VerbCLIP: Improving Verb Understanding in Vision-Language Models with Compositional Structures

Hadi Wazni, Kin Ian Lo, Mehrnoosh Sadrzadeh
University College London
{hadi.wazni.20, kin.lo.20, m.sadrzadeh}@ucl.ac.uk

Abstract

Verbs describe the dynamics of interactions between people, objects, and their environments. They play a crucial role in language formation and understanding. Nonetheless, recent vision-language models like CLIP predominantly rely on nouns and have a limited account of verbs. This limitation affects their performance in tasks requiring action recognition and scene understanding. In this work, we introduce VerbCLIP, a verb-centric vision-language model which learns meanings of verbs based on a compositional approach to statistical machine learning. Our methods significantly outperform CLIP in zero-shot performance on the VALSE, VL-Checklist, and SVO-Probes datasets, with improvements of +2.38%, +3.14%, and +1.47%, without fine-tuning. Fine-tuning resulted in further improvements, with gains of +2.85% and +9.2% on the VALSE and VL-Checklist datasets.

1 Introduction

Trained on extensive datasets of image-caption pairs, current vision-and-language models (VLMs) excel in various applications, yet stall in tasks that require structural knowledge and compositional reasoning (Thrush et al., 2022; Liu et al., 2023). Research by (Yuksekgonul et al., 2023; Lin et al., 2024) demonstrates some of the difficulties they face in understanding attributes, relationships, and order information. More specifically, (Hendricks and Nematzadeh, 2021) points out that VLMs often fail to distinguish between different verbs, instead relying predominantly on noun understanding. One possible reason for this issue is the inherent biases within the training datasets. These datasets host a limited number of examples where captions share similar contexts but differ in verbs. As a result, they focus on specific objects and subjects, with minimal emphasis on verbs. This tendency is a form of shortcut learning, a phenomenon in deep neural networks where models opt for simpler, superficial solutions over a deeper understanding (Geirhos et al., 2020).

Conversely, Compositional Distributional Semantic models (CDSMs) (Erk and Padó, 2008; Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Coecke et al., 2010) learn meaning representations of sentences by considering their compositional linguistic structures, such as the relationships between verbs and their subjects and objects. In the model proposed by (Baroni et al., 2014), verbs are represented as tensors that take lower-order word representations, typically vectors, as arguments. This means that intransitive verbs are represented as matrices, transitive verbs as cubes, and ditransitive verbs as hypercubes. These tensor-based representations have shown promising results in tasks such as verb disambiguation and sentence similarity (Kartsaklis and Sadrzadeh, 2013; Grefenstette et al., 2013). CDSMs have primarily been applied to text-only data and tasks, and have recently been used as text encoders for CLIP (Lewis et al., 2023).

The novel contribution of this paper lies in integrating VLMs with CDSMs within a framework called VerbCLIP to enhance verb understanding. We implement various methods for learning verb tensors on an image-caption matching task and evaluate these methods on VALSE, VL-Checklist, and SVO-Probes datasets. Our best tensor learning method achieves improvements of +2.38%, +3.14%, and +1.47% over CLIP. Beyond these quantitative improvements, a significant advantage of VerbCLIP is that it does not require training from scratch. Our code and data are available at https://github.com/lanlos-lab/verbclip.

2 Methodology

We present an overview of our framework, illustrated in Figure 1. It utilises frozen CLIP as the backbone. Initially, we input the original sentence

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 195–201
August 16, 2024

and image into CLIP's encoders to obtain a similarity score, reflecting the overall alignment between the general semantics of the text and the image. Simultaneously, we extract the subject-verb-object triplet from the sentence. These components are encoded separately: the subject and object as vectors, and the verb as matrices, forming a compositional text embedding that captures the detailed semantic relationships. We then calculate a similarity score between the compositional text embedding and the image embedding. We add the two scores to produce the final matching score.

2.1 Compositional Distributional Semantics Models (CDSMs)

We consider a number of compositional distributional semantics models, which have been proposed in past work but have not been applied to a visually grounded language setting. Table 1 represents the algebraic formulas used in our experiments.

Vector-based Models Following the work of (Mitchell and Lapata, 2008), vector-based models compute a sentence vector as a mixture of the original word vectors, using simple operations such as element-wise multiplication and addition. Multiplication can be seen as the intersection of features, while addition resembles the union. The main characteristic of these models is that they do not distinguish between the type-logical identities of different words. For example, an intransitive verb is considered of the same order as its subject (a noun), and both will contribute equally to the composite sentence vector.

Tensor-based Models Following the work of (Baroni and Zamparelli, 2010) and (Coecke et al., 2010), relational words such as verbs and adjectives are represented by multilinear maps (tensors). Meanings of words are composed through the application of these maps to vectors representing the arguments (usually nouns). These models offer a more linguistically motivated solution to the problem of composition, effectively addressing the 'bag of words' issue. However, a practical difficulty is that the creation and usage of third-order tensors can be computationally expensive. One solution is to first create a matrix presentation of the verb, which is then expanded to a tensor by applying the Frobenius coproduct (copying) map to either the left or right axis, resulting in the *Copy-Subject* and *Copy-Object* methods (Kartsaklis et al., 2012; Kartsaklis and Sadrzadeh, 2014). This map can be visualised as placing a matrix along a specific diagonal of a tensor. In this work, we propose a new method: *Copy-Add*.

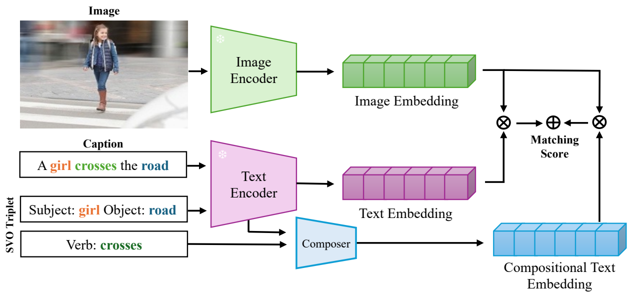

Figure 1: The VerbCLIP framework makes use of two types of text embeddings: the *Text Embedding*, which captures the meaning of the entire caption; and the *Compositional Text Embedding*, which captures the syntactically sensitive meaning by combining word-level embeddings of the subject, verb, and object.

Copy-Subject The semantic interpretation of a transitive sentence involves a two-step compositional process. Initially, the verb's meaning is applied to the object, creating an intermediate representation that highlights how the verb's action targets the object. This result is then applied to the subject, integrating the roles of both subject and object with the verb's action to construct the overall sentence meaning. This approach effectively combines the individual meanings to reflect the sentence's complete semantic structure.

$$\overrightarrow{subj\ verb\ obj} = \overrightarrow{subj} \odot \left(\overline{verb} \times \overrightarrow{obj}\right)$$

Copy-Object The meaning of a transitive sentence is derived by first applying the verb's meaning to the subject, and then combining the result with the meaning of the object. Similarly, this process helps form a coherent semantic output by sequentially engaging the subject and object with the verb.

$$\overrightarrow{subj\ verb\ obj} = \left(\overrightarrow{subj} \times \overline{verb}\right) \odot \overrightarrow{obj}$$

Copy-Add Combining the *Copy-Subject* and *Copy-Object* methods provides a more comprehensive representation of the verb and enhances the sentence meaning. Here the parameters α and β can be trained to balance and optimise the combination, reducing biases and improving overall semantic interpretation.

$$\overrightarrow{subj\ verb\ obj} = \alpha \left[\overrightarrow{subj} \odot \left(\overline{verb} \times \overrightarrow{obj}\right)\right] + \beta \left[\left(\overrightarrow{subj} \times \overline{verb}\right) \odot \overrightarrow{obj}\right]$$

Method	Algebric Formula
Add	$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + (\overrightarrow{s} + \overrightarrow{v} + \overrightarrow{o}) \cdot \overrightarrow{I_{img}}$
Mult	$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + (\overrightarrow{s} \odot \overrightarrow{v} \odot \overrightarrow{o}) \cdot \overrightarrow{I_{img}}$
Copy-Subject	$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + (\overrightarrow{s} \odot (\mathbf{V} \times \overrightarrow{o})) \cdot \overrightarrow{I_{img}}$
Copy-Object	$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + ((\overrightarrow{s} \times \mathbf{V}) \odot \overrightarrow{o}) \cdot \overrightarrow{I_{img}}$
Copy-Add	$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + (\alpha[\overrightarrow{s} \odot (\mathbf{V} \times \overrightarrow{o})] + \beta[(\overrightarrow{s} \times \mathbf{V}) \odot \overrightarrow{o}]) \cdot \overrightarrow{I_{img}}$

Table 1: Compositional methods used in this study with their corresponding algebraic formulas. We make use of element-wise product $\odot$, matrix multiplication $\times$, and $\cdot$ dot product. The vectors $\overrightarrow{s}$, $\overrightarrow{v}$, and $\overrightarrow{o}$ are text embeddings for the subject, verb, and object entities respectively. $\overrightarrow{T_{sent}}$ and $\overrightarrow{I_{img}}$ are holistic embeddings for the input text and image. By default, we let $\alpha, \beta = 1$.

2.2 Creating verb tensors

We review several proposals for constructing tensors for verbs and opt to use matrices in our work. Matrices often perform as well as, or even better than, full tensors, thereby reducing the number of parameters needed in our framework (Polajnar et al., 2014).

Kronecker In work of (Grefenstette and Sadrzadeh, 2011b), the verb matrix is created as the outer product[1] of the verb vector with itself:

$$\overline{verb} = \overrightarrow{verb} \otimes \overrightarrow{verb}$$

Relational Following ideas from the set-theoretic view of formal semantics, (Grefenstette and Sadrzadeh, 2011a) suggest that the meaning of a verb is the sum of the outer product of its arguments (subject, object) over all occurrences of the verb in a corpus. This is represented as:

$$\overline{verb} = \frac{1}{N} \sum_{i=1}^{N} \overrightarrow{subj}_i \otimes \overrightarrow{obj}_i$$

where N is the number of examples. The intuition is that the matrix encodes higher weights to the contextual features of subjects and objects that are frequently observed together.

Linear Regression Building on the concept introduced by (Baroni and Zamparelli, 2010) of creating adjective matrices, we propose a verb matrix A, when applied to the vector representation of a noun (as either a subject or object), yields a vector that effectively captures the distributional semantics of the combined subject-verb or verb-object phrase. For example, for the verb-object compound "eat food", we compute the verb matrix A_{eat}, such that $\overrightarrow{y} = A_{eat} \times \overrightarrow{food}$, where $\overrightarrow{food}$ represents the distributional vector of "food" and $\overrightarrow{y}$ reflects the semantic composition of "eat food". To find matrix A, we minimise the discrepancy between the predicted vectors and the actual distributional vectors. This optimisation can be achieved through gradient descent or analytically[2], $A_{eat}^T = (X^T X)^{-1} X^T Y$, where the rows of matrix X are vectors of objects found in the corpus as arguments of the verb, and the rows of Y are the vectors of the corresponding verb-object phrases. A similar procedure is used to create matrices for subject-verb phrases.

3 Experiment

We focus on the task of matching images with correct captions. An image is described by both a positive and a negative caption; the negative caption differs from the positive only by a verb. Our aim is to achieve a higher matching score between the image and the positive caption compared to the negative one.

Evaluation Datasets We test our methods on VALSE (Parcalabescu et al., 2022), VL-Checklist (Zhao et al., 2023), and SVO-Probes (Hendricks and Nematzadeh, 2021). Detailed descriptions of the datasets are in the above papers; however, for this study, we selected only those entries where the verb differs between the positive and negative captions, while the subjects and objects are the same. For the SVO-Probes, we create negative captions by substituting the verb in the positive caption with its corresponding negative form from the given negative (SVO) triplet. For example, given a positive caption *'a woman is **running** in the field'* and a

[1] It is the tradition in the literature to use the Kronecker product to form a vector in a tensor-product space. In this work we use the outer product to obtain a matrix instead.

[2] The analytical formula fails when X is not full rank. In such cases, the Moore-Penrose pseudoinverse shall be used.

Method	VALSE			VL-Checklist			SVO-Probes		
	Kron	**Rel**	**Reg**	**Kron**	**Rel**	**Reg**	**Kron**	**Rel**	**Reg**
Copy-Subject	74.76	74.29	74.29	59.53	58.80	58.49	78.74	**78.90**	69.28
Copy-Object	72.86	72.86	73.33	58.53	56.62	52.56	78.35	78.85	70.63
Copy-Add	**75.24**	72.86	75.24	**60.41**	57.85	59.53	77.30	78.44	69.27
Copy-Add FT	75.71	74.29	<u>77.62</u>	<u>66.47</u>	65.47	62.90	77.30	78.49	69.28

Table 2: Comparison of accuracy (%) across three datasets using tensor-based methods. Verb matrices are built with Kronecker (Kron), Relational (Rel), and Regression (Reg) methods using the ViT-B/32 variant of CLIP.

Method	VALSE	VL-Checklist	SVO-Probes
Add	**74.76**	**60.00**	77.64
Mult	73.33	57.83	**78.68**
CLIP	72.86	57.27	77.43

Table 3: The accuracies (%) of vector-based methods using ViT-B/32. For CLIP, image embeddings are generated by CLIP's vision encoder (ViT-B/32); and text embeddings are generated by CLIP's text encoder. We compute the dot product between the image and the text embeddings to obtain the matching score.

negative verb *'walk'*, the resulting negative caption would be *'a woman is **walking** in the field'*. Out of the 14,097 images in the SVO-Probes dataset, 11,769 images were accessible from the internet in February 2024.

Data We extracted all subject-verb-object (SVO) triplets associated with each verb in the three datasets from the March 2022 English Wikipedia dump, using the dependency parser in spaCy. Then, we removed entries with pronouns, stop-words, and tokens that were less than three characters long. We prioritised the triplets, selecting only the top 2,000 subject-object pairs based on the frequency of occurrence. We ensured that for each verb, there were sufficient corresponding entries to build high-quality representation matrices. Verbs that failed to meet all the criteria were dropped. We ended up experimenting with 100 unique verbs in 210 entries from VALSE, 274 unique verbs in 9,407 entries from VL-Checklist, and 290 unique verbs in 14,544 entries from SVO-Probes.

4 Results and Discussion

The compositional tensor-based methods significantly outperform CLIP and vector-based models, with Copy-Add showing the highest performance in most cases. Copy-Add appears capable of utilising information from the combination of subject-verb and verb-object, and incorporating further information from the object and subject. This highlights the importance of ordering and syntactic information in the compositional methods. Upon fine-tuning the weights, α and β, we noticed further improvement (+2.85% and +9.2% on the VALSE and VL-Checklist datasets respectively).

We noticed lower performance improvements on the SVO-Probes dataset compared to VALSE and VL-Checklist. This discrepancy is likely due to the nature of the SVO-Probes dataset, which contains sketchy samples and tends to be noisy, with significant issues such as object mismatches, as detailed in (Castro et al., 2023; Jiang et al., 2024).

In terms of learning verb matrices, regression methods demonstrated lower accuracies, whereas the Kronecker (Kron) and Relational (Rel) methods performed better. The fact that Kron requires no training data makes it an effortless choice for constructing verb matrices, while still providing competitive performance.

In terms of verb-type performance, the Copy-Add model significantly improved accuracy for interaction-based verbs such as "hang" (+12.5%), "hold" (+11.6%), "attached" (+3.7%), and "take" (+29.62%). However, while it struggled with some visually static verbs like "stand" (-5.8%) and "sit" (-6.0%), it showed improvement in others such as "observe" (+50%) and "look" (+10.87%). Furthermore, we tested sentence pairs where the subject and object nouns are swapped, such as *"A **man** lies on the **sofa**"* vs *"A **sofa** lies on the **man**"*. CLIP often misinterprets these as equally plausible, reflecting its approach of processing text as independent words, similar to a bag-of-words approach. In contrast, Copy-Add model correctly identifies *"A **man** lies on the **sofa**"* as the correct caption by capturing structured detailed semantics. Overall, VerbCLIP

The goat *stands* in the grass.	A baby *speaks* on the telephone.	A person *holding* ski-poles.	A man *threw* the ball.
The goat *lies* in the grass.	A baby *sits* on the telephone.	A person *crossing* ski-poles.	A man *holding* the ball.

	Positive	Negative
CLIP	28.71	**28.73** ✗
VerbCLIP	**37.28** ✓	37.12

	Positive	Negative
CLIP	28.01	**28.11** ✗
VerbCLIP	**36.51** ✓	36.06

	Positive	Negative
CLIP	28.65	**28.68** ✗
VerbCLIP	**35.16** ✓	34.87

	Positive	Negative
CLIP	18.50	**19.54** ✗
VerbCLIP	**5.095** ✓	4.759

Figure 2: Examples where CLIP pairs images with incorrect text captions, as indicated by higher similarity scores for negative captions. In contrast, our framework achieves more accurate matching. The positive captions (marked in green) and negative captions (marked in red) are semantically very close, with the verb being different.

incorporates syntactic and semantic structures, allowing it to better understand context and dynamic actions.

5 Limitations

Creating verb matrices or tensors is computationally intensive, which poses a significant challenge when scaling to very large pretraining datasets. Additionally, our approach assumes a fixed linguistic structure, typically the subject-verb-object format, which does not account for the varied and flexible ways verbs are used in natural language. However, tensors are natural components of quantum systems, and quantum computing resources can efficiently learn them. The Quantum Natural Language Processing (QNLP) framework (Lorenz et al., 2023; Wazni and Sadrzadeh, 2023), inspired by categorical quantum mechanics and the DisCoCat (Distributional Compositional Categorical) framework, uses string diagrams to translate grammatical structures into quantum processes. This advanced option could offer a promising solution.

6 Conclusion

The CLIP model is noted for its limited ability to understand verbs, often relying heavily on nouns. Our approach seeks to mitigate this issue by introducing verb-focused compositional methods, which have demonstrated enhanced performance across the SVO-Probes, VL-Checklist and VALSE datasets. Our framework can boost the zero-shot inference capability of other models, such as SLIP (Mu et al., 2021) and BLIP (Li et al., 2022), without the need for further training or fine-tuning. Scaling to longer and more complicated sentences with varied grammatical structures is a work in progress.

7 Acknowledgement

The authors gratefully acknowledge the two anonymous reviewers for their valuable comments. Hadi Wazni would like to thank the UCL CS department for the PhD scholarship that supported this work. Kin Ian Lo was supported by the Engineering and Physical Sciences Research Council [grant number EP/S021582/1]. Mehrnoosh Sadrzadeh is grateful to the Royal Academy of Engineering for the Research Chair/Senior Research Fellowship RCSRF2122-14-152 on Engineered Mathematics for Modelling Typed Structures.

References

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for composition distributional semantics. *Linguistic Issues in Language Technology*, 9.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Santiago Castro, Oana Ignat, and Rada Mihalcea. 2023. Scalable performance analysis for vision-language models. *Preprint*, arXiv:2305.18786.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Preprint*, arXiv:1003.4394.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical*

Methods in Natural Language Processing, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

E. Grefenstette, G. Dinu, Y. Zhang, M. Sadrzadeh, and M. Baroni. 2013. Multi-step regression learning for compositional distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 131–142, Potsdam, Germany. Association for Computational Linguistics.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011b. Experimenting with transitive verbs in a discocat. *Preprint*, arXiv:1107.3119.

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

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. 2024. Comclip: Training-free compositional image and text matching. *Preprint*, arXiv:2211.13854.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601, Seattle, Washington, USA. Association for Computational Linguistics.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. In Proceedings of the 11th workshop on *Quantum Physics and Logic,* Kyoto, Japan, 4-6th June 2014, volume 172 of *Electronic Proceedings in Theoretical Computer Science*, pages 249–261. Open Publishing Association.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *Proceedings of COLING 2012: Posters*, pages 549–558, Mumbai, India. The COLING 2012 Organizing Committee.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. 2023. Does clip bind concepts? probing compositionality in large image models. *Preprint*, arXiv:2212.10537.

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

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2024. Revisiting the role of language priors in vision-language models. *Preprint*, arXiv:2306.01879.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. 2023. Qnlp in practice: Running compositional models of meaning on a quantum computer. *Journal of Artificial Intelligence Research*, 76:1305–1342.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. *Preprint*, arXiv:2112.12750.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Tamara Polajnar, Luana Făgărăşan, and Stephen Clark. 2014. Reducing dimensions of tensors in type-driven distributional semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1036–1046, Doha, Qatar. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.

Hadi Wazni and Mehrnoosh Sadrzadeh. 2023. Towards transparency in coreference resolution: A quantum-inspired approach. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 15–27, Singapore. Association for Computational Linguistics.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? *Preprint*, arXiv:2210.01936.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *Preprint*, arXiv:2207.00221.