PRISM: A New Lens for Improved Color Understanding

Arjun R. Akula[*]
Google DeepMind

Garima Pruthi
Google

Inderjit S Dhillon
Google

Pradyumna Narayana
Google

Sugato Basu
Google

Varun Jampani[†]
Stability AI

Abstract

While image-text pre-trained models, such as CLIP, have demonstrated impressive capabilities in learning robust text and image representations, a critical area for substantial improvement remains—precise color understanding. In this paper, we address this limitation by introducing **PRISM**, a simple yet highly effective method that extends CLIP's capability to grasp the nuances of precise colors. PRISM seamlessly adapts to both recognized HTML colors and out-of-vocabulary RGB inputs through the utilization of our curated dataset of 100 image-text pairs, which can be effortlessly repurposed for fine-tuning with any desired color. Importantly, PRISM achieves these enhancements without compromising CLIP's performance on established benchmarks. Furthermore, we introduce a novel evaluation framework, **ColorLens**, featuring both seen and unseen test sets that can be readily repurposed to assess a model's precision in understanding precise colors. Our comprehensive evaluation and results demonstrate significant improvements over baseline models. Project page: https://prism-google.github.io

1 Introduction

Vision-language foundation models (VLMs), such as Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), learn transferable rich knowledge in a joint space for vision and language with remarkable zero-shot and few-shot capability in 2D visual recognition tasks such as classification (Zhang et al., 2021; Zhou et al., 2022b), detection (Gu et al., 2021), retrieval (Jia et al., 2021), and text-conditioned image generation (Rombach et al., 2022). Recently, several techniques have been proposed to improve the fine-tuning stability of CLIP, enabling it to adapt and generalize effectively to a variety of tasks (Zhou et al., 2022a; Paiss et al., 2023; Zhang et al., 2022). However, despite emerging as a robust representation learner for text and images, a notable gap remains—an inadequacy in precise color understanding, a fundamental component of visual information that has been relatively underexplored.

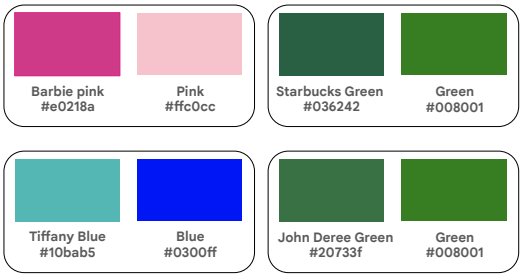

Figure 1: **Brand-Specific Colors versus Standard Colors**. This figure illustrates the contrasts between standard color shades and the unique, brand-specific shades used by well-known brands. The juxtaposition highlights the significance of precise color recognition in brand identity and consumer perception.

The significance of precise color understanding resonates profoundly in practical domains, particularly in advertising and branding, where it plays a pivotal role in establishing brand recognition and influencing consumer perceptions. Colors not only significantly influence consumer buying decisions, enhancing brand recognition and impacting visual appeal, but also evoke specific emotional and psychological responses crucial for brand identity. Several brands have invested significantly in establishing brand identity by designing unique (or non-standard HTML) color palettes, creating a visual language that is instantly recognizable worldwide, as shown in Figure 1. Failure to accurately recognize these unique shades in vision-language models would lead to significant shortcomings in downstream generation tasks (see (c) in Figure 2), such as automated advertising or brand-related content creation. Therefore, enhancing the color discernment capabilities of these models is not just a

[*] Correspondence to arjunakula@google.com.
[†] Work done at Google.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1659–1670
November 12-16, 2024

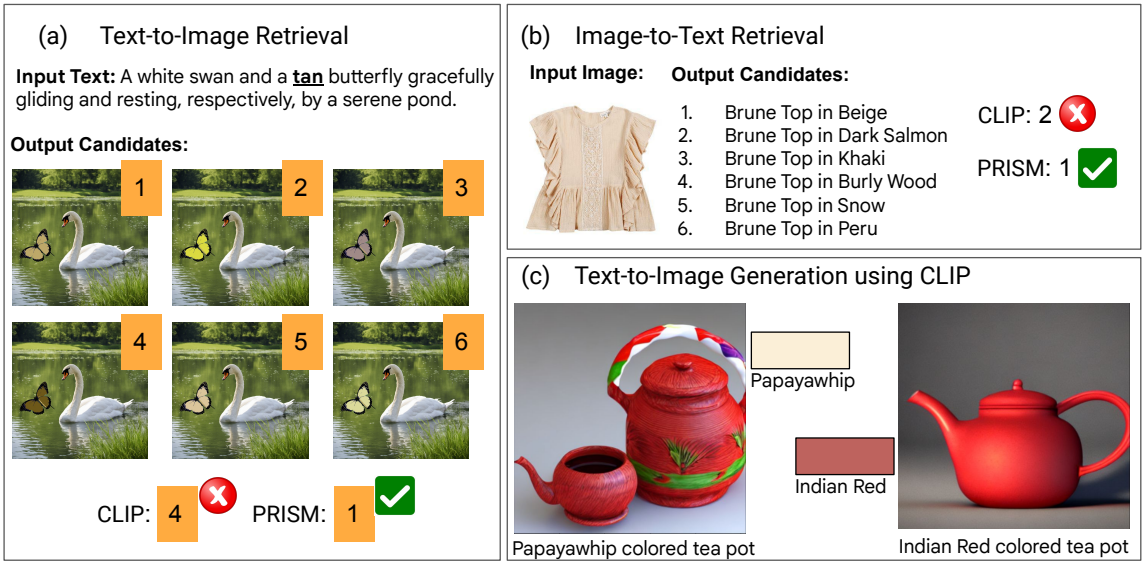

Figure 2: **Comparing CLIP and our proposed method PRISM:** (a) In image retrieval task, where precise RGB colors (e.g., D2B48C = tan color) are crucial, CLIP struggles in accurately retrieving images that match the specified color while PRISM excels at distinguishing and retrieving the correct color among subtle variations; (b) Similarly, in text retrieval, PRISM outperforms CLIP by achieving more precise matches between textual descriptions and corresponding images; (c) Few example images generated using Stable Diffusion 1.5 (with CLIP as text encoder) demonstrating noticeable discrepancy in accurately rendering desired color shades.

technical challenge, but a necessity for maintaining brand integrity in digital representations.

As illustrated in Figure 2, when tasked with retrieving images based on exact RGB colors (e.g., D2B48C representing tan color), CLIP frequently struggles to accurately retrieve images that align with the specified color, particularly when colors exhibit subtle resemblances. This limitation not only impacts the performance of image retrieval tasks but also extends to downstream applications reliant on VLMs, including image generation models, which face difficulties with generating images consistently adhering to the precise color palette.

The direct fine-tuning of VLMs for this purpose encounters inherent challenges, including the risks of overfitting and mode-collapse, primarily stemming from the limited availability of image-text pairs explicitly describing precise colors. In this work, we introduce **PRISM**, to address these limitations. At its core, our principal objective revolves around expanding the pre-trained representational domain, ensuring effective encapsulation of a one or more desired RGB color values, all the while retaining the VLM performance on established benchmarks. To achieve this, we meticulously construct a training set comprising 100 diverse and high-quality image-text pairs. We show that our curated training set can be seamlessly repurposed for fine-tuning, facilitating the implantation of any desired RGB triplet with remarkable ease.

In order to enhance the efficiency of fine-tuning, especially with the constraint of a relatively small set of examples, we introduce explicit hard negatives and encourage the learning of a disentangled embedding for the desired color. For RGB triplets not recognized as standard HTML colors, we employ a rare-token lookup in the vocabulary (Ruiz et al., 2023). Additionally, we construct a new benchmark **ColorLens** that can be readily repurposed to measure a model's precision in understanding precise colors. Our empirical findings demonstrate a significant enhancement over baseline models in retrieval tasks.

2 Related Work

Foundational vision-language (VL) models, designed to bridge image representation with text embedding, have achieved remarkable performance across a broad spectrum of uni-modal and multimodal applications (Chen et al., 2020; Kamath et al., 2021; Li et al., 2020; Lu et al., 2019). CLIP, as a widely acclaimed VL model, undergoes pre-training through a contrastive learning approach, leveraging a vast dataset of 400 million image-caption pairs sourced from the internet and revealed surprising capacities of open-vocabulary recognition and domain generalization (Radford et al., 2021; Zhou et al., 2022c). While CLIP and its

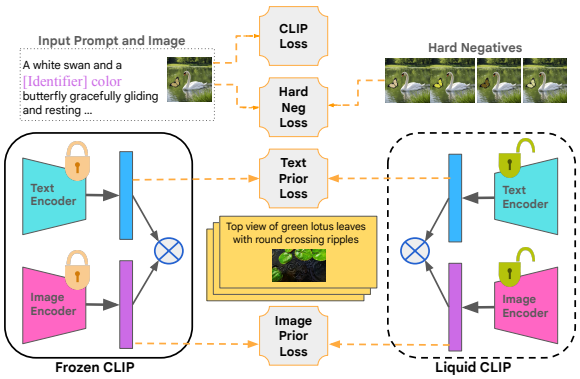

Figure 3: **Our proposed PRISM-based fine-tuning approach:** We implant unrecognized HTML colors using unique tokens, integrate hard negatives for disentangling color-relevant information, and employ regularization losses to preserve image and text embeddings, resulting in improved model performance. The overall loss function combines these elements to enhance the understanding of precise RGB colors in our fine-tuning process.

variants have received considerable attention in the context of prompt tuning (He et al., 2022; Zhou et al., 2022b) and continual fine-tuning (Garg et al., 2023; Ding et al., 2022), there has been no prior investigation dedicated to exploring the realm of precise color understanding.

CLIP Adaptation: Due to CLIP's versatility, several studies have adapted it for various purposes. Recent works such as Structured Vision Language Concepts (SVLC) (Doveh et al., 2023; Zhao et al., 2022) have demonstrated that using a 'bag of objects' in both images and text is sufficient for optimizing CLIP-Loss, resulting in a failure to differentiate fine-grained language nuances and comprehend structured concepts such as object attributes and relationships. Some works spot the limitations of CLIP in compositional reasoning and propose extensions to enhance the reasoning skills, rectifying object bias, and addressing associations (Liu et al., 2021; Yamada et al., 2022; Thrush et al., 2022). Another line of research has focused on improving methods for assessing both the perceived quality and abstract perception of images without task-specific training (Wang et al., 2023). This includes investigations into novel tasks like recoloring images to enhance specific emotions and providing textual rationales for such recoloring. However, there has been no prior work explicitly dedicated to improving the precise color comprehension capabilities of CLIP.

3 Method

In this section, we first describe the construction of our repurposable training and testing datasets, then present our fine-tuning paradigm in detail. Our primary objective is to enhance CLIP's nuanced understanding of colors by learning disentangled embeddings for the desired color using our curated small set of training examples, all while simultaneously preserving the semantic context of images and text.

3.1 Dataset Construction

While an abundance of paired image and text data exists, there is a lack of paired image-and-text data consisting of precise RGB colors of the objects depicted in the image. Therefore, to enable training and thorough evaluation of our proposed method, we undertook a systematic and controllable approach to synthesize the data splits leveraging the latest advancements in large language models, text-to-image generation, and object segmentation techniques.

We initiate the dataset creation process by harnessing the capabilities of GPT-4 (OpenAI, 2023). Our goal here is to generate text prompts that accurately describe objects while explicitly specifying their color attributes. The text prompts generated by GPT-4 subsequently undergo a manual review process. The aim is to ensure that the generated prompts are diverse, clear, and explicitly conveyed the color attributes of the depicted objects. Below are the sample instructions that we provide to GPT-4:

"Generate a series of descriptive text prompts for images, focusing on the precise depiction of objects with specific color values. Each prompt should:

1. ***Describe a Unique Object:*** *Choose an object for each prompt. This could range from everyday items like a fruit, a car, or clothing, to more unique or imaginative objects like a fantasy creature or futuristic technology.*

2. ***Specify Object Color*** *Include color for the key object. For example, 'A ripe banana with a red skin color resting on a wooden table'*

3. ***Provide Context and Detail:*** *Add descriptive details about the object's setting, texture, size, and any other relevant characteristics to create a vivid picture. For example, 'The banana is slightly curved, with small brown spots, indicating ripeness, and lies next to a steel knife'.*

4. ***Ensure Clarity and Simplicity:*** *While being detailed, keep the descriptions clear and*

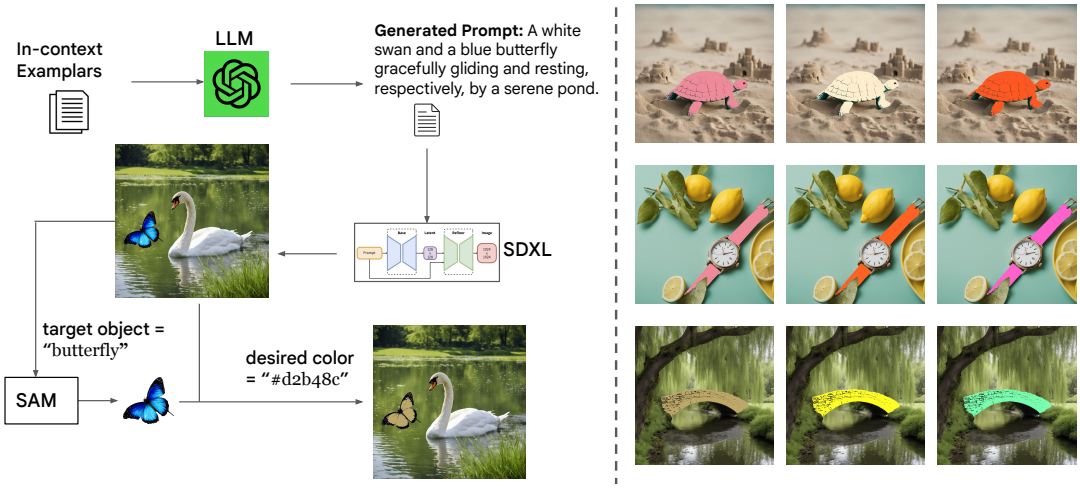

Figure 4: **Overview of our dataset construction process** highlighting the integration of GPT-4 for text prompt generation, Stable Diffusion XL for image synthesis, and SAM for segmentation, facilitating the creation of our train and eval splits. The left side illustrates the step-by-step pipeline for image generation, while the right side showcases diverse examples of images produced through our approach.

straightforward to facilitate accurate image generation. Avoid overly complex sentences or ambiguous descriptions.

5. ***Incorporate Interaction if Relevant:*** *If applicable, describe the object in interaction with other elements or characters to add dynamism to the scene. For example, 'A child in a bright green t-shirt holding the banana, ready to take a bite.' "*

For generating corresponding images that align with the curated text prompts, we employ Stable Diffusion XL (Podell et al., 2023). We used DreamStudio service (`https://beta.dreamstudio.ai/`) to generate images from the text prompts using SDXL 1.0. For each prompt, we generate between 4 to 8 images and then we manually select one image that most faithfully represent the intended text prompt. In total, we curated a training set consisting of 100 image-text pairs. To train our model to recognize any RGB color, we repurpose these curated images by segmenting object pixels using a object segmentation module (Kirillov et al., 2023) and modifying the segmented pixels to match specified RGB colors[1]. We used Segment Anything ViT-H model to identify object segmentation masks (`https://github.com/facebookresearch/segment-anything`). Figure 4 illustrates our dataset construction process.

[1]Using our controllable generation approach, we ensure a diverse range of prompts and control over object (color) modifiability.

For evaluating our model, we introduce **ColorLens**, comprising two critical evaluation settings: text-to-image retrieval and image-to-text retrieval. For the text-to-image retrieval setting, we create Test Seen and Test Unseen splits using the same pipeline discussed above, each with 50 image-text pairs and hard negatives. Seen split includes familiar objects with hard negatives, while unseen split involves unseen objects during finetuning allowing us to assess the model's generalization capabilities. In the image-to-text retrieval setting, we collect 100 image-text test pairs, consisting of 20 color shades. Images for the common colors are sourced from the extensive LAION-400M dataset (Schuhmann et al., 2021) and the rest are generated synthetically using the above pipeline. We add hard negatives by replacing the color name in the text prompts with the closest color shades.

3.2 PRISM

Rare-token Identifiers: In order to implant unrecognized HTML colors, we associate it with a unique token in the vocabulary. For instance, we use the format "chair in [identifier] color", where [identifier] serves as a distinct label linked to the desired RGB color. Following a similar approach as outlined in Ruiz et al. (2023), we conduct a rare-token lookup within the CLIP vocabulary to obtain three-letter unique identifiers (e.g., 'hta') that has no particular strong associations with specific concepts or meanings.

Disentangled Fine-tuning: In order to facilitate disentanglement of color-relevant information from

color-irrelevant details, we integrate hard negatives into our fine-tuning framework. In each fine-tuning step, we leverage the original ground truth image and its hard negative images, systematically generated by manipulating RGB channels. Alongside the original CLIP contrastive loss (L_{clip}) for text and ground-truth images, we incorporate a weighted hard negative loss (L_{hard}) with the specific aim of minimizing the CLIP similarity between the text description and the hard negative images.

Image and Text Prior Preservation: When we unfreeze all parameters in both the text-encoder and image-encoder, the model exhibits signs of overfitting to our limited training data, leading to language drift issues (Ruiz et al., 2023). To address this challenge, we adopt a strategy of sampling 5000 image-text pairs from the LAION 400M dataset, focusing on color-related content and encompassing a diverse range of objects. We then apply a regularization loss, denoted as $L_{\text{i\_prior}}$ for image embedding preservation and $L_{\text{t\_prior}}$ for text embedding preservation, designed to preserve the image and text embeddings for these 5000 pairs during fine-tuning. Below is the overall loss function (L) we use in fine-tuning and we illustrate the approach in Figure 3.

$$L = L_{\text{clip}} + \lambda_1 \cdot L_{\text{hard}} + \lambda_2 \cdot L_{\text{i\_prior}} + \lambda_3 \cdot L_{\text{t\_prior}} \quad (1)$$

4 Experiments

In the section, we present a comprehensive evaluation of our proposed method, PRISM, on retrieval tasks using our newly introduced ColorLens test splits specifically designed for assessing precision in color-based retrieval tasks. We evaluate our approach from both quantitative and qualitative perspectives. Through ablation studies, we systematically dissect the contributions of each component within our framework, highlighting their individual effectiveness in enhancing the model's performance. Our experiments consistently demonstrate the superiority of PRISM over state-of-the-art methods, including CLIP and ALBEF (Li et al., 2021), both in fine-tuning and adapter tuning settings [2]. Additionally, to provide a more comprehensive perspective, we extend the comparison to include models such as ViT-L-14 and ViT-B-32 for all the models.

In Tables 1 and 2 we compare PRISM against established methods across both seen and unseen ColorLens test splits. The evaluation is multi-faceted: The first column compares the original image-text matching performance using precision and rank metrics against a backdrop of 20 randomly selected negatives from the test set. The second column extends this challenge by using the entire test suite as potential negatives. The third column specifically targets color-based retrieval performance, introducing 'hard-negatives' that are identical in every aspect except for distinct differences in RGB color values of specific objects. These hard-negatives are crafted to vary in their deviation from the true color values, with smaller delta values ($\delta < \epsilon_1$) and larger ones ($\epsilon_1 < \delta < \epsilon_2$) to escalate the retrieval difficulty (see section A in supplementary for more details). Furthermore, in the final two columns, we increase the number of negatives from 27 to 64, testing the models' robustness under more challenging conditions.

As we can see, while most baseline models, both in their zero-shot and fine-tuned forms, exhibit strong performance in standard image-text matching (as evidenced in the first two columns of the results), there is a noticeable tendency for direct fine-tuning to lead to overfitting. This is particularly evident in the performance dip observed from CLIP to its fine-tuned variant (FT CLIP) in traditional matching tasks. Contrasting this, PRISM stands out by not only improving precision and ranking in the color-focused retrieval tasks but also maintaining robust performance in standard image-text matching. This clearly indicates PRISM's unique ability to enhance color-specific understanding while preserving the foundational semantics of the models. Further strengthening our findings, similar trends of PRISM's efficacy are observed in the unseen test split.

Ablations Table 3 presents an ablation of our PRISM model, specifically the Ours+FT B/32 variant. The significant difference in precision and recall between ablated versions and our method demonstrate the importance of the Prior Preservation Loss and Hard Negative Loss in our framework. Notably, the removal of the Prior Preservation Loss leads to enhanced performance in color-specific retrieval tasks, however it results in a notable decrease in performance for standard image-text matching. This suggests a pronounced risk of overfitting when trained on a limited dataset. On the other hand, omitting the Hard Negative Loss maintains the model's performance in standard image-text matching scenarios but significantly di-

[2] We ensure that the baseline models are appropriately fine-tuned to provide fair comparisons.

	20 Neg		*ALL Neg*		$\delta < \epsilon_1$ *(27)*		$\delta < \epsilon_2$ *(27)*		$\delta < \epsilon_1$ *(64)*		$\delta < \epsilon_2$ *(64)*	
Model	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓
CLIP B/32	98.0	1.02	98.0	1.06	8.0	10.80	14.0	10.28	2.0	24.90	12.0	21.96
CLIP L/14	**100.0**	**1.00**	98.0	1.04	8.0	10.14	18.0	8.50	4.0	22.80	12.0	18.30
ALBEF	90.0	2.59	87.0	2.76	4.0	15.10	7.0	16.55	1.0	27.91	6.0	28.70
ALBEF (MSCOCO)	91.0	2.40	88.0	2.60	4.0	14.89	8.0	16.54	1.0	27.90	7.0	26.89
ALBEF (Flickr30k)	92.0	1.97	90.0	2.00	5.0	13.34	9.0	14.34	1.0	27.60	8.0	25.70
CLIP Adapter B/32	**100.0**	**1.00**	98.0	1.04	10.0	10.28	22.0	8.14	4.0	21.84	16.0	16.88
CLIP Adapter L/14	**100.0**	**1.00**	**100.0**	**1.00**	12.0	10.34	18.0	8.64	6.0	23.42	14.0	18.70
FT CLIP B/32	**100.0**	**1.00**	96.0	1.06	12.0	10.50	16.0	8.76	6.0	22.82	10.0	18.68
FT CLIP L/14	**100.0**	**1.00**	98.0	1.02	10.0	10.44	16.0	8.64	6.0	23.54	10.0	19.30
Ours+Adap B/32	97.0	1.06	97.0	1.06	20.0	4.82	**52.0**	3.82	10.0	9.00	**38.0**	6.28
Ours+Adap L/14	**100.0**	**1.00**	98.0	1.02	10.0	8.24	22.0	6.78	7.0	15.08	20.0	14.70
Ours+FT B/32	**100.0**	**1.00**	98.0	1.04	24.0	4.24	40.0	**3.46**	**20.0**	**7.30**	36.0	**5.04**
Ours+FT L/14	**100.0**	**1.00**	98.0	1.04	**30.0**	**4.04**	40.0	3.90	18.0	7.54	34.0	5.96

Table 1: Evaluation of PRISM and baseline models on the ColorLens seen test split, demonstrating PRISM's enhanced precision and rank in color-based retrieval (last four columns) and consistent Performance in standard image-text matching (first two columns).

	20 Neg		*ALL Neg*		$\delta < \epsilon_1$ *(27)*		$\delta < \epsilon_2$ *(27)*		$\delta < \epsilon_1$ *(64)*		$\delta < \epsilon_2$ *(64)*	
Model	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓
CLIP B/32	**100.0**	**1.00**	**100.0**	**1.00**	10.0	7.32	20.0	7.74	8.0	16.80	16.0	15.32
CLIP L/14	**100.0**	**1.00**	**100.0**	**1.00**	12.0	8.00	22.0	7.60	10.0	17.78	18.0	16.80
ALBEF	89.0	2.50	88.0	2.52	6.0	14.10	10.0	13.90	5.0	19.70	7.0	21.12
ALBEF (MSCOCO)	90.0	2.45	89.0	2.50	5.0	13.76	11.0	13.01	5.0	19.52	7.0	21.01
ALBEF (Flickr30k)	92.0	2.00	91.0	2.01	7.0	11.00	14.0	11.23	6.0	18.00	10.0	19.00
CLIP Adapter B/32	98.0	1.02	98.0	1.02	12.0	6.58	32.0	5.36	8.0	14.12	22.0	11.46
CLIP Adapter L/14	**100.0**	**1.00**	**100.0**	**1.00**	10.0	7.92	24.0	7.58	10.0	18.14	20.0	16.84
FT CLIP B/32	**100.0**	**1.00**	**100.0**	**1.00**	10.0	8.86	18.0	8.12	6.0	20.28	12.0	18.06
FT CLIP L/14	**100.0**	**1.00**	**100.0**	**1.00**	8.0	8.52	20.0	7.78	8.0	19.54	12.0	18.08
Ours+Adap B/32	**100.0**	**1.00**	96.0	1.06	28.0	4.32	46.0	3.50	20.0	7.74	36.0	6.20
Ours+Adap L/14	**100.0**	**1.00**	**100.0**	**1.00**	14.0	5.02	28.0	5.58	11.0	13.14	22.0	12.76
Ours+FT B/32	**100.0**	**1.00**	**100.0**	**1.00**	**34.0**	**3.06**	50.0	**2.64**	**30.0**	**5.54**	40.0	4.58
Ours+FT L/14	**100.0**	**1.00**	**100.0**	**1.00**	24.0	3.32	**60.0**	2.74	16.0	5.96	**50.0**	**3.52**

Table 2: Performance of PRISM and baseline models on the ColorLens unseen test split.

minishes its effectiveness in distinguishing subtle color differences, indicating that while it effectively preserves the integrity of semantic representations, it struggles in the nuanced task of color differentiation.

4.1 Image-to-Text Retrieval

In the image-to-text retrieval setting, our evaluation strategically focuses on testing the generalization capabilities of our proposed PRISM method with real images. The images in this split of ColorLens stands in contrast to synthetic images used previously in the text-to-image retrieval test splits. From the LAION-400M dataset, we specifically mine images corresponding to 20 HTML color shades. When certain shades are not present in LAION-400M, we generate additional images using the pipeline detailed in section 3.1. We fine-tune the model using the PRISM method with our repurposable train images generated for each of these 20 shades and conduct comparative evaluations against the baseline models. The experimental setup for this task involves matching the given image with the correct text caption, emphasizing the precision of color identification.

The results, summarized in Table 4, demonstrate that PRISM significantly outperforms all baseline models. This remarkable performance indicates that our synthetic training dataset is highly effective in enhancing performance on real images. Furthermore, the results of our ablated model, displayed in Table 5, reaffirm the critical role of the Prior Preservation Loss and Hard Negative Loss in our framework. These components are instrumental in maintaining the balance between color-specific accuracy and overall image-text matching perfor-

	20 Neg		*ALL Neg*		$\delta < \epsilon_1$ *(27)*		$\delta < \epsilon_2$ *(27)*		$\delta < \epsilon_1$ *(64)*		$\delta < \epsilon_2$ *(64)*	
Model	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓
Ours+FT B/32	100.0	1.00	100.0	1.00	34.0	3.06	50.0	2.64	30.0	5.54	40.0	4.58
w/o Prior Loss	91.0	1.32	91.0	1.36	40.0	2.10	70.0	1.28	38.0	3.30	68.0	1.98
w/o HN Loss	100.0	1.00	100.0	1.00	4.0	9.40	24.0	8.04	4.0	20.20	14.0	16.94

Table 3: **Ablation study** of the PRISM model (Ours+FT B/32 variant) on ColorLens unseen test split, showing the impact of prior preservation loss and hard negative loss on color differentiation capabilities.

	Neg		*Hard Neg*	
Model	p@1 ↑	rank ↓	p@1 ↑	rank ↓
CLIP B/32	**100.0**	**1.00**	8.0	9.20
CLIP L/14	**100.0**	**1.00**	11.0	10.00
ALBEF	93.0	1.85	6.0	16.10
ALBEF (MSCOCO)	94.0	1.80	6.0	16.00
ALBEF (Flickr30k)	96.0	1.50	7.0	15.05
CLIP Adapter B/32	98.0	1.02	11.0	10.02
CLIP Adapter L/14	**100.0**	**1.00**	**11.0**	**10.00**
FT CLIP B/32	**100.0**	**1.00**	**9.0**	**10.90**
FT CLIP L/14	**100.0**	**1.00**	**10.0**	**9.50**
Ours+Adap B/32	**100.0**	**1.00**	25.0	5.00
Ours+Adap L/14	**100.0**	**1.00**	22.0	6.05
Ours+FT B/32	**100.0**	**1.00**	**31.0**	**4.06**
Ours+FT L/14	**100.0**	**1.00**	28.0	4.70

Table 4: Performance of PRISM and baseline models on the ColorLens image-to-text retrieval test split. The column Neg quantifies the performance of standard image-text matching, while the last two columns are dedicated to color-based retrieval - assessing the models' proficiency in accurately identifying and matching specific color shades with their corresponding text descriptions.

	Neg		*Hard Neg*	
Model	p@1 ↑	rank ↓	p@1 ↑	rank ↓
Ours+FT B/32	100.0	1.00	31.0	4.06
w/o Prior Loss	90.0	1.85	38.0	3.80
w/o HN Loss	100.0	1.00	5.0	18.50

Table 5: **Ablation Study** of the PRISM Model (Ours+FT B/32 Variant) on the ColorLens image-to-text retrieval test split, demonstrating the importance of both Prior Preservation Loss and Hard Negative Loss components on the model's ability to discern and match specific color shades.

mance, as evident from the substantial difference in results with and without these elements in our model.

5 Conclusion

We have presented PRISM, a novel and effective framework designed to address the critical challenge of precise color understanding. Leveraging a carefully curated training dataset comprising 100 image-text pairs, PRISM enables the seamless implantation of any desired RGB color value while preserving the core performance of CLIP on established benchmarks. Through the incorporation of explicit hard negatives, disentangled color embeddings, and rare-token lookup mechanisms, we have ensured the robustness and generalization of our approach. Furthermore, we introduced the ColorLens benchmark, encompassing both seen and unseen test sets, which provides a comprehensive evaluation of a model's ability to understand precise colors. Our empirical results demonstrate significant quantitative and qualitative improvements over baseline models in color-based retrieval tasks. We believe that PRISM has the potential to foster enhanced color-aware applications in various practical domains, from advertising to image generation.

References

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 *European conference on computer vision*, pages 104–120. Springer.

Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. 2022. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. 2023. Tic-clip: Continual training of clip models. *arXiv preprint arXiv:2310.16226*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. 2022. Cpl: Counterfactual prompt learning for vision and language models. *arXiv preprint arXiv:2210.10362*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

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. *Advances in neural information processing systems*, 34:9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. 2021. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*.

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. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

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. *arXiv preprint arXiv:2111.02114*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563.

Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. 2022. When are lemons purple? the concept association bias of clip. *arXiv preprint arXiv:2212.12043*.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562.

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 preprint arXiv:2207.00221*.

Chong Zhou, Chen Change Loy, and Bo Dai. 2022a. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022c. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

In this supplementary material, we provide additional details on our data collection, experiments and evaluation to supplement main paper.

A More Results

Figure 5 shows few qualitative results for text-to-image retrieval comparing CLIP and PRISM. PRISM accurately matches the specific shades in prompts such as 'A green bicycle and a golden retriever puppy with a slate blue ball', demonstrating its fine-tuned color differentiation, which CLIP struggles with.

We showcase qualitative results for image-to-text retrieval in Figure 6. While both CLIP and PRISM show proficiency in identifying standard HTML colors like red and violet, CLIP noticeably struggles with more nuanced shades such as Indian red and lawn green. This distinction underscores PRISM's superior ability in color discernment.

B Text-to-Image Retrieval

For our experiments, we generated hard negative images systematically by manipulating RGB channels. Specifically, we reduce individual color channels (R, G, or B) by a specified delta value, creating hard negatives that closely resemble the original images while differing only in color. Hard-negatives are crafted to vary in their deviation from the true color values, with smaller delta values ($\delta < \epsilon_1$) and larger ones ($\epsilon_1 < \delta < \epsilon_2$) to escalate the retrieval difficulty. In all our experiments, we used $\epsilon_1 = 30$ and $\epsilon_2 = 70$, where each of the RGB color values range between 0 to 255.

C Image-to-Text Retrieval

In the image-to-text retrieval setting, we focus on evaluating the generalization capabilities of our proposed PRISM method with real images and fine-tuning with multiple colors simultaneously. From the LAION-400M dataset, we specifically mine from 5-10 images corresponding to 20 HTML color shades. Specifically, we considered the following 20 HTML colors in our evaluation: `red`, `tomato`, `coral`, `indian red`, `light coral`, `green`, `lawn green`, `forest green`, `lime`, `lime green`, `cyan`, `light cyan`, `dark turquoise`, `turquoise`, `pale turquoise`, `plum`, `violet`, `orchid`, `fuchsia`, and `pink`. There are only a few standard colors in this selected list such as red, green, violet and cyan. When certain color shades are not

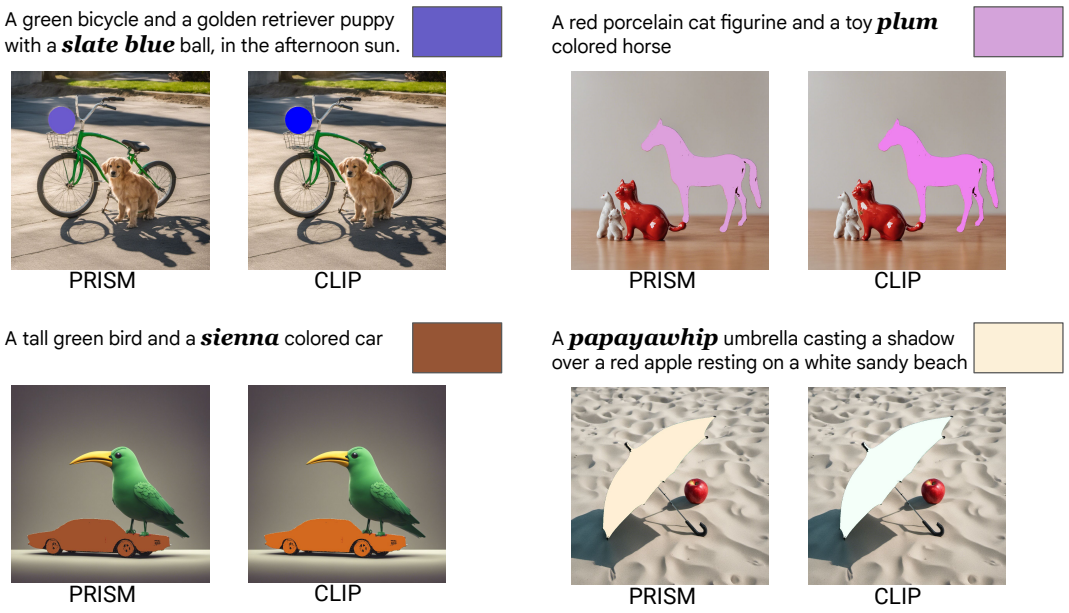

Figure 5: **Comparative visualizations of text-to-image retrieval results using CLIP and PRISM** for color-specific prompts. The examples illustrate PRISM's enhanced ability to accurately match detailed color descriptions, such as 'slate blue ball' and 'papayawhip umbrella', demonstrating its advanced color understanding compared to CLIP.

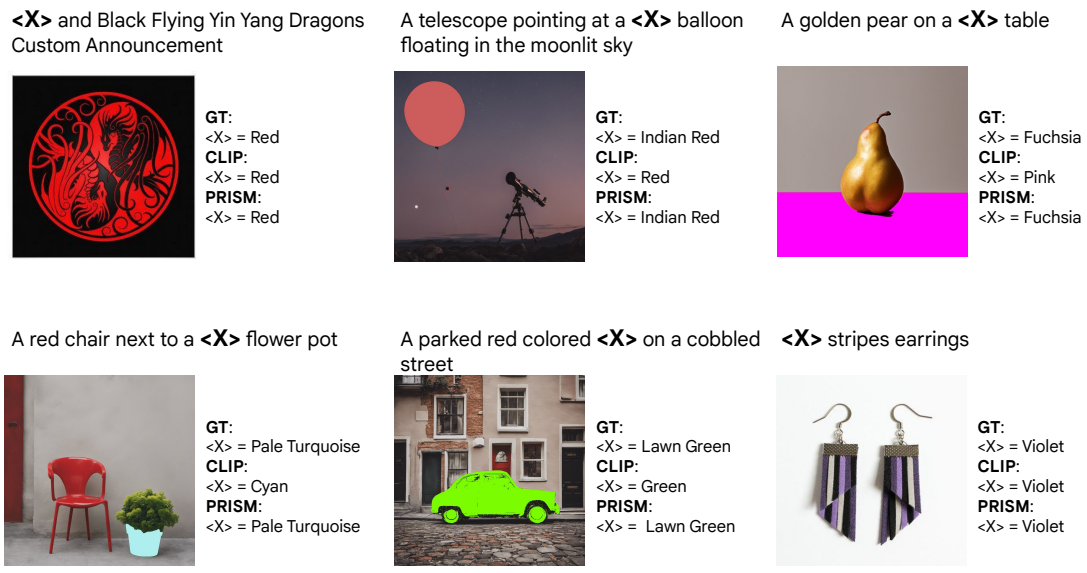

Figure 6: **Comparative visualizations of image-to-text retrieval results using CLIP and PRISM** for color-specific prompts. '<X>' represents specific color references. The corresponding ground-truth color used is denoted as 'GT'.

Model	*20 Neg*		*ALL Neg*		$\delta < \epsilon_1$ *(27)*		$\delta < \epsilon_2$ *(27)*		$\delta < \epsilon_1$ *(64)*		$\delta < \epsilon_2$ *(64)*	
	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓
CLIP B/32	98.0	1.02	98.0	1.06	8.0	10.80	14.0	10.28	2.0	24.90	12.0	21.96
CLIP Adapter B/32	100.0	1.00	98.0	1.04	10.0	10.28	22.0	8.14	4.0	21.84	16.0	16.88
FT CLIP B/32	100.0	1.00	96.0	1.06	12.0	10.50	16.0	8.76	6.0	22.82	10.0	18.68
Ours+FT B/32 (1 color)	100.0	1.00	98.0	1.04	24.0	4.24	40.0	3.46	20.0	7.30	36.0	5.04
Ours+FT B/32 (5 colors)	100.0	1.00	98.0	1.04	23.0	4.80	40.0	3.46	21.0	7.00	36.0	5.04

Table 6: Comparison of PRISM performance in ColorLens seen test when fine-tuned with 1 versus 5 colors in Text-to-Image Retrieval setting.

Model	*20 Neg*		*ALL Neg*		$\delta < \epsilon_1$ *(27)*		$\delta < \epsilon_2$ *(27)*		$\delta < \epsilon_1$ *(64)*		$\delta < \epsilon_2$ *(64)*	
	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓	p@1 ↑	rank ↓
CLIP B/32	100.0	1.00	100.0	1.00	10.0	7.32	20.0	7.74	8.0	16.80	16.0	15.32
CLIP Adapter B/32	98.0	1.02	98.0	1.02	12.0	6.58	32.0	5.36	8.0	14.12	22.0	11.46
FT CLIP B/32	100.0	1.00	100.0	1.00	10.0	8.86	18.0	8.12	6.0	20.28	12.0	18.06
Ours+FT B/32 (1 color)	100.0	1.00	100.0	1.00	34.0	3.06	50.0	2.64	30.0	5.54	40.0	4.58
Ours+FT B/32 (5 color)	100.0	1.00	100.0	1.00	34.0	3.06	49.0	2.85	30.0	5.58	40.0	4.50

Table 7: Comparison of PRISM performance in ColorLens unseen test when fine-tuned with 1 versus 5 colors in Text-to-Image Retrieval setting.

present in LAION-400M captions, we generate additional images using the pipeline detailed in Section 3.1 of main paper. For example, we couldn't obtain any caption containing the color "indian red" from LAION-400M (https://huggingface.co/datasets/laion/laion400m). In fine-tuning our model, we leverage our proposed 100 train samples in section 3.1. In other words, for each of the 20 colors we generate 100 train samples along with their corresponding hard negatives.

D Random Samples from Train Split

In Table 9, we show random selection (text prompt generated from gpt-4 and the corresponding image generated by sdxl) of the samples from our proposed train split of 100 samples.

E Results on Common Benchmarks

In Table 8, we show the zero-shot performance of CLIP and PRISM on CIFAR 10, CIFAR 100 and Caltech101 datasets. The results clearly indicate that our model with image and text prior preservation losses doesn't show any significant drop in the accuracy on these common benchmarks.

E.1 More Evaluations for Text-to-Image Retrieval

In this section, we compare the performance of PRISM fine-tuned with single color versus multiple colors. As shown in Table 6 and Table 7, we do not see any significant difference between model performance when fine-tuned with 1 and 5 colors.

Model	CIFAR10	CIFAR100	Caltech101
CLIP B/32	58.7	29.8	71.0
PRISM (**Ours+FT** B/32)	58.6	29.8	70.8

Table 8: Zero-shot accuracy comparison of PRISM and CLIP on common benchmarks.

	**Prompt** A yellow book next to a `red` vase
	**Prompt** Amidst a field of golden wheat a solitary `crimson` barn stands, its rustic appearance hinting at stories of the past
	**Prompt** A `blue` kite soaring high amidst fluffy white clouds, its tail trailing gracefully.
	**Prompt** A cherry tree in an orchard petals drifting gently to the ground with a `red` chair next to it
	**Prompt** A `red` wine barrel in a cool cellar, surrounded by aged bottles on wooden racks
	**Prompt** A sleek `green` violin resting on a satin cushion, with soft lighting

Table 9: Random examples from our proposed train split