ColorFoil: Investigating Color Blindness in Large Vision and Language Models

Ahnaf Mozib Samin[†◇*] **M Firoz Ahmed**[△] **Md. Mushtaq Shahriyar Rafee**[§]
[†]Queen's University, Kingston, Canada
[◇]University of Malta, Msida, Malta
[△]Shahjalal University of Science and Technology, Sylhet, Bangladesh
[§]Metropolitan University, Sylhet, Bangladesh
ahnaf.samin@queensu.ca[†], mfiroz.sust@gmail.com[△], rafee@metrouni.edu.bd[§]

Abstract

With the utilization of Transformer architecture, large Vision and Language (V&L) models have shown promising performance in even zero-shot settings. Several studies, however, indicate a lack of robustness of the models when dealing with complex linguistics and visual attributes. In this work, we introduce a novel V&L benchmark - ColorFoil, by creating color-related foils to assess the models' perception ability to detect colors like red, white, green, etc. We evaluate seven state-of-the-art V&L models including CLIP, ViLT, GroupViT, and BridgeTower, etc. in a zero-shot setting and present intriguing findings from the V&L models. The experimental evaluation indicates that ViLT and BridgeTower demonstrate much better color perception capabilities compared to CLIP and its variants and GroupViT. Moreover, CLIP-based models and GroupViT struggle to distinguish colors that are visually distinct to humans with normal color perception ability.

1 Introduction

Vision and language models (V&L) have exhibited improved performance for many V&L tasks in recent years (Lu et al., 2019; Su et al., 2019; Chen et al., 2020; Li et al., 2020; Radford et al., 2021; Dou et al., 2022). Thus, the current paradigm has now been shifting towards zero-shot learning, where models are evaluated without fine-tuning for specific tasks (Radford et al., 2021). Large-scale V&L models, in particular, show promise for task-independent zero-shot evaluation (Radford et al., 2021).

Several studies have been conducted to perform comprehensive evaluations of V&L models on a variety of tasks to identify their strengths and weaknesses (Agrawal et al., 2016; Jabri et al., 2016; Goyal et al., 2017; Shekhar et al., 2017; Agarwal et al., 2020). For instance, the VALSE evaluation benchmark has been proposed to assess the state-of-art V&L models for challenging linguistic constructs (Parcalabescu et al., 2021a). Therefore, five distinct tasks, including existence, plurality, counting, relations, actions, and coreference, have been introduced. In this benchmark, foils are generated from the existing V&L datasets for each of the tasks. A foil is referred to as a distractor or slightly incorrect example that is passed along with the correct example to the V&L model to assess the model's ability to correctly distinguish them (Shekhar et al., 2017; Parcalabescu et al., 2021a). Although the existing V&L benchmarks like VALSE help the community to test the capabilities of V&L models, there is still much work to be done to evaluate the robustness and generalizability of the models on numerous other tasks. It remains unknown how well the large V&L models can perceive colors from the visual content.

Color perception requires a human-like understanding of visual content. Thus, by evaluating the V&L models on color attributes, we can determine how closely the large V&L models perceive colors to humans. A V&L model can be biased towards detecting particular colors and perform poorly with others. Therefore, it is essential to investigate it in order to improve the explainability and interpretability of the models. By assessing the V&L models with their color-perception ability, we can ensure robustness in real-life applications.

In this study, we aim to shed light on the following research question: how well can the state-of-the-art large-scale V&L models perceive color-related attributes, such as red, green, yellow, etc.? Our contributions are mainly twofold:

- We introduce a novel V&L benchmark **ColorFoil** by creating foils from the MS COCO and Flickr30k datasets (Lin et al., 2014; Plummer et al., 2015) to investigate how well the

[*]Work performed while at University of Malta.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 294–300
April 30 - May 1, 2025 ©2025 Association for Computational Linguistics

models perceive and identify the color-related attributes.

- We perform a comparison between seven of the state-of-the-art V&L models including CLIP (Radford et al., 2021), ViLT (Kim et al., 2021), ViT (Dosovitskiy et al., 2020) and BridgeTower(Xu et al., 2022b) using our benchmark.

The outline of this paper is as follows. We provide a background study in Section 2. In Section 3, we describe the process of constructing ColorFoil from the MS COCO dataset. Experiments and results are discussed in Section 4. In Section 5, we discuss the limitations of our work. Ethical considerations are provided in 5. A conclusion and future scope is presented in Section 5.

2 Background

V&L Models The current state-of-art models are first pre-trained in a self-supervised way with a multi-task learning objective. The learning objectives can be predicting the masked texts or masked region in the images, determining whether or not the image and text corresponds, etc. The text and image input features can be concatenated together and passed to a Transformer encoder. This approach is known as single stream. Alternatively, the text and image inputs can be separately encoded to two different Transformers and then additional layers to merge them into multi-modal features.

CLIP Contrastive Language-Image Pre-training (CLIP) is a V&L model that is pre-trained with 400M image-text pairs with a contrastive objective (Radford et al., 2021). The model jointly trains a text encoder and an image encoder to maximize the cosine similarity of the image-text embeddings of real pairing while minimizing the cosine similarity of the embeddings of the incorrect pairings within a multi-modal embedding space. Each of the encoders are based on transformers. CLIP demonstrates the ability to perform zero-shot visual classification, object detection, and image generation tasks.

ViLT Vision-and-Language Transformer (ViLT) is pre-trained using a Transformer with more than 4M images with two objectives such as image text matching and masked language modeling (Kim et al., 2021). The text embedding and the image features are concatenated into a sequence and then fed into the transformer. Thus, ViLT is a single stream model. ViLT achieves competitive or better performance than other V&L models on downstream tasks while being 10 times faster due to simpler processing of visual inputs.

BridgeTower There is a visual encoder, a textual encoder and a cross-modal encoder with multiple lightweight bridge layers in the BridgeTower architecture (Xu et al., 2022b). The top layers of the unimodal encoders and each layer of the cross-modal encoder are connected with the bridge layers, thus enabling extensive interactions at each layer of the cross-modal encoder. Each of visual, textual and cross-modal encoders is transformer-based encoders. The model is pre-trained with 4M images with two common objectives: masked language modeling and image text matching. The model is found to outperform in all downstream V&L tasks with negligible additional computational cost.

ViT A Vision Transformer (ViT) is designed for image classification tasks, adapting the Transformer architecture from natural language processing (Dosovitskiy et al., 2020). It divides an image into fixed-size patches, linearly embeds each patch, and treats these embeddings as sequences akin to word tokens in text. Using self-attention mechanisms, the ViT captures global image context more effectively than convolutional networks, allowing for superior performance on large-scale image datasets. ViTs leverage transfer learning and pretraining for enhanced accuracy and efficiency.

GroupViT (Group Vision Transformer) is a variant of the Vision Transformer designed to improve efficiency and scalability in image classification tasks (Xu et al., 2022a). It enhances the standard ViT by introducing a group-wise processing mechanism, where the input image is divided into smaller groups of patches. Each group is processed independently through parallel self-attention layers, reducing computational complexity. The results from these groups are then aggregated to form a cohesive representation. GroupViT aims to retain the global context modeling capabilities of ViTs while optimizing resource usage, making it more suitable for large-scale and real-time applications.

Related Work Several V&L tasks include visual question answering (Goyal et al., 2017), visual reasoning (Suhr et al., 2018), image retrieval (Plummer et al., 2015), etc. Foiling is an approach that slightly edits the original captions to evaluate the robustness of the V&L models (Shekhar et al., 2017). Similar to our work, Shekhar et al. (2017) foiled the MS COCO dataset, and constructed the

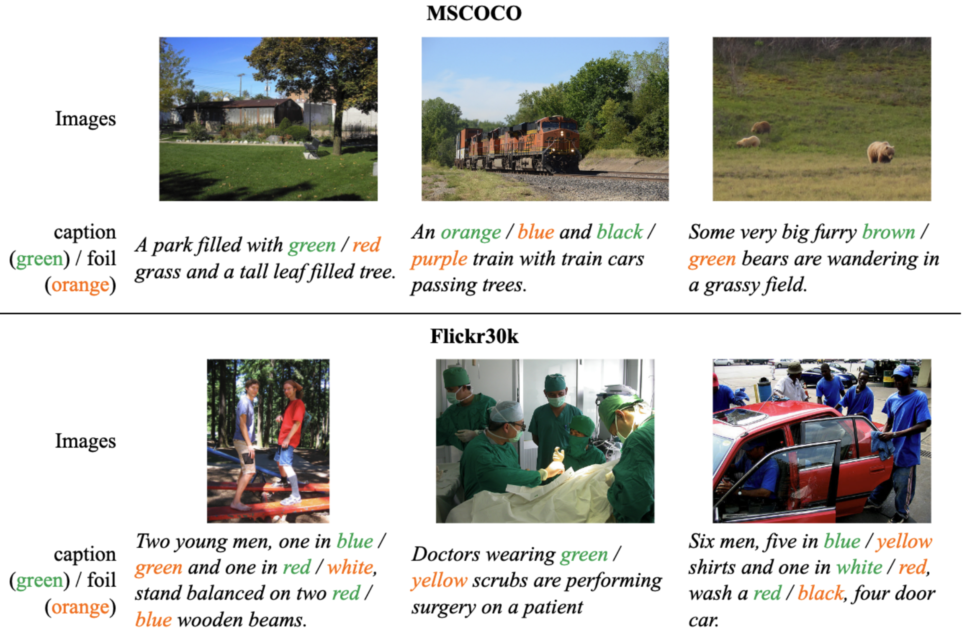


Figure 1: Examples from the ColorFoil benchmark where color-related attributes in the original captions have been modified to different colors.

FOIL-COCO dataset. However, their work did not focus on the perception of colors of the V&L models. Following the work of Shekhar et al. (2017), several studies have been performed that evaluated the V&L models (Shekhar et al., 2019; Gokhale et al., 2020; Bitton et al., 2021; Parcalabescu et al., 2021b; Rosenberg et al., 2021).

3 Construction of the ColorFoil Benchmark

The ColorFoil benchmark is automatically derived from the MS COCO (Microsoft Common Objects in Context) and Flickr30k dataset, which serves as a resource for studying image understanding, object recognition, image captioning, and visual question-answering tasks (Lin et al., 2014; Plummer et al., 2015). In the MS COCO dataset, textual annotations are provided solely for the train and validation (val) sets. To construct the ColorFoil, we obtain the images and annotations from the 2017 MS COCO validation set, resulting in a total of 5,000 image-text pairs. Among these instances, each of 2,511 pairs includes at least one word related to color. For Flickr30k dataset, we use the standard val and test sets to prepare the ColorFoil benchmark.

Our aim is to foil only the color name from the textual input, leaving the original image and the rest of the text input as it is. For example, given a caption like *A **blue** bus driving down a street past a park.* We foil the color-related word, resulting in a modified sentence like - *A **brown** bus driving down a street past a park.* If there are multiple color attributes in a caption, we foil all of them.

We utilize the **webcolors 1.3** python package to determine whether a substring within a caption corresponds to a color (Webcolors, 2023). This package encompasses a total of 147 colors. Our filtering process involves excluding captions that lack color names and selecting solely those containing at least one color name.

When replacing the original color name with a foiled alternative, we consider the most widely used colors. The chosen target colors for foiling consist of "blue", "black", "red", "pink", "yellow", "grey", "orange", "white", "green", and "brown." So, rather than utilizing the complete list of 147 colors from the **webcolors** package, we opt for a narrower selection of common colors for foiling. This decision is based on the fact that numerous colors in the package have limited practical usage (e.g. medium blue, mint cream, etc.). The target color for foiling is selected randomly from the 10 common colors. If the original color in the caption is one of the common colors, we randomly select any other common color for foiling except for the one found in the caption.

After excluding four instances of two-dimensional grayscale images due to compatibility issues with certain models, our resulting dataset

Models	1 Foil				2 Foils				4 Foils			
	MSCOCO		Flickr30k		MSCOCO		Flickr30k		MSCOCO		Flickr30k	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
ALIGN	86.03	93.32	87.70	93.45	79.47	88.75	81.43	89.76	71.03	83.06	74.57	85.43
AltCLIP	84.89	91.82	82.69	90.52	77.29	87.19	73.68	84.84	69.08	81.71	64.39	78.34
BridgeTower	97.31	98.63	96.83	98.32	95.71	97.81	94.46	97.15	92.61	96.16	90.81	95.18
CLIP	84.42	91.55	85.24	92.07	76.19	86.49	76.26	86.53	67.37	80.50	68.33	81.18
CLIPSeg	83.05	91.09	82.01	90.12	74.00	85.42	72.97	84.37	64.56	78.45	63.07	77.35
GroupViT	82.73	91.67	81.64	89.89	73.10	83.98	71.77	83.57	63.80	77.89	62.12	76.63
ViLT	95.69	97.79	94.29	97.06	92.83	96.28	91.85	95.35	88.74	94.04	87.38	93.27

Table 1: **Experiment results.** We evaluate seven of the state-of-the-art V&L models on the MS COCO and Flickr30k subsets from ColorFoil. Accuracy (%) and F1-scores (%) are reported. We conduct three experiments in which the models are presented different number of foils (modified caption) along with the original caption. The V&L models tend to struggle in challenging conditions with more foils. BridgeTower and ViLT outperform other V&L models including CLIP and its variants and GroupViT by a large margin.

comprises 2,507 pairs of RGB images along with their captions and foils from MSCOCO and 2500 pairs of RGB image-caption pairs from Flickr30k. To ensure data integrity, we conduct manual validation on a significant number of image-text pairs randomly selected from the benchmark and find no anomalies. Examples of original captions and corresponding foils are illustrated in Figure 1.

4 Experiments

Experimental Setup: We pass the original caption, foil as well as the corresponding image to a V&L model. The model provides the logits for each of the caption and foil corresponding to the image. We take the softmax of the logits. Our hypothesis is that a model with a well-perceivable ability to distinguish colors is supposed to provide a higher probability for the original caption and a lower probability for the foil.

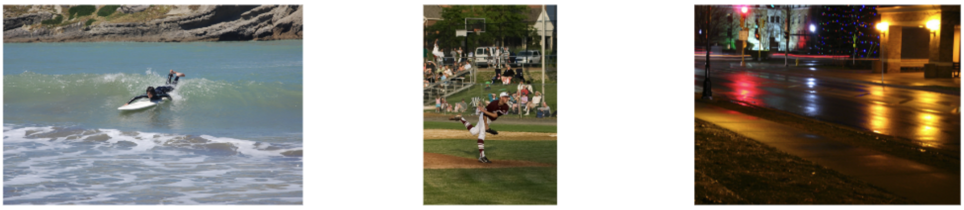
We evaluate all the models in a zero-shot setting. We utilize the HuggingFace transformer library to load the models (Wolf et al., 2019). These models are chosen due to the fact that they represent different architectural variants. CLIP has a text encoder and an image encoder, which are jointly trained with a contrastive loss. ViLT is a single-stream model. BridgeTower contains multiple bridge layers that connect the uni-modal encoders with the cross-modal encoder.

The evaluation metric employed in our study is accuracy and F1-score, which are widely used in similar contexts. To elaborate, if the model accurately identifies the foil in comparison to the original caption, the accuracy of that particular example is incremented.

Results: Table 1 shows the performance of different V&L models evaluated on the ColorFoil. All the models achieve much higher accuracy compared to a baseline random classifier with a 50% accuracy. CLIP obtains 83.1% accuracy while ViLT and BridgeTower get substantially higher accuracy of 95.6% and 97.2%, respectively on the 1-Foil experiment. It is worthwhile to mention that CLIP is pre-trained with 400M images, although this model is outperformed by both ViLT and BridgeTower pre-trained with only 4M images. BridgeTower architecture, which contains multiple bridges to make connections between the uni-modal encoders and the cross-modal encoder, achieves the highest accuracy.

The relatively poor performance of CLIP is also evident in its variants, including AltCLIP (Chen et al., 2023) and CLIPSeg (Lüddecke and Ecker, 2022). While the ALIGN model outperforms CLIP, it still lags behind BridgeTower and ViLT. GroupViT, similar to CLIP, struggles to achieve high performance. This performance trend is consistent across both MSCOCO and Flickr30k datasets, reinforcing our observations. When presented with more foils alongside the original caption, the models exhibit performance degradation. Nonetheless, BridgeTower and ViLT maintain strong performance even under these challenging conditions with more foils.

We present several examples for which the CLIP model incorrectly assigns higher probabilities to the foils (See Figure 2). These examples demonstrate that the CLIP model is unable to distin-



Images			
Captions	*A surfer is riding a wave in light blue water.*	*A man in black jersey pitching in baseball game.*	*A red traffic light at night next to a Christmas Tree.*
CLIP Predictions	*A surfer is riding a wave in light brown water.*	*A man in red jersey pitching in baseball game.*	*A white traffic light at night next to a Christmas Tree.*

Figure 2: Examples for which the CLIP model wrongly choose the foils instead of the captions.

guish between blue-brown, black-red, and red-white pairs, despite the fact that they are visually distinct to most humans.

5 Conclusion and Future Work

In this work, we introduce a novel benchmark, ColorFoil, derived from the MS COCO and Flickr30k datasets, to assess the perception ability of the cutting-edge V&L models to detect colors. To this end, we foil the colors from the original captions and feed both caption and foil along with the corresponding image to the model to observe whether it can provide a higher probability for the caption or not. Seven state-of-the-art V&L models, including CLIP, ViLT, ViT, and BridgeTower, have been benchmarked using the ColorFoil. While all models outperform a random classifier, ViLT and BridgeTower are much more capable to perceive colors compared to CLIP and ViT. This intriguing finding is seen using both MS COCO and Flickr30k datasets, which strengthens our analysis.

As part of our future work, we would like to evaluate the robustness of V&L models on additional tasks by constructing foils that swap gender (man -> woman), size (small -> large), emotions (smiling -> crying), and sentence negation (playing football -> not playing football), etc.

Limitations

We consider the 10 most common colors for our foils. However, our choice of common colors is subjective and there might be other frequently used colors that are not present in our foils.

Ethical Considerations

Training V&L models using images and corresponding texts that may contain gender bias, private data, or harmful content presents challenges in manual detection. To address this, we utilize the widely recognized MS COCO and Flickr30k datasets to create the ColorFoil benchmark, as it provides a reliable foundation (Lin et al., 2014; Plummer et al., 2015).

Ensuring reproducibility is a crucial aspect of scientific research. To foster open research practices, we will make our code publicly accessible, allowing others to reproduce and verify our findings.

Acknowledgments

Work supported by the Language and Communication Technologies program of the Erasmus+ project of the European Commission.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*.

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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. Altclip: Altering the language encoder in clip for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8666–8682.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 727–739. Springer.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021a. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021b. 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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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.

Daniel Rosenberg, Itai Gat, Amir Feder, and Roi Reichart. 2021. Are vqa systems rad? measuring robustness to augmented data with focused interventions. *arXiv preprint arXiv:2106.04484*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. 2019. Evaluating the representational hub of language and vision models. *arXiv preprint arXiv:1904.06038*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Webcolors. 2023. Python package index - webcolors 1.3.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022a. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. 2022b. Bridge-tower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*.