color and an object (CheckList<sub>S</sub>), such as *red fire hydrant*, and those that are longer, consisting of at least four tokens (CheckList<sub>L</sub>). In EQ-Kubric, not only the descriptions but also the images are generated. Generating both in tandem makes the images and their descriptions more consistent, as it eliminates effects of subjectivity in color naming.

Flickr30k and MS COCO are usually split into training, validation and test data according to Karpathy and Fei-Fei (2017). To prevent interference with the training data of models and because  is meant for evaluation purposes, we only collect descriptions from the respective test sets.

#### 4.2 Deriving Color Probes

To detect color mentions in the source datasets, we start out with a list of all HTML color names<sup>4</sup> and keep only those that occur at least 30 times

<sup>3</sup><https://www.flickr.com>

<sup>4</sup>[https://www.w3schools.com/colors/colors\\_names.asp](https://www.w3schools.com/colors/colors_names.asp)

in each dataset. This is done to ensure that all datasets share the same set of colors and that all of these colors occur with substantial frequency. The only exception to this is *orange*, which we keep even though it is not present in EQ-Kubric. Figure 3 shows the resulting set of colors on the RGB color wheel, which match the eleven basic colors described by Berlin and Kay (1969). In extracting descriptions that mention these colors from the source datasets, we perform the following normalizations: We replace British English mentions of *grey* with the American spelling as *gray*. *Orange* requires special treatment for two reasons: First, it starts with a vowel and should thus be preceded by *an* instead of *a* to prevent grammatical errors. We fix any such errors in the original data<sup>5</sup> and make sure to adapt a preceding determiner from *an* to *a* whenever we replace *orange* with a different color, and vice versa.<sup>6</sup> Second, orange can also refer to the fruit and thus occur as a noun rather than a color. After experimenting with different part of speech taggers, we found it more reliable to manually screen the data to exclude these cases.<sup>7</sup>

We further exclude descriptions that contain objects described with more than one color, e.g. *a green and white shirt*, because such sentences would not result in a specific enough probe. We thus remove all descriptions that either match the pattern  $\langle \text{color1 color2} \rangle$  or  $\langle \text{color1 and color2} \rangle$ . From the collected set of image descriptions, we construct two types of probes: color replacements and color swaps.

**Color replacements** For every occurrence of one of our eleven colors of interest, we systematically derive ten descriptions where this color is replaced with each of the other ones in turn. In this way, we are creating ten mismatched descriptions. These do not only allow to test how sensitive models are to color manipulations in general, but also to examine the pattern that emerges across replacement pairs: As humans, we sometimes do not agree on whether something might still be *orange* or already *red*, but agreement should be fairly high when it comes to distinguishing *yellow* from *blue*. The systematical setup of replacing every color mention with all

<sup>5</sup>6 in Flickr30k, 8 in MS COCO, 8 in CheckList<sub>L</sub>.

<sup>6</sup>Failing to do so may otherwise present as a clue to the model: It could recognize *an* and *orange* occurring together as a marker of a matching sentence, and one appearing without the other as indicative of a manipulated one.

<sup>7</sup>20 in ARO, 41 in MS COCO, 76 in CheckList<sub>S</sub>, 51 in CheckList<sub>L</sub>.

	Flickr30k	MS COCO	EQ-GEBE	ARO	CheckLists	CheckList <sub>L</sub>	EQ-Kubric
<b>Text Source</b>	Manual	Manual	Manual	Generated	Generated	Generated	Generated
<b>Image Type</b>	Photographs	Photographs	Video Frames	Photographs	Photographs	Photographs	Generated









							
<i>A woman in a <b>green</b> tank top looking at a drill while a crowd looks on.</i>	<i>A <b>yellow</b> wall living room with a large and bright <b>white</b> window.</i>	<i>Man in <b>green</b> hoodie spit after he brushed his teeth.</i>	<i>The brick ground and the <b>black</b> umbrella</i>	<i><b>Green</b> pants</i>	<i><b>Gray</b> circle on <b>white</b> mug</i>	<i>The <b>red</b> coffee mug is located behind the <b>black</b> boot</i>	
<b># Images</b>	681	2016	2382	4114	34103	4112	9649
<b># Sentences</b>	1434	3358	2386	18123	53574	6348	9649
<b># Tokens (θ)</b>	16.6	11.0	15.0	7.1	2.1	4.4	12.5
<b># Color</b>							
– Mentions	1968	3890	3553	22827	53577	6540	14848
– Swaps	287	379	758	4675	–	127	4138

Table 1: Source datasets of  and how many images, sentences, color mentions and color swaps these contribute to the benchmark. For details on how often which color is mentioned in which dataset and how often which color pair appears in the color swaps, see Tables 5 and 6 in the Appendix, respectively.

other colors allows us to assess to which extent the predictions of a model exhibit similar tendencies.

**Color swaps** Referring back to the example image in Figure 1, a model might not necessarily reject the description on the bottom left, where *red* is replaced with *blue*, because the jacket is not blue. Instead, it could merely reject it because there is not much *blue* anywhere in the image. To therefore ensure that a model is not just squinting at the image and finding mismatches because it does not see the colors that are mentioned, we build a second set of probes. For these, we take descriptions that contain exactly two colors and swap them so that *a red jacket and black pants* become *a black jacket and red pants*. Since both *red* and *black* are in the image and it is merely the order of the words that changes, this becomes a test of whether the model is able to recognize that the specific objects have to appear in the respective color.<sup>8</sup>

### 4.3 Evaluation


We now turn towards the evaluation setting and metric we employ. Each probe consists of an image  $i$  and two descriptions  $d_1$  and  $d_2$  of it. While  $d_1$  matches the respective image,  $d_2$  is a mismatch. The two descriptions are processed as two separate tuples  $(i, d_1)$  and  $(i, d_2)$  in a binary classification setup. This means that an individual binary decision is made regarding each of them.

<sup>8</sup>For ARO, this is a subset of the original dataset, which also probes word order. ARO is however not restricted to color, as it tests attribute understanding in general.

#### 4.3.1 Adjacent and complementary colors

With our color probes, we intend to create misaligned examples by replacing color mentions with different colors. However, there are cases where the same hue may be appropriately described by two different color words, e.g. a *red* object that could arguably also be described as *orange*. The replacement of *red* with *orange* may thus in a certain proportion of the examples lead to a description that is in fact not a mismatch. Since we are interested in whether this is reflected in the models predictions, we pay special attention to colors that are **adjacent** in the color wheel. This is indicated by  $\sim$  in our experiments. For these color pairings, it may be appropriate for the model to accept both the original and the modified descriptions. In the same vein, we also consider the other extreme by taking a look at **complementary** colors, indicated by  $\dagger$  in our experiments. For these, we can be certain that the shade that was initially described with one color word can not also be described by the respective replacement color. For an overview of adjacent and complimentary colors see Appendix 7.

#### 4.3.2 Metric

The probes in  consist of two image-description tuples that share the same image. We obtain a separate binary classification for each tuple. Straightforward evaluation metrics to apply would thus be accuracy, precision and recall. However, this does not take into account the paired

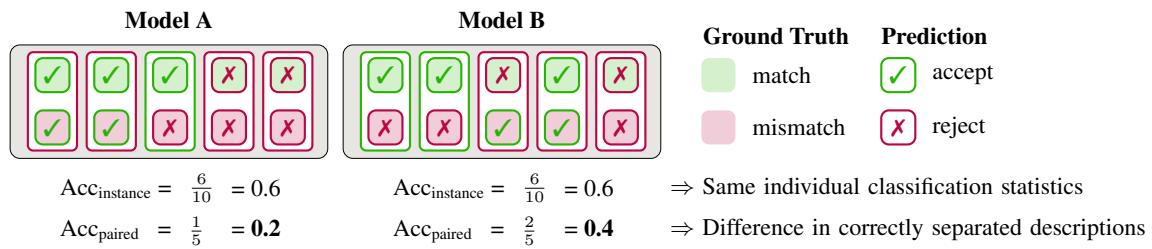


Figure 4: Each square represents a tuple of an image and a description. The respective upper and lower tuple share the same image, but differ in its description of it. While the description in the upper tuple matches the image, the one in the lower tuple differs in a color word and is therefore a mismatch. The two hypothetical models both correctly recognize three matches and three mismatches. This means that they have identical accuracy, precision and recall. When accuracy is however calculated on the basis of pairs of tuples (white rectangles), it becomes apparent that model B is superior in separating the matching from the mismatched description of an image.

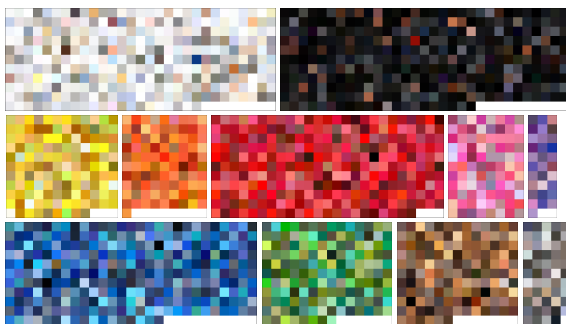


Figure 5: Hues in the Flickr30k subset that are described as *white*, *black* (top), *yellow*, *orange*, *red*, *pink*, *purple* (middle), *blue*, *green*, *brown* or *gray* (bottom).

nature of the tuples. This is visualized in Figure 4, where each square represents a tuple. Looking at the two hypothetical model predictions, both models achieve the same exact accuracy, precision and recall. To quantify how often a model both accepts the matching description *and* rejects the altered one, we calculate the **pair-wise accuracy**. This reveals a difference between the two models, as it shows model B to separate the tuples more accurately. Since this paired evaluation creates four possible outcomes for each probe, i.e. each pair of tuples, a random baseline would reach a pair-wise accuracy of .25.

## 5 Analysis

We first analyze the shades corresponding to color names in Flickr30k, then establish a language-only baseline, and finally probe a number of visio-linguistic models.

### 5.1 Color Naming in Flickr30k

As discussed in Section 3, the names people use to describe colors are not necessarily consistent with

the actual RGB values, because they can vary due to context or subjectivity. To gauge to what extent this is the case in our probes, we manually pick the RGB values of the objects that are described in the Flickr30k subset of [Rainbow](#). In doing this, we pick the color of a pixel that is representative of the overall appearance of the respective object. An overview of the resulting values is shown in Figure 5, which reveals the range of shades covered by the respective color names. Very similar or even the same shades sometimes appear in different patches of the Figure. This is partly due to context, i.e. the same hue appearing different to the human eye depending on the contrast in which it occurs. In other cases, this is the result of multiple annotators describing the same object in the same image with a different color name. We can therefore conclude that the aspects discussed in Section 3 do influence the color names that occur in human descriptions of photographs. For examples of the influence of white color constancy and subjective naming of the same hue see Figures 9 and 10 in the Appendix, respectively.

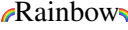
### 5.2 Language-Only Baseline

In our systematic replacement of colors, we can not control for the interplay of colors with the things they describe. There will be cases where these are linked, such as *blue sky* or an *orange safety vest*. Replacing these colors can skew the likelihood of a description matching an image, merely because the description alone is unlikely. To assess to which extent such language clues are present in [Rainbow](#), we calculate a language-only baseline. As visio-linguistic models usually build on the Transformer (Vaswani et al., 2017) architecture, we use a BERT (Devlin et al., 2019) model for this

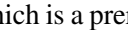
	Flickr30k	MS COCO	EQ-GEBC	ARO	CheckLists	CheckList <sub>L</sub>	EQ-Kubric
<i>Random Baseline:</i> .50							
Replacements	.73	.76	.74	.67	.54	.70	.67
Swaps	.67	.64	.69	.71	-	.71	.64

Table 2: Language-only baseline: The proportion of probes for which the matching description is deemed more likely than the mismatched one, determined by only considering the caption and never the image.

estimation (*bert-base-uncased* from HuggingFace<sup>9</sup>). To obtain language probabilities of descriptions, we employ the approach described by Salazar et al. (2020). They mask out tokens one-by-one and combine the results into a pseudo log-likelihood score.

We determine these likelihood scores for the two descriptions in each probe. Table 2 shows for which proportion of the probes the likelihood of the matching description exceeds that of the mismatched one, i.e. how often the language is biased towards preferring the matching description. We see a certain amount of such bias for each dataset in . This indicates that color words are sometimes associated quite strongly with the objects they refer to. For the color replacements, it must however be taken into account that for each strong association such as *blue sky*, there will be ten mismatched descriptions, one for each of the other colors. These are potentially all deemed less likely than the original matching description. Assessing how the language baseline varies across different color pairs reveals that there is no consistent preference of the matching over the mismatched description. Instead, it varies in direction and intensity. For a visualization of this, see Figure 8 in the Appendix.

### 5.3 Probing Visio-Linguistic Models

We test a range of visio-linguistic models with different properties. For a summary of them, see Table 7 in the Appendix. Each model has a pretrained binary classification head for image-text alignment, which is a prerequisite because  requires such binary predictions. We rely on the pretrained checkpoints released by the respective authors.

LXMERT (Tan and Bansal, 2019) uses object features that were pre-extracted using a Faster R-CNN (Anderson et al., 2018) and processes visual and textual input in a two-stream Transformer architec-

ture. UNITER (Chen et al., 2020) and VILLA (Gan et al., 2020) do the same in a single stream.<sup>10</sup> Further, we test a number of end-to-end models. SOHO (Huang et al., 2021) uses ResNet (He et al., 2016) to process images. ALBEF (Li et al., 2021), TCL (Yang et al., 2022) and BLIP (Li et al., 2022) use a Vision Transformer (Dosovitskiy et al., 2021). When it comes to object recognition, Vision Transformers have been shown to align more closely with human perception than Convolutional Neural Networks (Tuli et al., 2021). To include an example of the recent generative models, we also probe LLaVA-1.5 (Liu et al., 2023b,a). This model connects the CLIP (Radford et al., 2021) vision-and-language model and the large language model Vicuna (Chiang et al., 2023). To derive binary predictions from the textual output of LLaVA, we incorporate the image descriptions into a prompt: "Does the following sentence match this image?\n" + *description* + "\nPlease answer with either 'yes' or 'no'."

#### 5.3.1 Color replacements

Table 3 shows paired accuracy results for complementary and adjacent colors, macro-averaged over the individual color pairs.<sup>11</sup> Results are often below or not substantially higher than the chance baseline of .25. Performance is especially poor for SOHO. Even though LXMERT was pretrained on the same data as SOHO, it performs somewhat better. Among the three BLIP models it is actually the smaller one, trained with a lower volume of images, that has the overall best results. This shows that the addition of 115M more noisy samples from the LAION (Schuhmann et al., 2021) dataset into the training of BLIP<sub>B</sub> and BLIP<sub>L</sub> does more harm than good to the ability of the model to process the color names.

Across all models and datasets, there is a clear effect of complementary colors scoring higher than adjacent ones. This seems to reflect the greater hue difference in complementary color pairs. To gain more insight into how performance distributes over the individual color pairs, Figure 6 shows detailed results for two comparably well-performing models: The overall highest difference between performance on complementary and adjacent colors is achieved by TCL on the MS COCO dataset (Figure 6, left). Our enrichment of the Flickr30k

<sup>10</sup>Since the larger versions of UNITER and VILLA consistently outperform their smaller counterparts, we limit our results to these larger versions of the models.

<sup>11</sup>Results across all color pairs consistently fall between performance on adjacent and complementary color pairs.

<sup>9</sup><https://huggingface.co/bert-base-uncased>

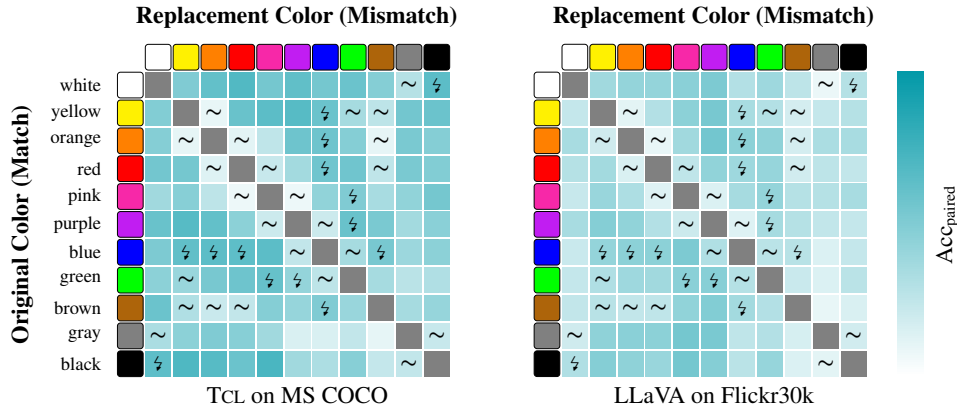


Figure 6: Detailed results for the model and dataset with the greatest overall performance difference between adjacent and complementary color pairs (top), and the model with the greatest performance difference on the RGB-annotated Flickr30k subset (bottom).  $\sim$  denotes adjacent colors,  $\⚡$  complementary ones.

	LXMERT	UNITER <sub>L</sub>	VILLA <sub>L</sub>	SOHO	ALBEF <sub>B</sub>	ALBEF <sub>L</sub>	TCL <sub>B</sub>	BLIP <sub>B-4M</sub>	BLIP <sub>B</sub>	BLIP <sub>L</sub>	LLaVA
<b>Flickr30k</b>											
$\⚡$ Complementary	.29	.38	.32	.12	.32	.22	.29	.33	.12	.07	<b>.42</b>
$\sim$ Adjacent	.27	.20	.15	.09	.17	.11	.13	.19	.06	.03	<b>.20</b>
<b>MS COCO</b>											
$\⚡$ Complementary	.28	.63	.59	.11	.59	.40	<b>.57</b>	.63	.27	.25	.48
$\sim$ Adjacent	.25	.34	.27	.05	.28	.16	<b>.22</b>	.35	.11	.09	.20
<b>EQ-GEBC</b>											
$\⚡$ Complementary	.25	.32	.29	.20	.24	.32	.33	.36	.32	.31	.46
$\sim$ Adjacent	.22	.21	.17	.13	.17	.24	.21	.25	.20	.19	.25
<b>ARO</b>											
$\⚡$ Complementary	.31	.49	.46	.18	.49	.53	.55	.55	.47	.44	.21
$\sim$ Adjacent	.27	.26	.23	.10	.34	.33	.34	.40	.27	.24	.08
<b>CheckList<sub>S</sub></b>											
$\⚡$ Complementary	.33	.55	.58	.31	.31	.43	.33	.34	.50	.42	.53
$\sim$ Adjacent	.29	.40	.39	.20	.26	.32	.26	.28	.34	.30	.32
<b>CheckList<sub>L</sub></b>											
$\⚡$ Complementary	.32	.49	.48	.21	.34	.40	.36	.38	.37	.39	.38
$\sim$ Adjacent	.28	.31	.28	.12	.25	.27	.25	.30	.23	.23	.21
<b>EQ-Kubric</b>											
$\⚡$ Complementary	.20	.23	.24	.21	.04	.21	.14	.22	.26	.24	.20
$\sim$ Adjacent	.18	.13	.13	.14	.03	.14	.10	.14	.17	.16	.09
<b>Average</b>											
$\⚡$ Complementary	.29	.45	.43	.19	.34	.36	.37	.41	.33	.31	.38
$\sim$ Adjacent	.25	.27	.24	.12	.22	.23	.22	.28	.20	.18	.19

Table 3: Paired accuracy, macro-averaged across all complementary vs. adjacent color pairs. Detailed results for boldface model-dataset combinations in Figure 6.

subset with RGB value annotations permits follow-up analyses that take into account the actual hue a model decision is centered around. Figure 6 (right) therefore also lists results of LLaVA on this data, as this model achieves the highest performance delta on Flickr30k.

Both matrices in Figure 6 convey the overall impression of symmetric effects. For both models, exchanging pink to purple has especially low paired accuracy. This can be traced back to a tendency to accept either name as a description of the respective hue. Such a pattern may emerge because these hues are candidates where humans would agree that the alternative color is also an appropriate description. However, it could also result from the model being unsure about the concept of the respective colors. We therefore compared the average hues for which a neighboring color term was accepted vs. rejected, which is depicted in Figure 11 in the Appendix. Results point towards the model being unsure what a certain color looks like.

### 5.3.2 Color swaps

Table 4 shows paired accuracy results for the color swaps. While they had achieved the best performances in the color replacement experiment, UNITER and VILLA are now among the lower scoring models. They are outperformed by ALBEF, TCL and BLIP. This indicates a superiority of these Vision Transformer-based models over models relying on pre-extracted image features, an effect especially pronounced for the MS COCO and ARO subsets. Although LLaVA performs well on many other benchmarks, it is among the lower performing models for the color swaps.



	LXMERT	UNITER <sub>L</sub>	VILLA <sub>L</sub>	SOHO	ALBEF <sub>B</sub>	ALBEF <sub>L</sub>	TCL <sub>B</sub>	BLIP <sub>B+LM</sub>	BLIP <sub>B</sub>	BLIP <sub>L</sub>	LLaVA
Flickr30k	.25	.23	.14	.05	.30	.21	.25	.32	.08	.06	.26
MS COCO	.22	.43	.33	.02	.52	.38	.54	.55	.20	.16	.28
EQ-GEBC	.21	.16	.12	.07	.17	.26	.22	.30	.20	.21	.34
ARO	.25	.29	.23	.06	.48	.52	.55	.55	.44	.43	.11
CheckList <sub>L</sub>	.31	.24	.20	.08	.18	.15	.19	.23	.12	.09	.35
EQ-Kubric	.17	.16	.15	.10	.04	.16	.14	.20	.20	.16	.10
Average	.24	.25	.20	.06	.28	.28	.32	.36	.21	.19	.24

Table 4: Paired accuracy results for the color swaps.

## 6 Conclusion

We present [Rainbow](#), a benchmark to probe color understanding of visio-linguistic models. Our annotation of one of its subsets with RGB values demonstrates how human perception influences color naming. Testing a set of visio-linguistic models showed them to pick up on which color terms describe neighboring hues. Swapping colors in descriptions revealed an inferior ability of models that rely on pre-extracted image features to ground color names to the objects they describe. Since the benchmark remains challenging for all models we probe, it will be exciting to see how future models fare on it.

## 7 Limitations

As we discuss in Section 3, color naming varies depending on culture and language. [Rainbow](#) consists of English image descriptions, and many of the datasets collect these descriptions from US American annotators. This means that it is centered around the eleven basic colors used in English and to some extent limited to a western-centric world view - different patterns may emerge for different languages and cultures.

Large volumes of images, usually collected from the internet, are used to pretrain visio-linguistic models. Therefore, a general problem of visio-linguistic probing is that models might have already seen some of the images during pretraining.

## 8 Ethics

**Data Privacy** All datasets [Rainbow](#) builds on consist of descriptions of images. These are either generated or given by people with no personal relation to the contents of the images. This makes it unlikely for descriptions to contain personal information or offensive content, and we did not en-

counter any in working with the data.

**Environmental Impact** While we do not fine-tune any models, we do use pretrained models for inference. All experiments were run on Nvidia Titan Xp and A40 graphics cards. For models that require pre-extracted image features (LXMERT, UNITER and VILLA), we had to first extract these features. [Rainbow](#) has descriptions of 54406 images. Extracting features for these images took a total of four GPU hours. [Rainbow](#) contains of 481,097 probes. The total inference time for all models was around 150 GPU hours.

**License** All datasets [Rainbow](#) builds on are released under a license that permits modification. We release [Rainbow](#) under [MIT license](#). This covers our annotation of the RGB values of color descriptions in Flickr30k and the code to derive [Rainbow](#) from the existing datasets.

## Acknowledgements

This work was partially conducted at “CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics” of the FernUniversität in Hagen, Germany.

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086. ISSN: 2575-7075.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: Visual Question Answering](#). In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. ISSN: 2380-7504.
- Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Univ of California Press.
- Marc H. Bornstein. 2007. [Hue categorization and color naming: Cognition to language to culture](#). In *Anthropology of color: Interdisciplinary multilevel modeling.*, pages 3–27. John Benjamins Publishing Company, Amsterdam, Netherlands.
- Zeyd Boukhers, Timo Hartmann, and Jan Jürjens. 2022. [COIN: Counterfactual Image Generation for Visual Question Answering Interpretation](#). *Sensors*, 22(6):2245. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. [Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [UNITER: UNiversal Image-TEXT Representation Learning](#). In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 104–120, Cham. Springer International Publishing.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [ImageNet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. [Large-Scale Adversarial Training for Vision-and-Language Representation Learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 6616–6628. Curran Associates, Inc.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating Models’ Local Decision Boundaries via Contrast Sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingham, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. 2017. [Color naming across languages reflects color use](#). *Proceedings of the National Academy of Sciences*, 114(40):10785–10790. Publisher: Proceedings of the National Academy of Sciences.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lisa Anne Hendricks and Aida Nematzadeh. 2021. [Probing image-language transformers for verb understanding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644.
- Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. [Seeing out of the box: End-to-end pre-training for vision-language representation learning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12976–12985.
- Leo M. Hurvich and Dorothea Jameson. 1957. [An opponent-process theory of color vision](#). *Psychological Review*, 64(6, Pt.1):384–404. Place: US Publisher: American Psychological Association.
- Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. 2022. [ComCLIP: Training-Free Compositional Image and Text Matching](#). ArXiv:2211.13854 [cs].
- Ivana Kajic and Aida Nematzadeh. 2022. [Probing representations of numbers in vision and language models](#). In *SVRHM 2022 Workshop @ NeurIPS*.
- Andrej Karpathy and Li Fei-Fei. 2017. [Deep Visual-Semantic Alignments for Generating Image Descriptions](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.
- Younghoon Kim, Kyle Thayer, Gabriella Silva Gorsky, and Jeffrey Heer. 2019. [Color Names Across Languages: Salient Colors and Term Translation in Multilingual Color Naming Models](#). *EuroVis 2019 - Short Papers*. ISBN: 9783038680901 Publisher: The Eurographics Association Version Number: 031-035.
- Akiyoshi Kitaoka. 2010. A brief classification of colour illusions. *Colour: Design & Creativity*, 5(3):1–9.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. [Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations](#). *International Journal of Computer Vision*, 123(1):32–73.

- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation](#). In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR. ISSN: 2640-3498.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. [Align before Fuse: Vision and Language Representation Learning with Momentum Distillation](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: Common Objects in Context](#). In *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham. Springer International Publishing.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#). In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. 2022. [Do vision-and-language transformers learn grounded predicate-noun dependencies?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1555, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. [Im2Text: Describing Images Using 1 Million Captioned Photographs](#). In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Jingjing Pan, Yash Goyal, and Stefan Lee. 2019. [Question-conditioned counterfactual image generation for vqa](#). *CoRR*, abs/1911.06352.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. [VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.
- Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. [Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks](#). In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.
- Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2018. [An Overview of Cross-Media Retrieval: Concepts, Methodologies, Benchmarks, and Challenges](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2372–2385. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.
- Ralph Pridmore. 2011. [Complementary Colors Theory of Color Vision: Physiology, Color Mixture, Color Constancy and Color Perception](#). *Color Research & Application*, 36:394–412.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. [Masked language model scoring](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.
- Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. [Are vision-language transformers learning multimodal representations? A probing perspective](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. [LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs](#). *NeurIPS Workshop Datacentric AI*, online, 14 Dec 2021 - 14 Dec 2021.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [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)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. [Vision and Language Integration: Moving beyond Objects](#). In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. **FOIL it! Find One mismatch between Image and Language caption**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2023. **From Show to Tell: A Survey on Deep Learning-Based Image Captioning**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Hao Tan and Mohit Bansal. 2019. **LXMERT: Learning Cross-Modality Encoder Representations from Transformers**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Guillaume Thierry, Panos Athanasopoulos, Alison Wiggett, Benjamin Dering, and Jan-Rouke Kuipers. 2009. **Unconscious effects of language-specific terminology on preattentive color perception**. *Proceedings of the National Academy of Sciences*, 106(11):4567–4570. Publisher: Proceedings of the National Academy of Sciences.
- Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Tom Griffiths. 2021. **Are convolutional neural networks or transformers more like human vision?** In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is All you Need**. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Vincent Walsh and Janusz Kulikowski. 1998. *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press.
- Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. **Equivariant Similarity for Vision-Language Foundation Models**. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11964–11974, Paris, France. IEEE.
- Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. **Vision-language pre-training with triple contrastive learning**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. **From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions**. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. **When and why vision-language models behave like bags-of-words, and what to do about it?** In *International Conference on Learning Representations*.
- Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. **VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations**. ArXiv:2207.00221 [cs].

## A Appendix

This appendix gives additional details that may be of interest to gain deeper insight into 🌈Rainbow🌈 and our analyses of it.

**Dataset** Tables 5 and 6 contain the distribution of color terms across the different subsets of 🌈Rainbow🌈. Figure 7 shows adjacent and complementary color pairs. Examples that demonstrate the effect of human perception on color names in Flickr30k are depicted in Figures 9 and 10.

**Experiments** Figure 8 shows more detailed results for our language-only baseline. Table 7 summarizes the models we probe. Figure 11 shows how the hues of elements in Flickr30k are related to the classification decisions of LLaVA, focusing on adjacent color pairs.

	Flickr30k	MS COCO	EQ-GEBC	ARO	CheckList <sub>S</sub>	CheckList <sub>L</sub>	EQ-Kubric	Σ
○ White	327	989	587	5369	16900	1781	3864	11136
● Yellow	130	241	106	539	3006	306	441	1457
● Orange	92	129	62	435	419	128	-	718
● Red	278	556	231	1060	5295	586	2401	4526
● Pink	95	118	121	359	305	129	408	1101
● Purple	33	56	32	205	150	41	630	956
● Blue	299	451	581	3697	6877	869	1024	5982
● Green	156	463	180	2772	6707	569	1319	4890
● Brown	144	266	71	3504	6921	609	1114	5099
● Gray	56	112	337	2217	4518	356	861	3583
● Black	358	509	1245	2670	2479	1166	2786	7568
Σ	1968	3890	3553	22827	53577	6540	14848	47086

Table 5: Overview of how often the individual colors occur in the different subsets of 🌈Rainbow🌈.

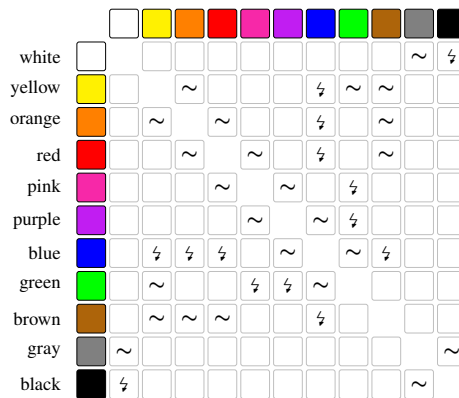


Figure 7: Matrix reflecting colors that are adjacent (~) or complementary (⚡) on the RGB color wheel.

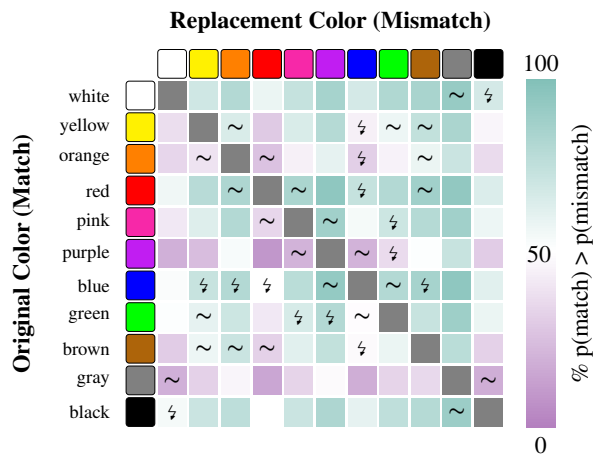


Figure 8: Language-only baseline results for the color replacement pairs, macro-averaged across all seven datasets and split into color pairs. Heatmap colors encode whether the original sentences are more likely than the manipulated ones (blue) or if it is the other way around (purple). ~ denotes adjacent colors, ⚡ complementary ones. There are indications of systematic effects for certain colors. Descriptions containing *gray* and *purple* are less likely than when these colors are replaced with a different one. Descriptions containing *red* tend to be more likely than when *red* is replaced with a different color. Results do not seem to correlate with colors being adjacent (~) or complementary (⚡) on the RGB color wheel.















































































































		Flickr30k	MS COCO	EQ-GEBC	ARO	CheckListL	EQ-Kubric	$\Sigma$	
white	- yellow	 	8	6	5	58	-	109	186
	- orange	 	5	3	5	55	4	-	68
	- red	 	10	34	21	118	9	477	660
	- pink	 	1	2	12	38	1	69	122
	- purple	 	1	4	-	25	1	102	132
	- blue	 	21	27	58	435	35	169	710
	- green	 	8	19	12	268	4	214	521
	- brown	 	10	19	7	457	5	180	673
	- gray	 	2	8	10	236	3	152	408
- black	 	28	44	131	345	16	395	943	
yellow	- orange	 	1	1	5	4	1	-	11
	- red	 	12	9	2	10	1	50	83
	- pink	 	1	1	-	5	-	-	7
	- purple	 	-	1	-	3	-	18	22
	- blue	 	7	6	16	58	1	7	94
	- green	 	4	6	-	20	6	33	63
	- brown	 	1	2	3	31	1	18	55
	- gray	 	-	3	5	27	1	22	57
	- black	 	10	4	22	26	-	73	135
orange	- red	 	2	3	1	4	-	-	10
	- pink	 	-	1	2	4	-	-	7
	- purple	 	-	-	-	1	-	-	1
	- blue	 	17	-	2	34	2	-	53
	- green	 	-	6	-	18	-	-	24
	- brown	 	1	-	-	29	-	-	30
	- gray	 	-	-	2	23	-	-	25
	- black	 	2	8	15	29	2	-	54
red	- pink	 	4	1	3	8	-	32	48
	- purple	 	-	-	-	2	-	63	65
	- blue	 	20	14	16	85	1	110	245
	- green	 	3	16	10	57	3	128	214
	- brown	 	4	6	5	69	2	106	190
	- gray	 	1	3	8	54	1	78	144
	- black	 	24	24	58	65	3	236	407
pink	- purple	 	2	1	2	3	-	16	24
	- blue	 	4	5	10	23	2	2	44
	- green	 	-	2	-	25	1	16	43
	- brown	 	1	2	-	23	-	16	42
	- gray	 	1	1	3	12	-	6	23
	- black	 	4	3	13	19	1	74	113
purple	- blue	 	1	-	4	9	-	38	52
	- green	 	2	-	-	12	-	43	57
	- brown	 	2	4	-	11	-	32	49
	- gray	 	-	1	1	8	-	26	36
	- black	 	1	5	9	14	1	71	100
blue	- green	 	5	18	7	198	2	85	313
	- brown	 	6	3	9	289	-	48	355
	- gray	 	3	3	33	172	2	35	246
	- black	 	14	9	112	188	5	151	474
green	- brown	 	5	11	-	251	3	70	337
	- gray	 	1	3	5	146	-	61	216
	- black	 	8	8	24	141	3	181	362
brown	- gray	 	2	1	4	136	-	57	200
	- black	 	12	15	12	183	1	137	359
gray	- black	 	5	3	74	111	3	132	325
$\Sigma$		287	379	758	4675	127	4138	10237	

Table 6: Overview of how often which color pair occurs in the color swaps.



Figure 9: Examples of white color constancy. The respective objects seem white in the context of the image. Picking their RGB color values shows that this is actually not the case. We pair these hues with very similar ones from different images, where they are described with a color name other than white. Overall, these are examples of how the terms humans use to describe color in photographs are not always consistent with how the respective color appears in isolation. This is what we describe as the effect of *context* in Section 3.



- > A large woman with long pink hair dressed in black [...]
- > A woman with pink hair dressed in black talks to a man.
- > A girl with bright red - hair and black clothes is posing [...]
- > A red - haired woman in black is posing for a man [...]



- > A man wearing a [...] neon green safety vest [...]
- > A young, male adult wearing [...] a green reflective vest, [...]
- > A worker in a yellow vest stands on train tracks.
- > A person in a bright yellow vest and hard hat [...]



- > Little girl in kitchen, kissing a fluffy orange cat.
- > The little girl is kissing the brown cat.
- > A young girl standing next to a yellow cat [...]



- > A child wearing a yellow shirt is jumping up and down.
- > A child wearing a yellow Doritos shirt jumps up [...]
- > A boy wearing an orange Doritos jersey jumps up in the air.
- > A boy wearing an orange shirt and brown shorts is jumping.
- > A boy wearing an orange doritos shirt looks like [...]



- > Four men leaning over a green fence and smiling .
- > Four men are outside looking down over the green bridge [...]
- > Four men [...] standing near a blue handrail smiling [...]
- > A group of men are standing beside a blue railing for a picture.



- > The brown dog is standing on the sandy beach .
- > Light brown dog running towards something at the beach .
- > A gray colored dog walks in wet sand at a beach .
- > The large gray colored dog is jumping on the beach .
- > A gray dog plays in the sand at the ocean .

Figure 10: Examples of annotators describing the same element in the same Flickr30k image with a different color name. This is the effect of what we describe as *subjectiveness* in Section 3.



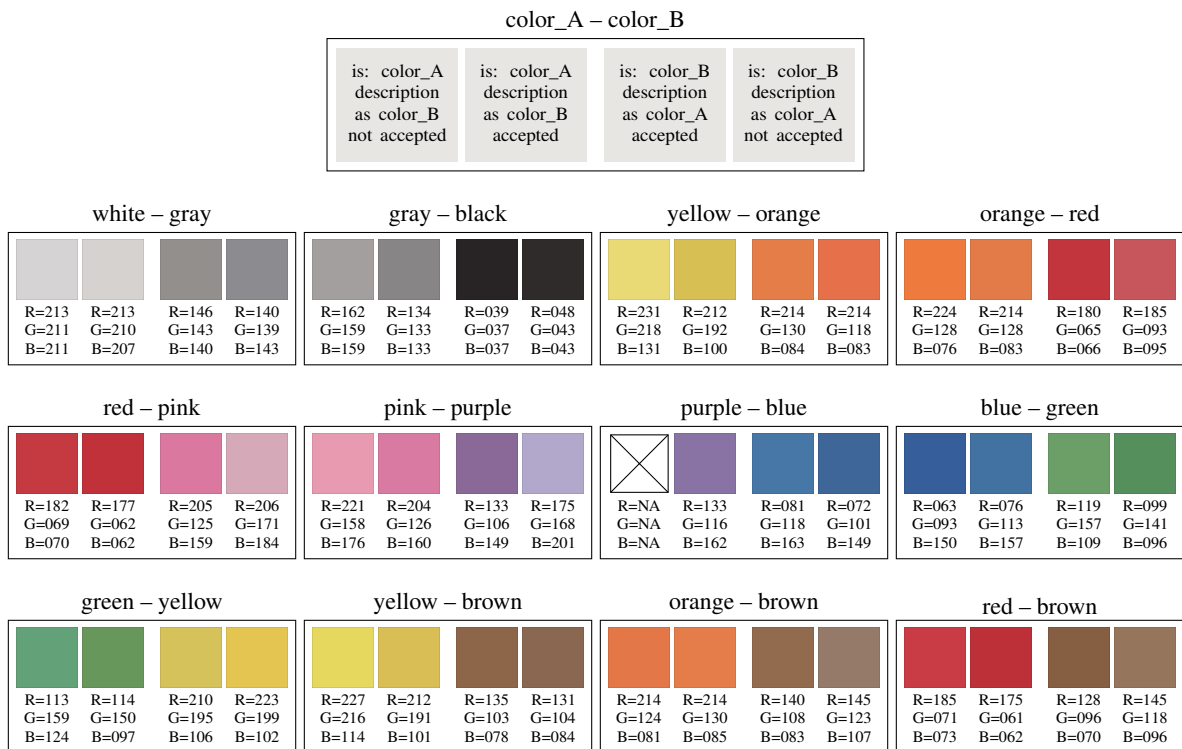



Figure 11: We annotated the hues corresponding to color names in Flickr30k. This permits the assessment of how these hues correspond to the classification decisions of models. LLaVA performed best on this dataset, which is why we give detailed results for this model here. We consider adjacent colors, and calculate the average of hues for which a description with a neighboring color term is/is not accepted. This is done to gauge whether these hues are plausibly describable with the alternative color term, or if the model is merely unsure about what exactly hues corresponding to a certain color term look like. The latter seems to be the case, as the overall appearances of the averaged hues for which the neighboring color term is accepted are not skewed towards the respective neighboring color. Consider for example the pair *green-yellow*: The average *green* hue for which a description as *yellow* is not accepted (rightmost square) is 'less *yellow*' than the averaged *green* hue for which a description as *yellow* is accepted (middle left square). Still, both shades are unambiguously green. This indicates that the model has a rather high level of ambiguity regarding where *green* ends and where *yellow* begins. Results therefore indicate that the observed pattern of lower performance for neighboring color terms is not due to the hues being somewhat ambivalent. Manual inspection of some of the images did not suggest this was due to context effects causing hues to appear differently either, which the model could have picked up from the pretraining data.

Model	Architecture	Visual Input	Datasets in Pretraining	#Images	
LXMERT (Tan and Bansal, 2019)	2-stream	Faster R-CNN	COCO, VG	0.2M	
UNITER (Chen et al., 2020)	base	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
	large	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
VILLA (Gan et al., 2020)	base	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
	large	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
SOHO (Huang et al., 2021)		end2end	ResNet	COCO, VG	0.2M
ALBEF (Li et al., 2021)	base	end2end	ViT-B/16	CC, SBU, COCO, VG,	4M
	large	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M	14M
TCL (Yang et al., 2022)	base	end2end	ViT-B/16	CC, SBU, COCO, VG	4M
BLIP (Li et al., 2022)	base <sub>14M</sub>	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M	14M
	base <sub>129M</sub>	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M, LAION	129M
	large	end2end	ViT-L/16	CC, SBU, COCO, VG, CC12M, LAION	129M
LLaVA (Liu et al., 2023a)		end2end	ViT-L/14	CC, SBU, COCO, VG, CC12M, LAION	129M

Table 7: Summary of the models we probe with . To judge their relative performances, one may want to take into account the data they were pretrained on. This is why we include the datasets models are based on, as well as the total number of images these datasets contain. Datasets are: Conceptual Captions (CC, Sharma et al. (2018)), SBU Captions (SBU; Ordonez et al. (2011)), MS COCO 2014 (COCO; (Lin et al., 2014)), Visual Genome (VG, Krishna et al. (2017)), Conceptual 12M (CC12M, Changpinyo et al. (2021)) and LAION (Schuhmann et al., 2021). Faster R-CNN (Anderson et al., 2018) is pretrained on Visual Genome (Krishna et al., 2017), ResNet (He et al., 2016) and all Visual Transformer (ViT, Dosovitskiy et al. (2021)) models are pretrained on ImageNet (Deng et al., 2009).