Deciphering the Role of Representation Disentanglement: Investigating Compositional Generalization in CLIP Models

Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah

Sharif University of Technology, Tehran, Iran
{reza.abbasi,rohban,soleymani}@sharif.edu

Abstract. CLIP models have recently shown to exhibit Out of Distribution (OoD) generalization capabilities. However, Compositional Out of Distribution (C-OoD) generalization, which is a crucial aspect of a model's ability to understand unseen compositions of known concepts, is relatively unexplored for the CLIP models. Our goal is to address this problem and identify the factors that contribute to the C-OoD in CLIPs. We noted that previous studies regarding compositional understanding of CLIPs frequently fail to ensure that test samples are genuinely novel relative to the CLIP training data. To this end, we carefully synthesized a large and diverse dataset in the single object setting, comprising attributes for objects that are highly unlikely to be encountered in the combined training datasets of various CLIP models. This dataset enables an authentic evaluation of C-OoD generalization. Our observations reveal varying levels of C-OoD generalization across different CLIP models. We propose that the disentanglement of CLIP representations serves as a critical indicator in this context. By utilizing our synthesized datasets and other existing datasets, we assess various disentanglement metrics of text and image representations. Our study reveals that the disentanglement of image and text representations, particularly with respect to their compositional elements, plays a crucial role in improving the generalization of CLIP models in out-of-distribution settings. This finding suggests promising opportunities for advancing out-of-distribution generalization in CLIPs. For more details and access to our dataset, please visit https://github.com/abbasiReza/CLIP-COoD.

Keywords: Compositional Out-of-Distribution (C-OoD) Generalization · CLIP · Disentanglement

1 Introduction

Out-of-Distribution (OoD) generalization which is the ability of a model to generalize to the data distributions differing from the training distribution is very important for most learning models [1]. In recent years, several studies suggested that some Vision-Language Models (VLMs) such as the CLIPs [2], exhibit OoD generalization [2, 3]. Specifically, several studies reported that CLIP

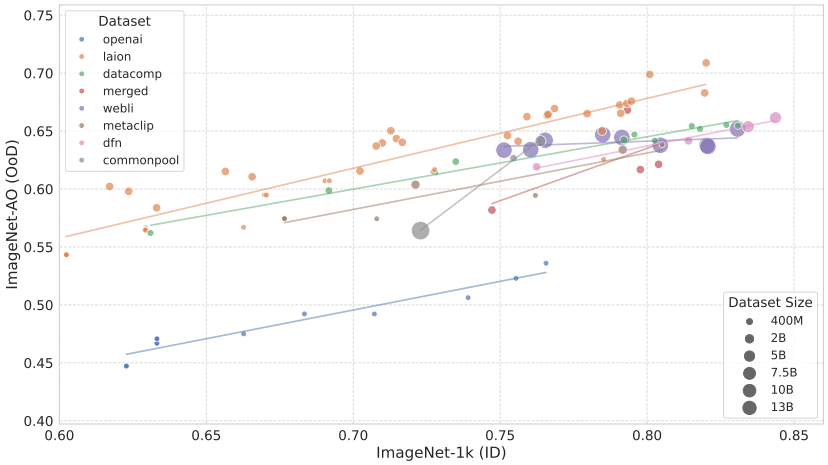

Fig. 1: Comparing zero-shot compositional out-of-distribution (C-OoD) generalization across diverse CLIP models and training sets. In-distribution (ID) performance is evaluated on the ImageNet validation set with object name labels, while the C-OoD generalization is assessed on our designed compositional dataset using attribute-object pair labels. Noticeably, CLIP models trained on the Common Pool dataset exhibit a steeper accuracy slope when transitioning from the ID to the OoD compositional setting compared to models trained on other datasets like WebLI. CLIPs trained on the LAION and DataComp datasets also show significantly higher C-OoD across ID accuracy. Despite improved in-distribution accuracy, models pretrained on WebLI do not demonstrate substantial gains in generalizing to the novel compositional out-of-distribution test cases.

models demonstrate enhanced zero- and few-shot accuracies on parallel versions of ImageNet, comprising images with various style shifts with respect to the original ImageNet [3, 4].

In particular, Compositional OoD (C-OoD) generalization is a main branch of the OoD generalization, focusing specifically on the ability of models to generalize to unseen combinations of known concepts or entities. Essentially, compositional generalization relates to human-like inductive biases that leads to more efficient learning via composing seen concepts [5]. Recently, some studies have worked on evaluating or improving compositional generalization in the NLP tasks [5–7]. However, C-OoD generalization for vision tasks is less explored since the unseen compositions of concepts can not be easily created visually for investigation.

In the recent years, evaluating the ability of VLMs in encoding objects, attributes, and their relations has recently received attention [8, 9]. Some benchmarks such as VL-Checklist [10], Winoground [11], and Attribute-Relation-Order (ARO) [8] have been introduced to assess the image-text matching ability of VLMs in compositional setups more exactly. VL-Checklist provides a benchmark to evaluate VLMs capabilities in three categories of objects, attributes, and re-

Fig. 2: Examples of images from our generated dataset. This dataset is created by combining attributes and objects that do not appear in the CLIP training sets, specifically designed for benchmarking compositional OoD generalization purposes.

lations. ARO showcases that the reordering of words in the text does not highly impact on the similarity of the text with the corresponding image. Some of these studies [8, 11] discussed shortcommings of VLMs in encoding the compositional relationships between objects and attributes and [9] showed that VLMs can compose concepts in a single-object setting including single attribute-object compositions. Nonetheless, most of the work around compositional reasoning [12–15] were more concerned about compositional understanding of the inputs, and less attention has been paid to the OoD generalization in which the generalization ability are evaluated against truly novel compositions with respect to the training set. In a nutshell, the literature suggests that compositional understanding in VLMs might be more feasible in the single-object setups. However, until now the C-OoD capability of CLIPs is unexplored. This makes us ask the question:

Do CLIPs really have nontrivial C-OoD generalization in the single-object setting? and where does this ability stem from in such models?

We propose a new benchmark to evaluate the C-OoD performance of CLIP models. Our approach involves generating a dataset, called ImageNet-AO (Attribute Object), distinct from the CLIPs training data. We gather comprehensive lists of objects and attributes, then generate images by combining these objects and attributes using a text-to-image model. The generated images undergo several filtering processes to ensure they are aligned with their intended and specified object-attribute description, and are novel compared to the combined CLIP training datasets both in the text and image domains. We then evaluate different

CLIP models on our OoD dataset to classify an input image into its composition constituents. Fig. Fig. 1 gives an overview of this result, in which certain CLIPs, such as the ones trained on the LAION and DataComp, yielded strong C-OoD performance.

Finally, we analyze the factors that contribute to better performance in our benchmark. We found that the CLIPs that show higher C-OoD generalization typically exhibit strong disentangled text representations with respect to the composition constituents. We backed this observation by assessing numerous disentanglement metrics, and the intrinsic dimensionality of the composition text embeddings. We found that CLIPs with strong C-OoD accuracy also enjoy a more disentangled image representation, albeit at a lower level compared to that of the text embedding. Based on these results, we hypothesize that the inherent disentanglement of the text is induced from the text representation space to that of the images through contrastive learning. We elaborate on this hypothesis in Sec. 4. Consistently, various disentanglement metrics of the text and image representations are observed to be highly correlated in CLIPs. We also repeat all these experiments in datasets that were previously designed for evaluating disentanglement, and contain factors at a more fine-grained level, and note that all these observations hold.

Our contributions are summarized as follows:

- Designing an image test dataset of attribute-object pairs that are unseen in common CLIP training datasets.
- Benchmarking the compositional generalization of various CLIPs in the carefully designed and controlled setting.
- Discovering that the CLIP representation space is decomposable into embedding of concepts (e.g., objects and attributes) especially for the embeddings obtained by the text encoder, and suggesting that it is the source of compositional generalization.
- Demonstrating a strong connection between CLIPs text/image disentanglements and better C-OoD generalization through different disentanglement metrics, on both our ImageNet-AO datasets and exisiting datasets designed previously for disentanglement evaluation.

2 Methodology

In this section, we explain how we conducted our study step-by-step. We first describe how we created our challenging benchmark dataset, ImageNet-AO, which involves finding new combinations and making images with text-to-image models (Sec. 2.1). Examples of images in ImageNet-AO are shown in Fig. 2. Then, we dive into how we test CLIP models in the zero-shot setting, and the chosen criteria to evaluate the models (Sec. 2.2).

2.1 ImageNet-AO Dataset Design

To rigorously evaluate the compositional generalization capabilities of vision-language models, we devised an innovative dataset featuring compositions that

are out-of-distribution with respect to the training datasets of these models. Our dataset is crafted to include rare and unique compositions, thus ensuring it presents novel challenges to the VLMs under study. The dataset construction process is meticulously designed and involves several key steps, as depicted in Fig. 3 and detailed below:

Selection of Objects (Nouns) Our initial step involved curating objects by extracting class names from the ImageNet dataset. This choice facilitates a direct comparison between the performance of models on our dataset and their performance on the well-established ImageNet validation set. By selecting a diverse array of class names, we aim to increase the complexity and richness of the generated compositional images.

Selection of Attributes (Adjectives) We then selected 140 adjectives from the Visual Attributes Words (VAW) dataset [16]. These adjectives span various categories, including color, material, and texture, allowing us to create a wide range of descriptive combinations for image generation. A complete list of the 140 adjectives used from the Visual Attributes Words (VAW) dataset is provided in Appendix .

Image Generation with Attribute-Object Prompts Utilizing the SD-XL Turbo, one of the most advanced and efficient text-to-image models available, we generated images based on combinations of the selected attributes and objects. By pairing 140 adjectives with 1,000 nouns, we created 140,000 unique prompts, which were then used to produce corresponding images, enriching our dataset with a vast array of compositional variety.

Filtering Process To guarantee the integrity and the intended OoD characteristics of our dataset, we implemented a meticulous three-step filtering process. This approach ensures that our dataset not only accurately represents the specified attribute-object combinations but also stands apart from existing datasets in terms of composition and novelty. The steps are as follows:

Step 1 - Initial Validation: Each generated image was subjected to an initial evaluation to verify its accuracy in depicting the intended attribute-object pair, exclusively through human assessment. During this process, evaluators were tasked with answering two critical questions: "Is this an image of [object]?" and "Does it exhibit [attribute]?" If at least one of these questions was answered with a "no," the image was removed from consideration. This step ensured that only images accurately representing the specified characteristics were retained for further processing.

Step 2 - Exclusion of Known Combinations: To ensure the exclusivity of our dataset, we conducted a comprehensive search across several datasets (LAION, CommonPool, YFCC, and CC) to identify and eliminate any attribute-object combinations already present. This was achieved through a relaxed match-

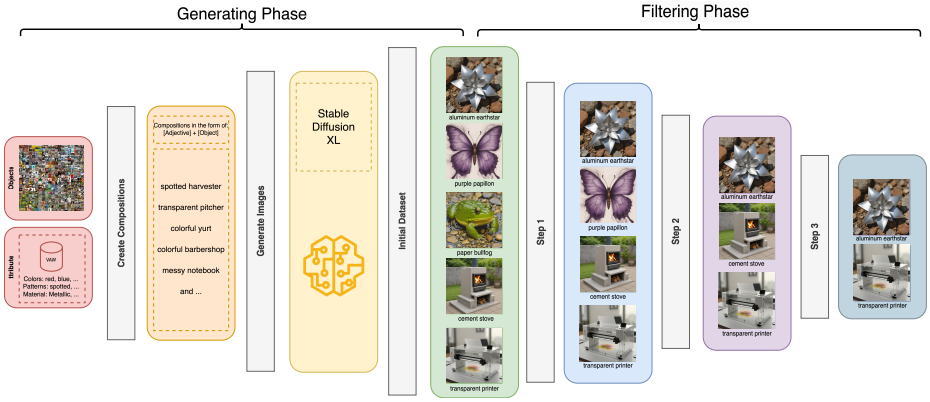

Fig. 3: Dataset Design Stages: The data design process involves a generation phase that makes the initial dataset from the whole set of the object and attribute compositions, and three distinct filtration steps. In the first filtration step, images where the target attribute or object lacks clear visibility are eliminated. In the second filtration step, the process removes images whose captions are already present in public datasets specifically curated for CLIP training. In the third filtration step, the faiss k-nearest neighbors algorithm is employed to identify and filter out images exhibiting similarities.

ing criterion, where combinations were removed if both the object and attribute appeared in a caption of an image, even if not in direct association.

Step 3 - Verification of OoD Status: The final step in our filtering process was to ensure the OoD nature of our dataset. We used the Faiss library [17] for a K-nearest neighbors search to compare our generated images against those in the LAION, CommonPool, YFCC, and CC datasets. Images were considered unique and retained in our dataset if no closely matching analogs were found, based on human evaluation. This rigorous approach ensured the novelty and uniqueness of our dataset by excluding combinations that had similar matches in the referenced datasets.

The dataset design process culminates in around 21,000 novel combinations of attributes and objects. The final generated dataset, after passing through the filtering process, comprises approximately 60,000 images representing 21,000 unique attribute-object combinations. Detailed properties and statistics about the dataset, including the list of attributes and objects used, can be found in the appendix. Additionally, another filtered version of the dataset is also available in the appendix.

2.2 Model/Data Zoo and Evaluation Criteria

In our experiments, we evaluate CLIP models trained on a diverse selection of datasets, including OpenAI's private dataset, LAION, YFCC15m, CC12m, DataComp, DFN-5B, WebLI, and CommonCrawl. These models leverage a variety of backbone image encoders such as ResNet50, ResNet101, ViT-B-32, ViT-B-16,

ViT-L-14, ViT-H-14, ViT-g-14, and ViT-BigG-14. Our evaluation also extends to new CLIP variations, including EVA CLIP, SigLIP, and CLIPA, allowing for a comprehensive assessment of their performance and generalization capabilities across different tasks and datasets.

3 Comparison of CLIP Models on ImageNet-AO

To evaluate the CLIP model performance in the classification tasks, we adopted the evaluation method developed by [18], similar to the zero-shot evaluation approach described in [2]. Our evaluation involves providing the model with the actual images and various captions, obtaining embeddings for both the images and texts, and calculating their cosine similarities. This allows us to estimate the relevance of the captions to the image content, similar to a classification task. Given that our dataset only provided class labels (attribute-object pairs) for images, we expanded on this by creating 80 captions per class using various templates. This approach, inspired by the methodology described in [2], allows for a more comprehensive representation of each class. We generated embeddings for these captions and averaged them to produce a final embedding for each class, which was then used in our zero-shot evaluation. For the test sets, all 1000 classes of ImageNet were used as the in-distribution set and expanded the number of classes to approximately 21000 for the OoD set. The CLIP evaluations are shown in Fig. 1.

While our results generally showed that models trained on larger datasets exhibited improved accuracy in both in-distribution and out-of-distribution settings, supporting the notion that larger training datasets can enhance compositional out-of-distribution generalization performance, it is crucial to note that dataset size alone does not directly predict model strength. The performance of models varied significantly with not only the dataset size but also the quality and curation of the data. For instance, CLIP trained on the unfiltered CommonPool-XL dataset performed weaker than CLIP trained on the CommonPool-XL dataset filtered using ClipScore, despite the unfiltered dataset containing an additional 7 billion images. This further reinforces that simply increasing dataset size does not necessarily lead to improved model performance, and carefully curating and filtering the data can be more effective than merely accumulating vast amounts of unfiltered data.

Additionally, as evident from Fig. 1, models with different configurations trained on various datasets exhibited different training slope trajectories. The models trained on CommonPool-XL with different data filtering techniques demonstrated particularly steep performance trends, suggesting that the combination of a large dataset and effective data curation can yield significant performance gains.

Interestingly, the SigLip (denoted as WebLI) models presented a unique case with a somewhat negative slope, indicating that while enhancements to the backbone architecture improve in-distribution data performance, they may adversely

affect out-of-distribution data performance. This highlights the nuanced relationship between architectural improvements and model generalization capabilities.

This extensive analysis, which encompasses the performance of diverse CLIP models across a broad spectrum of datasets, underscores the complexity of factors influencing model behavior and the pivotal role of dataset characteristics in achieving optimal performance in both in- and out-of-distribution settings. Further details on the performance evaluation of various CLIP models can be found in Sec. 7.4 of the Appendix.

4 Why CLIP has Compositional Generalization?

Having established the superior C-OoD performance of certain CLIPs, we next try to investigate the reasons behind these observations. It has been widely known that disentangled representations make meaningful construction of known concept mixtures in the embedding space feasible, hence resulting in better C-OoD generalization [19–21]. Here, disentanglement means assignment of separate and independent embedding dimensions to different factors of variations, which in this case are the objects and attributes.

We hypothesize that the discrete nature of the language, and large and diverse training datasets promote a more decomposable text representation. On the other hand, alignment of the text and image embeddings through contrastive learning in CLIPs induces this decomposability in the image domain. Based on these insights, we posit that representation decomposability is the key to the CLIP unseen compositional generalization. This claim is supported by two main arguments:

- Decomposability of the CLIP text embedding, measured through a comprehensive set of metrics, is correlated to the CLIP C-OoD generalization (Fig. 4, bottom row).
- Text representation disentanglement is induced in the image encoding, due to implicit maximization of the mutual information of text and image representations through contrastive learning. We elaborate on this claim empirically (Fig. 4, top row), and theoretically in what follows.

Why disentanglement is induced from one view to another in the contrastive learning? We next try to give some theoretical insight on why and how the disentanglement emerges in the CLIP vision encoder. Several studies have shown the relation between minimizing the contrastive loss and maximizing the mutual information [22]. Therefore, the CLIP training implicitly maximizes the mutual information between text and image embeddings. We claim that disentanglement in the text representation, which was evidenced previously, may encourage disentanglement in the image encoding. To see this, let y_1 and y_2 be the text embeddings for the objects and attributes, respectively. Let x_1 and x_2 be the corresponding image embeddings. Assuming a decomposable text embedding means $y_1 \perp y_2$, i.e. $p(y_1, y_2) = p(y_1)p(y_2)$. Now by minimizing the

contrastive loss, the mutual information $I(x_1, x_2; y_1, y_2)$ is maximized. By letting $x = (x_1, x_2)$, and $y = (y_1, y_2)$, we have:

$$\begin{aligned} I(x_1, x_2; y_1, y_2) &= \mathrm{D}_{\mathrm{KL}}(p(x, y) \parallel p(x)p(y)) \\ &= \mathrm{D}_{\mathrm{KL}}(p(x_1|x_2, y)p(x_2|y)p(y) \parallel p(x_1|x_2)p(x_2)p(y)) \\ &= \mathbb{E}_{x_1,x_2,y}[\log(p(x_1|x_2, y)/p(x_1|x_2))] + \mathbb{E}_{x_2,y}[\log(p(x_2|y)/p(x_2))] \\ &= \mathbb{E}_{x_2,y}[\mathrm{D}_{\mathrm{KL}}(p(x_1|x_2, y) \parallel p(x_1|x_2))] + \mathbb{E}_y[\mathrm{D}_{\mathrm{KL}}(p(x_2|y) \parallel p(x_2))] \end{aligned}$$

Maximization of the latter term makes x_2 and y dependent random variables, otherwise if $x_2 \perp y$, the expected KL divergence would be minimum (or zero), which is against maximizing the mutual information. Note that however, x_2 does not ideally depend on both y_1 and y_2, otherwise the two distributions in the KL divergence in the first term become similar, which is also against maximizing the mutual information. Putting these together, x_2 mostly depends on y_2 if the mutual information is maximized. Using a symmetric argument, x_1 mostly depends on y_1. Finally, because $y_1 \perp y_2$, we conclude that x_1 and x_2 tend to become independent. Therefore, maximizing $I(x_1, x_2; y_1, y_2)$ decomposes x if y is already decomposed.

5 Decomposable representation of CLIP Models

In this section, our primary objective is to leverage the generated dataset and other synthtic datasets to analyze our hypotheses, focusing on the decomposable CLIP representation space and its impact on the compositional OoD performance.

5.1 Attribute-Object Decomposition of Representation Space

In this section, we show that the representation space of the CLIP models on the proposed dataset can be decomposable into the representations of the objects and the attributes.

Disentanglement of Attributes and Objects Here, we aim to assess the level of embeddings disentanglement in various CLIPs on ImageNet-AO. We utilize some common disentanglement metrics, namely the Z-Diff Score [23], DCI [24] and Explicitness score [25] to quantitatively evaluate the embeddings. These metrics are typically employed for supervised disentanglement assessment and require access to the latent factors of data. Since we have a compositional text specifying the attribute and the object for each image, we can consider two super latent factors corresponding to attributes and objects respectively. More details about these disentanglement metrics and their formulas can be found in Appendix 7.5 .

We calculate these metrics for each CLIP model on our ImageNet-AO dataset. Subsequently, in Fig. 4 (bottom), we visualize the relationship between the C-OoD accuracy and the disentanglement metrics. Each point in the plot represents

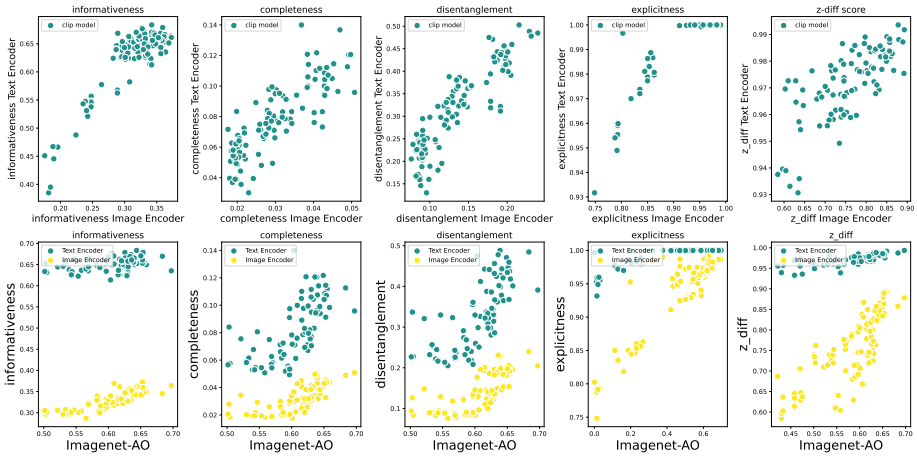

Fig. 4: Top: Representation disentanglments are correlated in text and image embeddings of CLIPs. Bottom: Disentanglment metrics vs. C-OoD Accuracy.

a CLIP model, with the x-axis denoting the C-OoD accuracy and the y-axis representing the disentanglement metric. As observed in bottom row of the plot, there is a discernible pattern where models with higher C-OoD accuracy tend to exhibit more disentangled text and image representations. This empirical observation aligns with our initial hypothesis. Notably, the disentanglement in the text embedding (blue points), is more pronounced compared to the image embeddings (green points). Additionally, in 4 (top), we show the correlation between the image encoder and the text encoder for different disentanglement metrics. This figure demonstrates that by increasing the disentanglement in the text encoder, the disentanglement in the image encoder also increases, indicating a correlation between them.

Intrinsic Dimensionality of the Composition Representations The previously reported metrics of disentanglement focus on the correspondence between embedding dimensions and latent factors, and hence often require training an auxiliary classifier, in which a given representation is classified into levels of any latent factor. One could alternatively take a training-free approach through measuring relative intrinsic dimensionality of the composition patterns. This could be achieved by measuring the soft rank of the embeddings of attribute-object pairs. The soft rank is defined by the number of singular values of a given matrix that are greater than a pre-specified positive threshold. The soft rank is then normalized and made comparable across CLIPs by being divided to the number of embedding dimensions. This way the soft rank measures the relative intrinsic dimensionality of the embedding space. If the representation is entirely disentangled, huge combinations of attribute-objects would only result in a small intrinsic dimensionality, i.e. sum of the intrinsic dimensionalities of object and

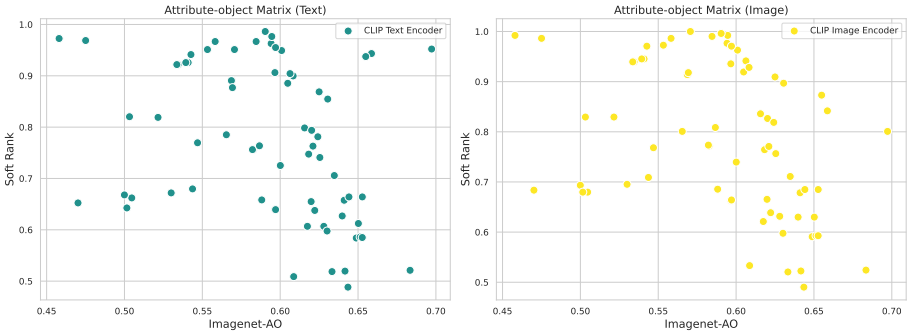

Fig. 5: The decrease in the soft rank of attribute-object representations relative to the embedding size correlates with improved C-OoD accuracy. This indicates that decomposing representations of attributes and objects results in a low dimensional representation of CLIPs that exhibits robust C-OoD performance. This highlights the representation disentanglement in CLIPs with strong C-OoD generalization.

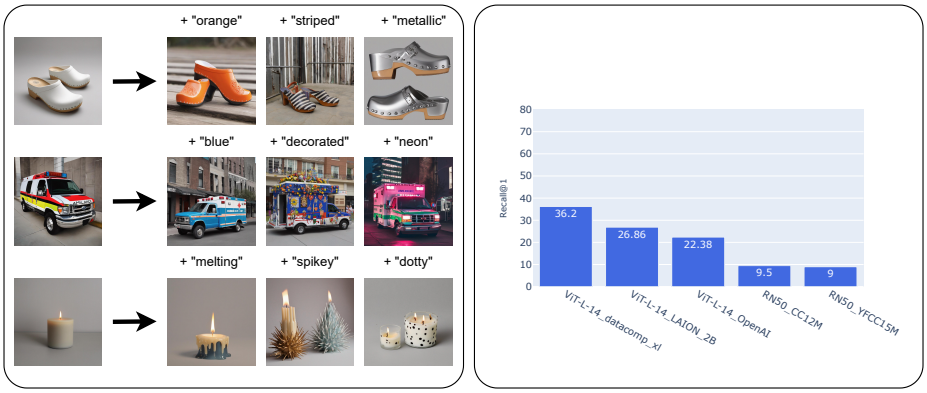

Fig. 6: The performance of various CLIP models in the task of image±text retrieval. A model's superior performance in this task indicates that its representation is more decomposable.

attribute spaces. Otherwise, each attribute-object embedding would appear to be *novel* with respect to other composition embeddings, resulting in a near full-rank space.

For this experiment, we use ImageNet-AO, which provides around 21,000 unique combinations of attributes and objects. We utilize their image embeddings, obtained from the CLIP image encoder, and caption embeddings, obtained from the CLIP text encoder, to calculate the soft rank with a threshold of 0.1. Fig. 5 shows that the intrinsic dimensionality is decreasing as the C-OoD accuracy increases, in both text and image domains.

Image retrieval with image±text queries Inspired by the work of [26], we designed an experiment to evaluate the compositional nature of embeddings learned by the CLIP models. Our primary objective is to assess the representation disentanglement of the CLIP models trained on diverse datasets. To accomplish this goal, we devised a test in which we input an image from our dataset into the image encoder of the model, and obtain its corresponding embedding. Next, we employed the text encoder of the model to compute the embedding of an adjective, ensuring that the adjective differed from those associated with the current image. These two embeddings were then combined through summation and used as a query in a process similar to the image retrieval. We then show the image closest to the generated query embedding. A total of 200 random images were used to conduct this test for each model.

In order to evaluate the accuracy of the models predictions, we consider the image that is most similar to the query as the correct prediction if it possess both the intended object and adjective. A higher level of accuracy in the image retrieval task indicates that the model embeddings are more disentangled. Model evaluations are demonstrated in Fig. 6. The Recall@1 performance of various models aligns with our expectations. Specifically, we anticipated that models excelling in C-OoD tasks would also exhibit more disentangled representations. We previously observed in Fig. 1 that CLIPs associated with LAION and DataComp datasets stand out as having highest C-OoD accuracies. These two CLIPs also performed best in this experiment.

5.2 Disentanglement of Fine-Grained Factors

In the field of Disentanglement Representation Learning, the concept of disentanglement is explored from two distinct perspectives: fine-grained factors at the dimension level and coarse-grained factors at the vector level [27]. Our initial investigation into CLIP models, utilizing our curated dataset, provided insights into coarse-grained disentanglement (e.g. separating attributes and objects as two factors) and revealed multifaceted evaluation metrics. Moving forward, we aim to delve into the realm of fine-grained disentanglement at the dimension level. However, our current dataset poses inherent limitations in segregating factors at such a granular level. Consequently, to facilitate a comprehensive evaluation of fine-grained disentanglement, it becomes necessary to adopt specialized datasets designed explicitly for disentanglement studies within this domain.

For our in-depth analysis of the fine-grained disentanglement, we selected two distinguished datasets: Sprites [28], Shapes3D [29] as they are specifically designed for disentanglement studies in image-centric models. Examples from these datasets can be seen in Fig. 7.

Since our focus extends beyond image-centric models to evaluate disentanglement in both the text encoder and image encoder components of CLIP models, we generated captions for each image based on the vector of factors associated with that image. This approach enables us to assess the disentanglement capabilities of CLIP models in both the visual and textual domains.

Fig. 7: Disentanglement datasets. a: Sprites dataset, consist of 6 factor and 54,000 images b: Shape3D, consist of 5 factor and 32,000 images

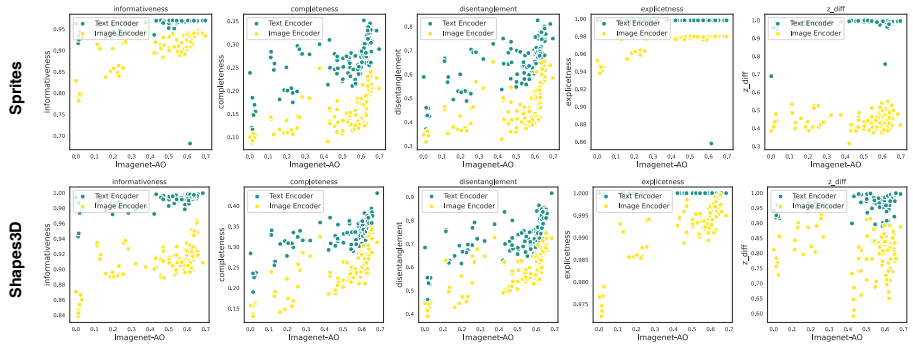

Fig. 8: Disentanglment metrics vs. C-OoD Accuracy on Sprites and Shapes3D dataset.

Figure 8 shows the text encoder exhibits higher disentanglement than the image encoder. As models improve on the C-OoD task, disentanglement tends to increase for both encoders.

More Analysis on decomposability of the representation space Using the Shapes 3D, we conducted two experiments to investigate the representation of factors more accurately.

In first experiment, we employ the 480,000 images of Shapes 3D dataset, each with specific latent factors such as floor hue, wall hue, object hue, scale, shape and orientation. We train a classifier to calculate the Z-Diff Score and utilize it to determine which dimensions are most critical for each latent factor. In the process of calculating the Z-Diff score, we train a classifier that can determine, for a group of data points that have a fixed specific value for one of the latent factors, what that factor is. By using this classifier, we can identify which dimensions are more important for determining each factor. Subsequently, we extract the top 100 important dimensions for each factor and calculate how many dimensions are common across factors. Our results, presented in Table

1, demonstrate that models with higher C-OoD accuracy tend to exhibit fewer common dimensions across factors. This finding suggests that improved C-OoD generalization is associated with more disentangled representations.

In the second experiment, we looked at the impact of disentanglement on zero-shot object color manipulation using two identical images except for the object color. We calculated the embeddings using the CLIP and used the classifier of the first experiment to identify the most important dimensions for detecting object color. By switching the top k dimensions between the two image embeddings, we tested the models' ability to detect captions matching the switched new color. The results are summarized in Table 1 showing that models with higher C-OoD accuracy require fewer dimension switches to achieve the color change, indicating that disentangled representations enable more effective zero-shot modifications.

Table 1: Number of common dimensions across factors and switching dimensions for color manipulation in the Shapes 3D dataset

Dataset	Architecture	C-OoD Acc.	# Com. Dims	# Sw. Dims
LAION	ViT-L/14	64.61%	2	40
LAION	ViT-B/16	61.55%	5	60
LAION	ViT-B/32	61.05%	7	90
OpenAI	ViT-L/14	52.28%	3	5
OpenAI	ViT-B/16	49.22%	4	10
OpenAI	ViT-B/32	47.07%	6	30
CC	RN50	26.64%	15	200
YFCC	RN50	12.23%	21	250

6 Conclusion

This study examines how well CLIPs can generalize to new compositions of objects and attributes. We created an authentic benchmark of compositional images that are truly novel with respect to the CLIP training sets, and found that CLIPs ability to decompose the text/images representation space (into the embedding of concepts) is crucial for the compositional generalization. We have assessed the decomposability through the lens of several well-known metrics, as well the composition representation intrinsic dimensionality. These experiments were conducted on a wide range of datasets, from our attribute-object dataset to the ones previously designed specifically to evaluate disentanglement. We also covered a wide variety of problem setups in this direction, ranging from factor classification, and image±text retrieval, to factor manipulation. All mentioned assessments consistently demonstrate a strong connection between text and image representation disentanglement and C-OoD generalization.

References

1. Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.
2. 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
3. Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.
4. Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.
5. Zi Wang and Daniel Hershcovich. On evaluating multilingual compositional generalization with translated datasets. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1687, Toronto, Canada, July 2023. Association for Computational Linguistics.
6. Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online, August 2021. Association for Computational Linguistics.
7. Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur P Parikh, and Emma Strubell. Improving compositional generalization with self-training for data-to-text generation. *arXiv preprint arXiv:2110.08467*, 2021.
8. 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 *The Eleventh International Conference on Learning Representations*, 2023.
9. Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models, 2023.
10. Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023.
11. 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, 2022.
12. Timothy Ossowski, Ming Jiang, and Junjie Hu. Prompting large vision-language models for compositional reasoning, 2024.
13. Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples, 2024.

14. Haoxiang Wang, Haozhe Si, Huajie Shao, and Han Zhao. Enhancing compositional generalization via compositional feature alignment, 2024.
15. Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023.
16. Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild, 2021.
17. Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.
18. Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.
19. Tao Yang, Yuwang Wang, Cuiling Lan, Yan Lu, and Nanning Zheng. Vector-based representation is the key: A study on disentanglement and compositional generalization, 2023.
20. Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021.
21. Zhenlin Xu, Marc Niethammer, and Colin A Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.
22. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
23. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.
24. Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.
25. Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. *Advances in neural information processing systems*, 31, 2018.
26. 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
27. Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning, 2023.
28. Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder, 2018.
29. Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.
30. Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain, 2022.

31. Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2020.
32. Milton Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, 35:10136–10149, 2022.
33. Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.
34. Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.
35. Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024.
36. Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36, 2024.
37. E Paxon Frady, Spencer Kent, Quinn Tran, Pentti Kanerva, Bruno A Olshausen, and Friedrich T Sommer. Learning and generalization of compositional representations of visual scenes. *arXiv preprint arXiv:2303.13691*, 2023.
38. Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. *arXiv preprint arXiv:2310.05327*, 2023.
39. Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023.