ColorSwap: A Color and Word Order Dataset for Multimodal Evaluation

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, Tristan Thrush
Stanford University
{jirayu, tthrush}@stanford.edu

Abstract

This paper introduces the ColorSwap dataset, designed to assess and improve the proficiency of multimodal models in matching objects with their colors. The dataset is comprised of 2,000 unique image-caption pairs, grouped into 1,000 examples. Each example includes a caption-image pair, along with a "color-swapped" pair. We follow the Winoground schema: the two captions in an example have the same words, but the color words have been rearranged to modify different objects. The dataset was created through a novel blend of automated caption and image generation with humans in the loop. We evaluate image-text matching (ITM) and visual language models (VLMs) and find that even the latest ones are still not robust at this task. GPT-4V and LLaVA score 72% and 42% on our main VLM metric, although they may improve with more advanced prompting techniques. On the main ITM metric, contrastive models such as CLIP and SigLIP perform close to chance (at 12% and 30%, respectively), although the non-contrastive BLIP ITM model is stronger (87%). We also find that finetuning on fewer than 2,000 examples yields significant performance gains on this out-of-distribution word-order understanding task.

1 Introduction

Recent years have seen remarkable developments in pretrained vision and language models (Radford et al., 2021; Li et al., 2022; Singh et al., 2022; Li et al., 2023, 2019; Rombach et al., 2021; Betker et al., 2023; Liu et al., 2023b). Their performance is exceptional in tasks such as visual question-answering (Liu et al., 2023b), text-to-image generation and manipulation (Minderer et al., 2022), and image captioning (Li et al., 2022, 2023).

Despite the success, recent work reveals that vision and language models often struggle to comprehend fine grained distinctions in images (Krojer et al., 2022) and compositional relationships, particularly in differentiating captions with the same words but different word orders (Thrush et al., 2022). The Winoground dataset from Thrush et al. demonstrates that, given two captions composed of the same words in a different order, the performance of well-known models in correctly matching these captions to their respective images is close to random chance. It is not well understood whether known multimodal architectures are even capable, in practice, of learning how to perform well at a balanced word-order-understanding task with small-scale finetuning data. Finetuning has been done on some balanced word-order datasets, but only in settings where half of the images are lacking, as far as we are aware (Yuksekgonul et al., 2023).

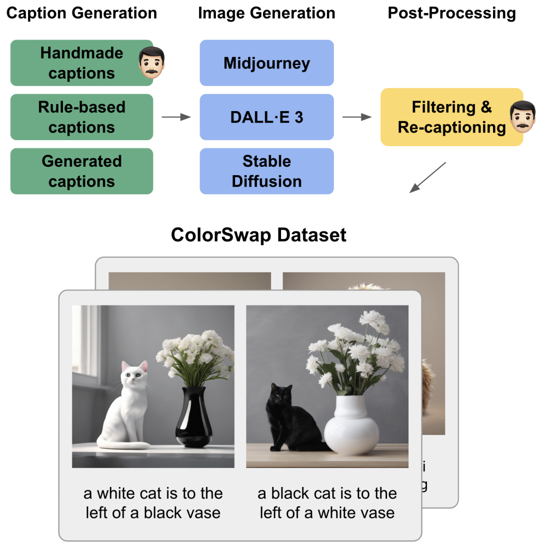

Figure 1: An overview of the ColorSwap dataset creation methodology. The human emoji marks components that require human annotator input.

To address such questions, we introduce the ColorSwap dataset, along with a data generation process that enables the quick creation of a larger dataset fitting the same schema as Winoground, as shown in Figure 1. Our dataset specifically focuses on a subset of Winoground-style examples, emphasizing the swapping of color words in captions for two main reasons: 1) accurately associating

Findings of the Association for Computational Linguistics: ACL 2024, pages 1716–1726
August 11-16, 2024 ©2024 Association for Computational Linguistics

colors with objects is of practical importance in AI-generated art, and 2) it is a conceptually simple and targeted word order understanding task - it is very clear what the human judgements are and why. One of the simplifying features of this task is that color words are often directly adjacent to the objects they modify. For example, "someone holding a [yellow umbrella] wearing a [white dress]", although it is important to note that it isn't always this simple: see Figure 2.

Figure 2: An image from DALL·E 3 (Betker et al., 2023) when given the caption "The key to the shed is blue" (we ensured that the caption was not rewritten by ChatGPT (OpenAI, 2022)). DALL·E 3 does not always make this mistake, but it is unreliable. Even though "the shed is blue" is a substring, the full sentence is saying that the key is blue. Our dataset does not target difficult cases where colors modify far objects in the string.

Finally, we conduct evaluations of several vision and language models using this dataset. We find that all models, even GPT-4V (OpenAI, 2023a), make a significant number of mistakes. Although, contrastive models such as CLIP and SigLIP struggle drastically. The dataset could be a useful benchmark for contrastive models in particular, but also for visual language models and even industry diffusion models such as Midjourney (Midjourney, 2023).[1] Further, we found no model which lacks the capability, in practice, to learn some level of generalizable word-order judgements from a fairly small set of naturalistic finetuning data.

2 ColorSwap

In this section, we introduce the ColorSwap dataset, comprised of 1,000 unique examples created by four expert annotators (each with two or more years or NLP experience and living in the United States) with the aid of generative models. Each example includes a caption paired with an image, along with a color-swapped version of the caption and image. We randomly select 700 examples for training and 300 for evaluation.

Our data collection methodology uses three key steps: 1) Caption Generation, where we generate a variety of initial captions using three different methods; 2) Image Generation, which involves employing various text-to-image models to create images corresponding to these captions; and 3) Post-Processing, which includes human review to ensure accuracy, maintain quality, and re-caption images. Post-processing is essential as current text-to-image models often mix up the colors. See Figure 1 for an illustration of our data collection process and Table 1 for a summary of the dataset's composition.

Caption	Image	# Pairs
Handmade	Midjourney	39
Handmade	DALL·E 3	167
Rule-based	Stable Diffusion	782
Rule-based	DALL·E 3	394
Generated	Stable Diffusion	212
Generated	DALL·E 3	406

Table 1: **Number of pairs per method.** Rule-based captions are rewritten by humans during post-processing. Generated captions come from Large Language Models.

2.1 Caption Generation Methods

Handmade. In this method, annotators manually create captions by creatively brainstorming scenarios and contexts that involve at least two objects. This process ensures a high degree of originality and diversity in the captions, but is time-consuming. Once we have some captions, we can use them to bootstrap the following two approaches.

Rule-based. The second method employs a systematic color swapping technique in predefined caption templates, using sets of objects and colors. It generates a broad range of color-object scenarios, though the captions lack creative variability. To tackle this, the post-processing stage involves human review and caption rewriting.

Generative model. This method leverages generative models, particularly GPT-4 (OpenAI, 2023a) and Claude-2 (Anthropic, 2023). We prompt these models with examples from our first method and some examples from the Winoground dataset.

[1]As with other targeted evaluations, there is an implicit confounder: the data is from a different distribution than the training data. ColorSwap is a test of word order understanding on out of distribution data, not just word order understanding.

These models then generate additional caption pairs based on these inputs.

2.2 Image Generation

We utilize diffusion models for image generation, a method increasingly used in multimodal dataset creation (Bitton-Guetta et al., 2023; Wu et al., 2023a; Lee et al., 2023). For diversity and cost-efficiency, various diffusion models are employed. Despite their limitations in accurately handling the color composition task, we sample multiple images for a given caption and later select the most suitable image. This approach enables us to create ColorSwap even with generative models that would find the task challenging. We use the open-source Stable Diffusion model (Podell et al., 2023), as well as stronger commercial models such as Midjourney (Midjourney, 2023) and OpenAI's DALL·E 3 (Betker et al., 2023). Details are in Appendix A.

2.3 Post-Processing

Filtering. During post-processing, annotators sift through images produced by diffusion models to ensure quality. Their task is to identify the image that most accurately aligns with its caption. If no image meets the criteria, the entire set is discarded. This step is crucial as it ensures that the final dataset is accurate and sensible to human evaluators.

Re-captioning. To ensure naturalistic and diverse captions, we do a manual re-captioning of image pairs if they were generated with rule-based captions. See Figure 3 for an example. See Appendix B for snapshots of post-processing interfaces.

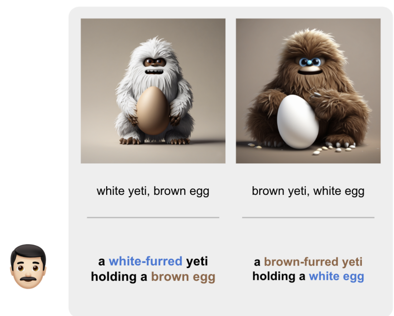

Figure 3: Illustration of re-captioning process.

3 Experiments

Here, we outline our experiments on the ColorSwap dataset. We evaluate both image-text matching (ITM) models and visual language models. All of these models make a significant number of errors in our color composition task, although contrastive ITM models in particular struggle substantially. Subsequently, we fine-tune these models using the training split of the dataset, aiming to understand whether minimal tuning can significantly improve their ability to understand word order.

3.1 Evaluation Metrics

To assess model performance, we adopt the three metrics introduced in Thrush et al. (2022), for Winoground, as our dataset has the same schema. The *text score* measures whether a model can select a correct caption given an image, while the *image score* is about selecting the correct image, given a caption. The *group score* combines both aspects.

3.2 Off-the-shelf Models

Image-text matching models. For ITM models, we evaluate CLIP (Radford et al., 2021), FLAVA (Singh et al., 2022), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023). FLAVA and BLIP offer two matching methods: 1) a contrastive method, and 2) using cross-modal parameters with an ITM head. CLIP and SigLIP only match text to images in the contrastive way. With these models, we gauage the competnence of these two alternative architectures on the task. Generally, we use the standard base versions of models. More details on model selection are included in Appendix C.

Visual language models. We assess LLaVAR (Zhang et al., 2023), LLaVA-1.5 (Liu et al., 2023a), and GPT-4V (OpenAI, 2023b). We follow the VLM Winoground evaluation methodology in (Wu et al., 2023b) where we obtain the text score by prompting them to select the correct caption from two options when provided with an image. Similarly, for the image score, we present them with a caption and two images, from which they must select the one that best corresponds to the caption. To avoid positional bias (Zheng et al., 2023), we further randomize the order in which captions or images are presented to these models. The scores here are not strictly comparable to those for the ITM models, which consider one image and one caption at a time and output floating point scores. More details on model selection and prompts are included in Appendix C.

Results. Table 2 outlines the performance of the models without finetuning. BLIP and SigLIP exhibit superior performance compared to CLIP and FLAVA, both of which are around the levels of random chance. Matching with an ITM head also improves the image-text matching performance, especially for BLIP. Also, GPT-4V, despite its status as a leading closed-source model, still exhibits genuine errors on the simple task posed by this dataset.

In Appendix E, we run an experiment to provide evidence that the models' poor performance is due to the compositional nature of this task and not simply because the images are generated by diffusion models (and so are out of the pretraining distributions). A selection of examples and model responses is included in Appendix F. We report confidence intervals for these results in Appendix G.

Model & Method	Text ↑	Image ↑	Group ↑
Image-text matching models			
Random chance	25.00	25.00	16.67
Contrastive matching			
CLIP	**35.67**	14.67	11.67
FLAVA	**35.33**	25.00	15.67
BLIP	**75.67**	**56.00**	**51.00**
SigLIP	**61.67**	**37.00**	**30.33**
ITM matching			
FLAVA	**36.33**	18.67	10.33
BLIP	**94.67**	**89.00**	**87.33**
Visual Language Models			
Random chance	25.00	25.00	6.25
LLaVAR	**27.67**	**25.67**	**8.33**
LLaVA-1.5	**69.67**	**54.33**	**42.00**
GPT-4V	**91.33**	**76.33**	**72.00**

Table 2: Performance of models on ColorSwap. Results above chance are **bold**. Note that random chance is different in the ITM versus visual language model (VLM) cases because VLMs output a binary value and ITM models output a float (an effectively continuous value).

3.3 Fine-tuning on ColorSwap

Winoground remains a challenging task, with even advanced models like GPT-4V, using chain-of-thought prompting, struggling to solve it effectively (Wu et al., 2023b). We are not aware of any demonstrations in other papers that provide an answer to whether multimodal models in practice can even be finetuned from a fairly small set of training data to understand any aspects of word order. So, we fine-tune the best performing off-the-shelf BLIP model on our dataset. Due to the continued popularity of CLIP, we also fine-tune the CLIP model. Training details are given in Appendix D.

Performance improvements post-finetuning. For the ColorSwap dataset, CLIP and BLIP significantly improve on the test set after finetuning on the train set. They are able to learn generalizable knowledge about word order from 1,400 training pairs. In the case of CLIP, performance increases by several times across all metrics. See Table 3.

Model & Method	Text ↑	Image ↑	Group ↑
Contrastive matching			
CLIP	**35.67**	14.67	11.67
CLIP fine-tuned	**72.00**	**69.33**	**63.00**
BLIP	**75.67**	**56.00**	**51.00**
BLIP fine-tuned	**86.33**	**82.67**	**79.67**
ITM matching			
BLIP	**94.67**	**89.00**	**87.33**
BLIP fine-tuned	**96.00**	**96.67**	**95.33**

Table 3: Performance improvements on the ColorSwap test set post-finetuning on the ColorSwap train set. Results above chance are **bold**.

Additionally, we extend our evaluation to the Winoground dataset (Thrush et al., 2022) and show the results in Table 4. Even though the finetuned models are able to learn a sensitivity to word order in our minimal color composition task, performance on more complicated compositional tasks remains largely unaffected. Given that there is no practical issue stopping our models from learning sensitivity to word order in the simpler case, compositional understanding for Winoground may simply be a matter of pretraining data or scale.

Model & Method	Text ↑	Image ↑	Group ↑
Contrastive matching			
CLIP	**31.25**	11.25	9.00
CLIP fine-tuned	23.00	9.75	6.25
BLIP	**37.75**	15.75	12.75
BLIP fine-tuned	**33.25**	17.50	13.25
ITM matching			
BLIP	**48.50**	24.50	**20.25**
BLIP fine-tuned	**46.25**	**26.75**	**26.75**

Table 4: Performance on Winoground before and after finetuning on the ColorSwap dataset. Results above chance are **bold**. There is no major difference.

4 Conclusion

We introduce the ColorSwap dataset, a collection of 2,000 unique image-caption pairs and 2,000 hard negative pairings. It is specifically designed to evaluate and improve minimal compositional color comprehension abilities of vision and language models. Our methodology for assembling this dataset involved the use of diffusion models for image generation and the incorporation of human input to ensure naturalness and accuracy. We show that popular off-the-shelf vision and language models exhibit extreme limitations in comprehending even this basic color composition task. However, minimal fine-tuning of these models on the ColorSwap dataset significantly improves their basic understanding of word order.

5 Limitations

All of the images in ColorSwap come from diffusion models. The poor performance of models on the dataset could come from the fact that the diffusion images are simply out of distribution, not because these models have issues with color-word compositionality - although we provide some evidence against this in Appendix E. Similarly, the captions may not be "in-distribution" for a variety of reasons. There is a risk of misattributing failure reasons. However, if a model scores well, then we can be assured that the model is able to correctly differentiate between color-swapped images in this particular setting. ColorSwap is additionally a static dataset, meaning that models evaluated on it must be vetted for training data contamination.

6 Ethical Considerations

For all diffusion models and datasets that we used, we believe that our use is consistent with their intended use and licenses. However, the diffusion models that we used are trained on a variety of opaque sources from the internet, and may make use of data without creators' consent.

References

Anthropic. 2023. `https://www.anthropic.com/index/introducing-claude`.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *ICCV*.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *ICML*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Midjourney. 2023. `https://www.midjourney.com/`.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers. *ECCV*.

OpenAI. 2022. ChatGPT.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis.

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 *ICML*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *CVPR*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023a. Better aligning text-to-image models with human preference. In *ICCV*.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. 2023b. The role of chain-of-thought in complex vision-language reasoning task.

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 *The Eleventh International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Diffusion Model Usage Details

Stable Diffusion (Rombach et al., 2022). We use Stability AI's Stable Diffusion XL Base 1.0 model for generating images for 994 captions from the rule-based approach. Each caption is suffixed with `", 4k"` to optimize the image quality, signaling the model to generate high-resolution images. Additionally, we incorporate `"ugly"` as a negative prompt, which guides the model to avoid outputting deformed images. We set the generation guidance scale to 7.5, the number of denoising steps to 50, the dimension of the generated images to 1,024 $\times$ 1,024 pixels, and the number of images per caption to 8. Although this method is cost-effective, allowing for local running of the model, it sometimes struggles to generate at least one correct image, resulting in a number of failures.

Midjourney (Midjourney, 2023). To tackle harder handmade captions, we explore Midjourney, known for its detailed and artistic image-generation capabilities. We access this model through its Discord server, which offers a straightforward and interactive platform. To ensure the accuracy of the generated images, we often revise and refine the captions. However, this iterative approach resulted in extended processing times, leading to the inclusion of only 39 images from Midjourney in our final dataset. The image dimensions are 512 $\times$ 512 or 1,024 $\times$ 1,024 pixels.

DALL·E 3 (Betker et al., 2023). Finally, we utilize OpenAI's DALL·E 3 for its advanced capabilities in generating highly realistic and detailed images. This model is particularly adept at handling complex and nuanced captions, making it an ideal choice for our more challenging captions from the handmade and generative model approach. While DALL·E 3 is preferred for its high performance and scalability through its API, it comes with a high cost of $0.04 per image. Additionally, DALL·E 3 still makes mistakes, which adds to the expense. Our dataset includes a total of 955 images generated using this model. The image dimensions are 1,024 $\times$ 1,024 or 1,024 $\times$ 1,792 pixels.

B Annotator Interfaces

In this section, we provide snapshots of the annotator interfaces. See Figure 6 and Figure 7 below.

C Experiment Configurations

Models selection. For CLIP, we select the base model that utilizes a ViT-B/32 Transformer architecture as an image encoder. For FLAVA, we select the full model that also has a ViT-B/32 as its encoders. For BLIP, we choose their base model trained on COCO dataset. For SigLIP, we select the base model pre-trained on WebLi at resolution 224x224. The Hugging Face model names for each of the models are listed in Table 5.

Model	Hugging Face Model
CLIP	`openai/clip-vit-base-patch32`
FLAVA	`facebook/flava-full`
BLIP	`Salesforce/blip-itm-base-coco`
SigLIP	`google/siglip-base-patch16-224`
LLaVAR	`truehealth/LLaVar`
LLaVA-1.5	`liuhaotian/llava-v1.5-13b`

Table 5: **Selected models and Hugging Face model names**

Visual language model evaluation prompts. We obtain the text score by prompting visual language models to select the correct caption from two options (Text prompt) and the image score by prompting them to select the correct image from two images (Image prompt). Figure 4 shows these evaluation prompts. For LLaVAR and LLaVA-1.5 where their model interfaces do not directly support multiple images in the input, we horizontally concatenate the images instead.

Text prompt

```
{image} Does this image present (A)
{caption_1}, or (B) {caption_2}?
Note, you must choose one of the two
options.
```

Image prompt

```
{image_1}, {image_2} Which image better
aligns with the description {caption}?
The first image or the second image?
Note you must choose one of two
options.
```

Figure 4: **Visual language model evaluation prompts.** We replace `{image}` with an image and `{caption}` with an appropriate caption.

D Finetuning Details

We train CLIP on the training split of the ColorSwap dataset for 100 epochs. The initial learning rate is $2 \cdot 10^{-5}$ with a linear decay schedule. We employ the Adam optimizer with a weight decay of 0.1 during training, and the batch size is configured to 64. We also fine-tune the BLIP image-text matching model for 100 epochs. The initial learning rate is $1 \cdot 10^{-5}$ with a linear decay schedule, the weight decay is 0.05, and the batch size is 32.

E Is Low Performance Caused by OOD Diffusion Images?

To better understand whether the poor performance on the ColorSwap dataset is due to the images being AI-generated or the models' inability to distinguish compositional color relationships, we run an additional experiment. Instead of choosing between an image-caption pair and a color-swapped version, we assess whether the models can distinguish between an image-caption pair and another randomly selected image-caption pair from our dataset. If the models' poor performance is entirely due to their inability to interpret diffusion-generated images, we would expect them to perform poorly on this task as well. We evaluate some of the worst-performing models from our main experiments, along with BLIP, on this task.

Model & Method	**Text ↑**	**Image ↑**	**Group ↑**
Random chance	25.00	25.00	16.67
Contrastive matching			
CLIP	**100.00**	**99.67**	**99.67**
FLAVA	**99.33**	**100.00**	**99.33**
BLIP	**100.00**	**100.00**	**100.00**
ITM matching			
FLAVA	**82.67**	**97.67**	**81.00**
BLIP	**100.00**	**100.00**	**100.00**

Table 6: Performance of models on distinguishing matching ColorSwap image-caption pairs from randomly selected non-matching pairs. Results above chance are **bold**.

Table 6 suggests that the models can distinguish between matching and non-matching image-text pairs even when the images are AI-generated. This provides some evidence that the poor performance on the ColorSwap dataset is not due to a specific issue with the out-of-distribution nature of AI-generated images.

F Qualitative GPT-4V Evaluation

In Figure 5, we include three examples from our visual language model evaluation on GPT-4V.

G Confidence Intervals

We provide confidence intervals for the overall model results on ColorSwap in Table 7.

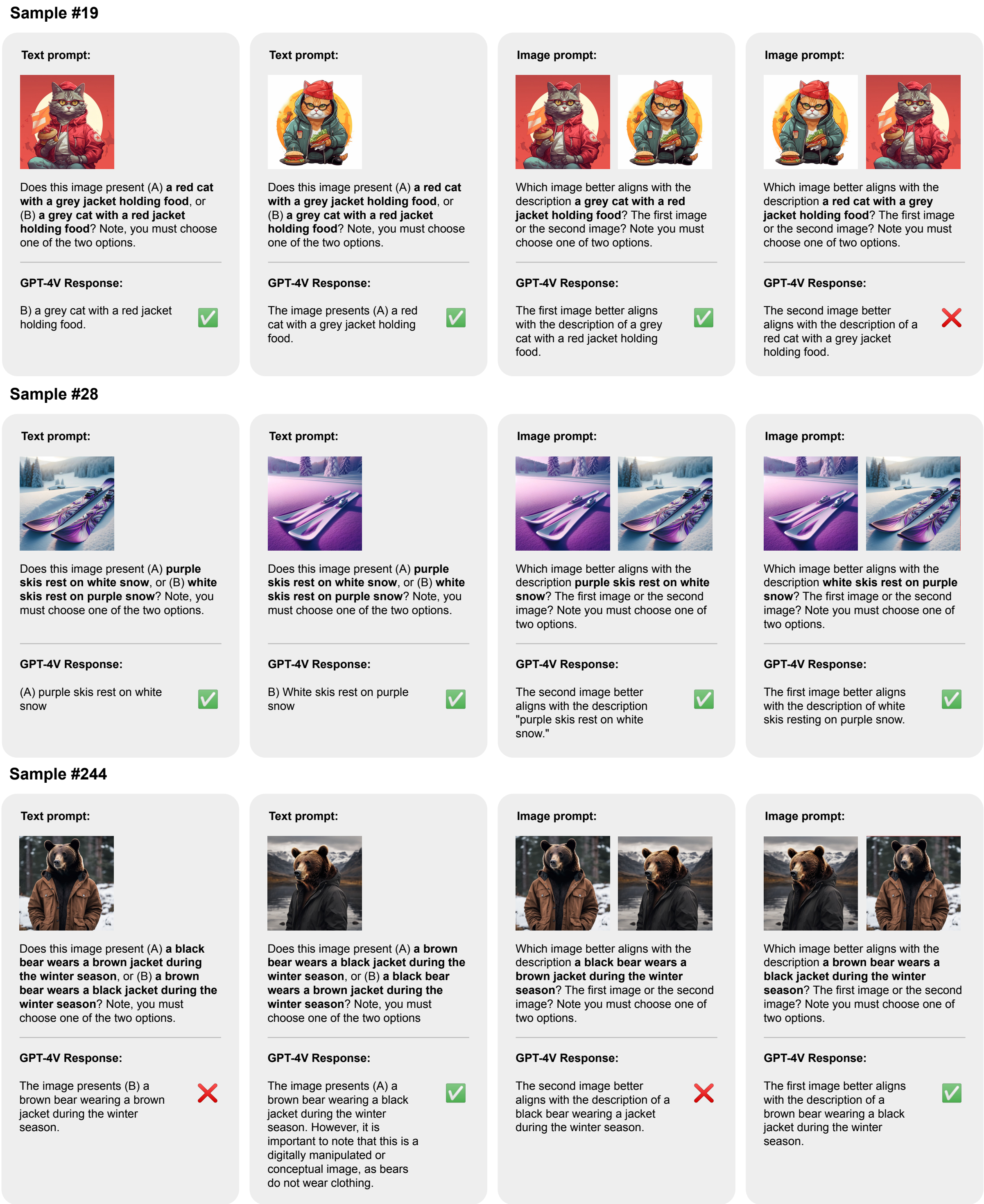

Figure 5: Example 19, 28, and 244 of the ColorSwap dataset. The responses are generated by GPT-4V given different captions and images.

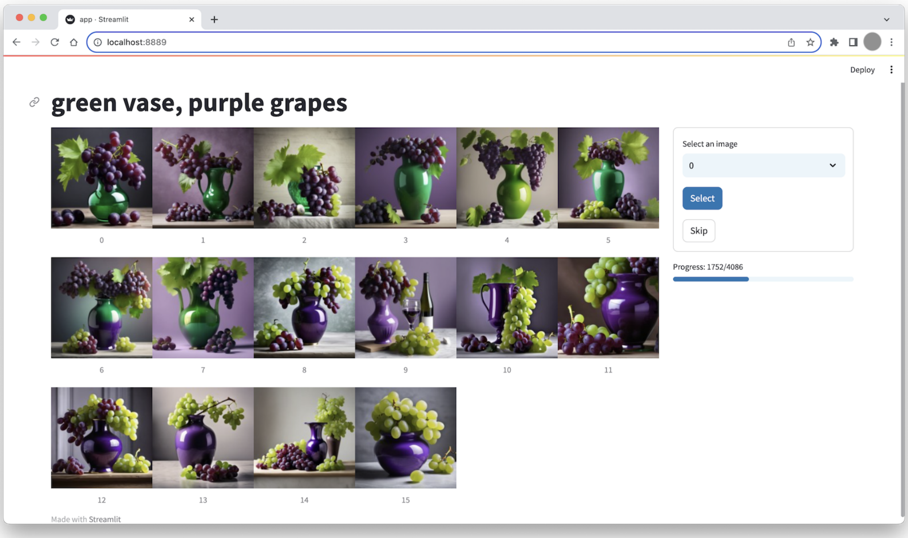

Figure 6: The interface for the selection and filtering process. It allows annotators to choose the image they believe best matches a given caption, with the option to skip if no image seems appropriate. The images presented to the annotators include those generated from the correct and incorrect captions within the same example. This approach is based on the understanding that diffusion models can produce accurate images even from the wrong caption (e.g. a diffusion model could generate an image of a green vase with purple grapes from the caption "purple vase, green grapes", which would be correct for the other caption in the pair).

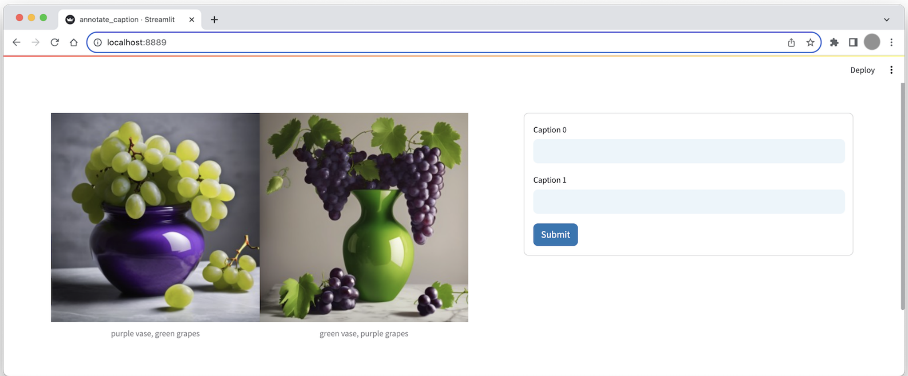

Figure 7: The re-captioning interface. The annotators are provided with two selected images to rewrite the captions. Both images are shown simultaneously so the annotators can infer common things in both pictures and add them to the captions for more nuanced examples.

Model & Method	**Text ↑**	**Image ↑**	**Group ↑**
	Image-text matching models		
Random chance	25.00	25.00	16.67
Contrastive matching			
CLIP	**35.67** [**30.33**, **41.00**]	14.67 [11.33, 19.33]	11.67 [8.67, 16.00]
FLAVA	**35.33** [**30.33**, **41.33**]	25.00 [20.33, 30.00]	15.67 [11.67, **20.00**]
BLIP	**75.67** [**70.67**, **80.33**]	**56.00** [**50.33**, **61.67**]	**51.00** [**45.33**, **56.67**]
SigLIP	**61.67** [**56.00**, **67.00**]	**37.00** [**31.67**, **42.67**]	**30.33** [**25.33**, **35.67**]
ITM matching			
FLAVA	**36.33** [**31.33**, **42.00**]	18.67 [15.00, 23.67]	10.33 [7.67, 14.67]
BLIP	**94.67** [**91.67**, **97.00**]	**89.00** [**85.00**, **92.00**]	**87.33** [**83.00**, **90.67**]
	Visual Language Models		
Random chance	25.00	25.00	6.25
LLaVAR	**27.67** [**22.67**, **33.00**]	**25.67** [**21.00**, **31.00**]	**8.33** [5.67, **12.00**]
LLaVA-1.5	**69.67** [**64.33**, **74.67**]	**54.33** [**48.36**, **59.67**]	**42.00** [**36.33**, **47.33**]
GPT-4V	**91.33** [**87.67**, **94.33**]	**76.33** [**71.33**, **81.00**]	**72.00** [**66.67**, **77.00**]

Table 7: Performance of models on ColorSwap with confidence intervals. Results above chance are **bold**. Note that random chance is different in the ITM versus visual language model (VLM) cases because VLMs output a binary value and ITM models output a float (an effectively continuous value).