Investigating Anisotropy in Visual Grounding Under Controlled Counterfactual Perturbations

Gabriele Lombardo, Luigi Maiorana, Liliana Lo Presti, Marco La Cascia
University of Palermo, Italy

CVPR JUNE 3-7, 2026 DENVER COLORADO

OVERVIEW

VISUAL GROUNDING

Localize a specific object in an image given a natural language description

COUNTERFACTUALS

Benchmarks do not contain **counterfactual captions**, i.e., describing object that are not present in the image

Donut with hole

APPROXIMATION BEHAVIOR

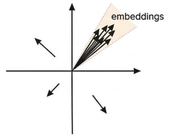
Models often fallback to approximation: predicting a plausible bounding box that satisfies only part (e.g., the object category) of the expression

Donut with hole on the left

MODELS

- We analyze two transformer-based models: **TransVG [1]** and **SwimVG [2]**
- Such architectures are not designed to output confidence or counterfactual rejection scores

ANISOTROPY



- Anisotropy is the concentration of embeddings along few dominant directions
- Contextual language embeddings often exhibit **strong anisotropy**
- This may reduce the discriminability of counterfactual captions

CONTRIBUTIONS

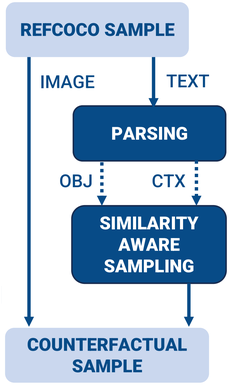
- Propose a general framework to quantify the impact of embedding geometry on approximation behavior by similarity metrics
- Provide empirical evidence that embedding anisotropy might not be the underlying cause of approximation

METHODOLOGY

PARSING CAPTIONS

We use **spaCy** to split each caption into components:

- **object**, the words referring to the category
- **context**, all remaining words describing attributes



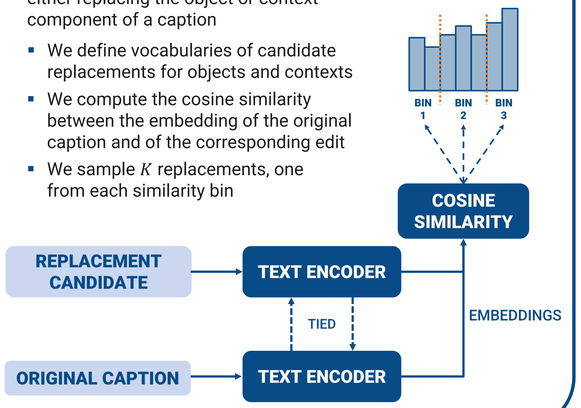
MEASURING ANISOTROPY

- We compute the empirical distribution of **cosine similarity** between the embeddings of uniformly sampled caption pairs **[3]**
- We partition this distribution using quantiles and obtain K similarity bins

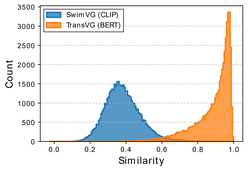
CONTROLLED SAMPLING

We build counterfactual datasets by either replacing the object or context component of a caption

- We define vocabularies of candidate replacements for objects and contexts
- We compute the cosine similarity between the embedding of the original caption and of the corresponding edit
- We sample K replacements, one from each similarity bin

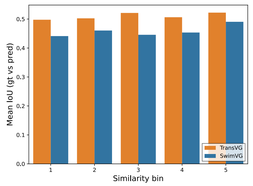


RESULTS



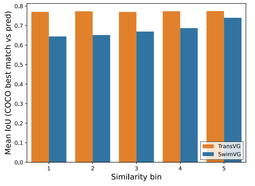
LANGUAGE ENCODERS ANISOTROPY

- $N = 50{,}000$ randomly sampled caption pairs, $K = 5$ bins
- BERT representations are more anisotropic
- CLIP embeddings are more directionally uniform because of **contrastive multimodal pretraining [4]**



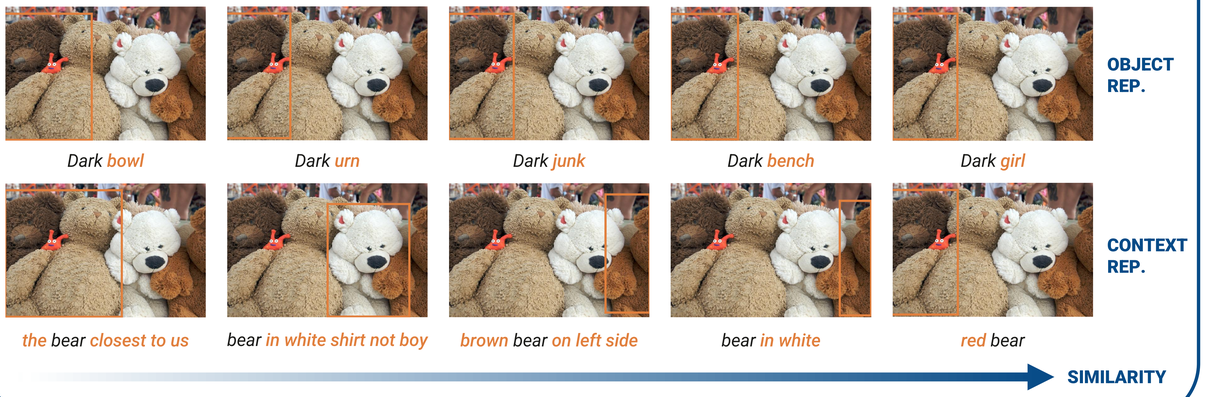
OBJECT REPLACEMENT

- We measure the IoU between the counterfactual prediction and **the original RefCOCO+ ground truth**
- A high IoU indicates that the model preserves alignment with the original context rather than responding to the altered object component
- **No meaningful correlation** for both models



CONTEXT REPLACEMENT

- We measure the maximum IoU between the counterfactual prediction and **all MS-COCO bounding boxes of the target category**
- A high IoU indicates that the model ignores the altered context and bases its prediction solely on the original object component
- **No meaningful correlation**, SwimVG exhibits a slightly higher correlation coefficient ($\rho = 0.125$)



CONTACT

Gabriele Lombardo
University of Palermo
gabriele.lombardo08@unipa.it

REFERENCES

[1] J. Deng *et al.*, «TransVG++: End-to-End Visual Grounding With Language Conditioned Vision Transformer», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, fasc. 11, pp. 13636–13652, nov. 2023.
[2] L. Shi, T. Liu, X. Hu, Y. Hu, Q. Yin, and R. Hong, "SwimVG: Step-Wise Multimodal Fusion and Adaption for Visual Grounding," *IEEE Transactions on Multimedia*, pp. 1–12, 2025 .
[3] K. Ethayarajh, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65.
[4] R. Wolfe and A. Caliskan, "Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3050–3061.