Evaluating Compositional Generalisation in VLMs and Diffusion Models

Beth Pearson
University of Bristol
beth.pearson@bristol.ac.uk

Bilal Boulbarss
University of Amsterdam
bilal.boulbarss@student.uva.nl

Michael Wray
University of Bristol
michael.wray@bristol.ac.uk

Martha Lewis
University of Amsterdam
m.a.f.lewis@uva.nl

Abstract

A fundamental aspect of the semantics of natural language is that novel meanings can be formed from the composition of previously known parts. Vision-language models (VLMs) have made significant progress in recent years, however, there is evidence that they are unable to perform this kind of composition. For example, given an image of a red cube and a blue cylinder, a VLM such as CLIP is likely to incorrectly label the image as a red cylinder or a blue cube, indicating it represents the image as a 'bag-of-words' and fails to capture compositional semantics. Diffusion models have gained significant attention for their impressive generative abilities, and zero-shot classifiers based on diffusion models have been shown to perform competitively with CLIP in certain compositional tasks. In this work we explore whether the generative Diffusion Classifier has improved compositional generalisation abilities compared to discriminative models. We assess three models—Diffusion Classifier, CLIP, and ViLT—on their ability to bind objects with attributes and relations in both zero-shot learning (ZSL) and generalised zero-shot learning (GZSL) settings. Our results show that the Diffusion Classifier and ViLT perform well at concept binding tasks, but that all models struggle significantly with the relational GZSL task, underscoring the broader challenges VLMs face with relational reasoning. Analysis of CLIP embeddings suggests that the difficulty may stem from overly similar representations of relational concepts such as left and right. Code and dataset are available at: github.com/otmive/diffusion_classifier_clip

1 Introduction

Compositionality is a fundamental part of how humans learn (Chomsky, 1957; Janssen and Partee, 1997). It allows us to take familiar concepts and combine them in new ways to interpret novel situations, learn from limited examples, and build increasingly complex ideas. Within formal semantics

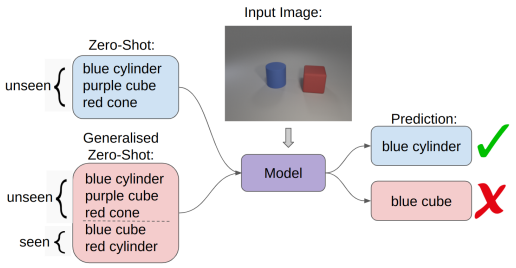


Figure 1: We evaluate the compositional generalisation of Vision-Language Models (VLMs) by assessing their ability to bind colours to objects and relations to objects in both zero-shot and generalised zero-shot settings across single-object, two-object, and relational scenarios

as in e.g. (Montague, 1973), compositionality is assumed in the formalism. However, vision-language models (VLMs) fall short in tasks requiring compositional understanding (Diwan et al., 2022; Yuksekgonul et al., 2022; Lewis et al., 2024). Even with advances in attention mechanisms (Vani et al., 2024) and positional encoding (Su et al., 2024), VLMs are unable to match the compositional reasoning skills of humans (Sinha et al., 2024; Hua et al., 2024b). VLMs such as CLIP (Radford et al., 2021) have been shown to treat captions as a bag-of-words (Thrush et al., 2022) and are not able to bind concepts to objects in the same way humans can. For example, given an image of a *red cube* and a *blue cylinder*, a VLM may misinterpret the image as containing a *blue cube* or a *red cylinder* (see Figure 1). Additionally, a VLM should be able to generalise learned concepts to new unseen combinations of attributes and objects: if a model learns the colour *cyan* through images of *cyan cone* and the shape *cube* through images of *green cubes*, it should also be able to recognise images of *cyan cubes* or *green cones*. In formal semantics, given *correct* representations of the words *green*, *cyan*, *cube*, and *cone*, this property would naturally occur.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 122–133

Diffusion Models have gained significant interest in recent years for their state-of-the-art performance on image generation (Ramesh et al., 2022; Dhariwal and Nichol, 2021) and editing tasks (Brooks et al., 2023). Their performance as zero-shot classifiers in vision tasks is a recent topic of exploration (Clark and Jaini, 2023; Krojer et al., 2023). On compositional benchmarks such as Winoground (Thrush et al., 2022) or the Concept Binding Benchmark from Lewis et al. (2024), their performance has been shown to be comparable to that of CLIP (Li et al., 2023; Clark and Jaini, 2023). However, Winoground has been argued to require commonsense and world knowledge rather than purely testing for compositional abilities (Diwan et al., 2022), and performance on the Concept Binding Benchmark can be at chance.

In this paper, we contribute to the understanding of the compositional abilities of diffusion model-based classifiers by comparing with transformer-based classifiers on compositional tasks. Specifically, we explore how these two types of models are able to compose attributes and relations—tasks VLMs particularly struggle with. We aim to assess whether Diffusion Classifier can offer new insights or improvements in handling these challenging aspects of compositional semantics.

We consider two settings for our experiments—zero-shot learning (ZSL) and generalised zero-shot learning (GZSL). In ZSL, the aim is to recognise only unseen classes whereas GZSL aims to train models that are able to discriminate between both seen and unseen classes during test time (Pourpanah et al., 2022; Xian et al., 2017). The GZSL setting is particularly important for real world scenarios as there may only be labelled data for a small number of classes and capturing every possible class in the training set is often impossible. Therefore, it is important for models to be able to generalise to unseen classes in the presence of labels that have previously been seen.

To probe these abilities, we extend the Concept Binding Benchmark from Lewis et al. (2024), which evaluates model performance on attribute-object binding and relational composition. We evaluate the performance of Diffusion Classifier—a classifier built from Stable Diffusion (Rombach et al., 2022)—comparing it with CLIP and ViLT (Kim et al., 2021). Despite the dataset being lightweight, it still proves challenging for the models, particularly in the important GZSL setting.

The main contributions of this work are threefold: **(1)** We compare Diffusion Classifier, CLIP, and ViLT on compositional tasks. Diffusion Classifier generalises best in single-object settings, however, ViLT has by far the best two-object performance. All models struggle to reliably compose relations with objects. **(2)** To provide a more robust evaluation of compositional generalisation, we present our extension of the Concept Binding Benchmark from Lewis et al. (2024). This extended benchmark consists of three datasets to test VLMs in both zero-shot learning (ZSL) and generalised zero-shot learning (GZSL) scenarios. **(3)** We analyse the effects of fine-tuning on compositional semantic understanding, showing that models fail to form correct representations for spatial relations.

2 Related Work

Benchmarking Compositionality in VLMs There is a growing interest in the ability of VLMs to reason compositionally, with several benchmarks being proposed in recent years (Yuksekgonul et al., 2022; Ma et al., 2023; Hsieh et al., 2024; Dumpala et al., 2024; Ray et al., 2024; Zhao et al., 2022; Huang et al., 2024; Thrush et al., 2022; Hua et al., 2024a). Compositional generalisation is an important ability for VLMs to have because it encourages the interpretability and data efficiency of models (Bommasani et al., 2021). However, it has been argued (Lewis et al., 2024; Hsieh et al., 2024) that various compositionality benchmarks are 'hackable', showing that in some cases it is possible to solve the benchmark simply by comparing prompts (Wu et al., 2023) and ignoring the image. SugarCrepe (Hsieh et al., 2024) is designed to deal with this problem, but is still prone to the issue that the correct caption is statistically more likely in the training corpus. Unlike benchmarks that use complex real-world images, we use simple, synthetic images to ensure no spurious correlations and to directly test compositional understanding. We argue that VLMs should be able to handle these simpler reasoning tasks before advancing to more complex, real-world images.

Improving Compositionality in VLMs Methods have been proposed to improve the compositional abilities of VLMs (Cascante-Bonilla et al., 2023; Doveh et al., 2023). Several works use hard negative sampling to fine-tune CLIP on batches of similar images e.g. "a black cat sitting on a desk" and "a black desk sitting on a cat" which force the model to learn more detailed representa-

tions of the data (Yuksekgonul et al., 2022; Shou and Lin, 2024; Sahin et al., 2024). Other methods include different representations for objects within images such as trees or graphs (Singh et al., 2023; Yellinek et al., 2025) and adaptations to the contrastive loss function of CLIP to include more compositional supervision (Pandey et al., 2023; Zhang et al., 2024). Despite advancements, VLMs still struggle with compositional reasoning (Hsieh et al., 2024; Dumpala et al., 2024). Our benchmark aims to investigate why VLMs struggle with compositional tasks by testing in GZSL settings using in-distribution and out-of-distribution images to identify potential biases.

Diffusion Model Classifiers Recently, methods have been proposed to leverage diffusion models as zero-shot classifiers (Chen et al., 2023; Li et al., 2023; Krojer et al., 2023; Clark and Jaini, 2023). Li et al. (2023) propose Diffusion Classifier, a model built from Stable Diffusion, which achieves a higher accuracy than CLIP on tasks requiring compositional reasoning such as concept binding. Krojer et al. (2023) use a similar method for using Stable Diffusion (Rombach et al., 2022) as a classifier but include a normalising value based on the noise prediction error calculated with no text guidance. He et al. (2023) use the attention scores between the image and text representations of Stable Diffusion to adapt it for image-text matching tasks. Clark and Jaini (2023) also propose a zero-shot classifier created from Google's Imagen, which shows some ability to bind attributes such as shape, size and colour where CLIP fails to do so. For our experiments we use the Diffusion Classifier from Li et al. (2023) as Stable Diffusion is open source with easily accessible fine-tuning methods.

3 Experiments

We base the design of our benchmark on the experiments from Lewis et al. (2024) where three datasets were created for exploring composition of attributes and relations with objects. While this setup reveals that models often struggle even with simple object compositions, our aim is to extend this evaluation to include both Zero-Shot Learning (ZSL) and Generalised Zero-Shot Learning (GZSL) settings. To enable this, we adapt and expand the original benchmark to support systematic and rigorous testing in both settings.

The images are generated using the generation script for the CLEVR dataset (Johnson et al., 2017)—using a Blender script (Community, 2018) 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 across the dataset splits. For the single and two-object datasets, we consider the following colours: *blue*, *brown*, *cyan*, *gray*, *green*, *purple*, *red*, and *yellow*. We define the label sets for the single and two-object datasets as follows:

Let C be the set of colours and S the set of shapes. For object classification, each object is identified by its colour–shape pair, and the label set is defined as:

$$\mathcal{Y} = \{(c, s) \mid c \in C,\ s \in S\}.$$

Each element of $\mathcal{Y}$ represents a unique object (e.g., *red square*, *blue circle*). In the two-object dataset, labels consist of two such tuples, e.g., $((c_1, s_1), (c_2, s_2))$. For the relational dataset, we define a set of spatial relations $R = \{\textit{left}, \textit{right}\}$. We exclude the relations *front* and *behind* which were included in Lewis et al. (Lewis et al., 2024) as we found these to be too ambiguous—distinguishing which shape is further forward is often difficult even for humans. The relational label set is then defined as:

$$\mathcal{Y}_{\text{rel}} = \{(s_i, r, s_j) \mid s_i, s_j \in S,\ s_i \neq s_j,\ r \in R\},$$

where each triple describes a relation between two distinct shapes—for example, $(\textit{circle}, \textit{left}, \textit{square})$. All datasets are partitioned into five subsets: training ($\mathcal{Y}^{\text{train}}$), in-distribution validation/test ($\mathcal{Y}^{\text{IDval}}, \mathcal{Y}^{\text{IDtest}}$), and out-of-distribution validation/test ($\mathcal{Y}^{\text{OODval}}, \mathcal{Y}^{\text{OODtest}}$). In-distribution subsets use the same label space as the training set, i.e. $\mathcal{Y}^{\text{train}} = \mathcal{Y}^{\text{IDval}} = \mathcal{Y}^{\text{IDtest}}$, while OOD splits are defined such that:

$$\mathcal{Y}^{\text{train}} \cap \mathcal{Y}^{\text{OODval}} = \emptyset, \quad \mathcal{Y}^{\text{train}} \cap \mathcal{Y}^{\text{OODtest}} = \emptyset,$$
$$\mathcal{Y}^{\text{OODval}} \cap \mathcal{Y}^{\text{OODtest}} = \emptyset.$$

This setup enables evaluation both within the training distribution and on novel combinations, to assess generalisation. We give the structure of our single and two-object datasets within Figure 2. 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. in *gray cube*) have both been seen during training in other combinations. The structure of the relational dataset is given in Figure 3.

Figure 2: Single and Two-Object dataset design. Class labels belonging to each dataset split: train and in-distribution are highlighted in green, OOD validation in yellow, and OOD test in red.

	Train	ID Val	ID Test	OOD Val	OOD Test
Single-Object	1360	340	340	400	1100
Two-Object	7440	1860	1860	600	3700
Relational	440	110	110	250	400

Table 1: Our extended benchmark statistics for the three datasets showcasing the number of images within each of the splits.

For both ZSL and GZSL tasks, models are fine-tuned on images and labels from the training split of the data. In the ZSL setting, at test time, models must pick the correct label for an image from a set $\mathcal{S}$ of unseen labels, i.e. $\mathcal{S} \subseteq \mathcal{Y}^{\text{OODtest}}$. In the GZSL task, at test time, models must pick the correct label for an image from a set of both seen and unseen labels, i.e. $\mathcal{S} \subseteq \mathcal{Y}$ or $\mathcal{S} \subseteq \mathcal{Y}_{\text{rel}}$. This setup evaluates the ability of models to generalise colours or relations learned during fine-tuning to new unseen shape combinations. Because of this, the single and two-object train split contains at least one class containing each shape and each colour. Similarly, the relational train split contains at least one of each shape.

We only use positive examples when fine-tuning CLIP rather than both positive and negative examples to keep consistent with the DreamBooth fine-tuning method for Stable Diffusion which only accepts positive training examples. In addition, to further align with DreamBooth, we fine-tune CLIP with a small number of samples from each class (20-40 per class).

3.1 Single-Object

The single-object task tests the ability of models to recognise attribute-object pairs and is used as a baseline for analysing which combinations the models can recognise before experimenting in a two-object setting. Examples from the single-object dataset are shown in Figure 4 a) and b). In the single-object setting, we evaluate only on the GZSL task, and require models to select the correct label for the image from all possible label combinations, i.e. from the whole of $\mathcal{Y}$. Following convention, the class labels are given in the form of a prompt "a photo of a <class>".

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*. We follow Lewis et al. (2024) and present the model with labels for individual objects whereby the true label correctly describes one of the objects in the image and the others are incorrect. In comparison to giving the model a full description of the image (e.g. *green cone and purple cylinder*), this is a challenging setup which minimises the use of shortcuts by the model, for example if the model can recognise green cones correctly but not purple cylinders. As an example, the images in Figure 4 c) may have the true label *green cylinder* and hard negatives *green cone* and *purple cylinder*.

In the ZSL setting, models are given one correct label and two distractors from the same (unseen) split. For example, an image of a yellow cube may be paired with gray cylinder and brown sphere as distractors (see Figure 4 column d)).

In the GZSL setting, models choose from five labels: the true label, two standard distractors, and two hard negatives created by swapping attributes and shapes (e.g., yellow cone, cyan cube for Figure 4 d)). This makes the task more challenging and tests whether models prefer familiar (seen) classes over novel ones.

3.3 Relational

The relational dataset tests compositions of the relations *left* and *right* between two-objects in an image. The two-objects are always two distinct shapes, that is, we don't consider cases such as *sphere left sphere*. As with the two-object dataset, each image has two possible true labels. For instance, the images in Figure 4 column e) would have the true labels *cube left sphere* and *sphere right cube*. Again, we consider a ZSL and GZSL setting. In the GZSL setting, models choose from five options: the true label, two randomly selected labels, and two hard negatives. One hard negative alters the spatial relation (e.g., cube left sphere $\rightarrow$ cube right sphere), while the other swaps object order (e.g., cube left sphere $\rightarrow$ sphere left cube). The hard negatives require the model to recognise the specific relation in the image and not just recognise which two shapes are present—a task at which a

bag-of-words model would fail.

left	sphere	cube	cylinder	cone
sphere		sphere left cube	sphere left cylinder	sphere left cone
cube	cube left sphere		cube left cylinder	cube left cone
cylinder	cylinder left sphere	cylinder left cube		cylinder left cone
cone	cone left sphere	cone left cube	cone left cylinder	

right	sphere	cube	cylinder	cone
sphere		cube right sphere	cylinder right sphere	cone right sphere
cube	sphere right cube		cylinder right cube	cone right cube
cylinder	sphere right cylinder	cube right cylinder		cone right cylinder
cone	sphere right cone	cube right cone	cylinder right cone	

Figure 3: Relational dataset design. Class labels belonging to each dataset split: train and in-distribution are highlighted in green, OOD validation in yellow, and OOD test in red.

4 Results

We conduct experiments comparing frozen and fine-tuned CLIP, ViLT, and Diffusion Classifier (DC) on three datasets: single-object, two-object, and relational. Experiments are carried out in a Linux environment using an RTX 2080 GPU for both training and inference.

4.1 Single-Object

We test models' ability to compose single attribute-noun pairs. For each of the models we fine-tune with three different seeds and report the mean and standard deviation of each. Fine-tuning details and hyperparameters for each dataset are provided in Appendix A.

Model	ID Validation	ID Test	OOD Validation	OOD Test
Frozen CLIP	$85.29^{0.00}$	$80.59^{0.00}$	$67.75^{0.00}$	$87.36^{0.00}$
CLIP-FT	$95.29^{3.01}$	$95.59^{2.92}$	$93.57^{3.81}$	$91.21^{6.54}$
Frozen ViLT	$51.47^{0.00}$	$50.0^{0.00}$	$34.5^{0.00}$	$44.91^{0.00}$
ViLT-FT	$95.88^{0.00}$	$94.71^{0.00}$	$63.5^{0.00}$	$77.18^{0.00}$
Frozen DC	$40.80^{0.89}$	$40.98^{0.37}$	$58.0^{0.50}$	$60.0^{1.08}$
DC-FT	$97.74^{1.6}$	$97.16^{0.78}$	$99.50^{0.12}$	$99.47^{0.87}$

Table 2: Accuracy of models on the single-object task.

Results We see in Table 2 that CLIP has the best accuracy of the frozen models on this task. However, after fine-tuning, DC has the best overall accuracy. Both CLIP and DC show a strong performance on ID and OOD splits indicating that in the simple single-object setting they are able to generalise to unseen colour-shape combinations. In contrast, fine-tuned ViLT showcases strong performance only on the ID splits and shows a drop in accuracy to 63.5% and 77.18% on the OOD splits. ViLT frequently makes errors such as predicting *blue cone* for *cyan cone* or *gray cube* for *gray cylinder*—failing to generalise from familiar components seen during training (such as the colour cyan with a sphere, or the shape cylinder with other colours like red, green, or purple). Fine-tuned CLIP and DC are able to generalise in the single-object setting but ViLT's lower OOD performance shows even in simple settings composing unseen combinations can be difficult for VLMs.

4.2 Two-Object Zero-Shot

The two-object experiment tests whether models can correctly bind attributes to their corresponding objects, rather than simply recognising which shapes and colours are present. We report the average accuracy with the standard deviation for all models as shown in Table 3.

Model	ID Validation	ID Test	OOD Validation	OOD Test
Frozen CLIP	$83.71^{0.00}$	$85.27^{0.00}$	$93.0^{0.00}$	$69.51^{0.00}$
CLIP-FT	$90.13^{0.55}$	$90.39^{0.01}$	$99.39^{0.75}$	$80.15^{1.11}$
Frozen ViLT	$72.78^{0.56}$	$73.80^{0.73}$	$70.0^{0.00}$	$66.82^{0.32}$
ViLT-FT	$99.78^{0.00}$	$99.89^{0.08}$	$99.5^{0.00}$	$99.26^{0.18}$
Frozen DC	$61.18^{0.00}$	$64.53^{0.00}$	$91.83^{0.00}$	$58.3^{0.00}$
DC-FT	$82.59^{3.34}$	$83.21^{3.59}$	$93.89^{2.49}$	$72.80^{2.06}$

Table 3: Accuracy of models on the ZSL two-object task.

ViLT-FT has the highest accuracy achieving over 99% on all dataset splits. This is particularly surprising given its lower performance in the single-object task. ViLT may benefit from the reduced label space in the two-object ZSL experiment compared to having the full range of prompts in the single-object setting. CLIP-FT and DC-FT both show a decrease in performance on OOD test but not on OOD val. We believe the high OOD val accuracies are due to the reduced size of the OOD val split meaning there are only 4 very distinct prompts to choose from. The drop in performance of all models on the OOD test split further highlights that VLMs lack robust compositional understanding, even for the simpler zero-shot case. Current pre-training strategies rarely require models to explicitly learn compositional knowledge, suggesting that adjustments to pre-training may be necessary.

4.3 Two-Object Generalised Zero-Shot

In the GZSL two-object task, models must compose attributes with objects while also handling previously seen labels, providing a more rigorous test of generalisability. We report the accuracies

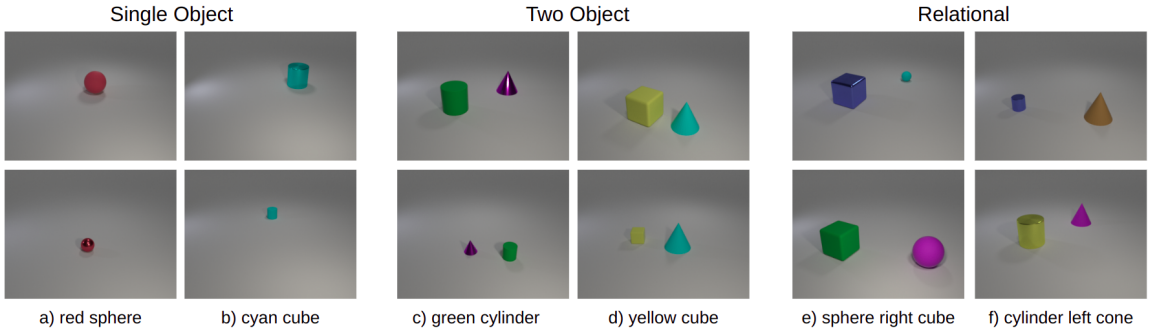


Figure 4: Samples from our extended benchmark with two example classes displayed from each dataset—single, two-object, and relational.

and standard deviations for the two-object GZSL experiment in Table 4.

Model	ID Validation	ID Test	OOD Validation	OOD Test
Frozen CLIP	$23.33^{0.00}$	$21.56^{0.00}$	$35.33^{0.00}$	$34.27^{0.00}$
CLIP-FT	$78.82^{3.05}$	$76.40^{0.86}$	$55.50^{5.92}$	$23.38^{5.28}$
Frozen ViLT	$31.56^{0.12}$	$32.71^{0.29}$	$47.83^{0.00}$	$29.1^{0.21}$
ViLT-FT	$99.71^{0.07}$	$99.86^{0.03}$	$91.67^{0.00}$	$83.46^{0.06}$
Frozen DC	$33.58^{0.00}$	$34.64^{0.00}$	$38.46^{0.00}$	$39.32^{0.00}$
DC-FT	$53.06^{3.20}$	$51.86^{3.41}$	$57.06^{5.03}$	$72.97^{2.05}$

Table 4: Accuracy of models on GZSL two-object.

Again ViLT-FT has the strongest performance for all dataset splits significantly outperforming other models. This suggests it is less biased towards seen labels as evidenced by the relatively stable performance across ZSL and GZSL. However, it does still exhibit a small drop in performance on the OOD splits indicating some limitations in generalising. CLIP-FT experiences a substantial drop in performance on the OOD splits especially OOD test, showing it has overfit to the training data. DC-FT interestingly shows the reverse pattern to the other models and has the highest accuracy on OOD. We hypothesise that this is due to the composition of the test split—for example, challenging colours like yellow and brown, which DC frequently confuses, constitute a smaller proportion of the OOD labels. While the high OOD test accuracy is particularly notable in the challenging GZSL setting, DC's lower accuracy on the ID splits (53.06% and 51.86%) suggests it lacks consistent attribute-object binding ability. Even ViLT-FT, the best-performing model overall, has a reduced performance on the OOD splits, highlighting limitations in the way models represent and combine attributes and objects.

In table 5 we show the percentage of total predictions made by the models which fall into each error category on the GZSL two-object task for the ID and OOD test splits. The *Colour* column is the percentage of predictions where the model correctly identifies the shape but chooses the colour of the second object in the image, the *Shape* column is the percentage of predictions correctly guessing the colour but choosing the second object's shape. The *Other* column is the predictions from the other two non-hard negatives.

Model	ID test			OOD Test		
	Colour	Shape	Other	Colour	Shape	Other
Frozen CLIP	$35.97^{0.0}$	$37.42^{0.0}$	$5.06^{0.0}$	$22.05^{0.0}$	$30.03^{0.0}$	$13.65^{0.0}$
CLIP-FT	$7.15^{1.58}$	$8.30^{0.53}$	$8.16^{1.28}$	$34.17^{3.64}$	$34.78^{2.95}$	$7.72^{1.29}$
Frozen ViLT	$35.57^{0.4}$	$26.54^{0.47}$	$5.18^{0.51}$	$35.98^{0.05}$	$28.56^{0.19}$	$6.35^{0.02}$
ViLT-FT	$0.03^{0.04}$	$0.0^{0.0}$	$0.0^{0.0}$	$8.76^{0.02}$	$7.54^{0.02}$	$0.17^{0.04}$
Frozen DC	$21.46^{0.0}$	$26.25^{0.0}$	$17.64^{0.0}$	$14.70^{0.0}$	$18.18^{0.0}$	$27.81^{0.0}$
DC-FT	$18.53^{0.55}$	$16.79^{1.24}$	$12.81^{1.63}$	$16.51^{0.71}$	$10.53^{1.75}$	$0.0^{0.0}$

Table 5: Breakdown of errors in two-object GZSL.

Both frozen and fine-tuned CLIP have a roughly even distribution of errors on colour and shape mistakes showing both types of composition are equally challenging. Frozen ViLT makes slightly more errors on colour, but after fine-tuning, errors across all categories drop to near zero with a slight tendency for colour errors remaining in OOD. Frozen DC makes slightly more mistakes on shape but after fine-tuning finds colour slightly more difficult especially in OOD. All incorrect predictions made by DC-FT in the OOD split correspond to hard negative labels, highlighting that binding colours to the correct objects is particularly challenging.

4.4 Relational Zero-Shot

The relational experiment tests how well models can compose spatial relations with objects, specifically we test the composition of the relations 'left' and 'right' with the object's shape eg. 'cube'. We show the relational ZSL results in Table 6.

Model	ID Validation	ID Test	OOD Validation	OOD Test
Frozen CLIP	$56.36^{0.00}$	$56.60^{0.00}$	$38.40^{0.00}$	$68.00^{0.00}$
CLIP-FT	$99.39^{0.86}$	$99.31^{0.57}$	$68.00^{13.91}$	$94.08^{3.86}$
Frozen ViLT	$74.55^{1.48}$	$68.52^{0.87}$	$42.40^{0.00}$	$64.67^{0.31}$
ViLT-FT	$78.18^{2.57}$	$76.04^{1.98}$	$70.53^{0.19}$	$65.0^{0.35}$
Frozen DC	$68.18^{0.00}$	$69.44^{0.00}$	$30.70^{0.00}$	$65.25^{0.00}$
DC-FT	$89.09^{4.64}$	$92.94^{1.18}$	$51.86^{2.31}$	$87.18^{9.18}$

Table 6: Accuracy of models on the ZSL relational task.

All models except ViLT-FT have a lower accuracy on OOD validation than OOD test. This could be due to the smaller size of the validation split, which limits prompt diversity making the distractor labels more likely to share shapes with the shapes in the true label. Both DC and CLIP only show slight drops in performance between OOD test and the ID splits demonstrating the capacity to recognise unseen object-relation combinations in ZSL settings. ViLT, while having overall lower accuracies, shows less variation across dataset splits, showing some capacity to generalise. All models show a substantial drop in performance in the relational ZSL compared with the two-object ZSL showing that systematically combining objects with relations is harder for these models than combining colour-object pairs. The difficulty the models have with relational information suggests they are focusing on recognising objects in the image rather than compositions between objects. While VLMs can often rely on these shortcuts and still achieve a strong performance, tasks that require relational reasoning reveal that they lack a full understanding of visual scenes.

4.5 Relational Generalised Zero-Shot

In the relational GZSL experiment, models must bind spatial relations to objects and predict previously unseen combinations of relations and objects in the presence of previously seen labels. We show the performance of the models on the GZSL relational task in Table 7 reporting the mean and standard deviation for the fine-tuned models.

Model	ID Validation	ID Test	OOD Validation	OOD Test
Frozen CLIP	$27.27^{0.00}$	$27.43^{0.00}$	$18.00^{0.00}$	$25.0^{0.00}$
CLIP-FT	$62.12^{0.43}$	$72.22^{3.97}$	$42.80^{18.39}$	$34.75^{16.33}$
Frozen ViLT	$13.94^{0.43}$	$16.55^{0.65}$	$22.53^{0.19}$	$26.5^{0.35}$
ViLT-FT	$16.55^{0.65}$	$22.53^{0.19}$	$26.5^{0.35}$	$25.50^{1.08}$
Frozen DC	$24.55^{0.00}$	$21.53^{0.00}$	$10.00^{0.00}$	$24.50^{0.00}$
DC-FT	$32.73^{2.57}$	$34.72^{2.60}$	$41.20^{4.57}$	$38.25^{4.02}$

Table 7: Accuracy of models on relational GZSL.

In the GZSL relational setting, CLIP-FT performs reasonably well on the ID splits with 62.12% and 72.22%, however, there is a significant drop in performance for the OOD splits with 42.80% and 34.75% on validation and test respectively. CLIP therefore seems to overfit to the training data and is not able to generalise to unseen labels. ViLT struggles with this task, with even the fine-tuned model hardly performing better than chance at 20%. Interestingly, DC has a lower accuracy on the ID splits than the OOD splits. Given DC's reasonable accuracies of 89.09% and 92.94% in the ID ZSL experiment, it appears DC is particularly confused by the presence of hard negative labels showing it is lacking fine-grained understanding. All models have a drop in performance from the ZSL task showing they struggle to compose relational con-

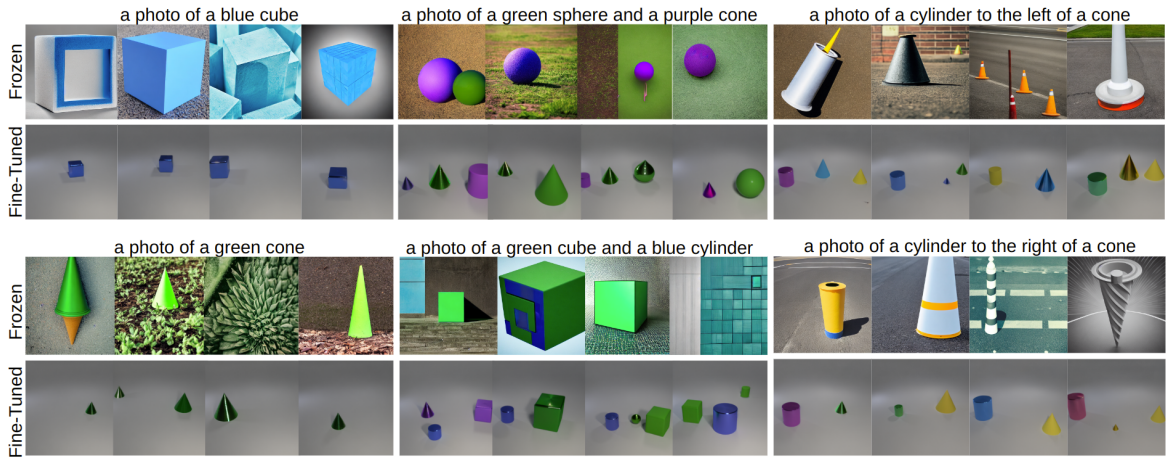


Figure 5: Images generated by Frozen and Fine-Tuned Diffusion-Classifier using prompts from the single, two-object and relational, shown from left to right. The top two rows are generated by labels from the train set and the bottom two from the test set.

cepts and especially cannot distinguish between hard negatives in the GZSL setting such as *sphere left cube* and *sphere right cube*. This suggests that the models are relying on object recognition rather than understanding relational positions.

We show the percentage of total predictions made by the models which fall into each error category for the ID and OOD test splits in Table 8. The column Left/Right shows the percentage of predictions which choose the hard negative where only the relation is incorrect e.g. *cube left sphere* instead of *cube right sphere*. The Shape column displays the percentage of predictions where the shapes are correct but in the incorrect order e.g. *sphere left cube* instead of *cube left sphere*. The Other column is the predictions from the other two non-hard negative labels.

Model	ID test			OOD Test		
	Left/Right	Shape	Other	Left/Right	Shape	Other
Frozen CLIP	$26.04^{0.0}$	$28.13^{0.0}$	$18.41^{0.0}$	$20.0^{0.0}$	$25.5^{0.0}$	$29.5^{0.0}$
CLIP-FT	$10.07^{1.99}$	$17.59^{3.46}$	$0.12^{0.16}$	$30.50^{10.5}$	$29.0^{5.75}$	$5.75^{4.13}$
Frozen ViLT	$31.25^{0.75}$	$36.34^{0.65}$	$15.86^{1.28}$	$8.0^{0.2}$	$40.67^{1.0}$	$24.83^{1.05}$
ViLT-FT	$25.87^{0.19}$	$24.40^{0.75}$	$27.20^{0.33}$	$9.08^{1.0}$	$36.42^{1.04}$	$28.83^{0.72}$
Frozen DC	$25.0^{0.0}$	$30.56^{0.0}$	$22.92^{0.0}$	$22.25^{0.0}$	$22.5^{0.0}$	$30.75^{0.0}$
DC-FT	$24.08^{1.99}$	$36.92^{1.34}$	$4.28^{2.78}$	$22.33^{1.48}$	$28.92^{5.60}$	$10.5^{4.30}$

Table 8: Breakdown of errors in relational GZSL.

Frozen CLIP and DC have fairly evenly distributed errors across the 3 categories while ViLT on the OOD test split has a larger proportion of shape errors. After fine-tuning, DC and CLIP have a reduced proportion of errors in the other category however ViLT still makes a considerable number of predictions where the two shapes in the image are not correctly identified. For fine-tuned CLIP and DC, the hard negative captions present the most difficulty with both types of hard negative being frequently predicted instead of the true label.

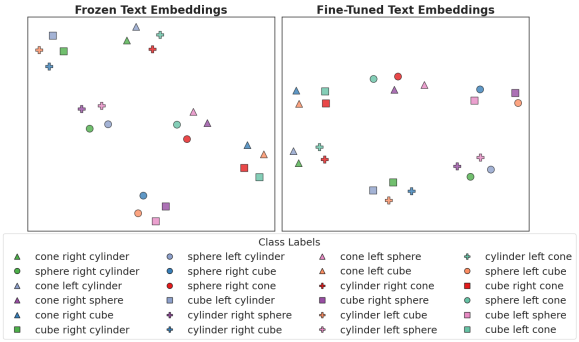
5 Model Understanding

Stable Diffusion Images We compare images generated by frozen and fine-tuned Stable Diffusion to evaluate what features Diffusion Classifier is able to learn from fine-tuning on each dataset. We use a guidance scale of 7 and 50 inference steps. Examples using prompts from each dataset are shown in Figure 5. Frozen Stable Diffusion is generally very poor at generating images in alignment with the specified prompt, except in the single-object case. Interestingly, the two-object and relational fine-tuned Stable Diffusion generate three objects fairly frequently showing some pre-training bias and knowledge is still preserved. The relational fine-tuned model fails to understand the difference between the left and right relations with the prompts "a cylinder to the left of a cone" and "a cylinder to the right of a cone" both resulting in images of a cylinder on the left—the class seen during training.

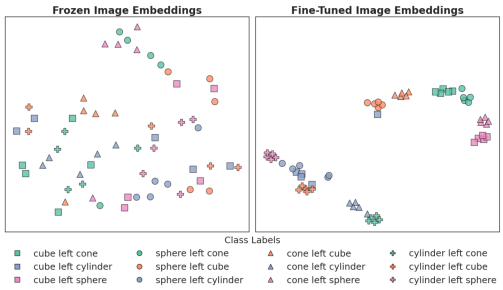
CLIP embeddings We show t-SNE visualisations of image and text embeddings from relational dataset examples for frozen and fine-tuned CLIP. For images, we show the embeddings of 5 samples from each class and only consider classes containing *left* since corresponding classes containing *right* use the same images.

Figure 6a shows the text embeddings which are clearly clustered into quadruples corresponding to prompts where the object shapes are the same, with no clear separation between prompts corresponding to different arrangements of objects. For example, the closest neighbours of *cube left sphere* are *sphere left cube*, *cube right sphere* and *sphere right cube*. Fine-tuning (right-hand plot) fails in most cases to overcome this clustering of similar prompts. An exception is the cluster of prompts containing *sphere left cube* and *cube right sphere*, which have been moved closer together, and are visibly distinct from *sphere right cube* and *cube left sphere*. Other groups of prompts tend to cluster according to ordering of nouns (e.g. *cube left cone* and *cube right cone*), or by bag-of-words similarity (e.g. *cube left cylinder* and *cylinder left cube*). This inability to distinguish prompts corresponding to different arrangements of objects likely contributes towards CLIP's inability to correctly caption images with the same shapes but different relations.

The t-SNE visualisation of image embeddings presented in Figure 6b shows that images belonging to the same class are mostly well-clustered. However, there are a few instances of classes in the wrong cluster e.g. a *cube left cylinder* sample appears within the *sphere left cube* cluster. Notably, we observe that embeddings of images with reversed relational structures tend to occupy similar regions in the space—for instance, *cylinder left cone* and *cone left cylinder* appear close together at the bottom of the plot, while *cylinder left sphere* and *sphere left cylinder* are both near the left side of the plot. This spatial overlap may contribute to CLIP's difficulty with relational reasoning. OOD classes such as *cube left cylinder* and *sphere left cylinder*, which are not directly fine-tuned, appear slightly less well clustered.



(a) t-SNE visualisation of frozen and fine-tuned CLIP text embeddings for relational prompts. Best viewed electronically or in colour.



(b) t-SNE visualisation of frozen and fine-tuned CLIP image embeddings for relational prompts. Best viewed electronically or in colour.

Figure 6: t-SNE visualisations of CLIP text and image embeddings for relational prompts after fine-tuning.

6 Discussion

We extend the Concept Binding Benchmark from Lewis et al. (2024) to assess concept binding in zero-shot (ZSL) and generalised zero-shot (GZSL) settings. Using this extended framework, we compare the performance of the discriminative models CLIP and ViLT against a generative model, Diffusion Classifier, on single-object, two-object, and relational compositional tasks. Diffusion Classifier shows the highest generalisation accuracy on the single-object task. ViLT achieves state-of-the-art performance on both ZSL and GZSL two-object tasks, demonstrating strong compositional ability in binding attributes to objects even in GZSL settings. Diffusion Classifier shows some capacity to generalise in the two-object GZSL setting, however, it falls short of ViLT's performance.

On the relational composition task, all models perform poorly, showing considerable drops in performance on the GZSL from the ZSL task showing that hard distractors such as *cube left sphere* versus *cube right sphere* are a particular problem. Despite initial hopes that Diffusion Classifier's generative approach might better handle compositionality, relational reasoning remains a major challenge for all models tested.

On all our experiments, our fine-tuned CLIP model consistently outperforms the model from Lewis et al. on the OOD splits (Lewis et al., 2024). We attribute this to our fine-tuning strategy of only using positive examples unlike Lewis et al. who use both positive and negative examples. We hypothesise that the inclusion of negative examples exacerbates overfitting. This is due to prompts appearing as negative training examples which then appear as positive examples in the OOD splits, causing CLIP to suppress their prediction. Therefore our positive-only approach appears to lead to better generalisation and reduced overfitting.

The low performance on the GZSL relational task suggests current VLMs may rely too heavily on shortcuts such as object recognition rather than developing structured, compositional representations. Our analysis of image and text embeddings in CLIP further supports this: relational concepts (e.g., *left* vs. *right*) are not sufficiently disentangled, especially in the text embedding space, limiting the models' capacity to reason about spatial relationships. Potential avenues to address this are training on datasets with more explicit compositional objectives and developing better prompting or fine-tuning strategies that encourage attribute and relation disentanglement. Further work in this area also includes analysis of where exactly the models fail: do they fail in forming correct representations of individual words, or do they fail in how these words are combined? This could be tackled using a formal semantic approach that has been integrated with a vector-based semantics, such as in Coecke et al. (2010) or Baroni and Zamparelli (2010).

However, while these routes to improved compositional understanding are important, we argue that our results highlight an important limitation of the tested models as they stand: at present compositional understanding is clearly limited. Since there may be a number of aspects of composition that we require models to perform, these should be considered at the pre-training stage rather than expecting users to fine-tune for these fundamental semantic abilities.

Limitations

While our benchmark uses synthetic, simplistic images, we chose this design specifically to reduce the risk of spurious correlations (Wu et al., 2023) and enable precise compositional structures to be tested for. We view this benchmark as a diagnostic test for probing specific compositional generalisation properties in VLMs that may be masked in more complex, real-world scenarios. Future work could include expanding these experiments to test other attributes such as material or size. Another interesting avenue for future research would be to expand the experiments to include more than two objects.

Acknowledgements

The authors wish to acknowledge and thank the financial support of the UK Research and Innovation (UKRI) [Grant ref EP/S022937/1] and the University of Bristol. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol—http://www.bristol.ac.uk/acrc/

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, and 1 others. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, X. Yang, Chen-Dong Duan, Hang Su, and Jun Zhu. 2023. Robust classification via a single diffusion model. *ArXiv*, abs/2305.15241.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Kevin Clark and Priyank Jaini. 2023. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36:58921–58937.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen J Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1):345–384.

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

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *arXiv preprint arXiv:2406.11171*.

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

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36.

Hang Hua, Jing Shi, Kushal Kafle, Simon Jenni, Daoan Zhang, John Collomosse, Scott Cohen, and Jiebo Luo. 2024a. Finematch: Aspect-based fine-grained image and text mismatch detection and correction. In *European Conference on Computer Vision*, pages 474–491. Springer.

Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. 2024b. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob Hansen, Sivan Doveh, Victor Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, Chuang Gan, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. 2024. Conme: Rethinking evaluation of compositional reasoning for modern vlms. In *Advances in Neural Information Processing Systems*, volume 37, pages 22927–22946. Curran Associates, Inc.

Theo MV Janssen and Barbara H Partee. 1997. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

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.

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

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does CLIP bind concepts? probing compositionality in large image models. In *Findings EACL 2024*, pages 1487–1500, St. Julian's, Malta. Association for Computational Linguistics.

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

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pages 221–242. Springer.

Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023. Cross-modal attention congruence regularization for vision-language relation alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5444–5455, Toronto, Canada. Association for Computational Linguistics.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. 2024. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36.

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.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573.

Ziyi Shou and Fangzhen Lin. 2024. Enhancing semantic understanding in vision language models using meaning representation negative generation. In *Fourth Workshop on Knowledge-infused Learning*.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 869–893, Singapore. Association for Computational Linguistics.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A survey on compositional learning of AI models: Theoretical and experimental practices. *Transactions on Machine Learning Research*. Survey Certification.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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.

Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, and Aaron Courville. 2024. Sparo: Selective attention for robust and compositional transformer encodings for vision. In *European Conference on Computer Vision*, pages 233–251. Springer.

Chenwei Wu, Li Erran Li, Stefano Ermon, Patrick Haffner, Rong Ge, and Zaiwei Zhang. 2023. The role of linguistic priors in measuring compositional generalization of vision-language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 118–126. PMLR.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591.

Nir Yellinek, Leonid Karlinsky, and Raja Giryes. 2025. 3vl: Using trees to improve vision-language models' interpretability. *Trans. Img. Proc.*, 34:495–509.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13784.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37.

A Fine-tuning Details

Optimal hyper-parameters were selected by performing a search for each model. We consider the parameters: learning rate, images per class, epochs and LoRA parameters where applicable. We select final parameters based on averaged performance on the ID val and OOD val dataset splits.

Single-Object CLIP was fine-tuned using 40 images per class for 30 epochs, using an Adam optimiser with a learning rate of $1 \cdot 10^{-6}$, a batch size of 16, and a contrastive loss. For DC, we used DreamBooth to fine-tune Stable Diffusion's U-Net and text-encoder. We use 30 images per class for 4000 steps with a learning rate of $5 \cdot 10^{-6}$ and a batch size of 1. All inferences were performed using 200 noise samples. ViLT was fine-tuned on 80 images per class using LoRA with a learning rate of $1 \cdot 10^{-5}$ setting the LoRA rank (r) to 16 and the scaling factor (alpha) to 32.

Two-Object CLIP was fine-tuned using 40 images per class for 30 epochs, using an Adam optimiser with a learning rate of $1 \cdot 10^{-6}$, a batch size of 16, and a contrastive loss. For DC, we fine-tuned using 30 images per class for 4000 steps with a learning rate of $5 \cdot 10^{-6}$ and a batch size of 1. All inferences were performed using 200 noise samples. ViLT was fine-tuned using LoRA with a learning rate of $1 \cdot 10^{-5}$ setting the LoRA rank (r) to 16 and the scaling factor (alpha) to 32.

Relational CLIP uses the same parameters as the single-object model except using 20 images per class for 50 epochs. DC is fine-tuned on 40 images per class for 5000 steps with the remaining parameters the same as the previous two models. ViLT is fine-tuned on 40 images per class using LoRA with a learning rate of $1 \cdot 10^{-6}$ setting the LoRA rank (r) to 8 and the scaling factor (alpha) to 16.