The Role of Linguistic Priors in Measuring Compositional Generalization of Vision-Language Models

Chenwei Wu [◇*] **Li Erran Li** [♠] **Stefano Ermon** [♡♠] **Patrick Haffner** [♠]
Rong Ge[◇] Zaiwei Zhang[♠]
[◇]Department of Computer Science, Duke University
[♠]AWS AI, Amazon
[♡]Department of Computer Science, Stanford University

Abstract

Compositionality is a common property in many modalities including text and images, but the compositional generalization of multi-modal models is not well-understood. In this paper, we identify two sources of visual-linguistic compositionality: linguistic priors and the interplay between images and texts. We show that current attempts to improve compositional generalization rely on linguistic priors rather than on information in the image, as the strength of the language model in detecting sentences that are syntactically and semantically likely overwhelms the vision part of the model. We find in particular that a benchmark for compositionality mostly favors pure language models. Finally, we propose a new benchmark for compositionality without such linguistic priors.

1 Introduction

Compositional generalization is the ability to combine known concepts in a novel way. It exists in multiple modalities, and is believed to be a key ingredient to human intelligence. For example, after knowing the meanings of "a person wearing a shirt" and "a cat playing with a dog", one should be able to understand "a person playing with a cat". Similarly, after seeing a wood chair and a green apple, one could imagine a green chair. Human are known to have good compositional generalization ability. However, it is not clear whether neural networks have these capabilities, especially in a multi-modal setting.

On the one hand, text-to-image generative models, including DALL·E 2 [24], Parti [31], and Imagen [25], are good at generating novel combinations of known concepts, e.g., an avocado chair or a snake made of corn. On the other hand, previous research have shown that many multi-modal networks lack compositional generalization [29]. This seems contradictory.

Moreover, a multi-modal network obtains information from multiple modalities. It can utilize information from a single modality, and it can also benefit from combining information from multiple modalities. As for compositional generalization, it is not clear whether this ability comes from a single modality or multiple ones. All this raises many questions. Where does the compositional generalization of multi-modal models come from? How do we measure multi-modal compositional generalization and resolve the contradiction mentioned above?

[*]Work done during internship at Amazon

NeurIPS 2023 Workshop on I Can't Believe It's Not Better (ICBINB): Failure Modes in the Age of Foundation Models

To answer these questions, we analyze the multi-modal compositional generalization in a more fine-grained way. We discovered the capability of language models on vision-language compositional generalization benchmarks and found that language models can beat most vision-language models without looking at the images, demonstrating a potential bias of these benchmarks.

We then decompose the source of compositional generalization into two parts: the uni-modal part and the multi-modal part. We illustrate the contribution of these two parts for commonly-used models. We also propose a new metric "hard test accuracy" to quantify the contribution from linguistic priors into the overall compositional generalization ability of vision-language models.

2 Related Works

Compositional generalization is frequently studied in natural languages and is also discussed in other domains including images [27], videos [7], and programming languages [6]. It can be measured by various benchmarks about different aspects: In the realm of natural languages, notable benchmarks include SCAN [14], COGS [13], CommonGen [17], etc. For multi-modal domains including vision-language, it can be measured by different tasks, e.g., visual question answering [2, 3, 10], verb understanding [8], counting [21], image-text retrieval [26], and word ordering [29]. People have also proposed different metrics and fine-grained aspects [1, 11, 12] with benchmarks to measure compositional generalization ability.

Extensive analyses have been conducted using these benchmarks to understand compositional generalization. Studies [9, 15, 19] have revealed that models like BERT are generally insensitive to word ordering, while vision-language models [20, 32] often struggle with tasks involving word order, relationship recognition, or counting.

Furthermore, the interaction among modalities in multi-modal models has been a subject of research. Key findings include the language bias in solving the "FOIL it!" benchmark [18], the phenomenon of "unimodal collapse" [20], and the asymmetry of information flow between image and text modalities [5].

3 Preliminaries

3.1 ARO Datasets: A benchmark for multi-modal compositional generalization

The ARO dataset [32] is designed to measure the compositional generalization of visual-language models. This dataset consists of 4 parts: Visual Genome Attribution (VG_A), VG Relation(VG_R), COCO Ordering(CC_O), and Flickr Ordering(FL_O). Each data point in this dataset contains an image with multiple captions (2 captions for VG tasks and 5 for Ordering tasks). One of these captions is the true caption corresponding to the image, and the other ones are corrupted captions that were obtained by changing the attribution, relation, or word ordering of the true caption. Given the image and the captions, the model is asked to pick the true caption among all candidates. Examples of the data in the ARO dataset are shown in Table 1, and the details of this dataset can be found in [32].

3.2 Performance of Vision-language Models on ARO Datasets

In [32], the authors computed the prediction accuracy of some most popular vision-language models. Given a pair of image and text, the model computes a score between 0 (or -1) and 1 reflecting the alignment between the image and the text, i.e., the degree to which the text matches the image. Perfectly-aligned image and text have a score of 1. Following [32], we use CLIP [22], BLIP [16], and FLAVA [28] to replicate their results, as shown in the top half of Table 2. BLIP and FLAVA have multiple heads that could be used to match images with texts, and we compute the accuracy of all these heads.

As noted in [32], the compositional generalization performance of commonly-used vision-language models are usually better than chance, but far from satisfying. This indicates the lack of compositional generalization ability for these models.

Table 1: Example captions from ARO datasets. VG_A/R for VG Attribution/Relation, CC/FL_O for COCO/Flickr Ordering

DATASET	CAPTION EXAMPLES
VG_A	**T**: THE WOOD FLOOR AND THE BLACK BAG **F**: THE BLACK FLOOR AND THE WOOD BAG
VG_R	**T**: THE PLATE IS ON THE TABLE **F**: THE TABLE IS ON THE PLATE
CC_O	**T**: A FIRE HYDRANT ON A CITY STREET **F**: ON A CITY A FIRE HYDRANT STREET
FL_O	**T**: GROUP GATHERED TO GO SNOWMOBILING **F**: GROUP GO SNOWMOBILING GATHERED TO

4 Pure Language Models Perform Well on ARO Datasets

We may see from Table 1 that we can easily pick the true caption by only looking at these captions alone because the true captions usually are more likely to happen. For instance, it is very unlikely that a table is on a plate or a bag is made of wood. Besides, previous works showed that large language models are very powerful in various tasks [4]. Therefore, by utilizing the knowledge only from the text side, it is possible to perform pretty well on these datasets. To validate our hypothesis, we would like to compute the prediction accuracy of language models on these ARO datasets.

4.1 Prediction Method

Predicting true caption with pure language modeling: Given multiple captions of the same length, we compute their perplexity and predict the caption with the smallest perplexity as the true caption. Concretely, given an input sentence $x = (x_1, x_2, \cdots, x_n)$ that contains n tokens, its perplexity computed by a language model LM is defined as

$$PP_{LM}(x) := -\frac{1}{n-1}\sum_{i=2}^{n} \log \Pr_{LM}\left(x_i | x_1 \cdots x_{i-1}\right). \tag{1}$$

Here $\Pr_{LM}\left(x_i | x_1 \cdots x_{i-1}\right)$ is the predicted probability of the i-th token being x_i given by the language model LM conditioning on the previous sequence $x_1 \cdots x_{i-1}$.

Language Models Selection: Since the number of parameters in the text encoder of vision-language models is usually at the scale of 100M or above, we choose the smallest version of GPT-2 [23], which has about 117M parameters, to do a fair comparison. To further investigate the compositional generalization of language models, we also use larger language models with the number of parameters ranging from 1B to 6B. These models include GPT-2-XL [23], OPT-1.3B [33], and GPT-J-6B [30]. To reduce the variance in the prediction, we use an unweighted average version of GPT-2-XL, OPT-1.3B, and GPT-J-6B, i.e., for each caption, we use these three language models to compute the perplexities and take the average, and select the caption with the lowest average perplexity. We call this predictor "AVG_LM", and this predictor will be used in later sections to build our new metric.

Performance of Language Models on ARO Datasets: The prediction accuracies of the language models are shown in the bottom half of Table 2.

From Table 2, we notice that for the Ordering tasks, GPT-2 is almost perfect in selecting the correct caption among all 5 candidates. This is intuitive because the false captions are obtained by permuting the words of the true caption in a fairly random way and usually destroys the sentence structure. In VG Relation, although the accuracy of GPT-2 is not close to perfect, it still outperforms the vision-language models. In VG Attribution, GPT-2 is slightly worse than BLIP, but are significantly better than the other models. Therefore, with roughly the same number of parameters, GPT-2 can achieve better or on-par performance than vision-language models. If we increase the number of parameters of the language model, the performance can be further increased.

To summarize, in most cases, it is possible to beat the vision-language models in compositional generalization by simply looking at the texts. Therefore, it is natural to ask: Does the compositional

Table 2: Prediction Accuracies on ARO datasets. ITC/ITM/CONTR are ITC/ITM/contrastive heads of models. Language Model Details in Section 4.1.

	CC_O	FL_O	VG_A	VG_R
CLIP	47.73	59.12	61.35	59.79
CLIP_FT	31.18	40.96	62.41	60.01
BLIP ITC	13.21	17.60	80.00	41.59
BLIP ITM	20.83	25.02	**88.56**	53.35
FLAVA CONTR	7.03	18.66	60.37	29.13
FLAVA ITM	4.81	13.20	68.83	23.71
GPT-2	95.28	96.26	76.57	85.07
GPT-2 XL	95.79	96.68	80.64	85.07
OPT 1.3B	**98.00**	**98.32**	81.60	84.30
GPT-J 6B	94.71	95.82	82.66	85.36
AVG_LM	97.33	97.70	84.48	**85.67**

generalization ability of these vision-language models mostly come from the text side? Answering this question requires a fine-grained analysis, which we show in the next section.

5 Fine-grained Analysis of Multi-modal Compositional Generalization

The compositional generalization ability of multi-modal models consists of two parts: uni-modal compositional generalization and its multi-modal counterpart. In the case of ARO datasets, since the image is fixed for each datum, the only two possible sources of information are the text alone and the interplay between the image and the text. To investigate the contribution of these two sources to the model performances, we analyze the prediction accuracy of models conditioned on the properties of caption candidates.

Testbed Selection: In [32], the authors proposed a new method to fine-tune CLIP on COCO. For each image in the COCO training set, they generate negative captions by randomly swapping two content words or phrases with the same part of speech as in the true caption.[2] The resulting model is called NegCLIP. They also fine-tuned CLIP on COCO normally without any additional hard negatives, resulting in the model CLIP_FT. Table 3 shows the VG Attribution accuracies of these three models, and NegCLIP clearly outperforms CLIP_FT and CLIP. Since these models are obtained in very similar ways, we decided to use our framework to understand where this improvement comes from.

5.1 Prediction Accuracies of CLIP variants based on language prior

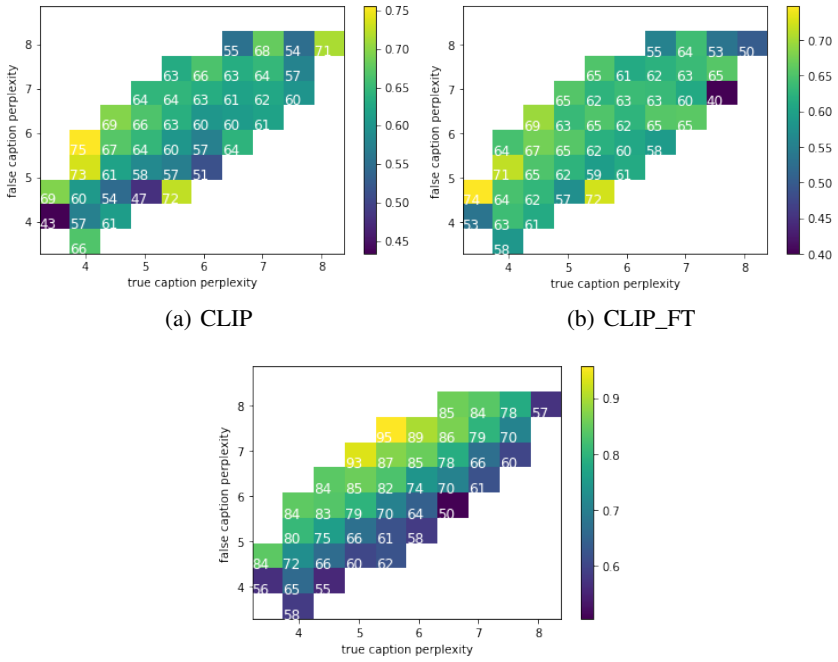
Similar to Section 4, we use the perplexity computed by AVG_LM to represent the possibility of a caption, and compute fine-grained model performance for input data that have similar perplexities for the caption candidates. Specifically, for VG Attribution, we divide the range of true/false caption perplexities evenly into 10 intervals[3], and this divides the input data into 100 blocks. We only compute the accuracy when there are at least 10 samples in a block.

Figure 1 shows the prediction accuracies of pre-trained CLIP, CLIP_FT, and NegCLIP based on linguistic priors. The smaller the true caption perplexity is, or the larger the false caption perplexity is, the easier it is for language models to select the true caption between the two candidates. In other words, the top-left parts are the easiest for language models, and the bottom-right parts are the hardest.

From Figure 1 we could notice that the prediction accuracies of both CLIP and CLIP_FT are relatively stable for different blocks, meaning their compositional generalization performance is

[2]They also did hard negative mining for images, which we do not include in this paper. However, our model trained with only hard negative text mining still matches their final performance. In this paper, NegCLIP refers to our replicated version which doesn't involve hard image negative mining.

[3]Note that the true caption has a slightly smaller average perplexity so the x-axis and y-axis are not perfectly aligned.



(a) CLIP (b) CLIP_FT

(c) NegCLIP

Figure 1: Prediction Accuracies of CLIP variants based on language prior. The x and y axis are the perplexities (computed by AVG_LM) of true and false captions, separately, and the number in each square shows the prediction accuracy (in percentile) of the vision-language model on data belonging to that particular region.

mostly independent from the perplexities of the caption candidates. However, for NegCLIP, the prediction accuracies for upper-left blocks are much higher than those of bottom-right ones. Therefore, the performance of NegCLIP has a strong correlation with the linguistic prior. Furthermore, despite the improvement in overall accuracy, the performance of NegCLIP for bottom-right blocks are similar to CLIP and CLIP_FT. Hence we hypothesize that NegCLIP mostly learns linguistic priors during its fine-tuning process.

5.2 Hard Test Accuracy and Linguistic Gap

To further quantify the alignment of these models with linguistic priors, we propose a new metric "hard test accuracy" to measure the dependency of the benchmark performance of models on linguistic priors of the captions.

To eliminate the effect of linguistic priors, we only selected the samples which are hard for the language models to classify, and evaluate the performance of vision-language models on these hard instances. For ARO datasets, we define "hard instances" to be the data where AVG_LM made mistakes, i.e., where the true caption has a higher perplexity than false ones. The number of hard instances is about 15% of the original dataset. We define "hard test accuracy" as the accuracy of a model on these hard instances, and define "linguistic gap" as the difference between the overall accuracy and the hard test accuracy.

The linguistic gap can reflect the dependency of the model's performance on linguistic priors. A model whose prediction is independent of the perplexities of the candidate captions should have a constant accuracy over hard and easy samples, yielding zero linguistic gap. A model relying on linguistic priors will be more accurate when the true caption has a lower perplexity than the false captions, so it will have a positive linguistic gap. Therefore, for models with a fixed overall accuracy, a larger linguistic gap implies a stronger dependency on linguistic priors.

Table 3 shows the overall and hard test accuracies of CLIP variants on VG Attribution. Similar to our intuition from Figure 1, the linguistic gap for CLIP and CLIP_FT is significantly smaller than that of NegCLIP. We also notice that the hard test accuracy of NegCLIP is almost the same as CLIP_FT, so the effect of the extra negative captions only helps the model learn linguistic priors and does not help with the interplay between image and text.

Table 3: Accuracies and Linguistic Gap on VG Attribution

	TOTAL ACC	HARD ACC	LING. GAP
CLIP	61.35	57.92	3.43
CLIP_FT	63.17	60.43	2.74
NEGCLIP	74.90	61.66	13.24
BLIP ITC	80.00	76.63	3.37
BLIP ITM	88.56	88.08	0.48
FLAVA CONTR	60.37	58.91	1.46
FLAVA ITM	68.83	71.10	-2.27

5.3 Hard Test Accuracy on the VG Attribution Dataset

Using this framework, we also evaluate the hard test accuracies for commonly-used vision-language models, including different heads of BLIP and FLAVA, as shown in Table 3. Although different models and different heads vary in their overall prediction accuracies, these models do not have a large linguistic gap, so their prediction on VG Attribution is relatively independent of the linguistic prior. This framework can be applied on other ARO datasets as well. The conclusions are similar to VG Attribution, and the detailed results are in Appendix A.

6 Conclusion and Future Work

In this paper, we studied the multi-modal compositional generalization ability of vision-language models. We showed that language models with similar sizes could achieve better performances than vision-language models on vision-language compositional generalization benchmarks without even taking the images as inputs, demonstrating the important role of linguistic priors in vision-language compositional generalization. We then decomposed the model performance on these benchmarks into a uni-modal part and a multi-modal part, and proposed a new metric to quantify these two parts.

For future directions, we would like to apply our framework to more models, benchmarks, and modalities. It will also be interesting to explore better ways of generating hard negatives for these models. Finally, it is more important to dig deeper into the reasons why multi-modal networks have this compositional generalization behavior. The reasons may lie in the shared embedding space for different modalities.

References

[1] Jacob Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

[2] Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*, 2021.

[3] Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. Covr: A test-bed for visually grounded compositional generalization with real images. *arXiv preprint arXiv:2109.10613*, 2021.

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

[5] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

[6] Yujian Gan, Xinyun Chen, Qiuping Huang, and Matthew Purver. Measuring and improving compositional generalization in text-to-sql via component alignment. *arXiv preprint arXiv:2205.02054*, 2022.

[7] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021.

[8] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

[9] Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, 2021.

[10] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[11] 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.

[12] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*, 2019.

[13] Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*, 2020.

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

[15] Karim Lasri, Alessandro Lenci, and Thierry Poibeau. Word order matters when you increase masking. *arXiv preprint arXiv:2211.04427*, 2022.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[17] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*, 2019.

[18] Pranava Madhyastha, Josiah Wang, and Lucia Specia. Defoiling foiled image captions. *arXiv preprint arXiv:1805.06549*, 2018.

[19] Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. When classifying grammatical role, bert doesn't care about word order... except when it matters. *arXiv preprint arXiv:2203.06204*, 2022.

[20] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

[21] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*, 2020.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[26] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

[27] Zhan Shi, Hui Liu, Martin Renqiang Min, Christopher Malon, Li Erran Li, and Xiaodan Zhu. Retrieval, analogy, and composition: A framework for compositional generalization in image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1990–2000, 2021.

[28] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[30] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

[31] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[32] 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 e-prints*, pages arXiv–2210, 2022.

[33] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Hard Test Accuracy on other ARO datasets

We also apply the framework of hard test accuracy on the other ARO datasets. However, since language models perform almost perfectly on the Ordering datasets, the number of hard instances is very small (400+ for COCO ordering and less than 100 for Flickr Ordering). As for VG Relation, when evaluating hard test accuracy, we only keep the relationships("hard relationships") that contain at least 10 hard instances, which reduces the number of relationships by 80%. Therefore, for VG Relation, we compute both the prediction accuracy on all data in hard relationships (hard total accuracy) and on hard instances in hard relationships (hard test accuracy). The results are shown in Table 4. The results for Flickr Ordering are very close to COCO Ordering and we omit them due to space constraints. Similar to VG Attribution, both CLIP and CLIP_FT have much smaller linguistic gaps than NegCLIP, and NegCLIP even has the worst hard test accuracy, meaning it aligns so much with the linguistic prior that the prediction power on hard instances got reduced.

Table 4: Overall and Hard Accuracies on COCO Ordering and VG Relation, format: "overall accuracy / hard test accuracy" for COCO Ordering, "overall accuracy / hard overall accuracy / hard test accuracy" for VG Relation.

	COCO_ORDER	VG RELATION
CLIP	47.73 / 43.56	59.79 / 53.52 / 55.90
CLIP_FT	31.18 / 27.10	60.01 / 53.11 / 50.48
NEGCLIP	91.74 / 71.56	80.50 / 60.94 / 48.62
BLIP ITC	13.21 / 17.51	41.59 / 52.02 / 58.31
BLIP ITM	20.83 / 23.95	53.35 / 62.26 / 64.80
FLAVA CON	7.03 / 4.79	29.13 / 38.17 / 36.00
FLAVA ITM	4.81 / 4.79	23.71 / 37.20 / 33.91

B More Experimental Details

Image-Text Matching Score Computation: For models that output embedding vectors, we compute the inner product of the normalized image and text embeddings and use this number as the matching score. For models that directly compute a score for image-text matching, we use this score directly. After getting the scores, for each datum in the ARO datasets, since the image is fixed, we predict the text with the highest matching score as the true caption and compute the prediction accuracy.

Model Details: We obtain CLIP [22], BLIP [16], and FLAVA [28] models using the same way as in [32], and download all language models from huggingface.