Diffusion Models for Improved Compositional Generalisation in VLMs

Beth Pearson
beth.pearson@bristol.ac.uk

Michael Wray
michael.wray@bristol.ac.uk

Martha Lewis
martha.lewis@bristol.ac.uk

University of Bristol

Abstract

Vision-language models (VLMs) have shown significant advancements in recent years, however, they struggle with concept binding and generalising to unseen labels. Diffusion models have gained a lot of interest due to their impressive ability to generate realistic images, and their applications in downstream tasks are only just being explored. In recent years, classifiers built from diffusion models have been proposed which have shown promise in compositional tasks by performing better than CLIP in certain settings. We propose a new benchmark: Concept Binding 2 (CoBi 2), to assess the performance of foundational VLMs in tasks requiring compositional understanding. Our dataset focuses on attribute-object binding in both zero-shot and generalised zero-shot settings, aiming to test the models' ability to generalise learned concepts to new scenarios. We evaluate the performance of Diffusion Classifier, a model built from Stable Diffusion, and compare it with state-of-the-art VLMs such as CLIP. We find that while CLIP excels in single-object and zero-shot tasks, the Diffusion Classifier outperforms CLIP in generalised zero-shot settings, demonstrating the ability to perform compositional generalisation.

1. Introduction

Vision-language models [1, 10] (VLMs) have made remarkable progress in recent years and are used in a variety of different applications such as zero-shot classification, visual question answering, and image captioning. However, they still fall short in tasks requiring compositional understanding [8, 13]. For example, given an image of a red cube and a green sphere, a VLM such as CLIP may misinterpret the image as containing a red sphere or a green cube. Additionally, VLMs should be able to generalise learned concepts to new unseen combinations of attributes and objects, for example, if a model learns the colour *cyan* through images of *cyan spheres* and the shape *cube* through images of *green cubes*, it should also be able to recognise images of *cyan cubes* or *green cylinders*.

Diffusion models [5] have gained a lot of interest recently due to their ability to generate highly detailed images. However, their application to downstream tasks is still an emerging topic. Recently, methods to create zero-shot classifiers from diffusion models have been proposed which have been shown to perform competitively with CLIP in certain compositional tasks [2, 7, 9]. Many existing zero-shot learning research tests only unseen classes during training—overlooking the more realistic generalised zero-shot learning setting where both seen and unseen classes are present during evaluation.

To address this, we propose the **Co**ncept **Bi**nding Benchmark 2 (CoBi 2) that evaluates model performance on attribute-object binding in both zero-shot and generalised zero-shot settings. We use this dataset to evaluate the performance of the Diffusion Classifier—a classifier built from Stable Diffusion—comparing it with state-of-the-art vision-language models.

Our initial results show that while CLIP initially exhibits the best performance as a frozen pre-trained model, it tends to significantly overfit to the training data during fine-tuning. However, Diffusion Classifier demonstrates the ability to perform concept binding and can generalise to novel combinations of shapes and colours exhibiting better compositional generalisation.

2. Related Work

CLIP has been shown to treat captions as a 'bag-of-concepts' [13]. Accordingly, it doesn't model word-order relationships and is unable to successfully bind concepts to objects within an image. Recently, methods have been proposed to leverage diffusion models as zero-shot classifiers. Li et al. [9], propose Diffusion Classifier, a model built from Stable Diffusion which achieves a higher accuracy than CLIP on tasks requiring compositional reasoning. Krojer et al. [7] use a similar method for using Stable Diffusion as a classifier but include a normalising value based

on the noise prediction error calculated with no text guidance. He et al. [4] use the attention scores between the image and text representations of Stable Diffusion to adapt it for image-text matching tasks. Clark and Jaini [2] also propose a zero-shot classifier created from Google's Imagen, which shows the ability to bind attributes such as shape, size and colour where CLIP fails to do so. Lewis et al. [8] compare CLIP with a set of compositional distributional semantics models on attribute binding tasks using a generalisation split of the data. However, neither CLIP nor the CDSMs are able to generalise to unseen class labels. Therefore, we investigate whether diffusion models are capable of using compositional generalisation to recognise unseen classes using Diffusion Classifier for our experiments.

3. Concept Binding Benchmark 2 (CoBi 2)

We base the design of our benchmark on the previous benchmark from Lewis et al. [8] where three datasets were created for exploring three different types of compositions: single objects, distinguishing paired objects, and the subject-relation-object setting. In this paper, we focus on single object and paired object concept binding tasks. We extend the dataset of [8] and structure the dataset to allow for assessment in a zero-shot setting, as well as a generalised zero-shot setting.

The images are generated using the generation script for the CLEVR dataset [6]—using a Blender script [3] to render 3D shapes. The original code included only three shapes *cubes*, *cylinders*, and *spheres* which we extend with an additional shape, *cones*, to increase the diversity in train, validation, and test splits. We consider the following colours: *blue*, *brown*, *cyan*, *gray*, *green*, *purple*, *red* and *yellow*. Given the set of colours C and the set of shapes S, we thus define the set $\mathcal{Y}$ of all possible labels of images as the Cartesian product between them: $\mathcal{Y} = \{(c, s)|c \in C, s \in S\}$. We design each dataset to separate the data into train, validation, and test splits such that the class labels within each are unique and do not intersect with other subsets. Given the set of labels $\mathcal{Y}$, the splits are defined as $\mathcal{Y}^{\text{train}}, \mathcal{Y}^{\text{val}}, \mathcal{Y}^{\text{test}} \subseteq \mathcal{Y}$, with the following conditions: $\mathcal{Y}^{\text{train}} \cap \mathcal{Y}^{\text{val}} = \varnothing$, $\mathcal{Y}^{\text{train}} \cap \mathcal{Y}^{\text{test}} = \varnothing$, and $\mathcal{Y}^{\text{test}} \cap \mathcal{Y}^{\text{val}} = \varnothing$. We give the breakdown for our dataset within Figure 1. The label 'red cube' is in the test set, meaning that it is not seen during training, but 'red' (e.g. in 'red sphere') and 'cube' (e.g. 'gray cube') have both been seen during training in other combinations. We test models on zero-shot (ZS) and generalised zero-shot (GZS) tasks. In both tasks, models are trained on images and labels from the training split of the data. In the ZS task, at test time, models must pick the correct label for an image from a set $\mathcal{S}$ of unseen labels, i.e. $\mathcal{S} \subseteq \mathcal{Y}^{\text{test}}$. In the GZS task, at test time, models must pick the correct label for an image from a set of both seen and unseen labels, i.e. $\mathcal{S} \subseteq \mathcal{Y}$.

This setup evaluates the ability of models fine-tuned on the train split to generalise colours learned from shapes observed during training to objects that have not been seen in that colour during training. Because of this, the train split contains at least one class containing each shape and each colour.

	red	green	purple	cyan	gray	blue	brown	yellow
sphere								
cube								
cylinder								
cone								

Figure 1. Class labels belonging to each dataset split: train is highlighted in green, validation in yellow, and test in red.

	Train	**Val**	**Test**
Single-Object	1700	400	1100
Two-Object	9300	600	3700

Table 1. Dataset statistics

3.1. Single-Object

The first dataset contains images containing only one object. This tests the model's ability to recognise attribute-object pairs (e.g. red sphere) and is used as a baseline for analysing which colour-shape combinations the models can recognise before experimenting in a two-object setting. An example of an image from the single-object dataset is shown in Figure 2. In the single-object setting, we evaluate only on the GZS task, and require models to select the correct label for the image from all possible label combinations, i.e. from the whole of $\mathcal{Y}$

The class labels are given in the form of a prompt 'a photo of a <class>'. Since the frozen models chosen are pre-trained on image captions typically in the form of a complete sentence, longer more sentence-like prompts usually yield better results. As an example, the class 'green sphere' will have a matching prompt 'a photo of a green sphere'.

3.2. Two-Object

The two-object dataset contains images of exactly two objects which differ in *both shape and colour*. For example the dataset contains images of a blue cube and a red sphere but not of a blue cube and a blue sphere. The purpose of this dataset is to test the model's ability to bind attributes to specific objects. For example, in Figure 3, will the models correctly predict the labels 'cyan cone' or 'yellow cube' or will they misinterpret the image as 'yellow cone' or 'cyan cube'? The two-object dataset follows a similar setup to the two-object experiment in Lewis et al. [8], where the model is presented with single object labels whereby one of the labels matches one of the objects in the image but the others

are incorrect.

The models are tested on ZS and GZS tasks. In both settings, we train the models on the train split of the data. In the ZS setting, at test time, we give the models one correct label for the image and two unseen distractor labels. For example, Figure 3, taken from the validation split, has a true label 'cyan cone' and distractor labels 'brown sphere' and 'gray cylinder'.

In the GZS experiment, the models are given 5 labels to choose from which are either unseen or designed to be 'difficult distractors': specifically, the shape and attribute are swapped. In this case, Figure 3 will have the label options 'cyan cone', 'brown sphere', 'gray cylinder', 'yellow cone' and 'cyan cube' where 'cyan cone' is the true label and 'yellow cone' and 'cyan cube' are the hard negatives. This is a harder task, as the labels are deliberately difficult to choose between, and moreover will have been seen in training. This allows us to test whether the model prefers before-seen classes over new ones. It is expected that the fine-tuned models will prefer classes from the train split.

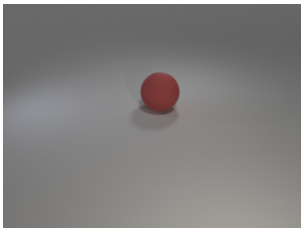

Figure 2. Single-object data, true label: 'red sphere'

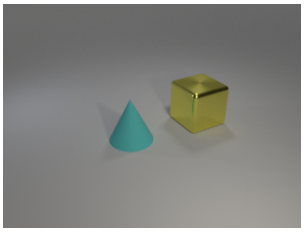

Figure 3. Two-object data, true labels: 'cyan cone' and 'yellow cube'

4. Implementation and Results

Implementation Details We test four VLMs on the single object dataset, using pretrained models only. We examine CLIP [11], BLIP [10], FLAVA [12], and the Diffusion Classifier [9]. Due to lower performance on the single object dataset, we opted against assessing FLAVA and BLIP on the two-object dataset. Rather, we focus on comparing Diffusion Classifier's compositional understanding with CLIP's. Diffusion Classifier requires estimating the noise prediction error over a number of noise samples. We vary the number of samples from 25 to 1000 and find that 200 samples yields the best performance.

For the two-object dataset, we test the performance of pretrained CLIP and Diffusion Classifier, as well as fine-tuning these two models. We fine-tune these models on the single-object dataset to test generalisability.

The DreamBooth method was used to fine-tune Stable Diffusion on the train split of the dataset. We fine-tune the U-Net and the text-encoder with a learning rate of $5 \cdot 10^{-6}$.

Model	Accuracy	Colour	Shape
CLIP	99.5	0	0
BLIP	65.5	11.5	23.0
FLAVA	67.3	8.3	24.0
Diffusion Classifier	55.5	15.3	26.5

Table 2. Performance CLIP, BLIP, FLAVA and Diffusion Classifier on the validation split of the single-object dataset. The 'Colour' column indicates the percentage of images where the model correctly identified the colour but not the shape, while the 'Shape' column represents the percentage of images where the model correctly identified the shape but not the colour.

A parameter search was carried out to determine the optimal training strategy for fine-tuning. This was done by fixing the number of training epochs whilst varying the number of training images per class and vice versa. We then create new Diffusion Classifier models from the fine-tuned Stable Diffusion instances. All models were fine-tuned with a learning rate of $5 \cdot 10^{-6}$and batch size 1. We varied the number of images per class from 10 to 50 in a 10-step interval and fine-tuned for 8000 epochs saving at each 1000-step interval. We found that the optimal performance to come from training on 30 images per class for 4000 epochs. We again use 200 noise samples. We fine-tune the ViT-B/32 variant of CLIP on 30 images from each class of the train split, keeping the same from Diffusion Classifier. We use the Adam optimiser, cross-entropy loss, batch size 8 and a learning rate of $1 \cdot 10^{-5}$. We experimented with models trained from 0 to 30 epochs and found 18 epochs to yield the best results.

Single-Object Results We compare the performance of frozen Diffusion Classifier, fine-tuned Diffusion Classifier, frozen CLIP, fine-tuned CLIP, FLAVA, and BLIP on the single-object dataset. Firstly, the frozen models are evaluated for their performance on choosing the correct attribute-object pair represented in the image. The models are given a list of 32 possible labels (combinations of the 4 shapes and 8 colours).

Classification accuracies for Diffusion Classifier, CLIP, FLAVA, and BLIP on the validation split of the single-object dataset are shown in Table 2. CLIP has by far the best performance with an accuracy of 99.5% on the validation split. BLIP and FLAVA have a similar performance while Diffusion Classifier has the lowest accuracy of all the models. All models except CLIP make more errors on predicting the correct colour of the image than shape. This could be due to there being more variance in colours seen during pre-training for the models where shapes have more consistent features.

CLIP has an almost perfect classification accuracy while

Model	Single Object	Two-Object ZS	Two-Object GZS
Frozen CLIP	99.5	93.0	35.3
CLIP-FT	100.0	99.7	17.2
Frozen DC	55.5	90.5	41.0
DC-FT	100.0	97.5	70.5

Table 3. Performance of frozen and fine-tuned CLIP models on the single-object task, the two-object zero-shot (Two-Object ZS) and two-object generalised zero-shot (Two-Object GZS) tasks.

BLIP and FLAVA fail to recognise as many classes. BLIP makes errors on 'gray cylinder', struggling to discern the shape from the gray background and FLAVA often mistakes brown as yellow. Diffusion Classifier makes similar mistakes on classes 'brown sphere' and 'gray cylinder', leading to a low overall accuracy.

Both CLIP and Diffusion Classifier reach an accuracy of 100% on the single-object dataset after fine-tuning (first col. Tab. 3) demonstrating that in single-object settings both models are able to generalise learned attributes from the training data to new instances.

Two-Object Table 3 shows the results for frozen and fine-tuned CLIP and Diffusion Classifier on all dataset tasks. CLIP, both frozen and fine-tuned achieves a higher accuracy than Diffusion Classifier in the zero-shot two-object experiment though all models achieve an accuracy above 90%. Fine-tuning the models improves the models attribute recognition, particularly in images containing gray objects, for instance frozen Diffusion Classifier only gets 16% accuracy on images of gray cylinder and yellow cube but after fine-tuning gets 98%. In the generalised task, the frozen models have a similar performance and are unable to bind the colour to the correct object with both models making errors by choosing the hard negative labels. However, after fine-tuning, Diffusion Classifier significantly outperforms CLIP achieving an accuracy of 70.5% compared to CLIP's 17.2%. We find the same results as Lewis et al. [8] that CLIP overfits to the training data and is unable to generalise to unseen combinations of colours and shapes. Diffusion Classifier demonstrates the ability to bind concepts to specific objects and is less fooled by hard negative labels but is still far from Zero-Shot performance.

5. Discussion

We introduce CoBi 2, a novel dataset aimed at benchmarking concept binding in zero-shot and generalised zero-shot environments. We hope CoBi 2 will be a useful resource for both training and evaluating foundational Vision Language Models. We find Diffusion Classifier is able to generalise to unseen combinations of labels where CLIP fails. This highlights the potential of Diffusion Models in performing concept binding, a task currently challenging for state-of-the-art vision language models. We plan to expand our dataset and experiments to include spatial relations between objects. We further plan to use analysis techniques to understand the representations that are being built by the models and how these are being used in prediction.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1

[2] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 1, 2

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2

[4] Xuehai He, Weixi Feng, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Discriminative diffusion models as few-shot vision and language learners. *arXiv preprint arXiv:2305.10722*, 2023. 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[6] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2

[7] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In *NeurIPS*, 2023. 1

[8] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of EACL 2024*, 2024. 1, 2, 4

[9] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 1, 3

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 2022. 1, 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[12] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 3

[13] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 1