Pretraining Frequency Predicts Compositional Generalization of CLIP on Real-World Tasks

Thaddäus Wiedemer [1,2,3*] **Yash Sharma** [1,2,3*] **Ameya Prabhu** [2,3]

Matthias Bethge [2,3,4] **Wieland Brendel** [1,2,4]

[1] Max-Planck-Institute for Intelligent Systems [2] Tübingen AI Center
[3] University of Tübingen [4] ELLIS Institute Tübingen

thaddaeus.wiedemer@gmail.com, ysharma1126@gmail.com

Abstract

We investigate the success conditions for compositional generalization of CLIP models on real-world data through performance prediction. Prior work shows that CLIP requires exponentially more pretraining data for linear performance gains on individual concepts. This sample-inefficient scaling could be mitigated if CLIP systematically understood new inputs as compositions of learned components, allowing rare observation to be mapped to common concepts. To explore CLIP's compositional generalization ability, we filter retrieval corpora for samples with object combinations not present in the pretraining corpus. We show that CLIP's performance on these samples can be accurately predicted from the pretraining frequencies of individual objects. Our findings demonstrate that CLIP learns to disentangle objects observed in its pretraining data and can recompose them straightforwardly. Additionally, we are the first to show how this ability scales with pretraining data. For data curation in practice, our results suggest that balancing object occurrences improves generalization, which should benefit CLIP's efficiency and accuracy without scaling data volume.

1 Introduction

Vision Language Models (VLMs) like CLIP [28] have seen widespread adoption for downstream tasks like classification, image retrieval, and image generation due to their transferability and impressive zero-shot performance. However, CLIP models are data-hungry, requiring exponentially more pretraining data for linear performance gains on downstream samples [36]. Similar sample-inefficient scaling has been reported for Large Language Models (LLMs) [15, 2], raising doubts about the feasibility of improving zero-shot performance of foundation models by scale alone.

A systematic way to overcome inefficient scaling is thought to be compositional generalization—the ability to understand and form novel combinations of learned concepts [6, 12, 38]. A model that can generalize in this way should more effectively combine learned concepts to understand new inputs, ultimately leading to increased zero-shot performance. Yet, the compositional abilities of large VLMs, such as CLIP, remain poorly understood: Existing studies in the visual domain are either theoretical, operate on synthetic data, or fail to verify whether compositions used for evaluation are truly novel given the pretraining data (see Sec. 2). Moreover, *the relationship between a VLM's compositional abilities and its pretraining data is entirely uncharacterized*.

*Equal contribution.

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

To address this gap, we aim to investigate CLIP's compositional generalization on real-world data as a function of its pretraining corpus. Specifically, we make the following contributions:

- We leverage the scalable concept-extraction pipeline proposed by Udandarao et al. [36] to curate text-to-image (T2I) and image-to-text (I2T) retrieval test sets for compositional generalization (Sec. 3.1). Each test set is created for the pretraining corpus of the tested model such that it contains only samples with novel combinations of known object classes.
- We show that CLIP models perform consistently well on our curated test sets, regardless of the architecture or scale of the backbone.
- We demonstrate that across architectures, parameter counts, and pretraining data scales, CLIP's ability to compose objects can be accurately predicted from the independent pretraining frequencies of each object in the composition (Sec. 3.3).

Our results are the first to establish a firm connection between the compositional generalization of a VLM and its pretraining data. The nature of this connection shows that CLIP obtains an independent understanding of object classes from web-scale data.

2 Related Work

Theoretical Works & Synthetic Data A growing body of works [24, 23, 30, 19, 38, 39, 27, 14] provides significant theoretical understanding of compositional generalization results in the vision domain. Similar works exist for compositionally in language [6, 12, 3], often under the more specific term *systematicity* [6, 12, 3]. In the language domain, promising progress has been made [18], but results in both domains nonetheless remain confined to synthetic datasets [17, 16]l; Sun et al. [33] actively questions the transferability of insights to real-world data. In contrast, our work analyzes compositional generalization using real-world retrieval datasets.

VLM Benchmarks & Contamination Several compositionality benchmarks have been proposed for VLMs [35, 19, 43, 42, 21, 10, 29, 37, 1]. However, these studies do not consider the overlap of concept combinations with web-scale pretraining data. Data contamination of this kind has been shown to significantly impact CLIP's zero-shot performance [22], making it difficult to distinguish between genuine generalization and mere memorization. A notable exception is Abbasi et al. [1], who generate test images of novel attribute-object pairs, but as a result, their benchmark resorts to synthetic data. Our work controls for data contamination by only considering combinations that do not occur in the pretraining data but do occur in real-world benchmarks.

3 Predicting Compositional Generalization from Pretraining Frequency

We adapt the pipeline from Udandarao et al. [36] in two steps to study the success conditions for compositional generalization. First, we use it to curate retrieval test sets that contain novel combinations of objects with respect to a pretraining set (Sec. 3.1). Second, we propose a simple modification to predict downstream performance in terms of samples rather than concepts (Sec. 3.2). Finally, we evaluate CLIP models with varying architectures, parameter counts, and pretraining data scales and show consistent scaling behavior (Sec. 3.3).

3.1 Curating Compositional Generalization Test Sets

We consider two standard retrieval datasets: Flickr-1K [41] and COCO-5K [20]. Both can be used for benchmarking text-to-image (T2I) or image-to-text (I2T) retrieval.

To measure compositional generalization, we follow Hupkes et al. [12] and retain only test samples containing multiple concepts $o_1, ..., o_n$, where

(i) the model has been familiarized with each concept o_i only in the absence of $o_{j \neq i}$,

(ii) the combination $o_1, ..., o_n$ is plausible.

Udandarao et al. [36] compile a list of 945 nouns in the text captions for these 2 retrieval datasets as possible concepts. The presence of a concept in a pretraining sample is established if it is part of the

caption (after lemmatization) *and* it is found in the image using RAM++ [11]. Consequently, our analysis of concepts is limited to tangible objects; more abstract concepts like actions or stylistic information are harder to annotate in the visual domain and can, therefore, not be reliably quantified using this pipeline. With this setup, the frequency $f_{\mathcal{D}}(o)$ of object class o in a pretraining corpus $\mathcal{D}$ simply counts the number of samples it occurs in.[1]

To address (i), we first consider how often objects $o_1, ..., o_n$ in a test sample x co-occur in the pretraining dataset. We call this quantity the co-occurrence frequency, formally given by

$$f_{\cap,\mathcal{D}}(x) = \|\{d \in \mathcal{D} \mid \text{for all } o \in x : o \in d\}\|\,. \tag{1}$$

Condition (i) above is then satisfied by test samples x which contain a novel combination of objects, i.e., $f_{\cap,\mathcal{D}}(x) = 0$, but each object has been observed at least once, i.e., $f_{\mathcal{D}} > 0$ for all $o \in x$.

Condition (ii) is hard to verify in general but is trivially met here since we filter real-world data.

Since all frequencies are dependent on the pretraining corpus, the size of our compositional generalization test sets differ. The number of samples in each test set (total and percentage) is shown in Figs. 1 and 2.

3.2 Per-Sample Prediction

Metrics We measure performance on each sample using Recall@k for $k \in \{1, 5, 10\}$ following Radford et al. [28]. Figs. 1 and 2 show results for Recall@10, other results are listed in App. A.

Sample Frequency We compute the average pretraining frequency f_{avg} of each test sample x as the geometric mean of the frequencies of the objects $o_1, ..., o_n$ in the sample x, i.e.,

$$f_{\text{avg}}(x) = \left(\prod_{o \in x}^{n} f(o)\right)^{\frac{1}{n}}. \tag{2}$$

The choice of geometric mean is motivated by the assumption that the model's performance on a combination of objects depends on the quality of the model's independent understanding of each object in the combination. For example, a simple retrieval engine might find samples containing two objects o_1, o_2 without considering their interaction by first finding samples containing o_1 and then filtering the results for samples containing o_2. In this case, the probability of retrieving a correct sample based on the prompt $\mathcal{P}$ can be written as

$$P(y = 1 | o_1, o_2 \in \mathcal{P}) = P(y = 1 | o_1 \in \mathcal{P}) P(y = 1 | o_2 \in \mathcal{P}). \tag{3}$$

The geometric mean reflects the multiplicative impact of each object on the whole [27].

Fitting a Predictor Our evaluation yields (y, f_{avg}) for each test sample, where $y = 1$ (0) indicates correct (wrong) retrieval. For each test set, we drop noisy outliers via IQR-removal on f_{avg}, following Kandpal et al. [15], Udandarao et al. [36]. We fit a logistic regression model with bootstrapping to predict $P(y = 1)$ given f_{avg}.

3.3 Results

Fig. 1 collects results for the T2I task, Fig. 2 for I2T. More results can be found in App. A.

Models We test CLIP models [28] with both ResNet [9] and Vision Transformer [5] architectures. Specifically, we evaluate ViT-B-16 [25] and RN50 [7, 26] trained on CC-3M [32] and CC-12M [4]; ViT-B-16, RN50, and RN101 [13] trained on YFCC-15M [34]; ViT-B-16, ViT-B-32, and ViT-L-14 trained on LAION400M [31]; and ViT-B-16 trained on SynthCI-30M [8]. We follow `open_clip` [13], `slip` [25] and `cyclip` [7] for implementation details.

Fraction of Compositional Generalization Samples Looking at the histogram percentages, $17 - 44\,\%$ of samples are unknown compositions of known concepts for the T2I task. For I2T, the fraction is much higher with $89 - 98\,\%$. The discrepancy stems from many images containing background objects that are inconsistently reflected in their captions. We also find that the fraction generally decreases with the size of the pretraining set.

[1] We use the term frequency instead of count since we only compare quantities for a given, fixed-size pretraining set, in which case normalization can be omitted.

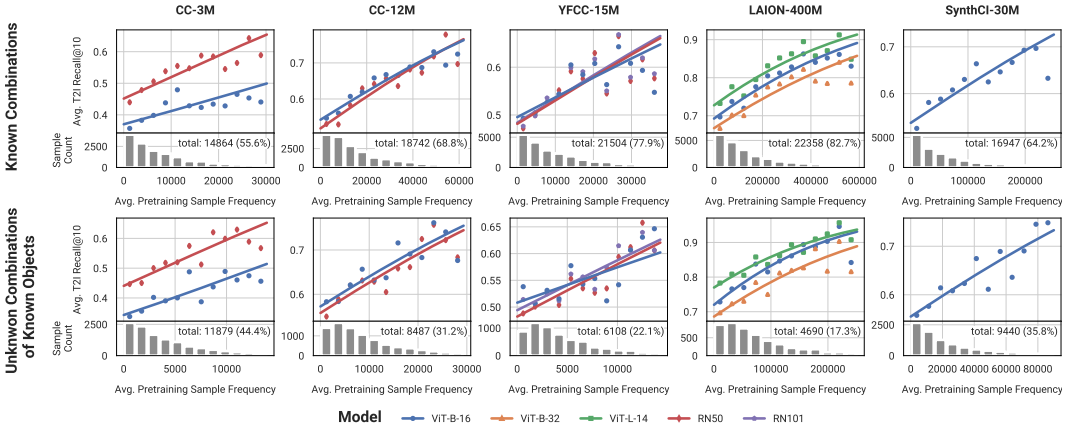

Figure 1: **T2I Recall@10**. CLIP's performance on unknown combinations (bottom) matches that on known combinations (top) and can be consistently predicted as a linear function of the average pretraining frequency of the constituent objects. All regression fits are significant at $p < 0.01$.

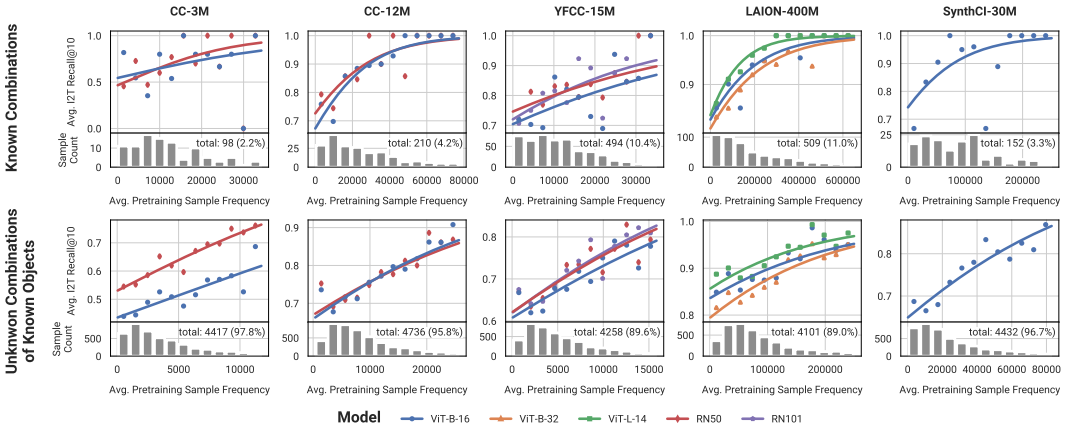

Figure 2: **I2T Recall@10**. CLIP's performance on unknown combinations (bottom) almost matches that on known combinations (top) and can be consistently predicted as a linear function of the average pretraining frequency of the constituent objects. All regression fits are significant at $p < 0.01$.

Overall Performance We find that CLIP's performance on the T2I task does not differ greatly between samples with known combinations (top row) and samples with novel combinations of known concepts (bottom row). The difference is slightly more pronounced on the I2T task, but performance is still high overall, indicating that CLIP generalizes well to novel object compositions.

Predicting Compositional Generalization We show a clear and consistent relationship between the average pretraining sample frequency f_{avg} and CLIP's retrieval performance, even on samples requiring compositional generalization. The relation is approximately linear, except for the best-performing models, where it flattens as retrieval recall approaches 1. Since the contribution of each object's pretraining frequency to the average pretraining sample frequency f_{avg} is multiplicative, this consistent relationship implies that underrepresented objects are the bottleneck for compositional generalization.

Control on Synthetic Data SynthCI-30M [8] consists of synthetically generated images designed to cover a diverse combination of concepts. Due to this process, we treat SynthCI-30 as a control to see if our results hold for a pretraining dataset sourced differently. We find that more test set combinations are unseen in SynthCI-30 than the pretraining sets derived from the real world, but the scaling of compositional generalization observed on real-world pretraining corpora also holds here.

Taken together, our results show that CLIP generalizes successfully to novel combinations of objects if it has observed the constituents sufficiently often during pretraining. Note that objects in the pretraining data do not occur independently. In fact, many training samples contain multiple objects, and some object classes co-occur much more frequently than others. The consistent scaling of CLIP's compositional generalization implies that the model can nonetheless disentangle objects and obtain an independent understanding of each object class.

For practitioners, our findings underline the importance of balancing object occurrences during data curation, as generalization is bottlenecked by the occurrence of each object.

4 Next Steps

Model Selection While we control for architecture, parameter count, and pretraining scale, our experiments could be extended to other CLIP variants [42], diffusion models, or even LLMs.

Type of Compositionality We only consider object compositions. We expect that our results may extend to attribute-object, foreground-background, texture-shape compositions in single-object scenes, since the independence assumption from Sec. 3.2 approximately holds. The bottleneck for these experiments is the concept-extraction pipeline. On the other hand, the scaling behavior of complex compositions, like attribute-binding with multiple objects, may not be as readily predictable.

Composition Granularity Our definition of the co-occurence frequency in Eq. 1 only considers whether *all* objects have jointly been observed during pretraining. For samples containing more than two objects, it might also be interesting to consider pair-wise object co-occurence and other partial co-occurences. How to integrate this information in the selection of test samples for compositional generalization remains is an open question.

5 Conclusion

Identifying conditions for successful real-world compositional generalization is a first step towards a future where models can be relied upon to generate new ideas, as *"an idea is nothing more nor less than a new combination of old elements"* [40]. The ability to forecast when such capabilities will be unlocked is valuable not only to understand the compositional abilities of existing models but also to guide the development and scaling of future methods. We take a first step in this direction by demonstrating how CLIP's ability to disentangle and recompose objects scales with the the frequency with which each object has been observed in the pretraining data.

Acknowledgements

We thank Vishaal Udandarao for helpful discussions and clarifying details regarding his work that we build our analysis on. We thank Rylan Schaeffer for helpful feedback on this manuscript. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1 and via the Open Philanthropy Foundation funded by the Good Ventures Foundation. WB is a Machine Learning Cluster of Excellence member, EXC number 2064/1 – Project number 390727645. This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting TW.

Author Contributions

The project was jointly led and coordinated by TW and YS. YS collected the data with input from TW and AP. TW and YS did the final analysis with input from AP. WB and MB participated in several helpful discussions during the project. TW, YS, and AP wrote the manuscript; TW made the figure with contributions from YS and AP.

References

[1] Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models. In *European Conference on Computer Vision (ECCV)*, 2024.

[2] Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*, 2024.

[3] Ian Berlot-Attwell, A Michael Carrell, Kumar Krishna Agrawal, Yash Sharma, and Naomi Saphra. Attribute diversity determines the systematicity gap in vqa. *arXiv preprint arXiv:2311.08695*, 2023.

[4] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[7] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719, 2022.

[8] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023.

[11] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310, 2023.

[12] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

[13] 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. URL `https://doi.org/10.5281/zenodo.5143773`.

[14] Whie Jung, Jaehoon Yoo, Sungjin Ahn, and Seunghoon Hong. Learning to compose: Improving object centric learning by injecting compositionality. *arXiv preprint arXiv:2405.00646*, 2024.

[15] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning (ICML)*, pages 15696–15707. PMLR, 2023.

[16] Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, 2020.

[17] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018.

[18] Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

[19] 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. *European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[21] 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.

[22] Prasanna Mayilvahanan, Roland S Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[23] Milton L. Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir JH Ludwig, and Gaurav Malhotra. Lost in Latent Space: Examining failures of disentangled models at combinatorial generalisation. In *Advances in Neural Information Processing Systems*, October 2022.

[24] 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 (ICLR)*, 2021.

[25] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision (ECCV)*, 2022.

[26] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022.

[27] Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. June 2023.

[28] 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 (ICML)*, 2021.

[29] 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.

[30] Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022.

[31] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[33] Kaiser Sun, Adina Williams, and Dieuwke Hupkes. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.19. URL https://aclanthology.org/2023.conll-1.19.

[34] 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.

[35] 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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance, 2024. URL https://arxiv.org/abs/2404.04125.

[37] Haoxiang Wang, Haozhe Si, Huajie Shao, and Han Zhao. Enhancing Compositional Generalization via Compositional Feature Alignment, February 2024.

[38] Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 6941–6960. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/15f6a10899f557ce53fe39939af6f930-Paper-Conference.pdf.

[39] Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7VPTUWkiDQ.

[40] James Webb Young and Keith Reinhard. *A technique for producing ideas*. NTC Business Books, 1975.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[42] 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. URL https://openreview.net/forum?id=KRLUvxh8uaX.

[43] 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. *arXiv preprint arXiv:2207.00221*, 2022.

A Additional Evaluations

Figs. 3 to 6 plot trends seen in Figs. 1 and 2 for Recall@5 and Recall@1. We observe similar trends.

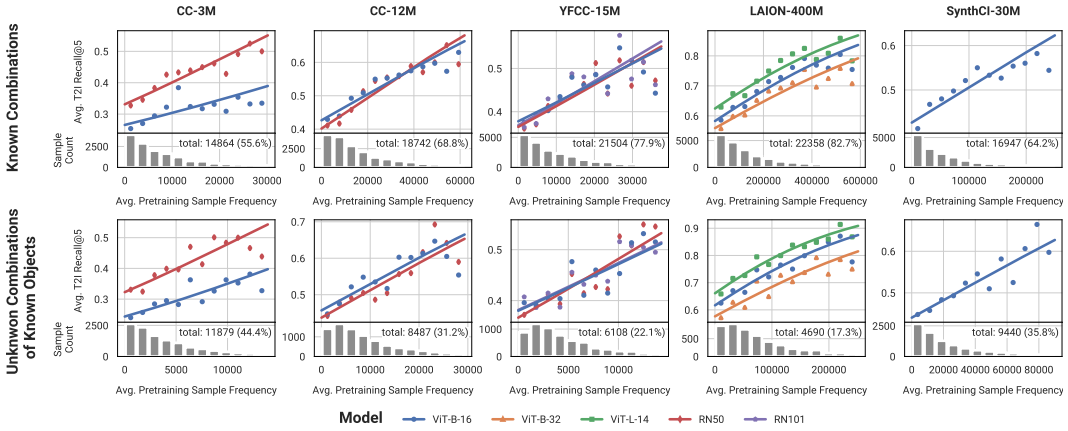

Figure 3: **T2I Recall@5** We see that on combinations that are both known and unknown to the model, across architectures and pretraining sets, there exists a predictive relationship between the sample frequency, i.e. the aggregated frequencies of objects in the combination, and the performance.

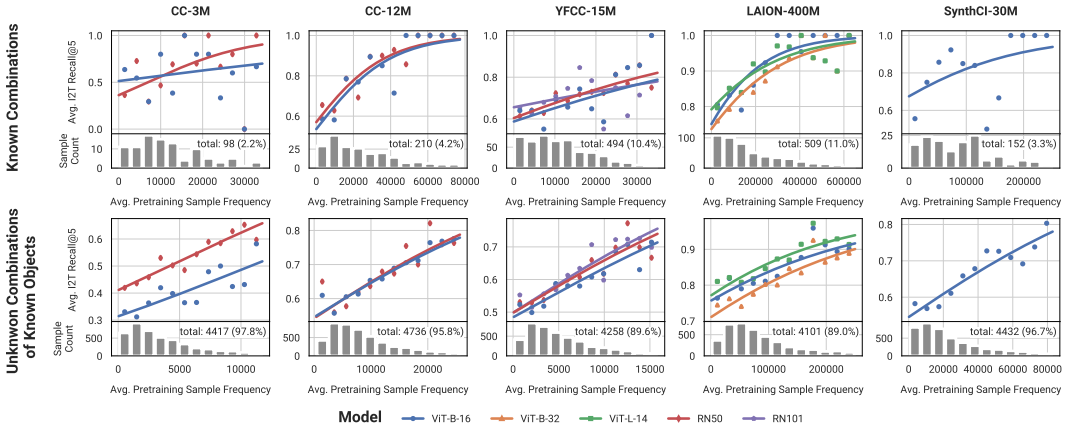

Figure 4: **I2T Recall@5** We see that on combinations that are both known and unknown to the model, across architectures and pretraining sets, there exists a predictive relationship between the sample frequency, i.e. the aggregated frequencies of objects in the combination, and the performance.

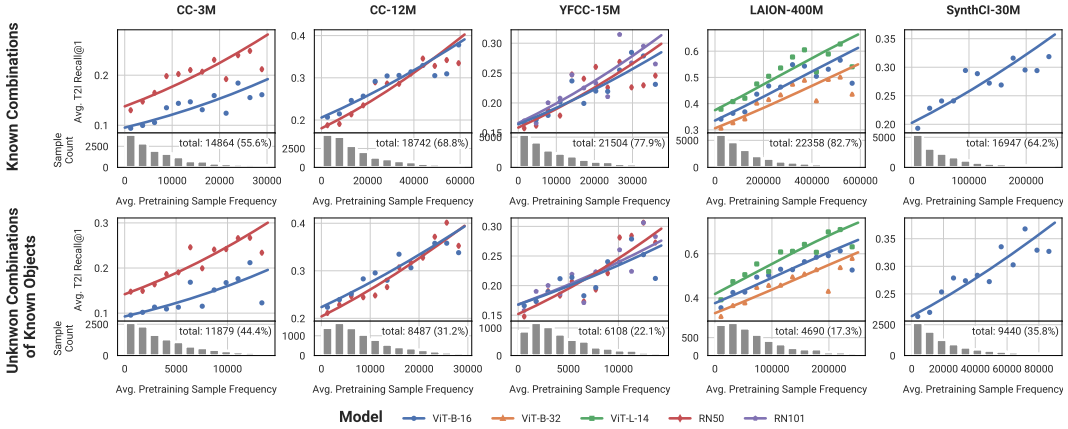

Figure 5: **T2I Recall@1** We see that on combinations that are both known and unknown to the model, across architectures and pretraining sets, there exists a predictive relationship between the sample frequency, i.e. the aggregated frequencies of objects in the combination, and the performance.

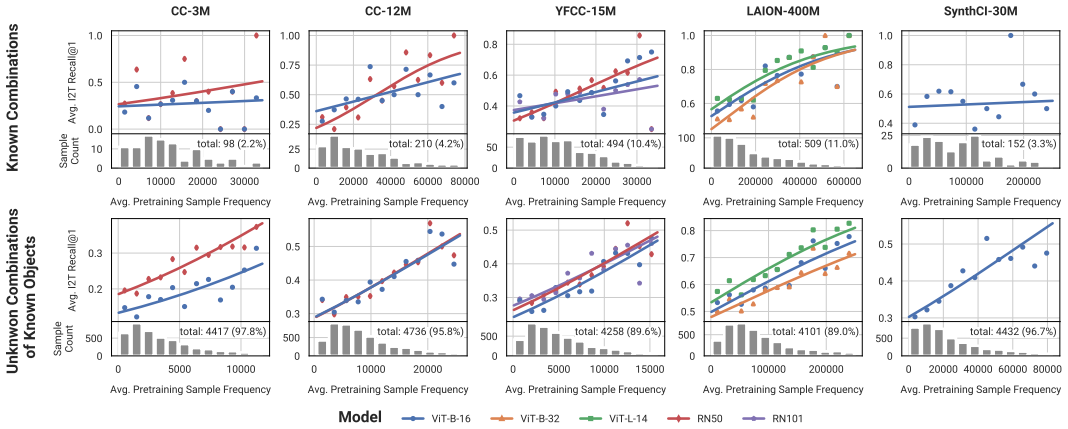

Figure 6: **I2T Recall@1** We see that on combinations that are both known and unknown to the model, across architectures and pretraining sets, there exists a predictive relationship between the sample frequency, i.e. the aggregated frequencies of objects in the combination, and the performance.