# On the Bias of CLIP for Object-Attribute Recognition

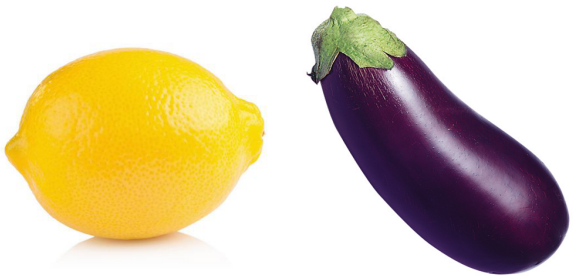
Yutaro Yamada[1*], Yingtian Tang[2*], Ilker Yildirim[1]
[1]Yale University
[2]University of Pennsylvania
{yutaro.yamada,ilker.yildirim}@yale.edu yingtian@seas.upenn.edu

## 概要

Large-scale vision-language models such as CLIP have shown impressive performance on zero-shot image classification and image-to-text retrieval. However, such zero-shot performance of CLIP-based models does not realize in tasks that require a finer-grained correspondence between vision and language, such as Visual Question Answering (VQA). We investigate why this is the case, and report an interesting phenomenon of CLIP, which we call the Concept Association Bias (CAB), as a potential cause of the difficulty of applying CLIP to VQA and similar tasks. CAB is especially apparent when two concepts are present in the given image while a text prompt only contains a single concept. In such a case, we find that CLIP tends to treat input as a bag of concepts and attempts to fill in the other missing concept crossmodally, leading to an unexpected zero-shot prediction.

## 1 Introduction

Recent large-scale vision-language models such as CLIP [1] and ALIGN [2] have shown remarkable performance on zero-shot classification and text-image retrieval tasks. These models are trained via cross-modal contrastive learning on web-scale image-text pairs and obtain powerful multimodal representations. Encouraged by these strong zero-shot capabilities, several recent papers explored CLIP for more complicated vision-language tasks. The initial attempt made by [3] reports near chance accuracy for zero-shot performance of CLIP on VQA-v2 [4], a common visual question answering benchmark. However, their approach simply uses "question: [question text] answer: [answer text]" as text input for the text encoder of CLIP, which makes the prediction harder than it should be. A subsequent work [5] proposes a better prompt generation method. They first convert a question into a masked prompt



CLIP: "In this picture, the color of the lemon is purple."

**図 1** When we ask CLIP the color of the lemon in the above image, CLIP answers "purple". The text prompt we use is "In this picture, the color of the lemon is [mask]", where CLIP picks one from [red, green, yellow, orange, purple].

(e.g. "What's in the bowl behind the cake" becomes "The [mask] is in the bowl behind the cake"), and filter impossible answers using a language model, which improves CLIP's zero-shot performance on VQA-v2.

However, the zero-shot performance of CLIP on VQA-v2 is still not state-of-the-art, which is achieved by task-specific models [3]. While investigating what makes CLIP hard to adapt to VQA, we discover an interesting phenomenon, which we call the Concept Association Bias (CAB).

To describe this phenomenon, we present a simple image containing a "lemon" and an "eggplant" to CLIP, and ask what color the lemon is, as shown in Figure 1. Surprisingly, CLIP predicts "purple" with high confidence. When we instead ask for the color of the eggplant, CLIP answers "yellow". To cross-check this phenomenon, we formulate a binary zero-shot image classification task on the same image where the two labels are "yellow lemon" and "purple lemon", and find that CLIP predicts "purple lemon" with high confidence.

We hypothesize that this phenomenon comes from the discrepancy between what is described in the image and text input, where CLIP attempts to fill in the missing con-

cept. The association between "purple" and "eggplant" is strong, so when asked to fill in the mask in "[mask] lemon", predicting "purple" instead of "yellow" makes more sense for CLIP, because the text description of "purple lemon" is aligned with the image that contains both a lemon and an eggplant more faithfully than "yellow lemon", which only describes the lemon in the image. In fact, when we randomize the color of the lemon and eggplant (e.g. "red" for lemon and "green" for eggplant), we find that this bias disappears, and CLIP picks the color almost randomly between the two.

Vision-language models such as CLIP are being deployed for increasingly broad range of downstream applications [6, 7, 8, 9]. However, the concept association bias suggests caution in such efforts.

## 2 Related Work

**Vulnerability of vision and language models** There are a number of papers that study the robustness of vision and language models. Some prior work [10] shows that Transformer trained via Masked Language Modeling [11] is insensitive to word orders, suggesting that the success of BERT largely depends on learning higher-order word co-occurrence rather than learning syntactic and semantic abstractions. Many benchmarks are proposed to evaluate robustness of ImageNet models towards various perturbations including common corruption [12], image style change [13], and different viewpoints [14]. Our work differs from these studies that are purely based on language or vision, because CAB is a cross-modal phenomenon, which occurs when both image and language data are used. [15] tests compositional generalization of vision and language models. [16] introduced a probing dataset called Winoground, which evaluates visuo-linguistic compositionality of vision and language models. They evaluate a diverse range of state-of-the-art vision and language models, including CLIP, but all of them perform close to or below random chance. Our work also reveals brittleness of CLIP through the lens of CAB, which has been overlooked in the past.

**Peculiarities of CLIP** In the image generation community, it has been reported that state-of-the-art models such as DALL·E 2 [6] struggle with composionality [17]. One of the potential causes of such failure has been attributed to the use of CLIP-based image encoder [6]. In fact, image generation models that do not use CLIP such as Imagen and Parti are known to be better at generating images that require compositional reasoning [18, 19]. However, few works go into depth to analyze the behavior of CLIP in zero-shot image classification and visual question answering. Our analysis based on CAB offers a new perspective on the weakness of CLIP-based models for compositional reasoning.

## 3 The Concept Association Bias

The zero-shot image classification of CLIP is remarkable for images that contain a single concept. However, when there are multiple concepts in the image but the text input does not cover all of them, the zero-shot classification of CLIP can be significantly biased towards the missing concept(s). We call this bias the Concept Association Bias (CAB). We first showcase this bias using color recognition tasks.[1)] For this analysis, we use the Natural-Color Dataset (NCD) [20], which is a dataset of vegetables and fruits with a white background. We take the following objects: *banana, brinjal, broccoli, carrot, cherry, corn, cucumber, lemon, orange, plum, pomegranate, strawberry, tomato*. We then randomly sample two images with different vegetable types and place the two objects side-by-side, resulting in 494 images in total. Examples are shown in Figure 2.

**図 2** Example images from Natural-Color Dataset (NCD) [20], modified for our color recognition tasks so that each image contains two different objects.

For zero-shot transfer from CLIP to our color recognition task, we ask for the color of one of the objects in the image. The labels we use are "red", "yellow", "purple", "green", and "orange", so it is a 5-way color recognition task. When there is a single object in the image, we use the following text prompt: "In this picture, the color of the object is [mask]." When there are two objects in the image, we specify one of these objects in the prompt. For exam-

1) For all experiments in the main text, we use the ResNet50-x4 backbone for CLIP.

ple, if there is a lemon and another object in the image, the prompt takes the following format: “In this picture, the color of the lemon is [mask].”

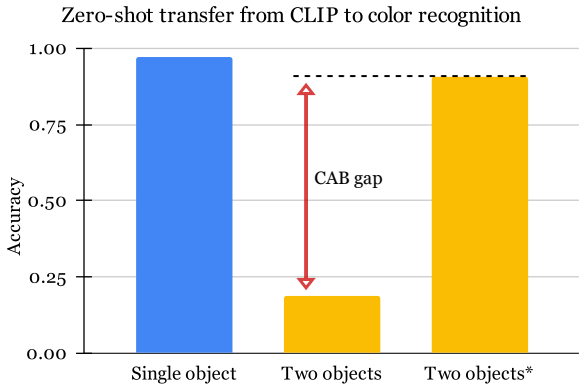



**図 3** Zero-shot performance of CLIP on color recognition tasks using NCD [20]. CLIP achieves almost perfect accuracy when there is a single object in the image, but the accuracy significantly drops when there are two objects. “Two object*” refer to the case in which we instead measure the accuracy of predicting the color of the object B when it is asked for the color of the object A, where we see 80% zero-shot accuracy. We claim this gap between Two objects and Two objects* is a result of the Concept Association Bias (CAB).

The results are shown in Figure 3. We first note that the zero-shot performance of CLIP on our color recognition task is almost perfect when there is a single object per image (“Single object” in Figure 3). However, the classification performance considerably degrades to below chance when there are two objects per image (“Two objects” in Figure 3).

How does this happen? We suggest that CLIP does not have a mechanism that stores object-centric representation that correctly binds the object’s name and its attribute. In another words, CLIP processes its input as a “bag of concepts”.

To inspect this possibility, we look at what kind of mistakes CLIP makes when there are two objects A and B. We find that many mistakes are derived from a common source. That is, when asked for the color of object A, CLIP often predicts the color of object B in the image. In fact, when we measure the accuracy of predicting the color of the object B when in reality it is asked to predict the color of the object A, we see that the zero-shot transfer performance of CLIP is much higher (“Two objects*” in Figure 3), approaching the single object accuracy.

To understand this phenomenon, we find it helpful to consider two variables per object, where each variable represents the object’s name in the image and the color attribute of the object, as shown in Figure 4. When the colors are natural (Figure 4 (a)), both the object “lemon” and its attribute “yellow” in the image are fully explained by the word “lemon” in the text prompt, resulting in the concept of the eggplant remaining. When CLIP performs zero-shot color recognition, we see that placing the color “purple” in the prompt can most faithfully explain the remaining concept of the eggplant in the image (Figure 4 (b)).

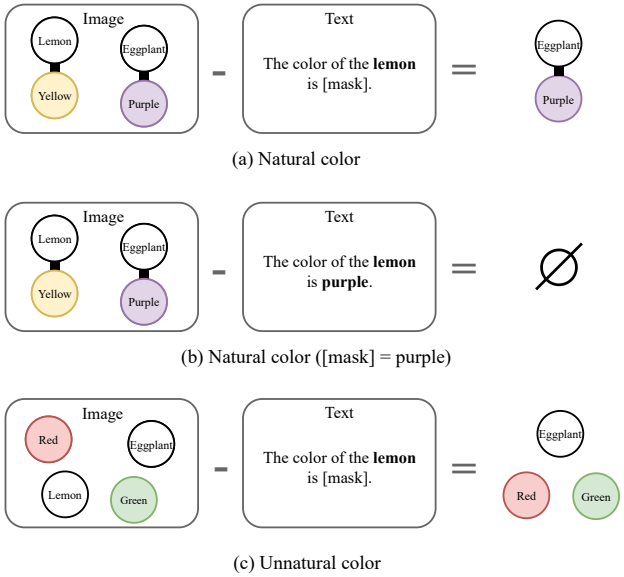



**図 4** The concept binding diagram. Two variables per object represent the object name and its attribute (e.g. color), respectively. We suggest that the text prompt and the image are represented as two separate “bags of concepts” in CLIP. When a pair of object-attribute concepts are naturally associated with each other, then both concepts can be accounted for by including in the prompt either of the object or the attribute. When only some of the concepts in the image are included in the text, this leaves other concepts in the image unaccounted for.

**図 5** Examples from UNCD. Single object (Top) and Two objects per image (Bottom).

The above explanation suggests that there is a strong association between the color “purple” and the object “eggplant” in CLIP to the point where “purple” can partially explain the concept of the eggplant. What if we break this strong association? Does the gap between Two objects and Two objects* disappear?

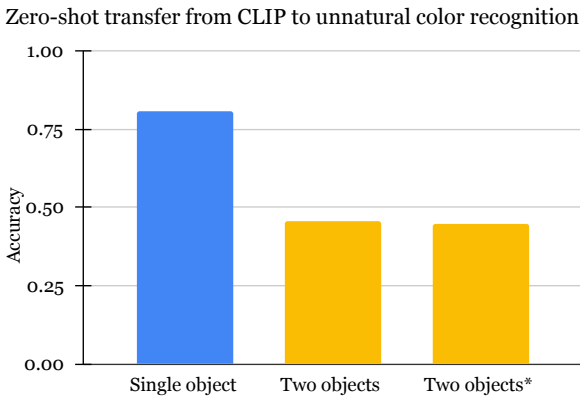



**図 6** Zero-transfer performance of CLIP to color recognition on UNCD, where we assign non-associated color to each vegetable. CLIP achieves 80% accuracy when there is a single object in the image. While the accuracy drops for Two objects, the drop is not as significant as the NCD case. Furthermore, the gap between Two objects and Two objects* vanishes, compared to the NCD case.

To test this, we create a version of NCD, which we call UNnatural-Color Dataset (UNCD), where we artificially change the color of each fruit and vegetable to non-associated color. Examples are shown in Figure 5. We repeat the same experiment on UNCD. The results are shown in Figure 6. We see that the zero-shot performance for a single object is still high, suggesting that CLIP can pick up the color attribute even if the color is not strongly associated with the object itself. However, for the two object cases, we see that there is almost no difference between Two objects and Two objects* tasks. In other words, CLIP predicts the two non-associated colors in the image with almost equal chance.

Why does the CAB gap disappear when objects are paired with random attributes in images? This result arises from a common mechanism that impacts both the Two objects and Two objects* tasks. To see this, we go back to our diagram in Figure 4 (c). When the colors are unnatural (*e.g.*, a lemon in red color and an eggplant in green color), then the remaining bag of concepts that are yet to be explained by the text include "red", "green", and "eggplant". This is because the color "red" is not associated with the concept of "lemon", and therefore the word "lemon" in the text prompt cannot explain the color "red", unlike the case that uses natural color. As a result, CLIP can choose either "red" or "green" for color recognition. And indeed, surprisingly, CLIP randomly chooses between the two – it does not associate the concept of "red" with the lemon, even though in the image the lemon unambiguously appears in red. Likewise, for the Two objects* task (in which the correct prediction is defined as the color of object B when asked for object A), CLIP essentially randomly picks one of the two colors present in the image, despite the fact that each object has their own very distinct color.

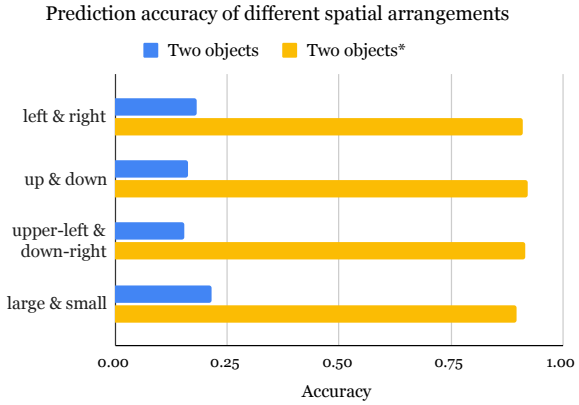



**図 7** The Concept Association Bias (CAB) remains regardless of the spatial configurations such as "left & right", "up & down", "upper-left & down-right", and "large & small". We use the same subset of NCD as in Figure 3.

### 3.1 The spatial arrangement has almost no effect on CAB

In our earlier experiments on NCD and UNCD, two objects are positioned side-by-side. To see if CAB is robust to the positioning of objects, we vary the spatial arrangement of the two objects in the image. Concretely, we test the following spatial configurations: left & right, up & down, and upper-left & down-right. We also vary the size of the two objects for left & right, which is denoted as "large & small". As Figure 7 shows, CAB is not affected by either spatial arrangements or the object size.

## 4 Conclusion

Every object has a set of concepts that are roughly associated with it. For instance, the object "lemon" can be associated with "yellow", "fruit", and so on. Such concept association is automatically learned in vision-language models such as CLIP, to the point where the word "yellow" can partially explain the object "lemon" in certain cases. We establish that the Concept Association Bias (CAB) exists for CLIP through a series of experiments. CLIP is increasingly popular in both computer vision and natural language processing. We hope our work raises awareness of the brittleness of CLIP as we develop new models on top of CLIP.

## Acknowlegments

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used for the experiments in this paper.

## 参考文献

[1] 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. Learning Transferable Visual Models From Natural Language Supervision. In **Proceedings of the 38th International Conference on Machine Learning**, pp. 8748–8763. PMLR.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In **Proceedings of the 38th International Conference on Machine Learning**, pp. 4904–4916. PMLR.

[3] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How Much Can CLIP Benefit Vision-and-Language Tasks?

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. pp. 6904–6913.

[5] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6088–6100. Association for Computational Linguistics.

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents.

[7] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point Cloud Understanding by CLIP. pp. 8552–8562.

[8] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. CLIP-Forge: Towards Zero-Shot Text-To-Shape Generation. pp. 18603–18613.

[9] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. pp. 3835–3844.

[10] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2888–2913. Association for Computational Linguistics.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics.

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations.

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization.

[14] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In **Advances in Neural Information Processing Systems**, Vol. 32, pp. 9453–9463.

[15] Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 9824–9846. Association for Computational Linguistics.

[16] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. pp. 5238–5248.

[17] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models.

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

[19] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation.

[20] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image Colorization: A Survey and Dataset.