Language Plays a Pivotal Role in the Object-Attribute Compositional Generalization of CLIP

Reza Abbasi
Sharif University of Technology
Tehran, Iran
reza.abbasi@sharif.edu

Mohammad Samiei
Sharif University of Technology
Tehran, Iran
mm.samiei@student.sharif.edu

Mohammad Hossein Rohban
Sharif University of Technology
Tehran, Iran
rohban@sharif.edu

Mahdieh Soleymani Baghshah
Sharif University of Technology
Tehran, Iran
soleymani@sharif.edu

Abstract

Vision-language models, such as CLIP, have shown promising Out-of-Distribution (OoD) generalization under various types of distribution shifts. Recent studies attempted to investigate the leading cause of this capability. In this work, we follow the same path, but focus on a specific type of OoD data - images with novel compositions of attribute-object pairs - and study whether such models can successfully classify those images into composition classes. We carefully designed an authentic image test dataset called ImageNet-AO, consisting of attributes for objects that are unlikely encountered in the CLIP training sets. We found that CLIPs trained with large datasets such as OpenAI CLIP, LAION-400M, and LAION-2B show orders-of-magnitude improvement in effective compositional OoD generalization compared to both supervised models and CLIPs trained with smaller datasets, such as CC-12M and YFCC-15M. Our results provide evidence that the scale and diversity of training data and language supervision play a key role in unlocking the compositional generalization abilities of vision-language models.

1. Introduction

The advent of large pre-trained models has significantly advanced the field of machine learning. Innovations such as GPT-3 [1], Chinchilla [2], PaLM [3], and CLIP [4] have broadened the horizons of generalization and underscored their exceptional capacity for zero-shot inference. The Out-of-Distribution (OoD) generalization of models like CLIP has been explored, revealing two differing perspectives on its origin: one attributing it to dataset diversity[5], the other to language supervision[6].

Most of the previous work studied the CLIP generalization under a certain type of out-of-distribution data, namely, distribution shifts [7, 8, 5]. However, there are other types of OoD generalization, including spurious correlation [9], and compositional generalization [10]. One has to note that each of these OoD generalization categories has a unique nature that should be studied separately.

This paper focuses on the compositional generalization, the ability of models to generalize new combinations of known concepts. Despite some shortcomings, it has been shown that Vision-Language Models (VLMs) can compose concepts in the single-object setting[11]. We explore if the compositional nature of VLMs impacts their compositional OoD generalization, hypothesizing that joint vision-language representation learning has enhanced CLIP's decomposability between objects and attributes in images containing single objects.

A significant challenge in evaluating OoD generalization is the unknown training distribution, as seen in models like CLIP where the training dataset has not been released. A novel benchmark design is proposed to assess CLIP models, involving a new compositional OoD dataset of unconventional attribute-object pairs distinct from the CLIP's training data called Imagenet-AO. We evaluate various CLIP models on this Imagenet-AO dataset to determine their performance and analyze contributing factors to the performance, offering insights into enhancing CLIP's generalization abilities. Our contributions include crafting an unseen attribute-object pair image test dataset called Imagenet-AO, providing a controlled benchmarking setting for various CLIP models using Imagenet-AO, and identifying the importance of compositional diversity in training captions for CLIP to demonstrate

decomposable representation and basic compositional generalization.

2. Related works

2.1. Robustness to Natural Distribution Shift

In certain applications, the test samples may exhibit different styles, colors, or contrasts compared to the training data. OoD generalization under such distribution shifts have extensively been studied, and it has been argued that training on diverse datasets is the most effective factor in increasing the robustness [12, 8], while combining various data modalities did not enhance the performance [5].

2.2. Compositional Generalization of CLIP

Compositional generalization, generalizing to unfamiliar compositions of familiar elements, poses challenges for models like CLIP. This includes associating attributes with objects, understanding object relationships, and extrapolating to unfamiliar concept combinations. Possible solutions to this problem include utilization of image scene graphs and augmentation framework for contrastive learning [13], leveraging LLMs to generate sentence-level descriptions for each compositional class [14], and fine-tuning the vocabulary for attributes and objects on seen classes, then recomposing the learned vocabulary in new combinations for the novel classes [15]. The emergence of concept representations within CLIP was studied in [16]. In [17], the authors examine VLMs struggles with relation, attribution, and order understanding. They propose a novel training procedure to improve these aspects. This work differs from the mentioned studies by investigating and comparing the power of CLIP's compositional generalization in a single-object setting, including attribute-object compositions, and creating a dataset with combinations of objects and unusual attributes.

3. CLIP Object-Attribute Compositional Generalization

Compositional OoD generalization refers to a model's ability to handle novel combinations of familiar concepts. This is critical in contexts like attribute-object images, where the goal is perceiving new compositions of objects and attributes.

Decomposable image representations that assign separate embedding dimensions to objects and attributes facilitate this generalization. Such representation makes meaningful construction of known concepts in the embedding space feasible. We hypothesize that large and diverse datasets reduce the dependency between attributes and objects, promoting a more decomposable understanding of images. Based on these insights, we posit that decomposability is the key to the CLIP model's unseen composition generalization. This claim is supported by two main arguments:

- Large and diverse datasets reduce entanglement between object and attribute tokens. In other words, they help to promote a more decomposable text representation (see sec. 5.2).
- Text representation decomposability is induced in the image encoding, due to implicit maximization of the mutual information. We elaborate on this claim in what comes next.

Why decomposability may arise in contrastive learning?

CLIP training maximizes the mutual information between text and image encodings. We claim that decomposability in the text representation, induces decomposability 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:

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

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.

4. ImageNet-AO: Dataset Desgin

To effectively assess the compositional generalization capabilities of models, we created a unique dataset of rare compositions, ensuring these were not present in the models' training data. This dataset was produced by creating compositional images via a text-to-image model, using an Attribute+Object template. Our process is as follows:

Selecting objects or nouns: We extracted class names from the ImageNet dataset, using these as objects (or nouns) in our structure to create a link between the generated images and ImageNet classes. This allows for comparison of

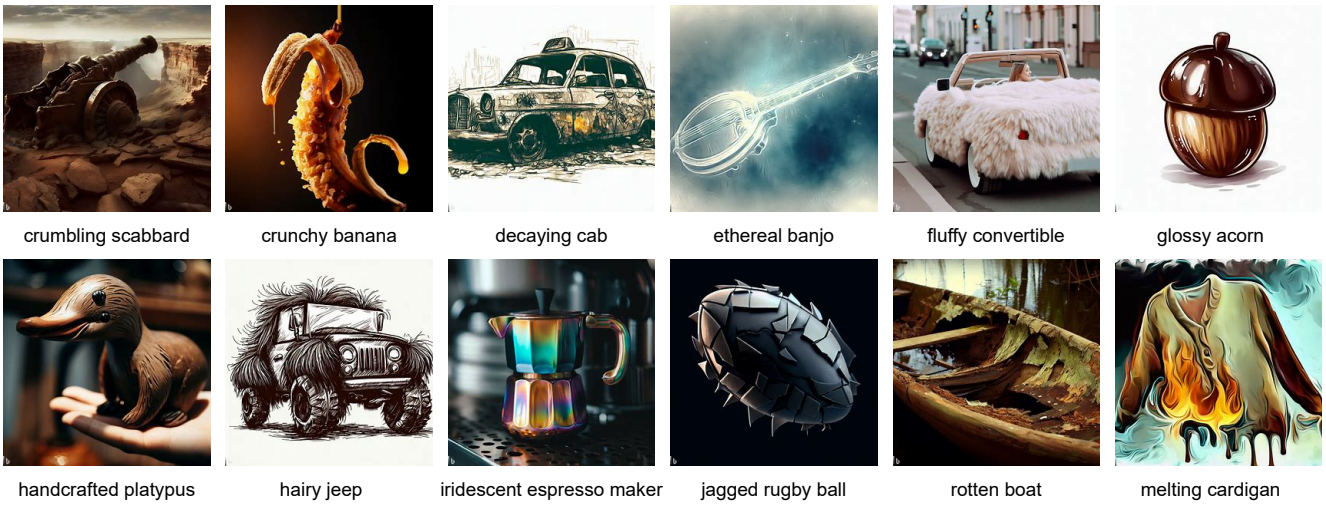


Figure 1. Examples of images from Imagenet-AO dataset. This dataset is created by combining attributes and objects that do not appear in the CLIP training sets, specifically designed for benchmarking OoD generalization purposes. More examples are in Figure 8.

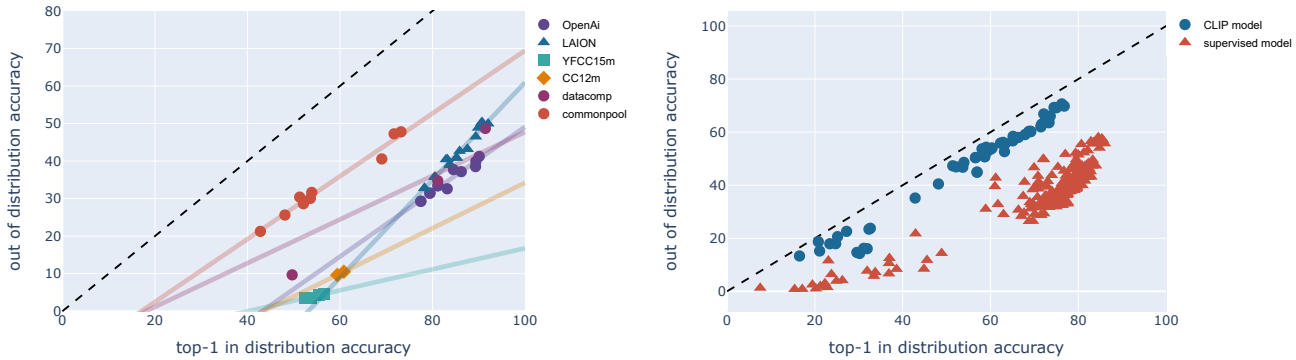


Figure 2. a) Comparing effective OoD generalization of CLIP models with diverse backbones and training sets in a zero shot setting, where no fine-tuning is performed on the target task. The in-distribution (ID) test set is the ImageNet validation split, with the labels being the object names, while the out-of-distribution test set is our designed compositional dataset, with labels being attribute-object pairs. Noticeably, there is a large gap between the performance of CLIPs that are trained on small datasets, e.g. CC15m and YFCC12m, and that of the CLIPs trained on gigantic datasets such as LAION and OpenAI. b) Comparing OoD generalization of the models trained with a supervised loss vs. CLIPs. ID and OoD test sets are the same as before, with the labels being the object names in both ID and OoD test sets, as the adjectives are not among the labels of the pre-trained supervised models. Despite being competitive on ID accuracy, the supervised models fall short of the OoD accuracy of the CLIP models.

model performances on the familiar ImageNet validation set. We aimed for a diverse set of class names to enhance the complexity of the generated images.

Selecting attributes or adjectives: The next step involved choosing 30 distinct adjectives that were relevant and could create unique combinations with the selected nouns, enhancing the diversity of our compositional images.

Selecting unseen (attribute, object) pairs: We combined the 30 adjectives with a pool of 1000 nouns, resulting in 30000 distinct pairs. These were given to the text-to-image model to generate corresponding images. To ensure these combinations were not present in the CLIP training set, we conducted a thorough search and removed any that were found.

Generating images for (attribute, object) pairs: The selected combinations were given to a text-to-image model for the image generation. Among various models, the Microsoft model powered by DALL-E proved to be the most powerful. However, it had limitations and some prompts were blocked for unknown reasons.

Validating the generated images: Lastly, human supervision was used to validate the generated images, with images not closely aligning with their prompts removed. After this process, around 12000 combinations remained, for which we successfully generated around 50000 accurate, high-quality images. An illustrative example of the diversi-

fied dataset generated through this process can be observed in Figure 1. This figure showcases a selection of images that exhibit various degrees of alignment with their corresponding prompts, highlighting the effectiveness of the validation procedure.

5. Experiments

In this section, we examine the effects of language supervision on compositional Out-of-Distribution (OoD) performance. We explore links between the training dataset characteristics, and CLIP OoD generalization. Specifically, we assess our hypothesis regarding the role of the training data quality and quantity in disentangling the object and attributes, and its consequences in compositional OoD generalization. In a nutshell, we found that CLIPs whose training sets consist of more diverse *caption compositions* would exhibit this property more than other CLIP models.

5.1. CLIP Models Comparison

We assessed CLIP model performance in zero-shot classification tasks using an evaluation method similar to that of [18] and [4] on ImageNet-AO dataset. We provided the model with images and captions, then calculated their cosine similarities to estimate the caption relevance to the image content. The models trained on the LAION 400m, LAION 2B, and DataComp 12.8B datasets showed similar performances on ImageNet-AO compared to the model trained on the OpenAI dataset. This indicates the potential efficacy of these datasets in training CLIP models for specific evaluated composition types. While larger training datasets typically resulted in enhanced accuracy, the CLIP model trained on YFCC15m displayed lower performance than the CC12m model, despite the former's larger dataset size. Additionally, experiments showed that models trained on Commonpool data filtered by LAION or CLIP scores outperformed the model trained on the full unfiltered Commonpool set, although the latter contained more data. This implies that various other factors can play a role in influencing the model behavior. To be more precise, the subsequent subsection discusses one of these factors that can significantly impact the model performance. To visualize the comparative performance of these CLIP models trained on different datasets, refer to Figure 2a.

5.2. Attribute-Object Tokens Mutual Information

We hypothesize that the use of datasets containing diverse, creative, and imaginary samples with less dependency between object and attribute during training is critical for enabling models to learn decomposable representations. To evaluate the degree of decomposability in the CLIP training data, we conducted an analysis by measuring the normalized mutual information (NMI) between object class and attributes, whose domains are defined based on the captions in

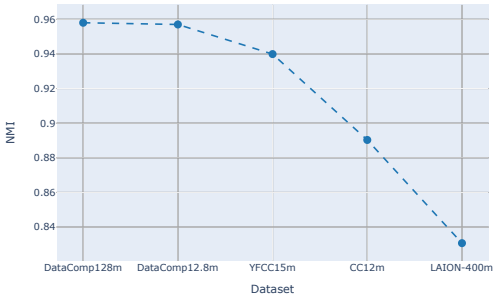


Figure 3. Normalized Mutual Information between attributes and objects in different CLIP training sets based on ImageNet-AO compositions

ImageNet-AO. The NMI is calculated based on the datasets on which CLIP was trained, enabling us to gauge the level of decomposability present in the training data. A lower NMI value indicates better disentanglement of attributes and objects.

The findings are depicted in Figure 3, which demonstrates that the LAION 400m dataset exhibits lower NMI values compared to the CC12m dataset. Similarly, the CC12m dataset displays lower NMI values compared to the YFCC15m dataset. These observations are aligned with the outcomes of our previous experiments on compositional OoD generalization.

When the mutual information between variables in a dataset is reduced, it indicates a diminished statistical dependence among those variables. In the context of decomposability, this implies that the factors of variation within the dataset are less entangled or intermingled. Additionally, the low values of NMI emphasize the diversity in the textual components of the dataset. This diversity is a crucial aspect for CLIP to attain high performance in effectively handling the OoD scenarios.

5.3. Comparison with Supervised Models

In this experiment, we investigated the impact of language supervision on CLIP models compared to supervised models under compositional OoD settings. We did not intend a direct comparison, but rather to explore if CLIP's language supervision improves the OoD accuracy. We assumed the object names as the class labels and evaluated the supervised models' accuracy on ImageNet-AO. For CLIP, we generated captions using only object names, removing adjectives, to align the evaluations.

Figure 2b shows CLIP models trained on OpenAI, LAION, and DataComp datasets consistently outperform supervised models on the OoD accuracy. This suggests that language supervision during CLIP training positively impacts the model representation decomposability, enabling generalization to detect unseen compositions.

6. Conclusion

This study examines the generalization of CLIPs to new object and attribute compositions. We created a benchmark dataset of compositional images and found that CLIPs training data quality is crucial for the compositional generalization. We showed that models trained on more diverse caption compositions perform better, and language supervision during training improves OoD generalization. The study highlights the importance of dataset diversity and decomposability in enhancing vision-language models' compositional generalization capabilities.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[4] 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. 1, 4, 7

[5] 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. 1, 2

[6] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022. 1

[7] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23664–23678. Curran Associates, Inc., 2021. 1

[8] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 1, 2

[9] Brian D Haig. What is a spurious correlation? *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2(2):125–132, 2003. 1

[10] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016. 1

[11] 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. 1

[12] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 2

[13] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *arXiv preprint arXiv:2305.13812*, 2023. 2

[14] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. *arXiv preprint arXiv:2305.14428*, 2023. 2

[15] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[16] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn primitive concepts? *arXiv preprint arXiv:2203.17271*, 2022. 2

[17] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2

[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. 4

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 7

[20] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 7

[21] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 7

[22] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 7

[23] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 7

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7

[26] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *arXiv preprint arXiv:2305.07017*, 2023. 7

A. Appendix.

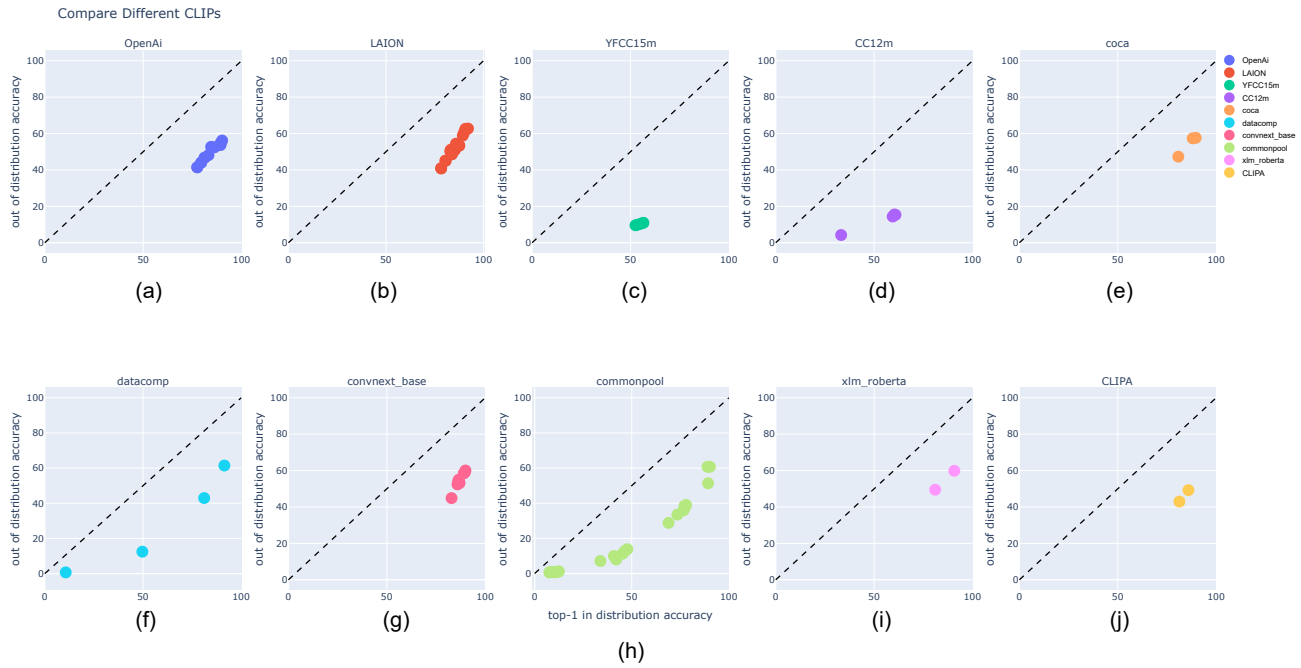


Figure 4. Evaluation OoD generalization of different CLIP models trained using various datasets. The evaluation involved testing these models on both in-distribution and out-of-distribution test sets.

Figure 4 shows the performance of different models on the our benchmark. The models are trained on different datasets or have special backbone, as follows:

- Figure 4.a:shows the performance of CLIP models trained on the OpenAI dataset [4].
- Figure 4.b: shows the performance of CLIP models trained on the LAION dataset [19] with 400 million or 2 billion image-text pairs.
- Figure 4.c: shows the performance of CLIP models trained on Yahoo-Flickr Creative Commons dataset with 15 million image-text pairs [20].
- Figure 4.d: shows the performance of models trained on CC12M dataset [21] with 12 million image-text pairs.
- Figure 4.e: shows the performance of the CoCa model [22] trained on the LAION dataset.
- Figure 4.f: shows the performance of CLIP models trained on the Datacomp dataset [23].
- Figure 4.g: shows the performance of ConvNeXt CLIP models [24] trained on the LAION dataset.
- Figure 4.h:shows the performance of CLIP models trained on the Common Pool dataset [23].
- Figure 4.i: shows the performance of CLIP models with a Roberta encoder [25] trained on the LAION dataset .
- Figure 4.j: shows the performance of CLIP models introduced by [26] trained on the LAION dataset.

A.1. Evaluation CLIP on Imagenet objects

Given that Imagenet-AO includes Imagenet objects, we conducted a comprehensive evaluation of the clip model on Imagenet objects. The results from this evaluation closely mirror the initial findings, with no significant deviations observed in the models' performance.

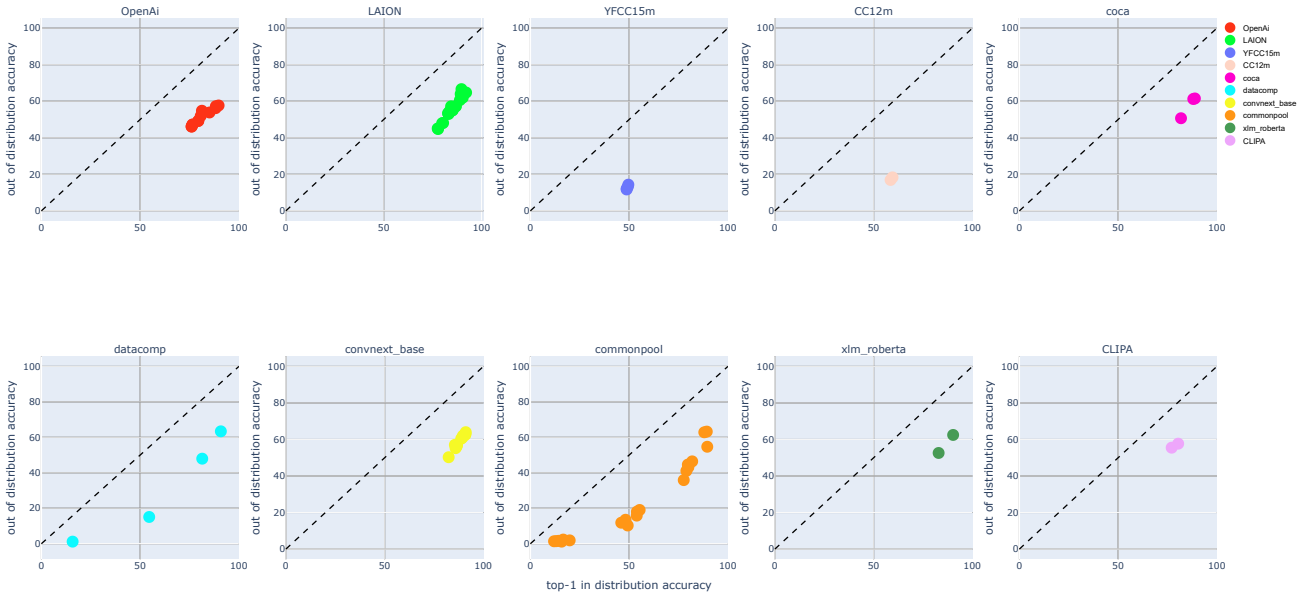
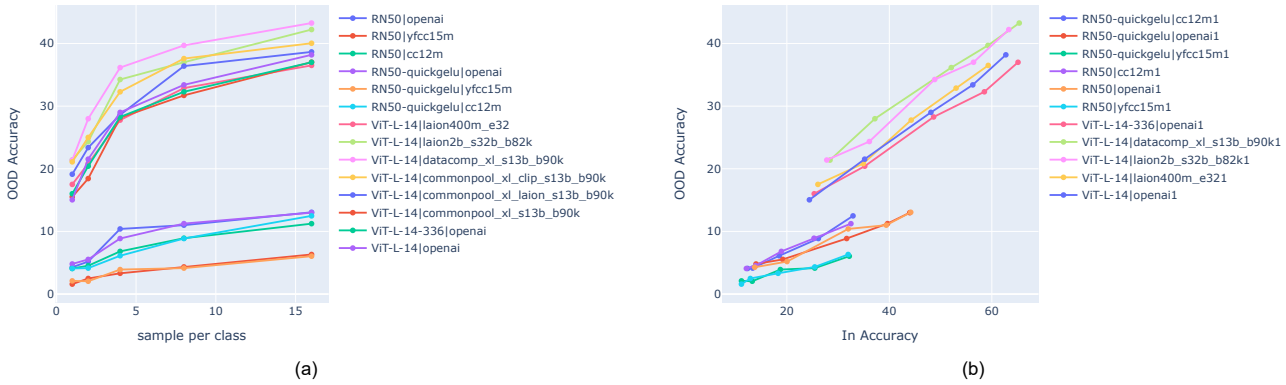


Figure 5. Evaluation of the CLIP models on Imagenet objects



(a) (b)

Figure 6. Comparison of OOD Accuracy in Various Few-Shot Settings for Different CLIP Models. This plot illustrates the out-of-distribution (OOD) accuracy performance across diverse few-shot scenarios, with the x-axis representing the number of samples used for fine-tuning, and the y-axis depicting the OOD accuracy.

A.2. Few-shot Evaluation

In this section, we conduct a few-shot evaluation of various CLIP models on Imagenet-AO. The objective is to fine-tune these models using 1, 2, 4, 8, and 16 samples per class and subsequently assess their performance. Few-shot learning is a critical aspect of CLIP's capabilities, as it allows the model to generalize effectively with limited training examples. During our few-shot evaluation, as shown in 6 we observed a noteworthy trend where models utilizing the Vision Transformer (ViT) image encoder exhibited higher performance compared to other configurations

A.3. Full finetune Evaluation

In this section, we present the results of our Full Fine-tune Evaluation, wherein we assess the performance of CLIP models with image encoders fine-tuned on the ImageNet dataset.

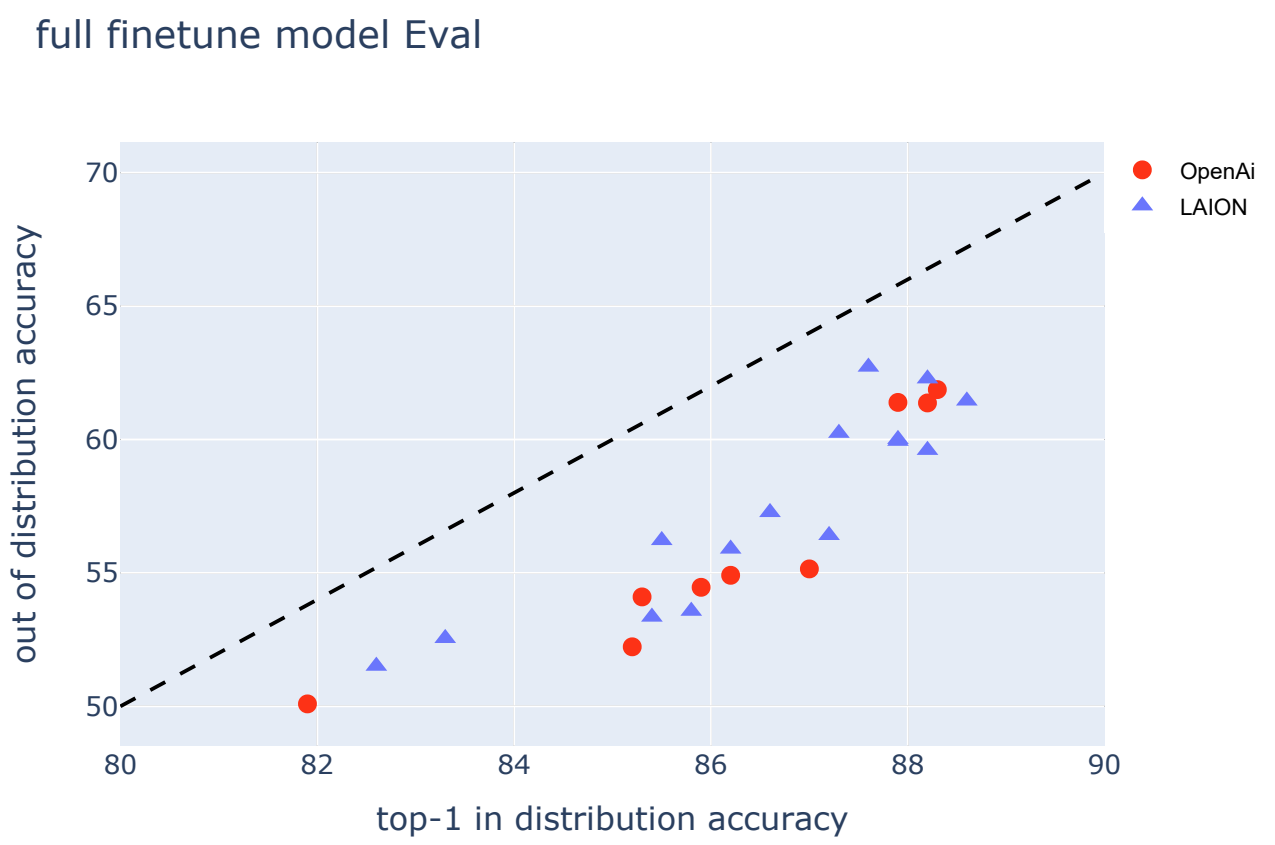


Figure 7. OOD Accuracy vs. ID Accuracy for Different CLIP Models Fine-Tuned on ImageNet.

A.4. Domain Shift

In this section, we delve into the Evaluation on Different Domain Shift, where we rigorously assess the performance of various CLIP models across distinct types of ImageNet datasets. Specifically, we evaluate the models on ImageNet-A, ImageNet-R, ImageNet-Sketch, as well as Imagenet-AO. Each of these datasets introduces specific domain shifts and challenges that differ from the standard ImageNet distribution. The result show in 1.

A.5. text-to-image Retrieval

In this section, we delve into the Text to Image Retrieval task and present a thorough evaluation of various CLIP models on our dataset. The objective of this evaluation is to examine how effectively each CLIP variant can retrieve relevant images based on textual queries, showcasing their ability to bridge the modal gap between language and vision. The result show in 2.

Table 1. Models performance on various datasets

Model	**ImageNet**	**ImageNet-v2**	**Imagenet-sketch**	**ImageNet-R**	**ImageNet-A**	**Imagenet-AO**
vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k	88.6	80.11	65.31	66.44	75.013	61.45
vit_large_patch14_clip_336.openai_ft_in12k_in1k	88.3	80.33	63.79	65.64	77.64	61.87
vit_huge_patch14_clip_224.laion2b_ft_in12k_in1k	88.2	79.24	65.77	66.56	69.91	62.28
vit_large_patch14_clip_336.laion2b_ft_in12k_in1k	88.2	78.87	59.74	59.74	68.84	59.6
vit_large_patch14_clip_224.openai_ft_in12k_in1k	88.2	79.07	61.83	61.4	71.12	61.37
vit_large_patch14_clip_224.openai_ft_in1k	87.9	79.26	62.52	63.47	70.85	61.39
vit_large_patch14_clip_336.laion2b_ft_in1k	87.9	78.35	63.3	63.91	61.7	59.94
vit_huge_patch14_clip_224.laion2b_ft_in1k	87.6	79.06	67.94	68.05	64.76	62.72
vit_large_patch14_clip_224.laion2b_ft_in1k	87.3	77.16	63.49	63.08	52.36	60.24
vit_base_patch16_clip_384.laion2b_ft_in12k_in1k	87.2	77.77	53.09	49.45	58.48	56.41
vit_base_patch16_clip_384.openai_ft_in12k_in1k	87	77.32	50.54	48.28	57.76	55.15
vit_base_patch16_clip_384.laion2b_ft_in1k	86.6	77.51	56.42	53.03	54.09	57.27
vit_base_patch16_clip_384.openai_ft_in1k	86.2	76.44	52.52	49.8	54.26	54.91
vit_base_patch16_clip_224.laion2b_ft_in12k_in1k	86.2	75.53	52.09	49.17	46.88	55.9
vit_base_patch16_clip_224.openai_ft_in12k_in1k	85.9	74.79	49.51	46.91	46.66	54.46
vit_base_patch32_clip_448.laion2b_ft_in12k_in1k	85.8	75.55	47.74	44.78	50.92	53.57
vit_base_patch16_clip_224.laion2b_ft_in1k	85.5	74.92	55.53	52.03	40.74	56.22
vit_base_patch32_clip_384.laion2b_ft_in12k_in1k	85.4	75.08	48.36	45.29	46.61	53.35
vit_base_patch16_clip_224.openai_ft_in1k	85.3	74.43	51.53	48.47	43.54	54.1
vit_base_patch32_clip_384.openai_ft_in12k_in1k	85.2	74.22	45.96	42.92	42.413	52.23
vit_base_patch32_clip_224.laion2b_ft_in12k_in1k	83.3	70.36	46.8	42.12	28.58	52.55
vit_base_patch32_clip_224.laion2b_ft_in1k	82.6	69.26	49.52	43.9	21.81	51.51
vit_base_patch32_clip_224.openai_ft_in1k	81.9	68.5	44.82	40.04	20.6	50.09

Table 2. Models performance on text-to-image Retrieval task

Model	R@1	R@5	R@10
RN50_openai	0.1628	0.4022	0.5318
RN50_yfcc15m	0.0359	0.0995	0.1484
RN50_cc12m	0.0627	0.1823	0.2673
RN50-quickgelu_openai	0.1628	0.4022	0.5318
RN50-quickgelu_yfcc15m	0.0394	0.1076	0.1569
RN50-quickgelu_cc12m	0.0687	0.1918	0.2774
RN101_openai	0.1856	0.4349	0.5708
RN101_yfcc15m	0.0404	0.1170	0.1670
RN101-quickgelu_openai	0.1856	0.4349	0.5708
RN101-quickgelu_yfcc15m	0.0431	0.1233	0.1767
ViT-B-32_openai	0.2011	0.4674	0.6020
ViT-B-32_laion400m_e31	0.2161	0.4818	0.6109
ViT-B-32_laion400m_e32	0.2158	0.4803	0.6097
ViT-B-32_laion2b_e16	0.2748	0.5700	0.7058
ViT-B-32_laion2b_s34b_b79k	0.2849	0.5751	0.7107
ViT-B-32_datacomp_m_s128m_b4k	0.0587	0.1612	0.2321
ViT-B-32_datacomp_s_s13m_b4k	0.0033	0.0122	0.0206
ViT-B-32-quickgelu_openai	0.2011	0.4674	0.6020
ViT-B-32-quickgelu_laion400m_e31	0.2437	0.5136	0.6468
ViT-B-32-quickgelu_laion400m_e32	0.2416	0.5158	0.6474
ViT-B-16_openai	0.2313	0.5105	0.6533
ViT-B-16_laion400m_e31	0.2727	0.5654	0.6943
ViT-B-16_laion400m_e32	0.2754	0.5637	0.6921
ViT-B-16_laion2b_s32b_b82k	0.3006	0.5964	0.7242
ViT-B-16_datacomp_l_s1b_b8k	0.2393	0.5218	0.6520
ViT-B-16_commonpool_l_clip_s1b_b8k	0.2189	0.4822	0.6154
ViT-B-16_commonpool_l_laion_s1b_b8k	0.2059	0.4582	0.5891
ViT-B-16_commonpool_l_image_s1b_b8k	0.1847	0.4265	0.5591
ViT-B-16_commonpool_l_text_s1b_b8k	0.1904	0.4278	0.5590
ViT-B-16_commonpool_l_basic_s1b_b8k	0.1683	0.3932	0.5173
ViT-B-16_commonpool_l_s1b_b8k	0.1307	0.3173	0.4264
ViT-L-14_openai	0.2818	0.5864	0.7243
ViT-L-14_laion400m_e31	0.3305	0.6298	0.7548
ViT-L-14_laion400m_e32	0.3310	0.6304	0.7543
ViT-L-14_laion2b_s32b_b79k	0.3790	0.6980	0.8108
ViT-L-14_datacomp_xl_s13b_b90k	0.4010	0.7028	0.8166
ViT-L-14_commonpool_xl_clip_s13b_b90k	0.3897	0.7004	0.8154
ViT-L-14_commonpool_xl_laion_s13b_b90k	0.3657	0.6777	0.7994
ViT-L-14_commonpool_xl_s13b_b90k	0.2775	0.5752	0.7154
ViT-H-14_laion2b_s32b_b79k	0.3659	0.6652	0.7785
ViT-g-14_laion2b_s12b_b42k	0.3628	0.6562	0.7720
ViT-g-14_laion2b_s34b_b88k	0.3653	0.6542	0.7726
ViT-bigG-14_laion2b_s39b_b160k	0.3757	0.6711	0.7893
coca_ViT-B-32_laion2b_s13b_b90k	0.2628	0.5498	0.6874
coca_ViT-B-32_mscoco_finetuned_laion2b_s13b_b90k	0.0006	0.0019	0.0040
coca_ViT-L-14_laion2b_s13b_b90k	0.3362	0.6381	0.7613
coca_ViT-L-14_mscoco_finetuned_laion2b_s13b_b90k	0.3626	0.6652	0.7823

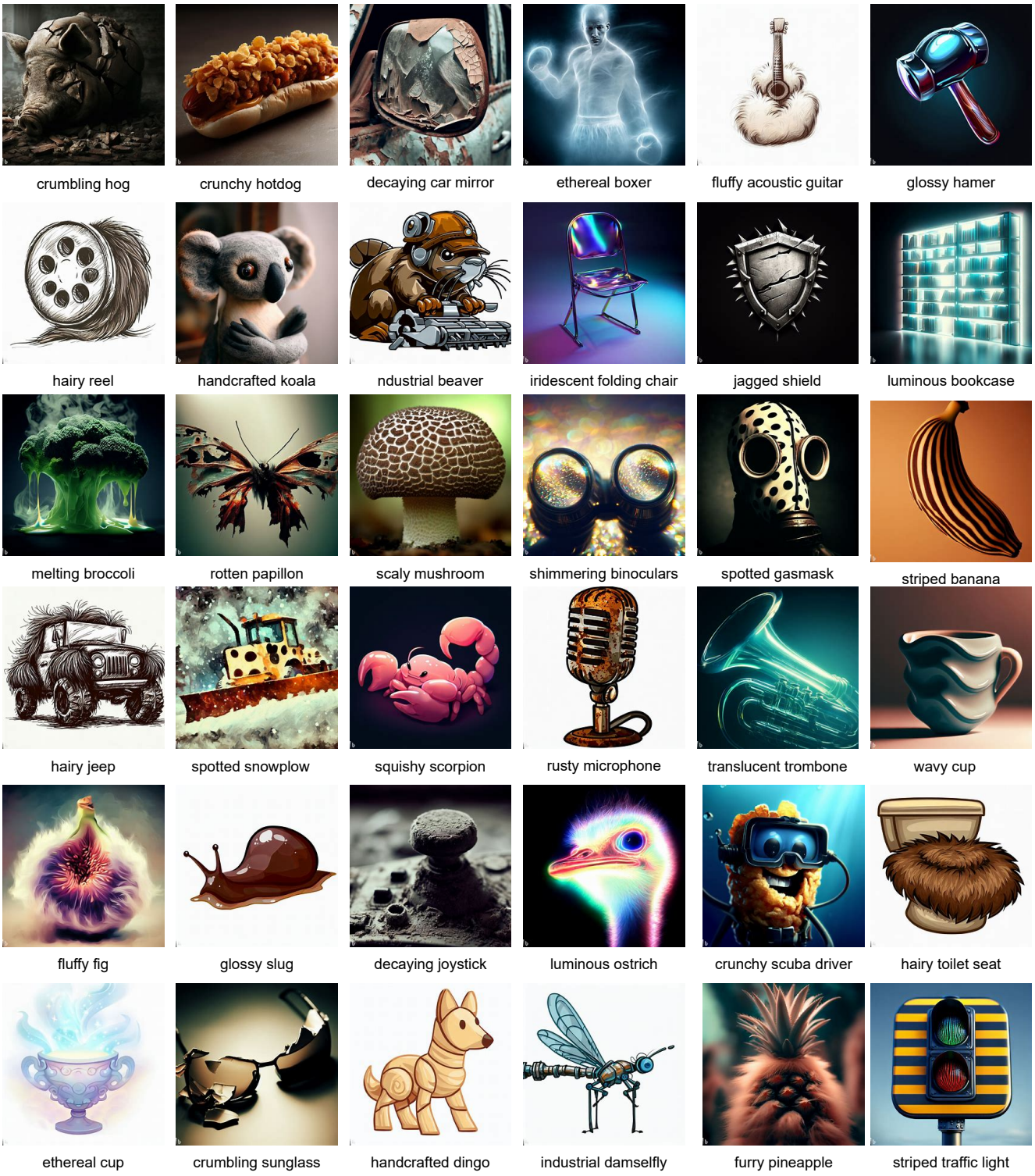


Figure 8. Examples of images from Imagenet-AO dataset.