In-Context Learning Improves Compositional Understanding of Vision-Language Models

Matteo Nulli [1] **Anesa Ibrahimi** [1] **Avik Pal** [1] **Hoshe Lee** [1] **Ivona Najdenkoska** [1]

Abstract

Vision-Language Models (VLMs) have shown remarkable capabilities in a large number of downstream tasks. Nonetheless, compositional image understanding remains a rather difficult task due to the object bias present in training data. In this work, we investigate the reasons for such a lack of capability by performing an extensive bench-marking of compositional understanding in VLMs. We compare contrastive models with generative ones and analyze their differences in architecture, pre-training data, and training tasks and losses. Furthermore, we leverage in-context learning as a way to improve the ability of VLMs to perform more complex reasoning and understanding given an image. Our extensive experiments demonstrate that our proposed approach outperforms baseline models across multiple compositional understanding datasets. The code is available here.

1. Introduction

Recent breakthroughs in foundation models (Radford et al., 2018; Dosovitskiy et al., 2021; Bommasani et al., 2021; Doveh et al., 2023a) are reaching human-level performance on many vision and language benchmarks. Particularly, Visual Language Models (VLMs) have shown impressive performance on many different downstream tasks, like image captioning (Yu et al., 2022), visual-question answering, image understanding (Liu et al., 2023), object localization (Dorkenwald et al., 2024) and others (Radford et al., 2021; Singh et al., 2022; Li et al., 2022). Nonetheless, some tasks, that are considered simple for humans, remain hard for VLMs. Among these lies the difficulty of correctly understanding the composition of objects in an image and their textual description. While humans can easily connect the images correctly with their corresponding descriptions, VLMs struggle to understand this sequential order of words and thus, perform poorly (Yüksekgönül et al., 2023). Recent advancements in computing and data scaling resulted in huge performance increases in many VL tasks, however, *compositional understanding* remains a challenging problem.

In this work, we study how state-of-the-art VLMs perform on benchmarks that evaluate their compositional understanding (Thrush et al., 2022; Yüksekgönül et al., 2023; Hsieh et al., 2023). Most works focus on contrastive pre-training objectives as the reason for bad performance on compositionality benchmarks (Yüksekgönül et al., 2023). The authors reason that employing a contrastive pre-training objective pushes VLMs to perform text-image retrieval without actually having any compositional understanding. This is the main reason why such models perform well on numerous benchmarks which do not require any compositional understanding. Differently from Yüksekgönül et al. (2023), we evaluate both contrastive (Radford et al., 2021) and generative models (Li et al., 2023a; Alayrac et al., 2022; Liu et al., 2023), to provide a deeper understanding of the underlying reasons for the lack of comprehension. Furthermore, we analyze their differences by looking at their pre-training strategy, data, and architectures. Next, we study the impact of different prompting strategies and introduce a new In-Context Learning (ICL) prompting method through synthetically generated, web images and captions. Specifically, we generate the synthetic data using GPT-4o[1] by instructing the model to prepare compositionally aware captions from a specified list of objects. We then give this as input to GPT-4o to generate an image matching the caption and a negative caption that distorts its compositional information, as shown in Figure 1. To simulate real images and captions, we randomly sample images from the COCO dataset (Lin et al., 2014) and manually annotate a positive and a negative compositional-aware caption for them. We use these examples as demonstrations for few-shot in-context learning for the generative models along with ICL prompting.

In summary, our contributions are the following: (i) We perform a comprehensive study on the behavior of generative and contrastive VLMs on several compositional understand-

[1]University of Amsterdam, The Netherlands. Correspondence to: <matteo.nulli@student.uva.nl>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

[1]https://openai.com/index/hello-gpt-4o/

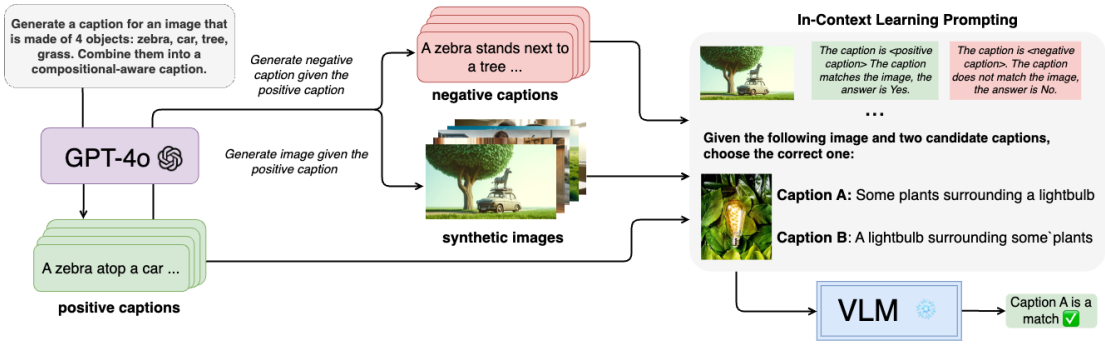

Figure 1. **Our in-context learning pipeline for compositional understanding of VLMs.** We instruct GPT-4o to generate a caption consisting of a few named objects such that they are compositionally defined. Using this positive caption we then instruct GPT-4o to generate an image and a negative caption that compositionally distorts the meaning. We feed these synthetic captions and images as few-shot examples to the VLMs. Afterwards, we instruct the model to predict between correct and wrong captions for images of compositional reasoning benchmarks while also using in-context learning prompting.

ing benchmarks. (ii) We introduce a ICL prompting framework by leveraging synthetic and real images and captions in a few-shot style. (iii) We demonstrate the advantages of our proposed ICL framework for compositional image understanding across various few-shot settings and datasets.

2. Related Work

Compositional understanding in VLMs Recent years have witnessed a significant increase in large-scale pre-trained VLMs within the field of multi-modal learning, enabling advanced tasks like image captioning. Foundation models such as CLIP (Radford et al., 2021) utilize large-scale image-text pairs for joint pre-training of image and text encoders using contrastive loss. Another notable architecture, BLIP (Li et al., 2022), leverages captions from noisy web data for comprehension and generation tasks. The (in)ability to grasp the underlying non-object notions of the images and text captions has led research into exploring compositional reasoning limitations of VLMs, revealing issues such as caption quality and density (Thrush et al., 2022; Yüksekgönül et al., 2023; Doveh et al., 2023b; Hsieh et al., 2023). Specifically, Doveh et al. (2023b) and Yüksekgönül et al. (2023) found that VLMs' output representations often resemble bags of words, neglecting several object attributes. To address these limitations, enhancements in caption quality and density were proposed (Urbanek et al., 2023; Touvron et al., 2023), including datasets with densely captioned images, significantly improving the compositional reasoning abilities of VLMs.

Enhancing reasoning through In-Context Learning In-Context Learning (ICL) is a framework through which Large Language Models can learn a new task by being presented with demonstrations of how it should be solved. As detailed by Dong et al. (2024) ICL has recently provided clear performance enhancements in Foundation Models, and this has come without the need of any parameter update. Indeed, ICL frameworks are easily interpretable as interacting with the model through examples is effortless and allows the model to gain understanding through comparison (Liu et al., 2021; Wu et al., 2022). Many studies have shown how in-context learning capabilities can be enhanced by additionally acting on the pre-training stages of models (Min et al., 2021; Dong et al., 2024). Regardless of the approach, ICL frameworks have shown to increase models performances on many downstream tasks and thus we set out to try its effectiveness also in compositional understanding.

3. Methodology

This paper explores how in-context learning prompts consisting of intermediate reasoning steps can help VLMs to be more compositionally aware. To that end, we introduce an in-context learning prompting framework, presented in Figure 1. The intermediate steps consist of images and corresponding captions in a few-shot style. We hypothesize that by doing this, VLMs can understand the complex compositional relations within an image through correct and incorrect examples of image-caption pairs. In the following sections we will describe the process in details.

Few-shot ICL prompting with synthetic images First, we generate a compositional aware caption with GPT-4o by prompting it in the following way:
Generate a caption for an image which is made of 4 objects: object 1, object 2, object 3, object 4. Can you combine them into a compositionally aware caption?
This will return the *positive* caption which is then fed back into GPT-4o, to generate a corresponding image with DALL·E. Finally, with GPT-4o, we also generate a compositionally-aware *negative* caption, that is in contrast with the correct caption, using the prompt of Appendix B. The same procedure is repeated five times and these exam-

True Caption: "*A zebra atop a car parked beneath a towering tree*"

Wrong Caption: "*A zebra stands next to a tree, while a car is parked on top of the tree's branches.*"

True Caption: "*A dog curled up on a chair with a book and a stuffed toy, enjoying a peaceful moment.*"

Wrong Caption: "*A dog playing with a stuffed toy on the floor, while a book lies open on the chair nearby.*"

True Caption: "*A cow grazes on grass while a key lies nearby, with a computer set up incongruously in the lush field.*"

Wrong Caption: "*A computer set up on the grass, while a cow stands nearby, a key hanging from its neck in the open field.*

True Caption: "*A person enjoying pastries at a table with a slot machine nearby, ready for a win.*"

Wrong Caption: "*A person playing a slot machine, with pastries stacked on the table nearby, savoring both luck and treats.*"

True Caption: "*A woman with a backpack and a tie, holding a basketball, ready for both business and play.*"

Wrong Caption: "*A woman in a tie playing basketball, her backpack resting nearby, blending work and leisure.*"

True Caption: "*Two cats sleeping on a purple blanket on top of a couch with two tv remotes next to them.*"

Wrong Caption: "*Two cats sleeping under an purple blanket below a couch with no tv remotes next to them.*

True Caption: "*A group of women talking while sitting at a table.*"

Wrong Caption: "*A group of women standing next to a table and talking.*"

True Caption: "*A room with computers on top of tables and some people working on them.*"

Wrong Caption: "A room with computers under the tables and no people working on them."

True Caption: "*A table full of salty and sweet food inside a cozy room.*"

Wrong Caption: "*A table without any food outside a cozy room.*"

True Caption: "*A bathroom with a sink on the bottom right, a cabinet in the bottom left.*"

Wrong Caption: "*A bathroom with a sink on the bottom left, a cabinet in the top left.*"

Figure 2. **Few-shot samples**. *First row*: The images and captions are synthetically generated with GPT-4o as seen in Figure 1. *Second row*: Images are manually captioned and retrieved from COCO dataset (Lin et al., 2014). We use these images to instill an understanding of the task within the generative models in a few-shot manner.

ples serve as few-shot examples for the model to understand the compositional nature of the images. The first row of Figure 2 shows the outputs of the described generation step. Along with these 5 examples, the model is also given a query input, as shown in Appendix B. This helps the model reason about its generation and clearly understand the task. Additionally, we randomly switch the correct caption for few-shot samples between *A* and *B*, to remove any bias toward choosing either of the characters, especially since *A* itself is a common token. Depending on the benchmark we use slight variations of the above prompt. More information is provided in Appendix D.

Few-shot ICL prompting with real images We also explore ICL prompting by employing real images. We use five images from COCO dataset (Lin et al., 2014) and caption them manually with a positive and negative caption, to be as compositionally-aware as possible. Similar to synthetic images, we use these image-caption pairs as few-shot examples with the same prompting template. From the *second row* of Figure 2 it can be seen that real ones provide a more complex and noisier context unlike synthetic images.

4. Experiments & Results

4.1. Datasets

We employ the following three datasets for evaluation of compositional understanding.

Attribution, Relation, and Order (ARO) (Yüksekgönül et al., 2023) evaluates four different aspects of compositionality. *Visual Genome Attributions* and *Visual Genome Relations* are testing the model's understanding of correct attributes and relations associated with objects within an image, and COCO Order and Flickr30k Order are testing the understanding of the correct ordering of words in a caption.

Winoground (Thrush et al., 2022), comprising an evaluation set of 400 pairs of two images and two captions. The pairs are very similar to each other, with slight linguistic

Model	Winoground			SugarCrepe							ARO				Avg.
	T	I	G	AO	AA	RA	RO	RR	SA	SO	C	F	VG-A	VG-R	
Contrastive Models															
ViT-B-16	32.7	12.5	10.0	89.2	78.0	85.9	95.1	69.0	67.5	63.0	18.4	34.3	57.4	22.1	52.5
ViT-B-32	34.3	10.8	7.5	87.1	77.9	82.6	93.8	68.9	67.4	60.2	32.6	39.5	61.8	46.1	55.0
bigG-CLIPA*[1]	34.0	13.0	10.3	**90.9**	**85.5**	84.3	**96.9**	71.1	75.2	62.6	30.4	33.9	59.3	38.8	56.2
SigLIP-256*[2]	36.7	15.2	13.5	90.5	82.9	85.7	96.1	69.9	77.0	68.2	33.7	40.0	62.9	37.7	57.9
SO-SigLIP*[3]	**38.0**	**20.3**	**15.8**	84.0	84.6	82.8	96.1	73.3	76.8	61.3	37.9	40.3	65.1	39.1	58.2
Generative Models															
LLaVA	20.1	1.00	0.01	78.1	62.8	89.7	94.1	81.2	69.3	**71.9**	78.9	83.4	61.0	54.1	60.4
CogVLM	32.3	1.00	0.70	79.7	74.2	**91.4**	96.7	**82.5**	**82.8**	71.5	**84.9**	**87.5**	**74.1**	**67.5**	**66.2**

Table 1. **Zero-shot performance of contrastive and generative models** on all the benchmarks and accompanying subscores. *Full Model names - 1: ViT-bigG-14-CLIPA, 2: ViT-L-16-SigLIP-256, 3: ViT-SO400M-14-SigLIP. **Winoground**: T=text, I=Image, G=Group. **SugarCrepe**: AO=add_obj, AA=add_att, RA=repl_att, RO=repl_obj, RR=repl_rel, SA=swap_att, SO=swap_obj. **ARO**: C=COCO, F=Flickr30k, VG-A=VG-Attribute, VG-R=VG-Relation.

differences in the captions.

SugarCrepe (Hsieh et al., 2023) benchmark tests various fine-grained compositional concept understanding aspects using COCO image-text pairs. An object, attribute, or relation is either replaced, swapped, or added to the original text such that the caption no longer matches the scene.

4.2. Baseline models

Contrastive Models Most state-of-the-art VLMs are trained with a contrastive loss (Radford et al., 2021; Chen et al., 2020; Zhai et al., 2023; Sun et al., 2023; Li et al., 2023b), and most of these use a Vision Transformer (ViT) (Dosovitskiy et al., 2021) as vision encoders, while some employ ResNet architecture (He et al., 2016). In our analysis, we use CLIP-based models from the OpenCLIP library (Ilharco et al., 2021). Some of them include EvaCLIP (Sun et al., 2023), SigLIP (Zhai et al., 2023), CLIPA (Li et al., 2023b), and CoCa (Yu et al., 2022).

Generative Models Generative models differ from CLIP-based models, not only in their ability to perform autoregressive generation but also in their training process. In Section 4 we analyze two generative models, namely LLaVA (Liu et al., 2023) and CogVLM (Wang et al., 2023).

4.3. Evaluation with contrastive VLMs

For contrastive evaluation, we select models from the OpenCLIP library (Ilharco et al., 2021) and assess all available options. Table 1 highlights the top five models with the best performance, distinguished by their pre-trained visual encoders. Table 1 shows a general pattern in the consistent out-performance of SigLIP models (ViT-SO400M-14-SigLIP, ViT-L-16-SigLIP-256) across most benchmarks and sub-scores with an average increase of 4% on Winoground and 6% on ARO. This is particularly relevant in tasks requiring the replacement or swapping of attributes/objects. The ViT-bigG-CLIPA model also performs competitively, especially on SugarCrepe where it achieves a slight increase concerning SigLIP.

The efficiency and ability of CLIP to perform on a wide range of tasks has been covered in Radford et al. (2021), as well as its limitations. One example is its poor performance on various fine-grained classification tasks that involve differentiating between different representations of objects. However, one of the most significant limitations is the restriction of only being able to choose among concepts from a given zero-shot classifier. This in effect, prevents it from generating novel outputs or combining existing concepts in new ways. Consequently, when a new instance is presented, CLIP is unable to accurately classify or generate a novel output which is a key component of compositional reasoning. Differently, the superior performance of SigLIP models over CLIPA is also attributed to the loss and training used for SigLIP. Indeed, CLIPA is using softmax normalization in the contrastive loss which therefore normalizes every positive pair with all negative ones leading to quadratic complexity. On the contrary, SigLIP reduces the calculation to a simpler sigmoid function and independently evaluates the positive-negative pairs in the batch. This allows SigLIP to be trained more efficiently and perform better at small batch sizes.

4.4. Evaluation with generative VLMs

Zero-shot performance We demonstrate the zero-shot performance of generative models in Table 1. It can be seen that CogVLM performs better than LLaVA, by increasing the Winoground text score by almost 12% and being consistently better with an average increase of 8% on ARO and similar yet overall better scores on SugarCrepe. The primary reason for this is its architectural and training advantages. CogVLM employs a vision expert module at each layer of the Large Language Model (LLM) comprising of

Method	Type of samples	Winoground			SugarCrepe							ARO			
		T	I	G	AO	AA	RA	RO	RR	SA	SO	C	F	VG-A	VG-R
Zero-Shot	-	20.1	1.00	0.01	**78.1**	**62.8**	**89.7**	**94.1**	**81.2**	**69.3**	**71.9**	78.9	83.4	61.0	54.1
1-Shot	Synthetic	21.0	2.00	0.70	71.1	61.9	77.7	89.5	73.2	64.7	59.3	79.3	85.3	58.8	58.7
	Real	**26.0**	**5.70**	**2.10**	72.9	61.1	75.6	89.5	73.3	65.0	60.9	81.8	84.9	63.6	61.5
5-Shot	Synthetic	25.0	5.00	2.00	75.8	58.9	81.2	91.2	80.1	65.4	59.7	**85.8**	88.8	**65.7**	**65.3**
	Real	25.5	4.70	**2.10**	74.7	60.2	84.0	92.7	80.5	65.4	61.7	85.7	**89.1**	62.7	63.1

Table 2. **Performance of LLaVA in a zero-shot and our in-context learning setting** using synthetic and real image and caption demonstrations. Synthetic demonstrations are generated as seen in Figure 1 and real images are taken from COCO and manually annotated. The meaning of each sub-score follows that of Table 1.

new *Query-Key-Value* and MLP weights initialized from the LLM. These are tuned for vision features while the original weights for the text remain frozen, allowing for using the already learned semantics of the LLM to make better use of image features. LLaVA instead makes simpler architectural choices tuning the LLM for image and text features together. Additionally, CogVLM uses both next-token prediction and object localization in its training, whereas LLaVA only uses next-token prediction.

Comparison to contrastive models Text encoders of CLIP-like models are order-agnostic (Yüksekgönül et al., 2023) and consequently do not perform well on the COCO-Order and Flickr30k-Order datasets of ARO. In contrast, generative models perform quite well on these tasks as the LLMs are pre-trained in a next-token prediction fashion. Regarding Winoground, generative VLMs show somewhat comparable performance on the text subscore but show quite degraded performance on image (and consequently group) scores. This could be explained by how these different model types match images and texts. Contrastive models match images and text by calculating the similarity of their logits, which is commutative, i.e. there is explicit direction from image to text or text to image. Generative VLMs on the other hand are commonly trained using image-captioning objectives and therefore have an explicit direction from image to text (i.e., describe the image shown to it) and thus, struggle when having to do the non-descriptive task of choosing between images given a caption. One could remedy this by comparing the caption sequence probability conditioned on different image inputs and matching based on these probabilities, but whether that truly reflects compositional understanding is unclear.

Few-shot performance In Table 2, we observe that both synthetic and real demonstrations improve the performance on Winoground and ARO, but decrease on SugarCrepe. One reason for this could be that the way we generate negative captions for both synthetic and real image demonstrations might not lend itself to SugarCrepe, leading to out-of-domain image-caption correspondences. Specifically, each sub-experiment within SugarCrepe creates negative captions by changing one or two aspects of the positive caption by adding, swapping, or replacing, objects or their attributes and relations. The negative captions of our demonstrations change multiple of these aspects at once. This style, however, aligns much closer to how positive and negative captions are used in Winoground and ARO, explaining the discrepancy between these benchmarks and SugarCrepe.

5. Conclusion

In this work, we explore the compositional understanding of contrastive and generative VLMs. Despite lower language understanding, contrastive models remain competitive due to their consistent evaluation method. Generative models face challenges such as asymmetric text-image relationships due to autoregressive training and reliance on frozen CLIP-like vision encoders. Furthermore, we introduce an ICL framework to examine the impact of synthetic and real images and captions as few-shot demonstrations. Our results show improved performance across diverse compositional understanding benchmarks, both when using synthetic and real images. This suggests potential benefits from using task-specific, few-shot examples for improving the capabilities of VLMs, such as compositional understanding.

Future work To improve compositional understanding, future VLMs could move away from contrastive vision encoders and make use of alternative training objective like patch-level prediction (Oquab et al., 2024; Yun et al., 2022) which has shown improved inter-patch understanding which could be useful for compositional understanding. To achieve similar results Densely Captioned Images have shown positive impact on compositional understanding (Urbanek et al., 2024), with further research possibly leading to substantial improvements. Alternatively, as compositional reasoning can be seen as a form of symbolic reasoning, transformer-based foundation models could be supplemented with logic components. Indeed, recent work has used neurosymbolic grounding to enable compositionally aware world models (Sehgal et al., 2023). As such, improving compositionality could be seen as falling under the larger umbrella of improving the reasoning capabilities of (multi-modal) foundation models, which might require more explicit symbolic components or finding non-symbolic architectures that can exhibit

stronger machine cognition characteristics.

References

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Dorkenwald, M., Barazani, N., Snoek, C. G. M., and Asano, Y. M. PIN: positional insert unlocks object localisation abilities in vlms. *CoRR*, abs/2402.08657, 2024. doi: 10.48550/ARXIV.2402.08657.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., and Karlinsky, L. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023a.

Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., and Karlinsky, L. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Hsieh, C., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. If you use this software, please cite it as below.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu,*

Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Li, X., Wang, Z., and Xie, C. An inverse scaling law for CLIP training. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sehgal, A., Grayeli, A., Sun, J. J., and Chaudhuri, S. Neurosymbolic grounding for compositional world models. *CoRR*, abs/2310.12690, 2023. doi: 10.48550/ARXIV.2310.12690.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15617–15629. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01519.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023. doi: 10.48550/ARXIV.2303.15389.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5228–5238. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00517.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288.

Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., and Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *CoRR*, abs/2312.08578, 2023. doi: 10.48550/ARXIV.2312.08578.

Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., and Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions, 2024. URL https://arxiv.org/abs/2312.08578.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogvlm: Visual expert for pretrained language models. *CoRR*, abs/2311.03079, 2023. doi: 10.48550/ARXIV.2311.03079.

Wu, Z., Wang, Y., Ye, J., and Kong, L. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *arXiv preprint arXiv:2212.10375*, 2022.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022, 2022.

Yüksekgönül, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. 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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Yun, S., Lee, H., Kim, J., and Shin, J. Patch-level representation learning for self-supervised vision transformers, 2022. URL https://arxiv.org/abs/2206.07990.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100.

A. Introduction

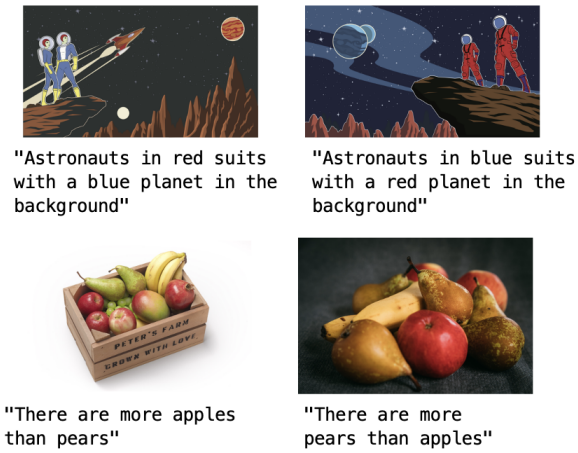

Figure A.1. **Compositional reasoning examples** from Winoground (Thrush et al., 2022), showing the close similarity between the pairs of images and text.

B. Method

Below we state the ICL prompting strategy used in our experiments,

USER: Does the image match the caption?
A. <CaptionA>
B. <CaptionB>
<image1>. The correct caption is: A/B
.
. *(We repeat the above 5 times for 5-shot in-context learning)*
.
USER: Similarly, given an image and two captions choose the correct caption. Think step-by-step and analyze the captions against the image. Begin by describing the key elements visible in the image. Then, compare these elements with the details mentioned in the captions. Clearly state your final answer only in a single character, either A or B.
<image>. The caption is:
A. <CaptionA>
B. <CaptionB>
ASSISTANT:

The prompting strategy used to generate the *wrong* caption corresponding to the correct one using GPT-4o is as below,

Generate counter caption to this one, with the same objects in a different position/attribute: 'correct caption'.

C. Appendix Experiments

The ICL prompting strategy used in SugarCrepe and ARO evaluation is as follows,

USER: <image> Given this image and two candidate captions (A and B), which caption is the better description of the given image? Clearly state your final answer only in a single character, either A or B.
A. <CaptionA>
B. <CaptionB>

The ICL prompting strategy used in Winoground evaluation is as follows,

After providing a brief explanation of your reasoning, clearly state your final answer as <Yes> or <No>.

Model	Vision encoder	LLM
LLaVA	CLIP ViT-L-336px	Vicuna1.5-7B
CogVLM	EVA-02-CLIP	Vicuna1.5-7B

Table C.1. **Generative VLMs** and the vision encoders and LLMs they use

D. Pipeline details

D.1. Contrastive evaluation pipeline

ARO and SugarCrepe For contrastive models, we evaluate ARO and SugarCrepe by first taking the positive and negative captions embeddings and comparing both with the embeddings of the image. We do this by computing the cosine similarity between each caption and the image embedding and increasing the number of correct predictions when the positive caption-image score is higher than the negative caption-image score. We adapt the code [2] by (Yüksekgönül et al., 2023).

Winoground For the Winoground benchmark, we follow (Ilharco et al., 2021) and perform a text and image encoding, for each image-caption pair. This results in two image feature representations and two caption feature representations. The final scores are then calculated by taking the cosine similarity score between the representations. This returns the real-valued outputs, which are then used to determine the text-image-group scores.

D.2. Generative evaluation pipeline

ARO and SugarCrepe For generative models, we evaluate ARO and SugarCrepe zero-shot by prompting the models using the ICL method shown in Appendix C. We then check the output of the model and increase the number of correct predictions if the model picks the correct caption choice. For 1-shot and 5-shot in-context learning, we use the prompt mentioned in Appendix B.

Winoground For the Winoground benchmark, we use a separate final instruction in the previous prompts as stated in Appendix C. If a *"yes"* character is found in the output, then the result of that corresponding pair is set to 1, if not it is set to 0. However, this evaluation strategy causes two major issues. First, the output is not always the same. Variations in the outputs result in both categorizing a correct caption as wrong if *"yes"* is never predicted and vice-versa if the predicted *"yes"* is not relating to the caption entailing the image/or choice but rather something else. To quantify this, consider the probability distribution $P(t)$ of token $t \in V$ (Vocabulary) across the sequence length s, derived from the logits L using the softmax function. Even if $P(t)$ is high, t might not be generated if another token has a higher probability. Secondly, given the binary value of 0/1, evaluating generative models on Winoground using the previous method results in having *text, image, group scores* to be all equal. To mitigate the aforementioned issues, we propose an alternative that relies on using the output logits of the desired word for evaluation. In this method, we first take the logits output tensor $L \in \mathbb{R}^{B \times S \times V}$, where B is the batch size (equal to 1 in this instance), S is sequence length and V is the vocabulary size. We take the token id *"yes"* (denoted as id_{yes}) in the third dimension, and compute the mean over the sequence length, $L_{yes} = \frac{1}{S}\sum_{s=1}^{S} L_{s,id_{yes}}$.

This results in a real-valued number $L_{yes} \in \mathbb{R}$, one per each caption-image pair given as input to the model. These values will then be compared in the same way as we do in contrastive evaluation to obtain the three accuracy scores. This technique is beneficial over the first one because it does not directly rely on generation, rather it focuses on the amount of "confidence" the model had about a specific token throughout the whole generated sequence.

[2] https://github.com/mertyg/vision-language-models-are-bows/blob/main/model_zoo/clip_models.py