ALIGNING VISUAL STRUCTURAL COMPOSITIONALITY IN HUMANS & VISION-LANGUAGE MODELS

Helena Balabin[1,2], Lauren Nicole De Long[3], Rohan Saha[2], Rik Vandenberghe[1], Marie-Francine Moens[1], Alona Fyshe[2]
[1]KU Leuven, 3000 Leuven, Belgium
[2]University of Alberta, T6G 2N8 Edmonton, AB, Canada
[3]Cancer Research UK Scotland Institute, EH4 2XR Edinburgh, UK

ABSTRACT

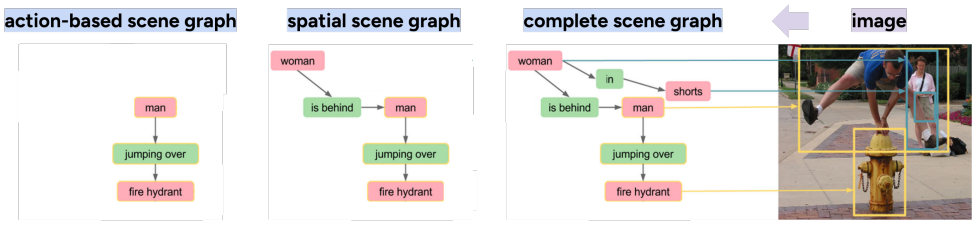
An open question across machine learning, neuroscience and cognitive science is whether current foundation models, in particular vision-language models (VLMs), learn representations that reflect human-like compositional processing. While linguistic compositionality is well-studied, the extent to which visual structural compositionality emerges in vision models remains under-explored. Here, we present a representational alignment probing framework that maps VLM embeddings to graph properties derived from human-annotated scene graphs in images and linguistic structures in text. Evaluating CLIP and several of its variants, we observe differences in alignment: While text encoders reliably reflect structural graph properties, vision encoders show limited alignment with visual relational structure. We then propose the GraphCLIP model architecture to more explicitly incorporate visual structural signal, but found no substantial performance improvements on our structural probing tasks.

1 INTRODUCTION

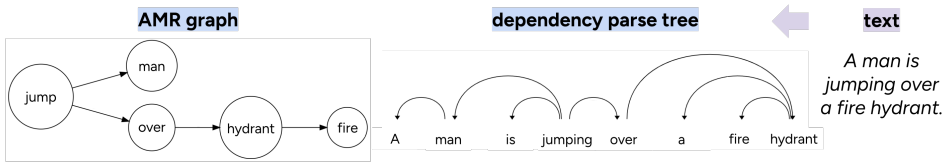
Despite substantial performance improvements on various tasks in recent years, it is not yet clear whether large language models (LLMs), and their multimodal counterparts, such as vision-language models (VLMs), form representations that reflect our human perception of the world. In particular, there is ongoing debate about whether compositionality (i.e., the ability to create new meaning from a complex expression based on the meaning of its individual parts and its combinatorial structure (Pelletier, 1994; Szabó, 2022)) naturally emerges from such model representations or not (Lake & Baroni, 2023; Lepori et al., 2023; Nichani et al., 2024; Song et al., 2025; Coopmans et al., 2022; Thomm et al., 2024; Dziri et al., 2023; Greff et al., 2020). By examining how compositional structures are encoded in models versus how they are interpreted by humans, one can not only identify where models do or do not align with human-like representational geometries, but also steer models' downstream behavior by explicitly incorporating such structural representations (Sucholutsky et al., 2025). Direct probes for compositional structures have been primarily applied to language models (Rogers et al., 2020; Arps et al., 2022), whereas most multimodal compositional probes have been performed more indirectly in the form of visual question answering or text-image matching tasks (Johnson et al., 2017; Thrush et al., 2022). This disparity may be largely attributed to existing well-defined linguistic structures such as dependency parse trees or Abstract Meaning Representation (AMR) graphs (i.e., graph structures that relate semantic roles and their interactions (Banarescu et al., 2013)), whereas similar structures tend to be less established in vision. However, scene graphs comprising humans or objects and relationships between them (Chang et al., 2023) provide an analogous framework to study the same phenomenon in vision.

Therefore, we leverage human scene graph annotations from the Visual Genome (VG) (Krishna et al., 2017) and Contrastive Language-Image Pretraining (CLIP)-based VLM representations (Radford et al., 2021) to address the following two research questions: (1) Is visual structural compositionality emergent in VLM representations to the same extent as textual compositionality? And, (2) can additional visual structural compositional signal in VLMs improve compositionality in VLM representations? By introducing novel probing tasks targeted at testing structural compositionality

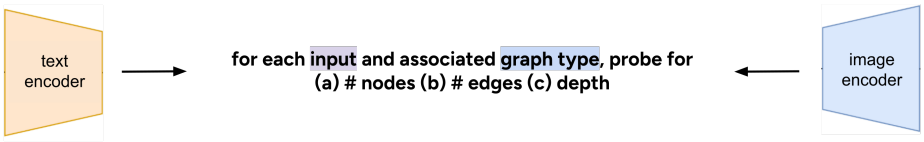
specifically within vision encoders of VLMs, we present evidence that (1) visual structural compositionality does not appear to emerge in the same human-like manner in vision as it does in language. However, (2) incorporating additional structural compositional signal in VLM representations within our proposed GraphCLIP model does not yield substantial performance gains on the probing tasks.



(i) Three graph types derived from scene graphs from images



(ii) Two graph types derived from text



(iii) Three probing tasks per graph type

Figure 1: **Probing setting.** To examine structural compositionality, we first leverage text-image pairs to derive three different types of scene graphs from images, and two graph types from the matching text. Then, by passing the text and images through the CLIP-based text- and image encoders, respectively, we set up three probing tasks for each input type and each associated graph type. The example shown in this figure is simplified, both image and text graphs can have a substantially higher number of nodes and edges. The edge "jumping over" is included in both the action-based and spatial scene graphs, as it includes both an action (jumping) and a spatial relation (over). *AMR, Abstract Meaning Representation.*

2 Related Work

Existing probing frameworks have provided evidence for emergent linguistic structures in language models by mapping transformer-based representations onto dependency parse trees (Hewitt & Manning, 2019; Rogers et al., 2020; Diego-Simón et al., 2024). While single elements of language (e.g., words) are connected through clear and discrete structures, visual elements like objects are often perceived as collections of separate, disconnected fragments. However, findings in cognitive science suggest that human vision relies on hierarchical and often action-centered structures similar to linguistic syntax (Võ, 2021; Hafri & Firestone, 2021; Bartnik & Groen, 2023). While previous work on vision transformers has examined how hierarchical or spatial structures may emerge from associations across different image patches (Fan et al., 2021; Jelassi et al., 2022), our study leverages human annotations on objects and their relationships in the form of scene graphs (Krishna et al., 2017) to probe visual structures in a manner that aligns more with linguistic structures and human cognition.

The introduction of benchmarks targeted at testing multimodal compositional processing such as Winoground (Thrush et al., 2022), Compositional REPresentation Evaluation (CREPE) (Ma et al., 2023) and SugarCREPE(++) (Hsieh et al., 2023; Dumpala et al., 2024) have highlighted that VLMs such as CLIP often fail to distinguish between similar text-image pairs with subtle differences in attribute binding (e.g., plants surrounding a light bulb versus a light bulb surrounding plants (Thrush

et al., 2022)). CLIP variants such as Structure-CLIP (Huang et al., 2024), TripletCLIP (Patel et al., 2024), LaCLIP (Fan et al., 2023), HyCoCLIP (Pal et al., 2025) and SigLIP (Zhai et al., 2023) aim to overcome this shortcoming, mainly by using or augmenting data with subtle differences in attribute binding to generate harder input examples for the models' the contrastive learning objective, resulting in improved benchmark performances. Nevertheless, by design, the aforementioned benchmarks focus on downstream tasks and hence present an indirect measure of the presence of compositional structure in VLMs. In contrast, our study aims at testing whether such structure is intrinsically present in vision encoders within CLIP and its variants.

3 METHODS

To investigate whether visual structural compositionality is emergent in VLM representations to the same extent as textual compositionality, we perform our experiments on parallel text-image data (see Figure 1). For each pair, we derive dependency parse trees and AMR graphs for text, and complete, spatial, and action-based scene graphs for images. Because compositional representations are based on their number of constituents, their relations, and resulting hierarchical depth, the linear decodability of these graph properties from VLM embeddings is serving as a probe of structural compositionality. Here, we focus on VLMs rather than on separate language and vision models, since their shared embedding space enables a more stringent cross-modal comparison of structural compositionality. Qualitative examples of the included tested graph structures are shown in Figure 1i and 1ii.

Data We base our probing tasks on a subset of overlapping entries of human-annotated scene graphs in the VG (Krishna et al., 2017) and captions in the Common Objects in Context (COCO) (Lin et al., 2014) dataset ($28,459$ images with ≈ 5 captions each $= 142,371$ text-image pairs). VG scene graphs were created based on images alone, without exposure to captions, using free-form object labels, attributes, and pairwise relationships, with multiple human annotations per image merged via normalization and canonicalization. We withhold the remaining overlapping VG-COCO entries ($115,212$ text-image pairs, $23,039$ images) to pre-train our proposed GraphCLIP variant (see below). To subsequently derive properties that reflect either textual or visual structural compositionality, we leverage a variety of different graph structures using either the text or image within each text-image pair.

For images, we start from the complete scene graph provided in the VG metadata and apply two filters to the nodes and edges in the complete scene graphs to derive two additional types of scene graphs that are focused on spatial relationships and actions, respectively (see Figure 1i). With respect to spatial scene graphs, we only include edges that contain prepositions, and nodes connected by such edges. Analogously, for action-based scene graphs, we filter by edges that describe actions (using by a fixed set of visual action verbs defined in the COCO-actions dataset (Ronchi & Perona, 2015)) and nodes connected by them. Both spatial and action-based scene graphs are therefore subsets of complete scene graphs, but not necessarily subsets of one another. Next, for text, we derive dependency parse trees and AMR graphs using the spacy and the amrlib python library based on the AMRBart model (Bai et al., 2022), respectively (see Figure 1ii). The resulting distribution of the (a) number of nodes, (b) number of edges and (c) depth for the resulting graphs of the five different graph structures is shown in Figure A1. Excluding graphs without any edges, we obtain a total of $142,359$ AMR graphs, $142,371$ dependency parse trees, and $137,960$, $100,328$ and $28,309$ for the complete, spatial and action-based scene graphs, respectively.

Baseline Models To test how textual and visual structural compositionality emerges in different VLMs, we include a total of five baselines.[1] Using CLIP (Radford et al., 2021) as a reference model, we first test various pre-existing CLIP variants that augment either text, images, or both in their training objectives. More specifically, our rationale for choosing CLIP-based models is their shared architecture and pre-training paradigm, enabling controlled comparisons in which observed differences can be attributed to specific changes to the common backbone. For instance, SigLIP-2 (Tschannen et al., 2025) uses grounded captioning, self-distillation and masked prediction to reconstruct missing image parts and align local image patches with global scene context, in addition to the sigmoidal loss proposed in the previous SigLIP model (Zhai et al., 2023). Conversely, LaCLIP (Fan

[1] Since the pre-trained model was unavailable for Structure-CLIP, it was not included.

et al., 2023) focuses on augmentations of text data through synthetic rewrites of image captions. TripletCLIP (Patel et al., 2024) refines text augmentations by generating hard negative captions that closely resemble the original captions but do not match the original images, and further generates hard negative images based on the hard negative captions. Lastly, HyCoCLIP (Pal et al., 2025) projects embeddings from Euclidean into hyperbolic space with an added entailment cone loss objective to enforce hierarchical part-whole relations between image regions and the whole image. To keep the CLIP variants comparable, the base model size is consistent for all tested models.

GraphCLIP To examine the effect of explicitly incorporating representations of visual structure into a CLIP-like model beyond the tested pre-existing CLIP variants, we propose GraphCLIP, which integrates Graphormer (Ying et al., 2021) representations alongside the CLIP (base model size) backbone. As a first step, we independently pre-train three Graphormer backbones, specific to spatial, action-based and complete scene graphs, (number of layers $L = 6$, embedding dimension $d = 512$) using a graph contrastive learning (GCL) objective implemented in PyGCL (Zhu et al., 2021) on $56,579$ VG examples that do not overlap with COCO data. Each model is trained for $2,500$ steps with a batch size of $b = 2048$ and a learning rate of $lr = 3e{-}4$ with linear decay to zero. Then, we train three GraphCLIP model variants by combining the pre-trained base CLIP and respective pre-trained Graphormer backbones and further train on the combined text, image and scene graph data from the withheld VG-COCO overlap data. Again, the scene graph data is specific to the employed graph type (action-based, spatial or complete scene graphs). We employ the following extended CLIP contrastive loss, with sim denoting cosine similarity, and x_i, z_i denoting image and graph embeddings, respectively:

$$\mathcal{L} = \frac{1}{4}(\mathcal{L}_{\text{text}\to\text{image}} + \mathcal{L}_{\text{image}\to\text{text}} + \mathcal{L}_{\text{graph}\to\text{image}} + \mathcal{L}_{\text{image}\to\text{graph}})$$

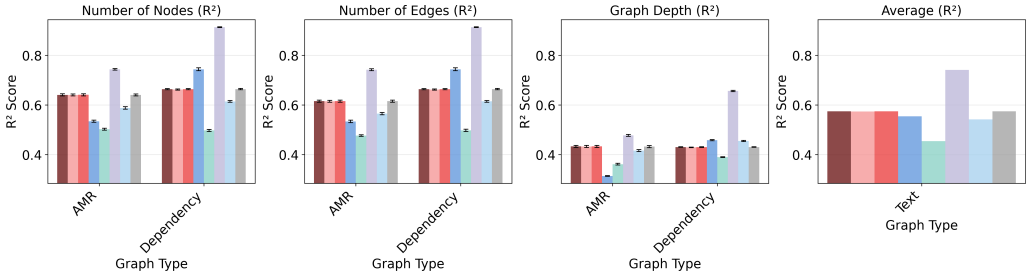
$$\mathcal{L}_{\text{graph}\to\text{image}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, x_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i, x_j)/\tau)}, \mathcal{L}_{\text{image}\to\text{graph}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(x_i, z_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i, z_j)/\tau)}$$

We train GraphCLIP for $2,500$ steps with $b = 512$ and a linearly decaying $lr = 3e - 4$, and a dropout rate of 0.1. To ensure stable multi-modal alignment, we use a custom curriculum, which freezes the CLIP image and text backbones for the first half of the total number of training steps, and then gradually unfreezes them layer per layer for the remaining steps.

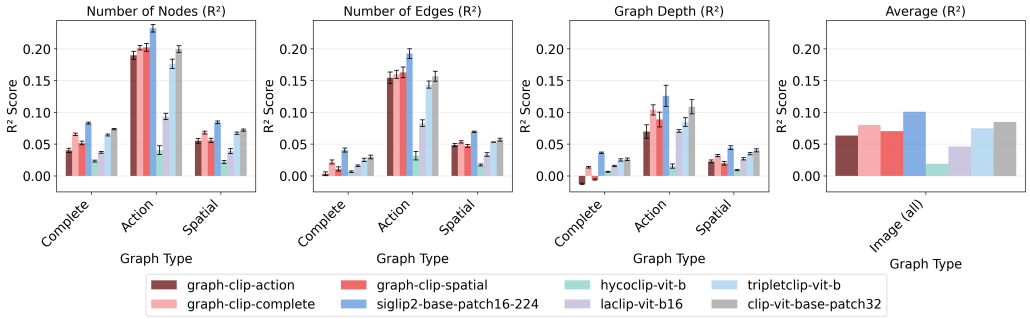
Linear Probing To quantify the extent to which VLM representations encode structural compositional information, we employ linear probing. i.e., linear models that learn to predict graph properties from VLM embeddings. Although compositional generalization requires mechanisms beyond representation alone, models that generalize compositionally rely on representations of structural primitives (Andreas, 2019). Linear probing tasks of structural graph properties thus allow us to directly examine whether the tested representations contain this necessary information. For each of the five graph types, we defined three regression tasks (see Figure 1iii): (a) predicting the number of nodes, reflecting compositional breadth, (b) predicting the number of edges, which captures relational density, and (c) predicting the depth (longest shortest path between any two nodes, as not all text or image graphs are necessarily also trees), which reflects hierarchical structure. For each VLM, we first extract frozen text and image embeddings by passing the input text and images through their respective encoders, then train ridge regression models to predict each property. Crucially, we match the embedding modality to the graph input type: text embeddings are probed for text-derived graphs, while image embeddings are probed for image-derived graphs, allowing for a direct comparison of how textual versus visual structural compositionality emerges in VLM representations. For evaluation, we use nested cross-validation: an outer 5-fold cross-validation for performance evaluation (80% train, 20% validation per fold) and an inner 5-fold cross-validation for hyperparameter selection, namely the regularization strength $\alpha \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. We report the coefficient of determination (R^2) averaged across the five outer folds and its standard deviation.

4 RESULTS

Figure 2i and 2ii show the differences in performances between text- and image-based probing tasks, respectively. Here, the R^2 score reflects how well the embeddings of a given model encode a given graph property, where $R^2 = 1$ indicates perfect linear predictability and $R^2 \approx 0$ indicates the property is not linearly accessible. While text-based probes yield higher overall performances, the



(i) Text-based linear probes based on AMR graphs and dependency parse trees



(ii) Image-based linear probes based on complete, action-based or spatial scene graphs

Figure 2: **Probing for Graph Structural Properties.** We report the coefficient of determination (R^2) for ridge regression models trained to predict (a) node count, (b) edge count, and (c) graph depth from text and vision encoders. Error bars indicate standard deviation across cross-validation folds. *AMR: Abstract Meaning Representation.*

relative rankings of model performances for text-based probing tasks differ from the relative rankings for image-based probing tasks. For instance, on the text-based probing tasks, LaCLIP (shown in purple in Figure 2i) consistently outperforms all other CLIP variants. All other CLIP variants, including the proposed GraphCLIP models, performed either on par or worse than the CLIP baseline on the text-based probing tasks. Moreover, dependency parse tree-based graph properties generally yielded higher probing performances compared to AMR-based graph properties. Conversely, for image-based probes, SigLIP-2 (shown in dark blue in Figure 2ii) stands out as the only model to outperform the CLIP baseline. The GraphCLIP models did not yield substantial performance improvements compared to the reference CLIP model, though the GraphCLIP model based on complete scene graphs performed slightly better than the spatial and action-based variants. Regarding the probes, across different image-based graph types (i.e., complete, spatial or action-based scene graphs), properties of action-based scene graphs yielded higher performances compared to spatial or complete scene graphs, which led to near-zero performances in some instances (see Figure 2ii, middle two panels). Across both text and image modalities, and consistently across models, probing for the number of nodes and edges results in overall higher R^2 scores compared to probing for the graph depth.

5 Discussion

Overall, our findings reveal differences in relative model performance rankings across modalities (see Figure 2), suggesting that the encoding of structural compositionality may be decoupled across encoders. This discrepancy might indicate a fundamental difference in the processing of compositional structure in text and image encoders, or it may reflect the characteristics of the underlying data representations (dependency/AMR trees versus scene graphs), despite our intention to create analogous probing tasks. Importantly, while AMR graphs and dependency parse trees were derived from automated parsers, the employed scene graphs for images were constructed from human annotations, which could have introduced some degree of ambiguity. Even within a given modality,

different graph types results in different probing task performances. Across models, textual probing tasks constructed from dependency parse trees resulted in higher R^2 scores compared to AMR graphs, suggesting that text encoders might capture surface-level syntax more reliably than abstract semantic features. For image-based probing tasks, action-based scene graphs yielded consistently higher R^2 scores compared to the other tested scene graph types, which can possibly be attributed to the high salience of visual actions in natural scenes. Moreover, the incorporation of explicit visual structural information within our proposed GraphCLIP model variants does not result in substantial performance improvements over the CLIP baseline. Across both modalities, the counts of nodes and edges might be more explicitly represented than compositional depth, reflected in consistently higher R^2 scores. This result suggests that while VLMs may successfully identify the presence and quantity of entities or relations, their embeddings may represent concepts in a flat rather than nested manner, and the deeper hierarchical organization of these elements may remain insufficiently represented.

One possible interpretation of our findings is that vision encoders (including our proposed GraphCLIP variants) may fail to encode the visual compositional structure defined by human-annotated scene graphs, possibly because they are constrained by the contrastive learning-based training objective; the effective incorporation of visual structural compositionality thus remains an open problem. Alternatively, the differences in relative model performance rankings across text- and image-based probes may instead reflect the fundamentally different roles of compositionality across modalities: Whereas linguistic meaning is tied to formal syntax, visual information may rely less on explicit compositional priors, and the ability of a given model to encode structure in one modality might not necessarily translate to the other. More specifically, SigLIP-2 is the only model to show substantial gains over the CLIP baseline in vision, possibly due to its reconstruction-based training objective focusing on connecting local and global contexts, suggesting that representational alignment may be learned locally (e.g., aligning specific image patches with specific graph properties). Conversely, LaCLIP yields the highest R^2 scores for text-based probes, potentially improving its linguistic structural representations through the use of synthetic augmentations. Finally, future work should systematically investigate non-linear probing settings as well as the effect of training data size, number of training steps and model size on probing task performances. Further, comparing probing task to downstream benchmark performances and incorporating human neural data may help to provide an additional reference for evaluating visual compositional structure.

ACKNOWLEDGMENTS

This research was supported by Fonds Wetenschappelijk Onderzoek (FWO) (grant number 1154625N).

REFERENCES

Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. Probing for Constituency Structure in Neural Language Models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6738–6757, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.502. URL https://aclanthology.org/2022.findings-emnlp.502/.

Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph Pre-training for AMR Parsing and Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6001–6015, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.415. URL https://aclanthology.org/2022.acl-long.415.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for Sembanking. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper (eds.),

Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2322.

Clemens G. Bartnik and Iris I. A. Groen. Visual perception in the human brain: How the brain perceives and understands real-world scenes. In *Oxford Research Encyclopedia of Neuroscience*. 2023. ISBN 978-0-19-026408-6. doi: 10.1093/acrefore/9780190264086.013.437. URL https://oxfordre.com/neuroscience/display/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-437.

Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A Comprehensive Survey of Scene Graphs: Generation and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, January 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3137605. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Cas W. Coopmans, Helen de Hoop, Karthikeya Kaushik, Peter Hagoort, and Andrea E. Martin. Hierarchy in language interpretation: evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience*, 37(4):420–439, April 2022. ISSN 2327-3798. doi: 10.1080/23273798.2021.1980595. URL https://doi.org/10.1080/23273798.2021.1980595. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2021.1980595.

Pablo Diego-Simón, Stéphane D'Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Rémi King. A polar coordinate system represents syntax in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105375–105396, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be36e50757bf9cd280aa74f89a7d1c23-Abstract-Conference.html.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: limits of transformers on compositionality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 70293–70332, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. pp. 6824–6835, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Fan_Multiscale_Vision_Transformers_ICCV_2021_paper.html.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the Binding Problem in Artificial Neural Networks, December 2020. URL http://arxiv.org/abs/2012.05208. arXiv:2012.05208 [cs].

Alon Hafri and Chaz Firestone. The perception of relations. 25(6):475–492, 2021. ISSN 1364-6613. doi: 10.1016/j.tics.2021.01.006. URL https://www.sciencedirect.com/science/article/pii/S1364661321000085.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-CLIP: towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, volume 38 of *AAAI'24/IAAI'24/EAAI'24*, pp. 2417–2425. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i3.28017. URL https://doi.org/10.1609/aaai.v38i3.28017.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. 35:37822–37836, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f69707de866eb0805683d3521756b73f-Abstract-Conference.html.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.215. URL https://ieeexplore.ieee.org/document/8099698/.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. ISSN 1573-1405. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, pp. 1–7, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06668-3. URL https://www.nature.com/articles/s41586-023-06668-3. Publisher: Nature Publishing Group.

Michael Lepori, Thomas Serre, and Ellie Pavlick. Break It Down: Evidence for Structural Compositionality in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42623–42660, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921. IEEE, 2023. doi: 10.1109/cvpr52729.2023.01050. URL https://ieeexplore.ieee.org/document/10205135/.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 38018–38070, Vienna, Austria, July 2024. JMLR.org.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models, 2025. URL http://arxiv.org/abs/2410.06912.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. 37:32731–32760, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/39781da4b5d05bc2908ce08e43bc6404-Abstract-Conference.html.

Francis Jeffry Pelletier. The Principle of Semantic Compositionality. *Topoi*, 13(1):11–24, March 1994. ISSN 1572-8749. doi: 10.1007/BF00763644. URL https://doi.org/10.1007/BF00763644.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html. ISSN: 2640-3498.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. 2020. URL https://doi.org/10.1162/tacl_a_00349.

Matteo Ruggero Ronchi and Pietro Perona. Describing Common Human Visual Actions in Images, June 2015. URL http://arxiv.org/abs/1506.02203. arXiv:1506.02203 [cs].

Jiajun Song, Zhuoyan Xu, and Yiqiao Zhong. Out-of-distribution generalization via composition: A lens through induction heads in Transformers. *Proceedings of the National Academy of Sciences*, 122(6):e2417182122, February 2025. doi: 10.1073/pnas.2417182122. URL https://www.pnas.org/doi/10.1073/pnas.2417182122. Publisher: Proceedings of the National Academy of Sciences.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B Tenenbaum, Katherine M Collins, Katherine L Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P O'Connell, Thomas Unterthiner, Klaus-Robert Müller, Mariya Toneva, and Thomas L Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, October 2025.

Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2022 edition, 2022. URL https://plato.stanford.edu/archives/fall2022/entries/compositionality/.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. Limits of Transformer Language Models on Learning to Compose Algorithms. *Advances in Neural Information Processing Systems*, 37:7631–7674, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0e797d5139ad94fc2dc2080c09119f29-Abstract-Conference.html.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00517. URL https://ieeexplore.ieee.org/document/9878945/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL http://arxiv.org/abs/2502.14786.

Melissa Le-Hoa Võ. The meaning and structure of scenes. 181:10–20, 2021. ISSN 0042-6989. doi: 10.1016/j.visres.2020.11.003. URL https://www.sciencedirect.com/science/article/pii/S0042698920301796.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in neural information processing systems*, volume 34, pp. 28877–28888, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01100. URL https://ieeexplore.ieee.org/document/10377550/.

Yanqiao Zhu, Yichen Xu, Qiang Liu, and Shu Wu. An empirical study of graph contrastive learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

6 APPENDIX

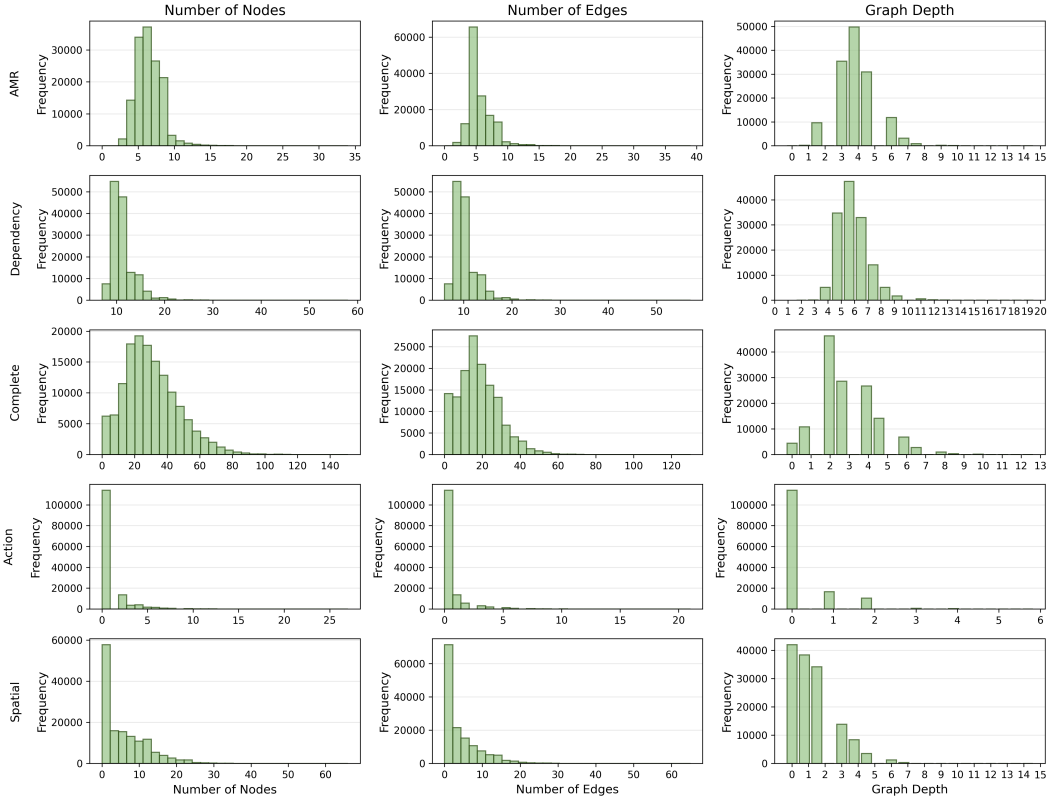


Figure A1: **Probe target variable distribution**. For each of the five defined graph types, this figure shows the distribution of the (a) number of nodes, (b) number of edges and (c) depth (longest shortest distance between any two nodes), respectively. Note that for a given text-image pair, the number of nodes (or edges or depth) in the image-based graphs do not necessarily correlate to the number of nodes (or edges or depth) in the text-based graphs.