Object-Centric Representations Generalize Better Compositionally with Less Compute

Ferdinand Kapl[1,2] **Amir Mohammad Karimi Mamaghan**[3] **Max Horn** **Carsten Marr**[4]
Stefan Bauer[1,2] **Andrea Dittadi**[1,2,5]

[1]Helmholtz AI [2]Technical University of Munich [3]KTH Royal Institute of Technology
[4]Institute of AI for Health, Computational Health Center, Helmholtz Munich
[5]MPI for Intelligent Systems, Tübingen

Abstract

Compositional generalization—the ability to reason about novel combinations of familiar concepts—is fundamental to human cognition and a critical challenge for machine learning. Object-Centric representation learning has been proposed as a promising approach for achieving this capability. However, systematic evaluation of these methods in visually complex settings remains limited. In this work, we introduce a benchmark to measure how well vision encoders, with and without object-centric biases, generalize to unseen combinations of object properties. Using CLEVRTex-style images, we create multiple training splits with partial coverage of object property combinations and generate question–answer pairs to assess compositional generalization on a held-out test set. We focus on comparing pretrained foundation models with object-centric models that incorporate such foundation models as backbones—a leading approach in this domain. To ensure a fair and comprehensive comparison, we carefully account for representation format differences. In this preliminary study, we use DINOv2 as the foundation model and DINOSAURv2 as its object-centric counterpart. We control for compute budget and differences in image representation sizes to ensure robustness. Our key findings reveal that object-centric approaches (1) converge faster on in-distribution data but underperform slightly when non-object-centric models are given a significant compute advantage, and (2) they exhibit superior compositional generalization, outperforming DINOv2 on unseen combinations of object properties while requiring approximately four to eight times less downstream compute.

1 Introduction

Object-centric learning has generated considerable interest due to its promise of enabling more compositional and generalizable representations (Greff et al., 2020; Locatello et al., 2020; Dittadi et al., 2022; Jiang et al., 2023; Brady et al., 2023). Despite these claims, the relationship between object-centric representations and compositionality remains largely untested in a systematic and principled manner. Compositionality itself has been explored in various domains, each with distinct definitions and evaluation protocols but sharing the core idea of recombining familiar components in novel ways. In text, it often refers to rearranging or recombining words and numbers (Lake & Baroni, 2018; Dziri et al., 2024). In images, it may involve combining known objects or recombining seen object properties to create novel objects (Kim et al., 2024; Haramati et al., 2024; Montero et al., 2024; Abbasi et al., 2024). Text-to-image generation tasks assess compositionality by increasing the complexity of visual combinations in a prompt (Wu et al., 2024; Li et al., 2024), while image-to-text tasks, such as Visual Question Answering, evaluate how well models can reason about new combinations of visual and linguistic elements (Hsieh et al., 2024; Ma et al., 2023; Yuksekgonul et al., 2022).

Given the aforementioned claims of object-centric learning as a potential solution and several preliminary indications (Yoon et al., 2023; Montero et al., 2024; Kim et al., 2024; Haramati et al., 2024), we now want to investigate them in greater depth. The flavor of compositionality we are most interested in is *object property composition* (Johnson et al., 2017; Abbasi et al., 2024; Montero et al., 2024; Kim et al., 2024), as it aligns closely with real-world scenarios and has been investigated the least. It is the ability of a model to generalize to novel combinations of previously seen object properties from a visual world. For instance, a model trained only on a red cube and a blue sphere should successfully handle a blue cube or a red sphere at test time.[1]

As this form of compositionality requires precise control over the factors of variation in the visual world, most works rely on synthetically generated images from a computer graphics tool (Kim et al., 2024; Montero et al., 2024) or a pretrained generative model (Abbasi et al., 2024). Although compositionality is often described as a core motivation for object-centric representations, its evaluation is typically limited to changing the number of objects (Johnson et al., 2017; Locatello et al., 2020; Karazija et al., 2021; Biza et al., 2023), a type of generalization known as *scene composition*. The works most similar to ours are Kim et al. (2024) and Montero et al. (2024), both investigating compositional generalization of object properties. However, Kim et al. (2024) use a generative formulation that only allows evaluation of generative models rather than general image representations and do not isolate which design choices contribute to better performance, while Montero et al. (2024) examine compositionality only under more limited settings.

In order to study the compositional generalization capabilities of visual representations for object property composition, we design our own benchmark. First, we use the generation pipeline from Kim et al. (2024) to generate CLEVRTex-style images by precisely defining the entire visual world. Specifically, we consider every combination of individual factors of variation—such as material, shape, and size—that characterize each object. Then, we allocate 80% of these object combinations to progressively smaller subsets, with the smallest containing only 10%. The remaining 20% constitutes the compositional generalization test dataset, ensuring that no test objects were encountered during training even though their individual properties (e.g., shape, material, and size) were observed many times. To evaluate this compositional generalization via VQA, we follow Mamaghan et al. (2024) by generating question–answer pairs for all images. This process results in five training datasets—CLEVRTex *"super easy"* (80%), *"easy"* (60%), *"medium"* (40%), *"hard"* (20%), and *"super hard"* (10%)—and one dataset dedicated to testing compositional generalization, called CLEVRTex *"COOD"*.

For our comparisons, we focus on pretrained foundation models and object-centric models that incorporate such foundation models as backbones, a leading approach in this domain. Specifically, we use DINOv2 (Oquab et al., 2023) as the foundation model and DINOSAURv2 (Seitzer et al., 2022) as its object-centric counterpart. To ensure a fair and comprehensive comparison, we account for differences in representation format by controlling for image representation sizes. We achieve this by employing a small cross-attention layer within the downstream VQA model to up- or downscale image representations as needed, ensuring that differences in compute allocation do not unfairly advantage one approach over another.

We then evaluate all models by training distinct downstream models on the VQA task under each training variant, testing on the respective in-distribution (ID) set as well as on the fixed compositional out-of-distribution (COOD) generalization set. Following the framework of Mamaghan et al. (2024), we vary the size of the downstream model and, additionally, the input size of the image representation. Finally, by carefully controlling the visual combinations that models are exposed to at train and test time, we can systematically adjust the hardness of the generalization task until even an oracle with access to ground-truth inputs struggles to generalize at test time.

Our main contributions can be summarized as follows:

- **Datasets**: We design our own compositional generalization benchmark based on the CLEVRTex dataset (Karazija et al., 2021; Kim et al., 2024; Mamaghan et al., 2024). To assess compositional generalization, we define a fixed held-out test set containing 20% of all object-property combinations, along with five progressively smaller subsets from the remaining combinations. The smaller the subset, the greater the challenge for generalization. The compositional test set ensures that no test objects appear during training, while all their individual properties—such as shape, material, and size—are encountered. Finally, we generate question–answer pairs for all images to evaluate all models on a VQA task.
- **Finding I**: For harder compositional generalization tasks, object-centric representations outperform DINOv2 at any compute budget, even when giving the downstream model for DINOv2 four to eight times more compute.

[1] This is a simplified version of the compositional generalization experiment in Johnson et al. (2017).

- **Finding II**: For the in-distribution setting, DINOv2 representations only begin to surpass the object-centric counterpart when the downstream model is given three times as much training compute. Even then, the improvement is less than 2% at the end of training with four times more compute.
- **Finding III**: In-distribution performance is very strongly correlated with compositional out-of-distribution performance. For easier generalizations, the relationship is almost perfectly linear, with only a small constant drop, but becomes sublinear as the tasks become more difficult.

2 Related Work: Compositionality

Compositionality has been defined and tested in a variety of ways. Below, we briefly summarize relevant approaches in text, images, text-to-image, and image-to-text settings.

Text. Lake & Baroni (2018) studied compositionality by training a model to decode natural-language commands into action sequences that feature novel combinations of concepts at test time. In a related line of work, Dziri et al. (2024) demonstrated that transformers can fail catastrophically on seemingly simple tasks (e.g., multi-digit integer multiplication) when test conditions differ slightly from training (e.g., more digits).

Images. Kim et al. (2024) explored compositionality without language annotations by constructing a visual world of objects with simple attributes (e.g., shape, texture). They controlled which portion of the combinatorial attribute space was shown during training and formulated a generative task where the model must learn and apply transformation rules (e.g., swapping shapes) to unseen combinations at test time. Haramati et al. (2024) probe, among other things, the compositional generalization of different components of their architecture in a reinforcement learning task that involves arranging objects in a specified way on a table with a robotic arm.

Text-to-image. Some recent work frames compositionality as a text-to-image generation task, prompting models with increasingly complex combinations of visual concepts to test that all mentioned concepts appear in the generated image (Wu et al., 2024; Li et al., 2024).

Image-to-text and VQA. The *SugarCrepe* benchmark evaluates compositional comprehension by presenting an image alongside a correct caption and a closely matched "hard negative", which can involve object swapping or replacement (Hsieh et al., 2024). The model must choose the caption that accurately describes the image, extending earlier approaches such as Ma et al. (2023).

Object-Centric Representations. In the context of reinforcement learning, Yoon et al. (2023) and Haramati et al. (2024) found that object-centric representations are mostly beneficial for tasks requiring relational reasoning with object interactions. Additionally, Haramati et al. (2024) also demonstrated that their agent can generalize compositionally to more objects than seen during training, both empirically and theoretically. Kim et al. (2024) provided some evidence that a slot-based State-Space Model improves compositional generalization, though the specific design elements driving this improvement remain unclear. Furthermore, Montero et al. (2024) show that a simple object-centric model reconstructs novel objects with hold-out ranges of properties (e.g., color or rotation) for a single object better than a non-object-centric alternative when the models have been trained on all combinations for the rest of the objects.

3 Problem Setup

3.1 Dataset Generation

Inspired by Kim et al. (2024), we create five CLEVRTex-style (Karazija et al., 2021) datasets labeled from super easy to super hard, each containing a progressively smaller share of all possible object types. Each object is defined by a triplet of properties (material, shape, size) chosen from 8, 8, and 3 possible values, respectively, yielding 192 unique object types. We render images using

Blender,[2] randomly selecting 3–6 allowed objects per scene in each dataset. Unlike Kim et al. (2024), who frame this as a generative task, we evaluate general visual representations via VQA instead. Specifically, we follow the question–answer generation approach of Johnson et al. (2017) and its CLEVRTex adaptation (Mamaghan et al., 2024), adding human-readable labels for each of the eight rendering materials (details in Appendix A). As a result, we obtain five training datasets—CLEVRTex *"super easy"* (80% of all possible objects), *"easy"* (60%), *"medium"* (40%), *"hard"* (20%), and *"super hard"* (10%)—each with 48k images (40k for training, 4k for validation, and 4k for in-distribution testing). A separate dataset, CLEVRTex *"COOD"*, consists of 4k images containing the remaining 20% of objects and is used consistently across all training variants to test compositional object property generalization. There are on average 42 question–answer pairs per image, resulting in roughly 170k QA pairs per test set and 1.7M QA pairs per training set.

3.2 Models and Evaluation

Starting from a strong pretrained vision model in DINOv2 (Oquab et al., 2023), we additionally pretrain an object-centric model for every CLEVRTex variant by reconstructing the self-supervised image features with a Slot-Attention bottleneck, corresponding to the DINOSAURv2 model (Seitzer et al., 2022; Didolkar et al., 2024). Architectural and hyperparameter details are in Appendix B.1. To train the downstream VQA model, we follow Mamaghan et al. (2024). Questions are encoded by a pretrained T5-base model (Raffel et al., 2020), and the answers as 28 distinct labels, which include "yes", "no", natural numbers up to the maximum number of objects, and all possible values of object properties. We feed the image features from DINOv2 or DINOSAURv2, together with the text embeddings into the downstream model (details in Appendix B.2).

DINOv2 and DINOSAURv2 produce image representations of different sizes, which strongly impacts the downstream model's FLOPs because the transformer encoder needs the biggest share of the compute and scales quadratically with sequence length (see Appendix B.3 for details). To enable a fair comparison, we introduce downstream model variants that include a single cross-attention layer immediately after the vision encoder output. This layer up- or downsizes the image representation to a target size, resulting in almost identical compute requirements for the DINOv2 and DINOSAURv2 variants, and is trained jointly with the downstream model. Other resizing methods were tested but proved less effective or introduced unnecessary complexity. For example, a cross-attention layer mapping to a target size of the same size as the input image representation, where we present some of the results in Section 4.1, but this did not result in significant and consistent improvements, so we excluded it from the detailed analysis in Section 4.2. More results are in Appendix C. We also experimented with replacing the slot-attention bottleneck in DINOSAURv2 with a similar cross-attention approach and substituting slot-attention with a cross-attention layer in end-to-end training. Both attempts yielded suboptimal results, suggesting that more careful consideration of architectural or hyperparameter choices is likely required.

Finally, to gauge dataset difficulty, we train two additional baselines: a lower question-only baseline using only the questions as inputs to the downstream model and a ground-truth oracle that supplies the true object properties of all visible objects in a scene as image representations for the downstream model. After training, each model is evaluated on its corresponding in-distribution test set and on CLEVRTex *"COOD"* to measure compositional generalization at every training checkpoint.

4 Experiments

In our experiments, we first show a strong correlation between in-distribution accuracy and compositional out-of-distribution performance. The relationship transitions from nearly linear to sublinear as generalization difficulty increases. Next, we factor in compute considerations by estimating FLOPs and find that DINOSAURv2, an object-centric model, achieves robust ID performance at significantly lower compute, whereas DINOv2 can surpass it only when granted substantially more resources. Finally, on harder compositional generalization tasks, DINOSAURv2 consistently outperforms DINOv2 at any compute budget, and in some settings, a small object-centric representation with a small downstream model beats a large DINOv2 variant with a larger downstream model, despite up to eightfold difference in compute.

[2] `https://www.blender.org/`

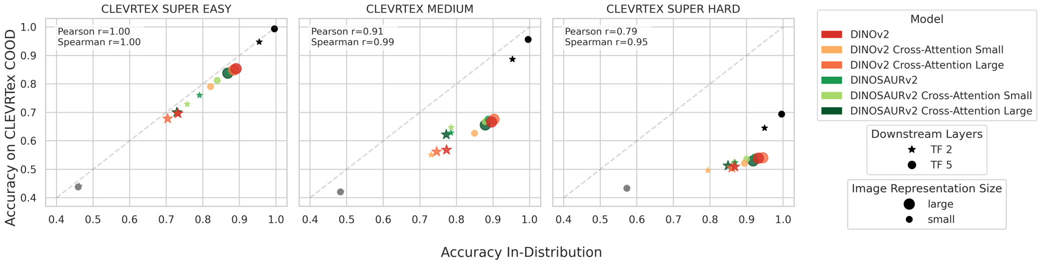


Figure 1: VQA in-distribution and compositional out-of-distribution accuracy are very strongly correlated (highly significant: p-value $< .01$). Performances for CLEVRTex *"super easy"*, *"medium"*, and *"super hard"* at the end of training (600k steps) with correlations and ground-truth oracle (upper right: black) and question-only baseline (lower left: grey).

First, we verify that our dataset and downstream model setup are suitable for testing compositional generalization by establishing that a model with the "right" representation is able to solve the task in-distribution (ID) but still might lack in compositional out-of-distribution (COOD) generalization as the hardness increases. We do this by training an oracle that uses the ground-truth object properties as image representation. As shown in Fig. 1, the oracle can achieve perfect ID test accuracy on all dataset variants, given a sufficiently large downstream model (for the full results, see Fig. 5). However, its compositional generalization drops notably when trained on smaller subsets of the full visual space. More precisely, while it still reaches 100% on *"COOD"* by training on *"super easy"*, it struggles to even reach 70% when training on *"super hard"*.

After confirming the validity of our setup, we turn towards learned image representations. For all vision encoders considered here, an obvious pattern emerges. As generalization difficulty grows, the benefit of a "better" representation shrinks where the in-distribution accuracies increase and the compositional out-of-distribution accuracies decrease. This is expected as the ID task gets easier by training and testing on fewer visual combinations. In contrast, the generalization becomes harder because a model has to generalize to the same fixed *"COOD"* dataset while being trained on fewer combinations.

4.1 Strong Correlation Between In-Distribution and Compositional Out-of-Distribution Performance

We find that in-distribution performance is very strongly correlated with compositional out-of-distribution performance across all dataset variants. For easier generalizations (from *"super easy"*, *"easy"*), the relationship appears almost perfectly linear, with only a small constant gap below the ideal performance curve for the super easy variant. In contrast, for harder generalizations (from *"hard"*, *"super hard"*), the trend remains monotonic but becomes sublinear.

Although using the larger downstream model is always better for both ID and COOD accuracies in Fig. 1 and Fig. 5, DINOSAURv2 variants already perform well with a smaller model, whereas DINOv2 benefits more from additional transformer layers. This seems to be mostly due to the image representation size, as the DINOSAURv2 variant with a large representation (with a cross-attention layer increasing the sequence length) often shows the same trend as for the large DINOv2 representation. However, there are exceptions, especially on *"medium"* in Fig. 1 middle, where for the small downstream model, the large DINOSAURv2 variant generalizes better than other large image representations from DINOv2. Notably, upscaling DINOSAURv2's image representations offers no ID or COOD benefit over the original small image representation and actually degrades performance.

Regarding best absolute performances, a DINOv2 variant consistently reaches the highest ID accuracy on each dataset. For the compositional generalizations the picture is more nuanced, for the easier variants DINOv2 is slightly better (<1%), while for medium and harder datasets a DINOSAURv2 variant is always better (<2%).

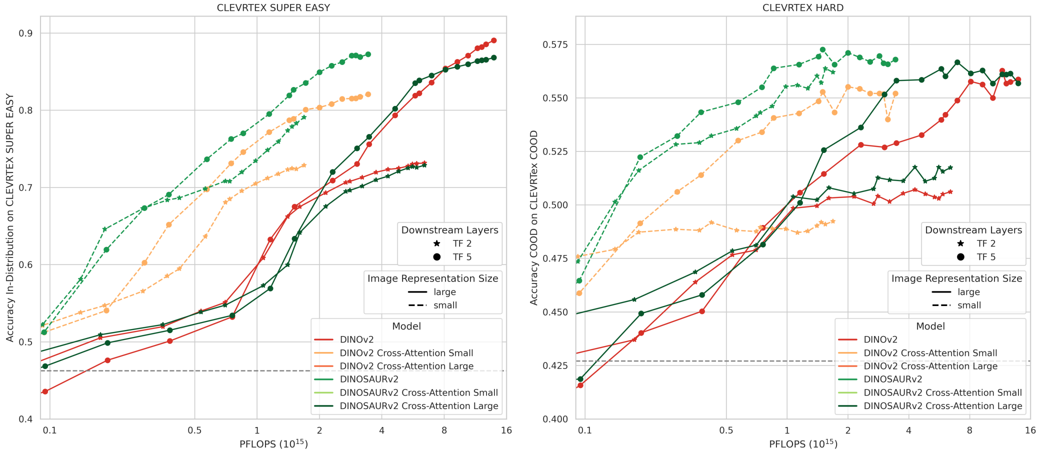


Figure 2: Object-centric representations slightly underperform on in-distribution tasks (left) but outperform on compositional out-of-distribution tasks (right), both at four to eight times lower compute. VQA ID accuracy for CLEVRTex super easy (left) and COOD for CLEVRTex hard (right) at different compute budgets with question-only baseline (dashed grey).

4.2 Object-Centric Representations Slightly Underperform In-Distribution At Significantly Lower Compute

In the previous section, we did not directly address the mismatch in downstream compute across different image representations. Instead, we compared end-of-training performance directly. Here, we provide a fairer comparison (for details see Section 3.2 and Appendix B.3) by estimating the FLOPs per training step for each downstream model[3] including only the model components affected by the different image feature sizes.[4]

First, we examine in-distribution test performance on CLEVRTex *"super easy"* across varying compute budgets (Fig. 2 left), as it presents the hardest ID setting with the largest observed differences between models (see beginning of Section 4). **For any compute budget up to around four PFLOPs (end of training for smaller image representations), DINOSAURv2's object-centric representation outperforms all others**. Interestingly, a smaller downstream model converges faster and plateaus earlier, making it advantageous in the early training phase.

To surpass DINOSAURv2's ID performance, one must invest over three times more compute with DINOv2, yet the resulting gain is under 2% at the end of training with four times the compute. As before, increasing the representation size of DINOSAURv2 offers no benefit whatsoever and worsens performance despite a higher compute budget. Similar trends hold for all other in-distribution settings but diminish for harder datasets, where the differences between models are generally smaller (see Appendix C).

Considering all models at the same time, using a small downstream model for best ID performance only makes sense for very constrained compute budgets under 0.5 PFLOPs for all datasets. Then, DINOSAURv2's small image representation with a small downstream model performs best.

4.3 Object-Centric Representations Excel At Harder Compositional Generalization With Much Less Compute

Moving on to compositional generalization, we find that models begin plateauing in accuracy and overfitting in cross-entropy increasingly early, going from *"super easy"* to *"super hard"* (see Figs. 2 right, 6 and 7).

On CLEVRTex *"medium"* (Fig. 6), at any given compute budget, the small DINOSAURv2 representation outperforms all alternatives; even granting DINOv2 four times more compute

[3] `https://github.com/facebookresearch/fvcore`

[4] We can cache image and text representations prior to downstream training.

fails to surpass it. This pattern holds for harder generalizations as well, i.e. CLEVRTex *“hard”* in Fig. 2 right and *“super hard”* Fig. 6.

Additionally, in the hard generalization scenario (Fig. 2 right), a small DINOSAURv2 representation coupled with either downstream model consistently beats any DINOv2 representation and downstream model combination for all compute budgets. **In short, a small object-centric representation with a small downstream model is better than a large DINOv2 representation with a large downstream model, even at nearly eight times the compute**.

5 Conclusion

In this work, we systematically evaluated the compositional generalization capabilities of object-centric representations in controlled, visually rich settings. By introducing a benchmark based on the CLEVRTex dataset, we demonstrated that object-centric models, specifically DINOSAURv2, exhibit superior compositional generalization compared to non-object-centric alternatives, such as DINOv2, while requiring significantly less compute. Our results further reveal that object-centric representations converge faster in in-distribution tasks but underperform slightly when giving non-object-centric models significantly more compute. Moreover, our analysis confirmed a strong correlation between in-distribution and compositional out-of-distribution performance.

These findings reinforce the potential of object-centric approaches for tasks requiring systematic compositional reasoning and highlight the need for further exploration into their applications beyond synthetic benchmarks. Future work may extend this by investigating the effectiveness of object-centric learning in real-world scenarios, incorporating more diverse datasets, and optimizing architectural choices to enhance performance across a broader range of vision tasks.

Acknowledgments

We would like to thank Max Seitzer for the helpful discussions and comments. This work was partially supported by the Helmholtz Foundation Model Initiative, supported by the Helmholtz Association. A.D. acknowledges support from G-Research. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUPITER — JUWELS (Jülich Supercomputing Centre, 2021) at Jülich Supercomputing Centre (JSC).

References

Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models. *arXiv preprint arXiv:2407.05897*, 2024.

Ondrej Biza, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin Fathy Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In *International Conference on Machine Learning*, 2023.

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. *arXiv preprint arXiv:2305.14229*, 2023.

Aniket Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Mike Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. Zero-shot object-centric representation learning. *arXiv preprint arXiv:2408.09162*, 2024.

Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in object-centric learning. In *International Conference on Machine Learning*, 2022.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Dan Haramati, Tal Daniel, and Aviv Tamar. Entity-centric reinforcement learning for object manipulation from pixels. *arXiv preprint arXiv:2404.01220*, 2024.

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.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *NeurIPS*, 2023.

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

Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A138), 2021. doi: 10.17815/jlsrf-7-183. URL `http://dx.doi.org/10.17815/jlsrf-7-183`.

Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. *arXiv preprint arXiv:2111.10265*, 2021.

Yeongbin Kim, Gautam Singh, Junyeong Park, Caglar Gulcehre, and Sungjin Ahn. Imagine the unseen world: a benchmark for systematic generalization in visual world models. *Advances in Neural Information Processing Systems*, 36, 2024.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

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*, pp. 10910–10921, 2023.

Amir Mohammad Karimi Mamaghan, Samuele Papa, Karl Henrik Johansson, Stefan Bauer, and Andrea Dittadi. Exploring the effectiveness of object-centric representations in visual question answering: Comparative insights with foundation models. *arXiv preprint arXiv:2407.15589*, 2024.

Milton L Montero, Jeffrey S Bowers, and Gaurav Malhotra. Successes and limitations of object-centric models at compositional generalisation. *arXiv preprint arXiv:2412.18743*, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

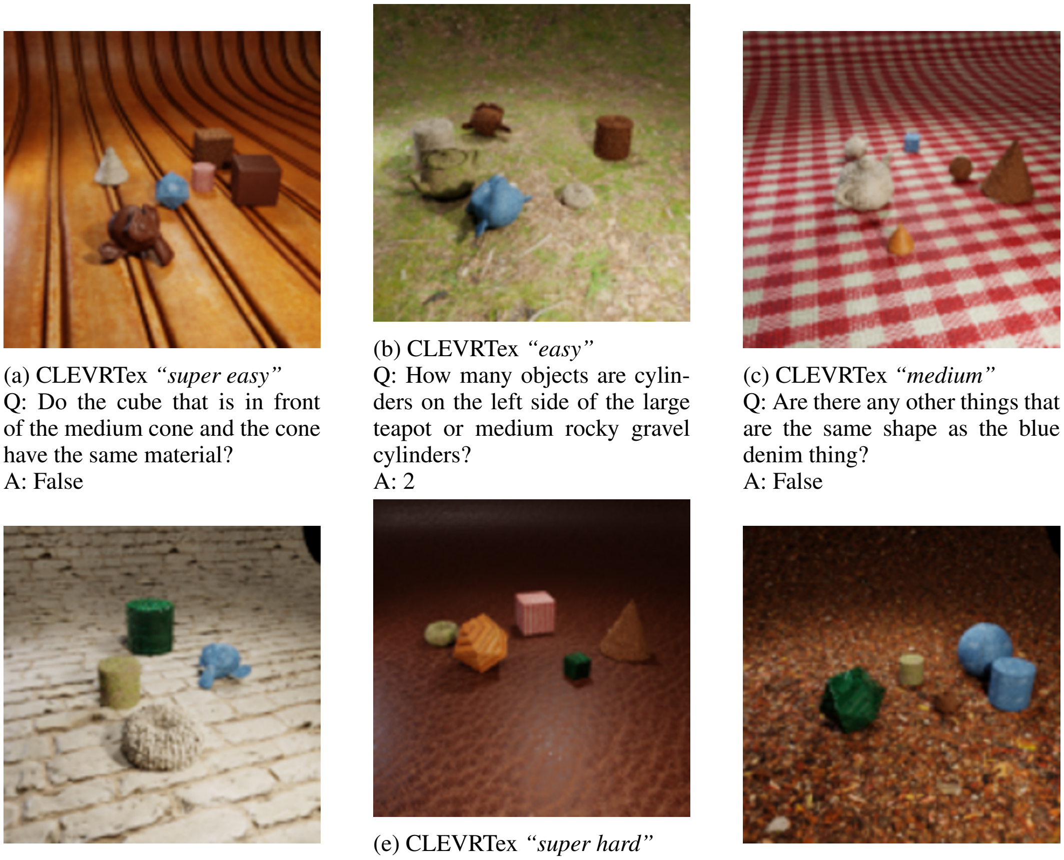
Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024.

Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning. *arXiv preprint arXiv:2302.04419*, 2023.

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? *arXiv preprint arXiv:2210.01936*, 2022.

A Data Generation

The used attributes for image generation are in Table 1, and some example images with question–answer pairs are in Fig. 3.



(a) CLEVRTex *"super easy"*
Q: Do the cube that is in front of the medium cone and the cone have the same material?
A: False

(b) CLEVRTex *"easy"*
Q: How many objects are cylinders on the left side of the large teapot or medium rocky gravel cylinders?
A: 2

(c) CLEVRTex *"medium"*
Q: Are there any other things that are the same shape as the blue denim thing?
A: False

(d) CLEVRTex *"hard"*
Q: Are there any medium things behind the big thing to the left of the white sandstone object?
A: False

(e) CLEVRTex *"super hard"*
Q: What size is the cube left of the block that is in front of the thing behind the green forest torus?
A: Large

(f) CLEVRTex *"COOD"*
Q: How many other objects are there of the same material as the small teapot?
A: 0

Figure 3: Dataset examples with question–answer pairs from each CLEVRTex variant.

Table 1: Attributes for the image and question generation.

Material	**Shape**	**Size**
green tiled	cube	small
blue denim	cylinder	medium
red fabric	monkey head	large
green forest	icosahedron	
red leather	teapot	
rocky gravel	sphere	
rusty metal	cone	
white sandstone	torus	

B Models

B.1 DINOSAURv2

The hyperparameters of the base configuration for the DINOSAURv2 versions can be found in Table 2.

Table 2: Hyperparameters of DINOSAURv2.

Hyperparameter		CLEVRTex Variants
Training Steps		300k
Batch Size		128
LR Warmup Steps		10k
Peak LR		0.0002
Exp. Decay Half-Life		100k
Feature Extractor		DINOv2_s
Patch Size		14
Feature Dim.		384
Gradient Norm Clipping		0.1
Image Size		224
Cropping Strategy		Full
Image Tokens		256
Decoder	Type	MLP
	Layers	4
	MLP Hidden Dim.	2048
Slot Attention	Iterations	3
	Number of Slots	7
	Slot Dim.	256
	MLP Hidden Dim.	1024

B.2 Dowstream VQA Model

Architecture We adopt a transformer-based architecture for VQA, following Mamaghan et al. (2024). We first project both image and text representations via separate linear layers (output size 126) with a dropout of 0.1, and augment them with a two-dimensional one-hot vector to indicate whether they originate from image features or text embeddings. We then add a sinusoidal positional encoding to the text embeddings. To perform classification, we use a trainable CLS $\in \mathbb{R}^{128}$ vector. We concatenate the image and text representations (plus the CLS token) and pass them through a transformer encoder with $d_{model} = 128$ and a hidden dimension of 128. The transformed CLS token is fed into a two-layer MLP (hidden dimension 128) with layer normalization, a dropout rate of 0.1, and a ReLU activation between layers. This MLP outputs a probability distribution over all possible answers.

Training For all CLEVRTex variants, we train the downstream models with a batch size of 128, a learning rate of 0.0001, and a cross-entropy loss for 600k steps. We use downstream model variants where we vary the number of layers of the transformer encoder, either 2 or 5 layers with 64 heads.

B.3 Compute

The base models DINOv2 and DINOSAURv2 produce for all CLEVRTex variants, with an image size of 224, representations of shape $[256, 384]$ and $[7, 256]$, respectively. This results in a huge compute mismatch for the downstream model in Fig. 4, where the FLOPs for the downstream model with the DINOv2 image representation are roughly four times that of DINOSAURv2 for both transformer sizes. To remedy that, we use a single cross-attention layer with four heads right after the vision encoder to map from the large, i.e. the shape for DINOv2, to the small image representation or vice-versa. We additionally considered mapping from the input size to the output size to match the compute of different variants even more, but this did not result in consistent improvements.

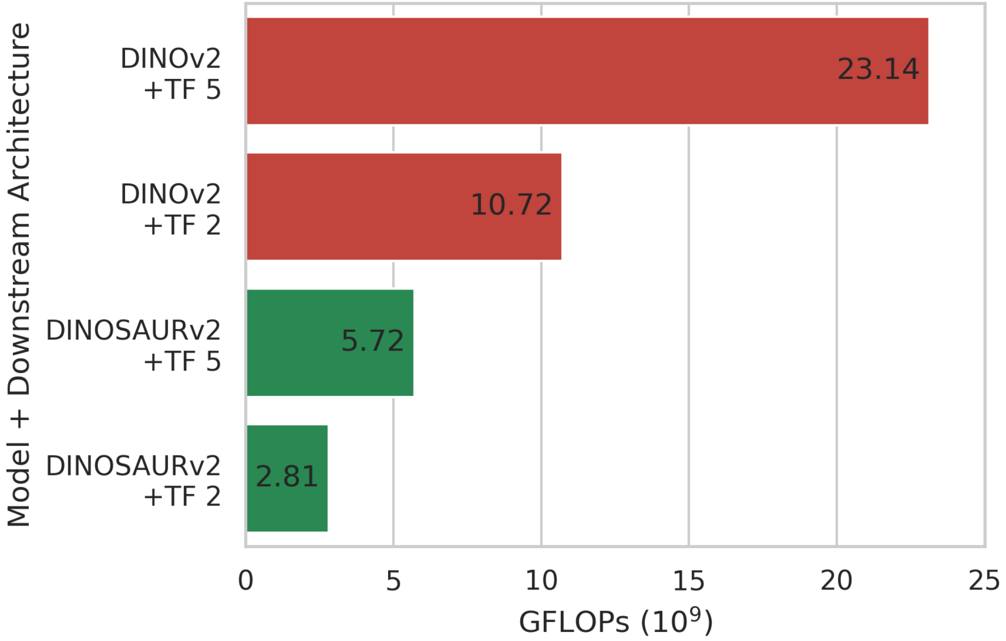


Figure 4: GFLOPs for one step of the downstream model for DINOv2 and DINOSAURv2 with both two or five layers for the transformer encoder (TF 2, TF 5).

C Additional Comparisons

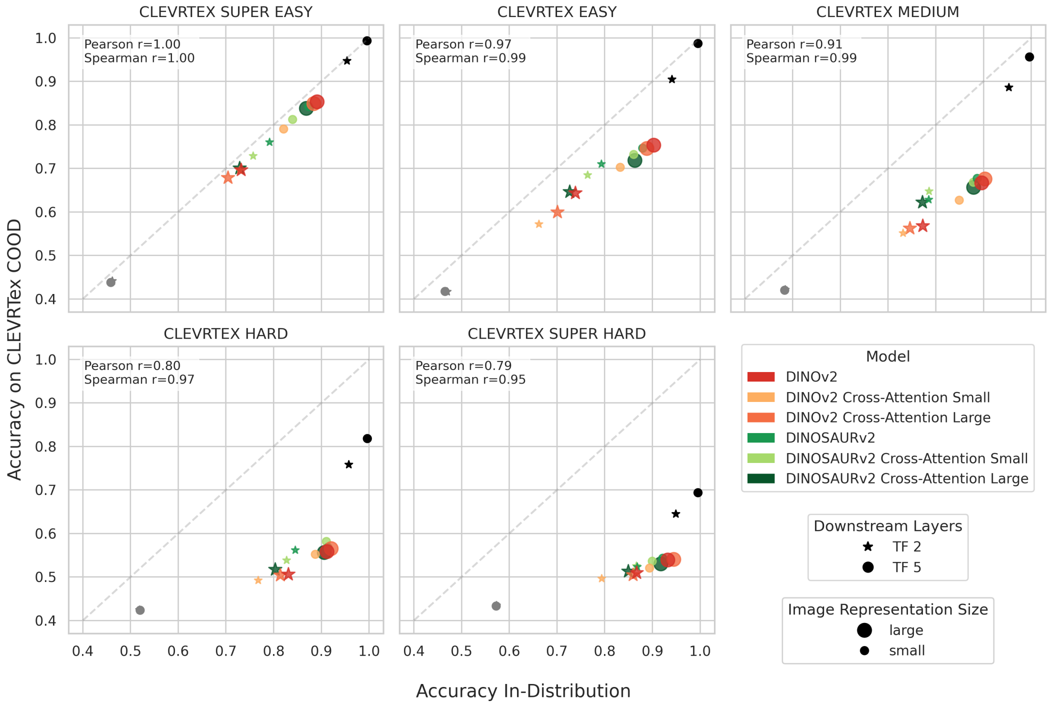


Figure 5: VQA in-distribution and compositional out-of-distribution accuracy are very strongly correlated (highly significant: p-value < .01). Performances for every CLEVRTex dataset variant at the end of training (600k steps) with correlations and ground-truth oracle (upper right: black) and question-only baseline (lower left: grey).

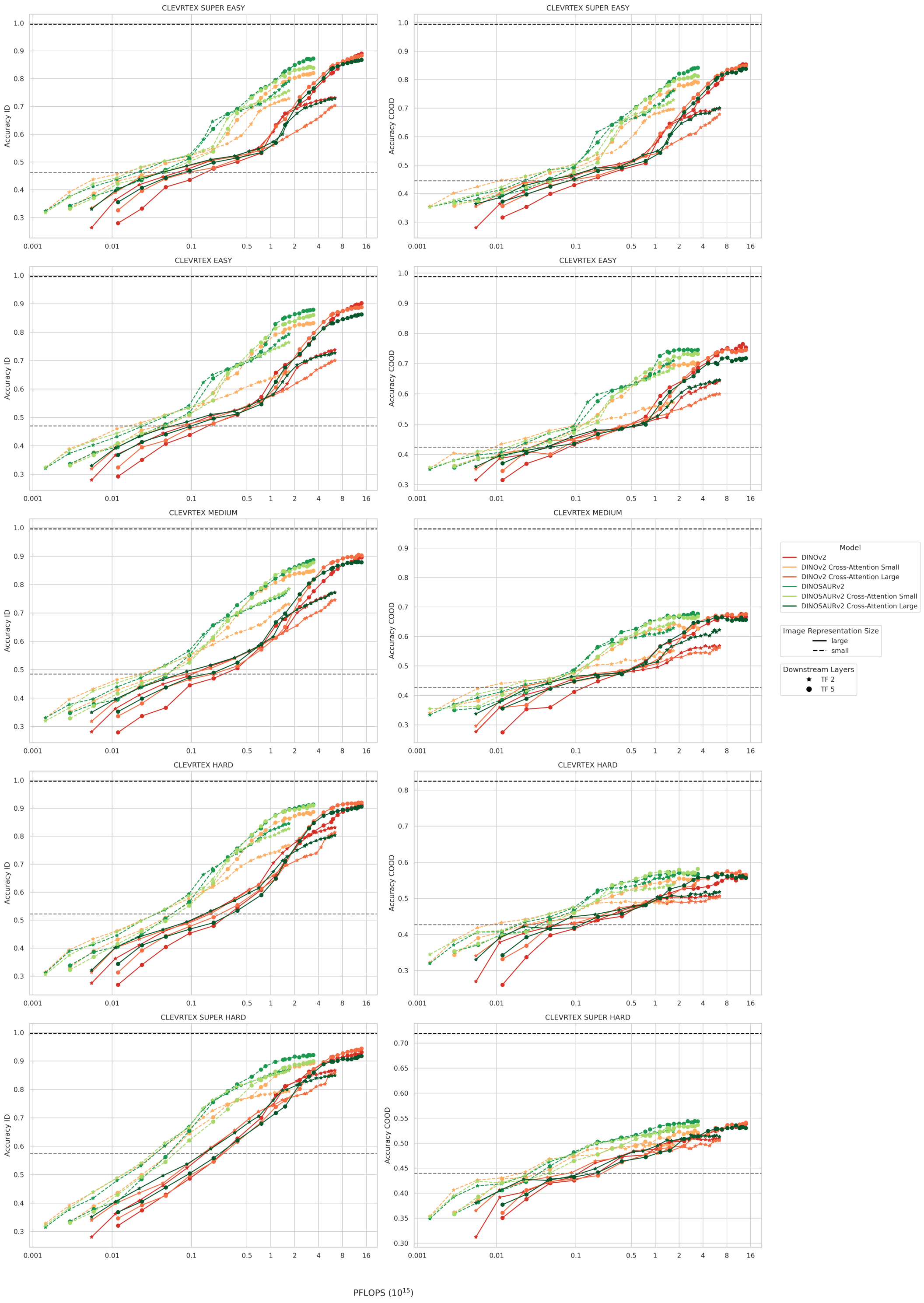


Figure 6: VQA in-distribution and compositional out-of-distribution accuracy for all CLEVRTex dataset variants with ground-truth oracle (upper: dashed black) and question-only baseline (lower: dashed grey).

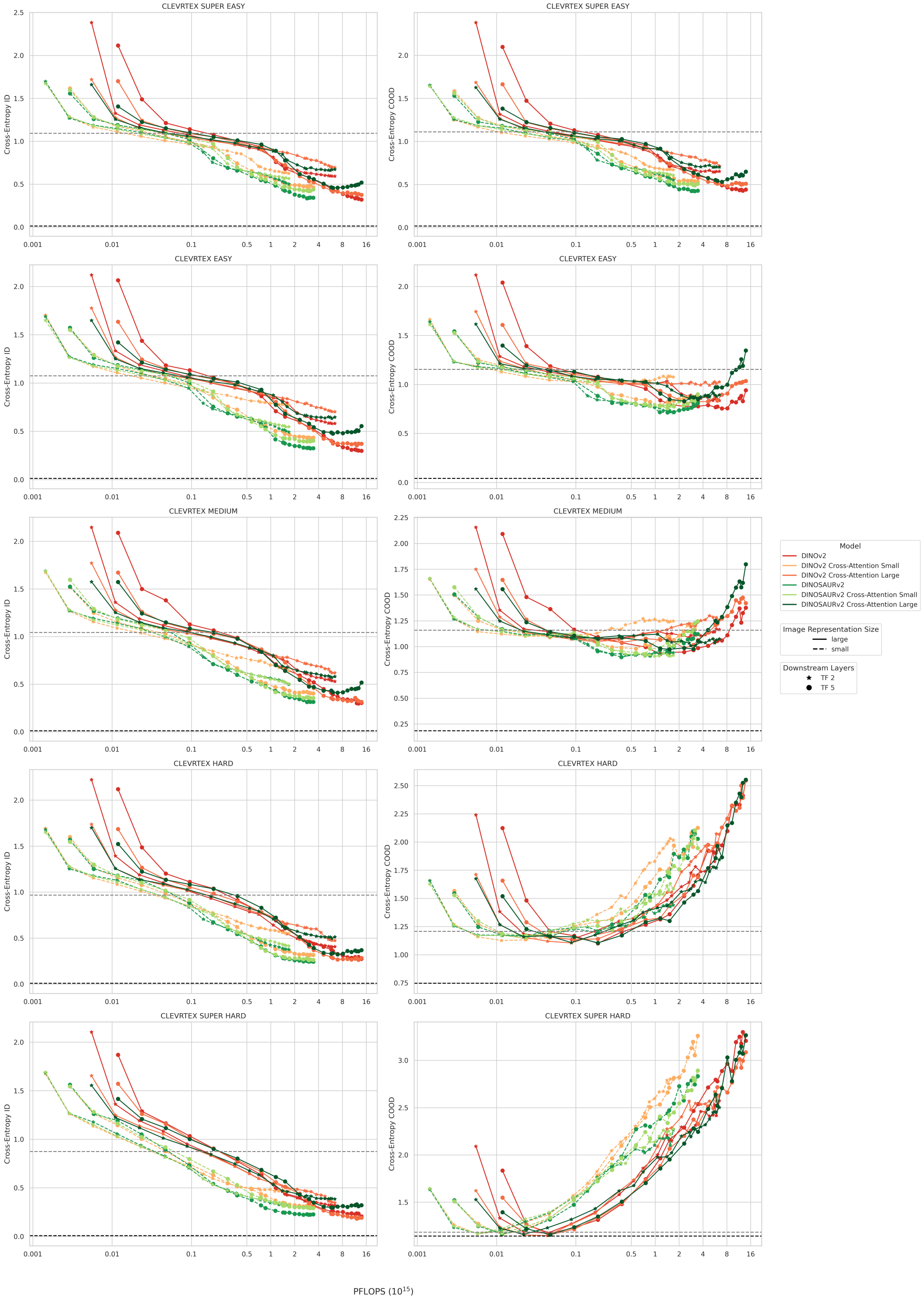


Figure 7: VQA in-distribution and compositional out-of-distribution cross-entropy for all CLEVRTex dataset variants with ground-truth oracle (lower: dashed black) and question-only baseline (upper: dashed grey).