Investigating Anisotropy in Visual Grounding under Controlled Counterfactual Perturbations

Gabriele Lombardo Luigi Maiorana Liliana Lo Presti Marco La Cascia
Department of Engineering
University of Palermo
Palermo, Italy
{gabriele.lombardo08, luigi.maiorana02, liliana.lopresti, marco.lacascia}@unipa.it

Abstract

*Visual Grounding benchmarks assume that the object described by a referring expression is always present in the image, and grounding models are therefore rarely evaluated under semantically mismatched captions. In such cases, models frequently exhibit approximation behavior, producing a plausible bounding box that satisfies only part of the expression (*e.g.*, preserving the original object while ignoring modified contextual cues). Because mismatched captions represent realistic edge cases, this behavior compromises reliability and raises concerns from an explainability perspective. Identifying its underlying causes is thus essential for improving model faithfulness and interpretability. Adopting a mechanistic interpretability viewpoint, this work examines whether embedding anisotropy contributes to counterfactual failures. A similarity-controlled counterfactual caption generation protocol is introduced to systematically perturb object or contextual components within predefined embedding similarity intervals, enabling a fine-grained analysis of grounding behavior as a function of alignment. Experiments on two Transformer-based models with markedly different embedding geometries (BERT-based TransVG and CLIP-based SwimVG) reveal no meaningful correlation between cosine similarity and approximation. These findings suggest that anisotropy alone does not account for counterfactual errors, and that robustness requires investigating finer-grained geometric properties of the embedding space.*

1. Introduction

Visual Grounding (VG), also referred to as Referring Expression Comprehension (REC), requires a model to localize an object in an image given a text description. Recent Transformer-based architectures have achieved strong performance on standard benchmarks such as RefCOCO [38], RefCOCO+ [38], RefCOCOg [20], and Flickr30K Entities [27]. Modern vision-language models [16] typically rely on large-scale pretrained models both for vision [1, 25] and language [4, 23] to operate in a shared multimodal embedding space.

Despite this progress, current evaluation protocols rely on a strong assumption: the object described in the referring expression is always present in the image. As a consequence, the ability of models to handle counterfactual or mismatched referring expressions is rarely assessed. Furthermore, most grounding architectures are not designed to output calibrated confidence scores or explicit rejection probabilities. When exposed to counterfactual captions, visual grounding models often exhibit *approximation behavior*, *i.e.* predicting a plausible bounding box that satisfies only part of the expression (*e.g.*, grounding the correct object category while ignoring modified contextual cues). An example of this is shown in Figure 1. This approximation behavior does not merely result in localization errors. From an explainability perspective, such predictions undermine the faithfulness of the model's output. In realistic edge-case scenarios, this lack of faithfulness compromises the reliability of the system, particularly when incorrect yet plausible predictions are produced without any indication of uncertainty or rejection.

In this work, we investigate whether the geometric structure of language embedding spaces contributes to approximation in visual grounding models. Prior studies in natural language processing [5, 8] have shown that contextualized embeddings produced by Transformer architectures exhibit strong anisotropy: representations tend to concentrate along a small number of dominant directions rather than being uniformly distributed in the embedding space. Such geometric biases may reduce discriminability between semantically distinct but lexically similar expressions, potentially affecting a model's ability to detect counterfactual descriptions.

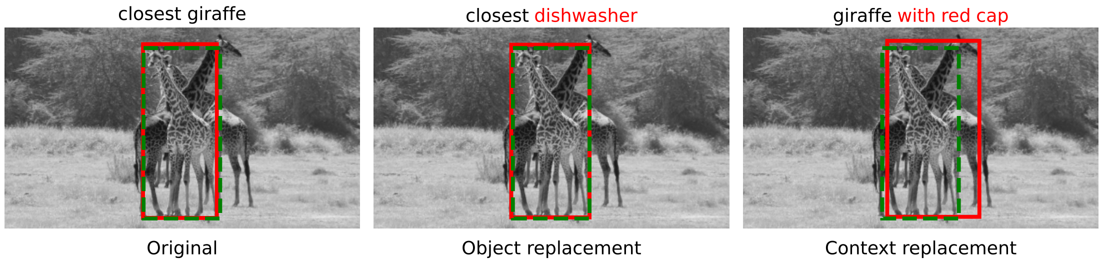


Figure 1. Example of approximation behavior under object replacement and context replacement. In both cases, the model predicts a bounding box that ignores counterfactual cues (highlighted red text) in the modified referring expression. The original RefCOCO ground-truth annotation is shown with a dashed green outline, while the model prediction is drawn as a red bounding box.

Since current benchmarks do not allow a systematic study of the impact of embedding geometry on grounding behavior, we construct controlled counterfactual datasets derived from RefCOCO+ by systematically replacing either the object or the contextual component of each caption. Importantly, replacements are sampled as controlled perturbations in embedding space to achieve predefined similarity levels with respect to the original caption. This design enables a fine-grained analysis of grounding performance as a function of embedding similarity.

We compare two state-of-the-art Transformer-based grounding models employing language encoders with different geometric properties. We first quantify anisotropy through the average cosine similarity between randomly sampled caption embeddings. We then measure approximation behavior using the Intersection over Union (IoU) between predicted bounding boxes and the original ground-truth annotations under counterfactual conditions.

If anisotropy in language embeddings corresponds to semantically different captions having similar embeddings, one might hypothesize that stronger anisotropy exacerbates approximation. However, our analysis reveals no clear correlation between embedding geometry and approximation, suggesting that this behavior may emerge from more complex multimodal interactions.

This paper's contributions are:

- We introduce a similarity-aware counterfactual caption generation procedure that explicitly accounts for the geometry of the underlying language embedding space.
- We propose a quantitative framework to measure the approximation tendency of visual grounding models under controlled counterfactual perturbations.
- We provide empirical evidence that embedding anisotropy might not be the main reason for the approximation behavior exhibited by visual grounding models in counterfactual settings.

2. Related Work

2.1. Visual Grounding

Visual Grounding [2, 3, 11, 14, 30, 36] is the vision-language task of localizing a specific object in an image given a natural language description. This task can be seen as the language-grounded extension of object detection, analogous to how Referring Expression Segmentation (RES) [10] extends object segmentation.

Early visual grounding models [18, 39] were closely related to conventional object detection architectures and are commonly referred to as two-stage methods.

State-of-the-art visual grounding approaches that are based on the Transformer architecture [31] are commonly referred to as one-stage methods [3, 30]. These models are trained end-to-end, and bounding box coordinates are predicted directly from multimodal token representations, typically through dedicated detection/regression heads. While these architectures achieve strong localization performance, they are not explicitly designed to deal with semantic inconsistencies or counterfactual descriptions.

2.2. Visual Grounding Benchmarks

Standard evaluation protocols for visual grounding rely on two families of datasets: Flickr30K [37] and its extension for visual grounding Flickr30K Entities [27], and RefCOCO [38], RefCOCO+ [38] and RefCOCOg [20], all based on the MS-COCO [17] dataset.

These benchmarks share a strong assumption: the object described in the caption is always present in the corresponding image. Consequently, the ability to reject mismatched captions remains largely unaddressed, preventing a proper assessment of the model's semantically faithful behavior under counterfactual perturbations.

Recent work has attempted to address this limitation by constructing datasets derived from the Ref-COCO family through random caption shuffling [12, 15] or by automatically editing captions to describe objects not present in the image [7, 40]. However, automatically generated counterfactual samples are prone to annotation noise and semantic inconsistencies, which limit their reliability for training and evaluating rejection-aware grounding models. This highlights the need for cleaner benchmarks explicitly designed to evaluate robustness to absent or mismatched referring expressions.

2.3. Automatic Parsing of Captions

Recent advances in multimodal dataset construction have been largely driven by the increasing capabilities of vision-language models (VLMs) to bridge visual and linguistic knowledge and generate detailed image descriptions, enabling scalable, automatic collection of annotations for a wide range of multimodal tasks.

In the context of referring expression comprehension and segmentation, the objective is to associate image regions with contextually grounded textual expressions. A common pipeline consists of generating image-level captions, extracting noun phrases via syntactic parsing, and aligning them with visual regions. The authors of [35] first produce both class-agnostic segmentation masks and rich image captions, from which noun phrases are extracted and subsequently matched to regions using CLIP [28]. Similarly, [32] generate segmentation masks using SAM [13] with bounding box prompts, where the bounding boxes are obtained by feeding parsed noun phrases into Grounding DINO [19].

Related works [26, 34] follow similar strategies to construct bounding box-text pairs, by combining caption generation, phrase extraction, and vision-language alignment.

2.4. Contextual Embeddings

Modern vision-language architectures [3, 30] are Transformer-based and often adopt dual-encoder or cross-encoder architectures, where visual and textual representations are aligned and fused through contrastive objectives or cross-attention mechanisms to enable object localization. However, the geometric structure of these pre-trained embedding spaces, particularly those learned via contrastive objectives such as CLIP [28], remains a critical factor influencing downstream performance.

Isotropy, *i.e.* a uniform distribution of embeddings across all directions of the representation space, has traditionally been regarded as a desirable property of static word embeddings, although it is not consistently observed in all static embedding techniques [22].

More recent work has highlighted that contextualized embedding spaces produced by Transformer architectures exhibit anisotropy [6, 9], meaning that representations concentrate along a small number of dominant directions rather than being uniformly spread. This geometric bias has been hypothesized to negatively impact semantic discriminability and downstream task performance. However, empirical evidence has questioned this assumption. In particular, [8] report that anisotropy in BERT [4] embeddings is unlikely to be the primary cause of degraded performance in several downstream tasks. Similarly, [5] show that Transformer embedding spaces exhibit a form of *local* isotropy and do not significantly benefit from post-hoc isotropy calibration after training.

Although anisotropy has been studied in the field of natural language processing, its implications for visual grounding behavior in counterfactual settings remain largely unexplored. Recent open-vocabulary grounding approaches [19, 32, 35] further highlight the importance of robustness to absent or mismatched queries, reinforcing the relevance of our counterfactual setting.

3. Methodology

3.1. Measuring Anisotropy

Given a visual grounding model, we aim to analyze if the anisotropy of the embedding space learned by its language encoder may influence the model's ability to distinguish counterfactual semantic perturbations.

Anisotropy is the concentration of the embeddings along few dominant directions in a hypercone of the embedding space. The greater the anisotropy, the narrower this cone.

In natural language processing, directions in embedding space are commonly interpreted as encoding semantic information. Accordingly, we adopt cosine similarity as a measure of proximity between representations, as it captures angular alignment independently of vector magnitude, is geometrically well-defined in high-dimensional spaces, and is computationally efficient to evaluate at scale.

Specifically, we compute the cosine similarity between the embeddings of N uniformly sampled pairs of captions. Following [6], anisotropy is quantified as the average cosine similarity across all sampled pairs. An average value close to 1 indicates that randomly selected captions are highly aligned, reflecting a strong concentration of representations within a narrow cone of the embedding space. Such a geometry may reduce directional discriminability, particularly under small semantic perturbations.

3.2. Building the Approximation Datasets

To gain insights into the behavior of visual grounding models under counterfactual samples, we construct new datasets derived from the validation and test sets of RefCOCO+. We select RefCOCO+ [38] among the RefCOCO variants because it excludes absolute location terms (*e.g.*, "left", "top"), thereby reducing explicit positional bias in the referring expressions.

We denote a sample from RefCOCO+ as a tuple $(\mathbf{I}, t, \mathbf{b}, c)$, where $\mathbf{I}$ is the image, t is the caption describing the target object, $\mathbf{b}$ is the bounding box annotation of that object, and c is the corresponding MS-COCO [17] category. In the context of this work, a caption is composed of two parts: (i) the *object*, *i.e.*, the word(s) referring to the category c, and (ii) the *context*, comprising all remaining words that describe the position, attributes, or relations of the referred object.

We build the approximation datasets by systematically replacing either the object or the context of a caption. In contrast to prior approaches [7, 40], this replacement process is conditioned on embedding similarity, allowing us to control the degree of perturbation in the language embedding space. These datasets will be used to assess the model's tendency to approximate its predictions under counterfactual inputs, *i.e.*, predicting a bounding box that aligns with either only the original object or the original context component of the input caption. The proposed protocol allows us to isolate the effect of semantic similarity while keeping the visual input fixed.

Dependency Parsing RefCOCO+ captions are natural language descriptions collected through an interactive game, and objects belonging to the same category are often referred to using different lexical terms (*e.g.*, the *person* category may appear as "man", "person", "boy", *etc.*). To extract the *object* component of a caption, we use dependency parsing with spaCy [24] to identify the semantic head corresponding to the target entity, also handling multi-word cases, *e.g*. "teddy bear". The *context* is defined as all remaining words that are not part of the object specification.

Building the replacement datasets, we discard captions for which the parser is not able to find a clear semantic head, or multiple candidates are found.

Sampling Replacements Directly sampling vectors from the language encoder's embedding space would not yield semantically valid captions. Instead, we adopt a discrete sampling strategy defined over structured candidate sets.

We first define two vocabularies $\mathcal{O}$ and $\mathcal{C}$, corresponding to candidate objects and contexts, respectively. In our experiments, $\mathcal{O}$ is built from the list of the original MS-COCO categories expanded with common synsets retrieved from WordNet [21] to increase diversity. The context vocabulary $\mathcal{C}$ consists of all context segments extracted from RefCOCO+.

For a given model, in order to obtain controlled perturbation levels, we partition the empirical distribution of cosine similarities between randomly sampled caption embeddings (as defined in Section 3.1) into K intervals (in our experiments, we set $K = 5$). To ensure an equal number of samples per interval, edges are determined using the distribution quantiles.

Given a caption t with object component o_t, we sample K replacements, one for each similarity interval. If no candidate is found for a given interval, the caption is discarded. For each candidate $o \in \mathcal{O}$, we compute similarity using one of the two following strategies:

- *Word-level similarity*: cosine similarity between the embeddings of o and o_t.
- *Sentence-level similarity*: cosine similarity between the embedding of the original caption t and that of the modified caption obtained by replacing o_t with o.

The different strategies for similarity computation allow us to highlight the sensitivity of the embedding space to different perturbations: lexical or contextual respectively. An example of object replacement is shown in Figure 1.

To sample a context replacement for caption t, we evaluate each candidate $c \in \mathcal{C}$ by computing the cosine similarity between the embedding of the original caption t and that of the caption obtained by substituting its context with c. A replacement is then sampled within the desired similarity interval. Figure 1 shows an example of a caption generated by context replacement.

In summary, we construct three datasets derived from RefCOCO+: (i) a word-level object replacement dataset, (ii) a sentence-level object replacement dataset, and (iii) a context replacement dataset. Each dataset isolates a different type of controlled perturbation in the caption, allowing us to analyze how object- and context-level modifications affect grounding behavior by keeping the image fixed.

3.3. Measuring approximation

From a faithfulness standpoint, a grounding model should modify its prediction when the caption semantics are altered. Predicting the annotated object under counterfactual conditions suggests an approximation behavior. To evaluate the tendency of a visual grounding model to ignore components of a counterfactual caption and produce a best-match prediction, we leverage both the original RefCOCO+ bounding box annotations and the full set of ground-truth bounding boxes provided by MS-COCO.

In the object replacement setting, we define approximation behavior as the model predicting the original RefCOCO+ bounding box despite the caption being counterfactual. To quantify this effect, we compute the Intersection over Union (IoU) between the bounding box predicted under the modified object component and the original RefCOCO+ ground-truth annotation. Since the model has no explicit mechanism to output a "not present" decision, it is forced to return a bounding box for every query. A high IoU in this setting indicates that the model fails to revise its prediction in response to the altered object semantics, thereby preserving alignment with the original caption reference rather than following the modified object in the counterfactual caption, which is not present in the image.

In the context replacement setting, approximation occurs when the model ignores the counterfactual context and bases its prediction solely on the (unchanged) object component of the caption. Since RefCOCO+ samples are constructed to distinguish among multiple entities of the same category, several ground-truth bounding boxes of the same MS-COCO category are present in each image. In this case, any of these bounding boxes constitutes an approximation behavior. Accordingly, we define the metric as the maximum IoU between the model prediction and all MS-COCO bounding boxes associated with the target category in the image. In this setting, selecting any instance of the correct category reflects a failure to incorporate contextual information rather than a localization error.

4. Results

4.1. Experimental Setup

We conduct our experiments on two state-of-the-art transformer-based models.

TransVG [3] adopts a dual-encoder structure, with separate encoders for both visual and linguistic inputs. The vision and language features are concatenated and fed into a *Vision-Language Fusion Transformer*. A special `[REG]` token is appended to the full sequence, and the output state of this token is used to directly predict the four bounding box coordinates using a regression head.

SwimVG [30] adopts instead a cross-encoder architecture following the early-fusion paradigm in which textual features are injected into the vision encoder to produce caption-aware visual representations. Multimodal interaction therefore occurs directly within the vision branch, removing the need for a separate fusion transformer. The `[REG]` token is appended to the input sequence of the vision encoder, and its final representation is used by a regression head to predict the bounding box coordinates.

4.2. Dataset Qualitative Analysis

During the counterfactual generation process, we apply two filtering stages:

1. if the parser is not able to find a clear semantic head, the caption is discarded.
2. if no replacement can be found for a given caption in any of the bins listed in Table 1, the caption is discarded.

The impact of the filtering pipeline on dataset size is detailed in Table 2. Since parsing is deterministic, the number of headless captions is identical across all four cases (969); the first row is therefore only reported for completeness.

As caption generation is fully automatic, concerns may arise regarding the quality of the edited captions. In particular, two potential issues can be identified:

- in the *object replacement* setting, the replacement category may still refer to objects present in the image;
- in the *context replacement* setting, the modified context may result in a caption that remains valid for the image. An example is shown in Figure 2.

We emphasize that the former case is explicitly avoided in our approach, as replacement candidates are selected by excluding all categories annotated in MS-COCO for the corresponding image.

To evaluate the impact of the latter issue, we conduct a manual validation on a subset of 1,200 samples, uniformly sampled from the two context replacement datasets. This analysis identifies 238 edited captions that remain plausible for the image, indicating that approximately 80% of the generated samples constitute valid counterfactuals.

Figure 2. An example of context replacement failure. In this case, the edited caption (bottom row) can still refer to a valid object in the image.

Model	1	2	3	4	5	6
SwimVG (CLIP) [30]	0.0	0.303	0.352	0.398	0.459	1.0
TransVG (BERT) [3]	0.0	0.813	0.898	0.943	0.968	1.0

Table 1. Edges for the $K = 5$ similarity bins computed using the quantiles of the distribution of the cosine similarity of random pairs of captions.

	SwimVG		TransVG	
	object	context	object	context
No semantic head	969 (13%)	969 (13%)	969 (13%)	969 (13%)
No replacement	3604 (55%)	795 (12%)	6 (0%)	2538 (38%)
Caption count	3005	5814	4071	6603
Sample Count	18,030	34,884	24,426	39,618

Table 2. Number of captions discarded after each filtering step. All settings start from 7,578 initial captions. The *Caption count* row reports the number of captions retained after the full filtering pipeline. Each retained caption generates $K + 1$ samples (K counterfactual replacements plus the original), resulting in the final sample counts reported in the last row.

4.3. Language Encoders Anisotropy

Figure 3 compares the cosine similarity distributions between caption embeddings produced by the two models,

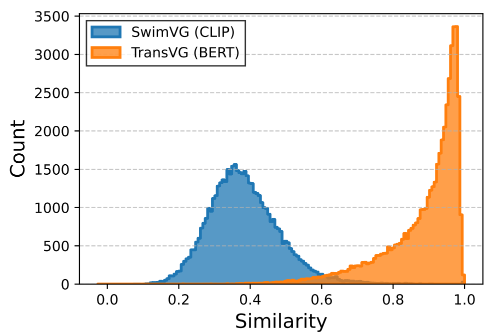


Figure 3. Cosine similarity distribution between the embeddings generated by the language encoders for randomly sampled pairs of captions.

Strategy	SwimVG		TransVG	
	ρ_p	ρ_s	ρ_p	ρ_s
Word-level	0.055	0.044	0.011	0.011
Sentence-level	0.063	0.038	0.030	0.019
Context-level	0.125	0.124	0.007	0.007

Table 3. Correlation coefficients between IoU score and similarity. The first two rows correspond to the object replacement setting, while the last row corresponds to the context replacement setting. ρ_p is the Pearson correlation coefficient, while ρ_s is the Spearman rank-order correlation coefficient.

computed over $N = 50{,}000$ randomly sampled caption pairs from RefCOCO+.

The average cosine similarity is approximately 0.38 for SwimVG (CLIP encoder) and 0.88 for TransVG (BERT encoder). Since perfect isotropy would correspond to an average cosine similarity close to zero, both embedding spaces exhibit anisotropy. However, the substantially higher average similarity observed for BERT indicates a much stronger concentration of embeddings within a narrow cone of the representation space [22]. In contrast, CLIP embeddings are more directionally dispersed due to contrastive multimodal pretraining [33]. This behavior is further reflected in the quantiles reported in Table 1, which show that BERT embeddings are concentrated in the similarity range $[0.813, 0.968]$. Overall, these results suggest that the BERT language encoder induces a significantly more anisotropic embedding geometry than CLIP. These differences imply that these two models can be used appropriately to assess whether global anisotropy affects a model's approximation behavior.

4.4. Visual Grounding Approximation

The results of evaluating the models on the object-replacement and context-replacement datasets are shown in Figure 4 and Figure 5, respectively. Importantly, each similarity bin contains the same number of samples by construction, ensuring a controlled comparison across similarity ranges. In our analysis, we exclude counterfactual captions derived from samples for which the model's prediction on the original caption achieved an IoU score lower than 0.5 with respect to the RefCOCO ground-truth bounding box. This filtering ensures that approximation is evaluated only as deviation from the correct grounding behavior.

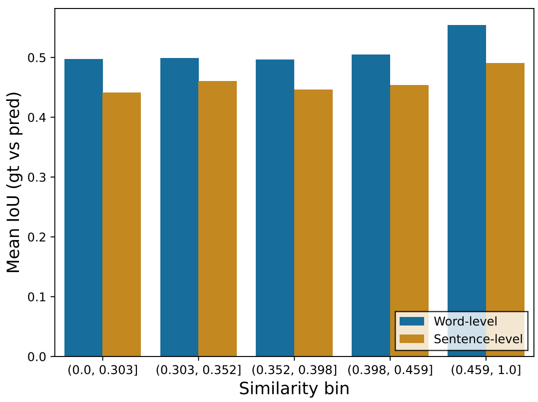
If embedding anisotropy influenced the emergence of approximation behavior, we would expect higher similarity bins showing higher IoU values (as illustrated in the first row of Figure 7). Instead, in the object-replacement setting (Figure 4), both TransVG and SwinVG exhibit consistent behavior under word-level and sentence-level similarity computations. As previously stated, TransVG relies on embeddings produced by BERT, whose representations are strongly anisotropic, with embeddings concentrated in a narrow cone of the vector space. In such a geometry, cosine similarity has reduced discriminative power across directions. By contrast, CLIP's representation space used by SwimVG, while not perfectly isotropic, is more directionally dispersed. Despite these geometric differences, neither model shows a meaningful correlation between cosine similarity and localization performance. This observation is quantitatively confirmed by the near-zero Pearson and Spearman coefficients reported in Table 3. Therefore, while embedding geometry differs substantially across models, anisotropy alone does not explain the visual grounding approximation behavior observed in our experiments. This is consistent with other works in natural language processing [5, 8] questioning the assumption that isotropy is a crucial property for contextual embeddings.

Comparing Figure 4 and Figure 5, we further observe that localization performance under context replacement is consistently higher than under object replacement across comparable similarity ranges. This suggests that grounding models are more sensitive to perturbations affecting the referential noun than to modifications of contextual information. In other words, altering the semantic anchor of the referring expression leads to stronger degradation than perturbing surrounding descriptive cues, suggesting that grounding decisions are primarily driven by object-level semantics, while contextual cues play a secondary role. From an explainability perspective, this indicates a dominance of noun-driven shortcuts over contextual reasoning.

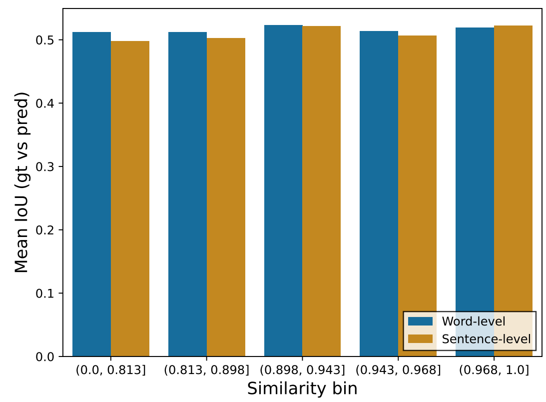
5. Qualitative results

Figure 6 and Figure 7 provide qualitative results for both models exhibiting approximation behavior.

In all but one case (middle top row of Figure 7) it is interesting to observe that the model prediction matches the ground-truth annotation under the lowest-similarity perturbation, while it changes under the highest-similarity re-

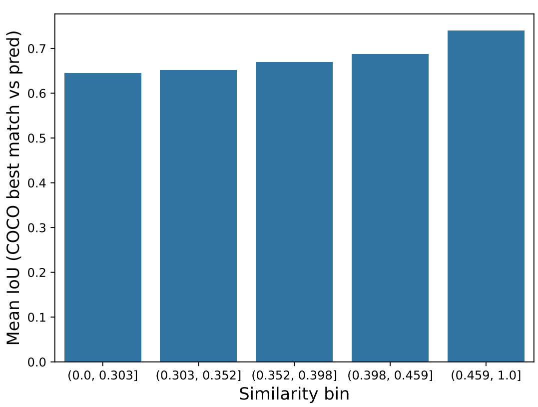


(a) SwimVG

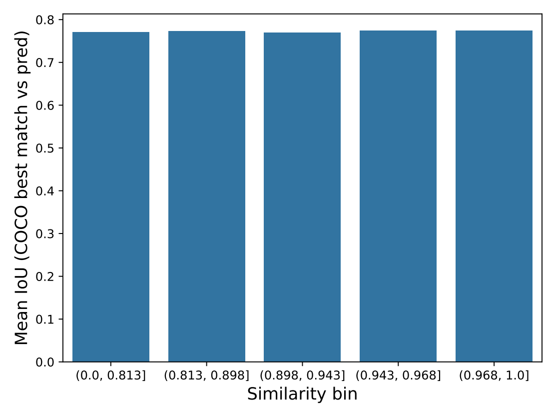


(b) TransVG

Figure 4. Comparison of the mean IoU score per similarity bin in the object replacement case, using both word-level and sentence-level similarity computation. IoU is computed between model predictions and the RefCOCO+ ground-truth bounding boxes.



(a) SwimVG



(b) TransVG

Figure 5. Comparison of the mean IoU score per similarity bin in the context replacement case. IoU is computed between model predictions and the best-matching bounding box of the referred category from MS-COCO.

placement. Moreover, the middle caption in the second row shows that the model can still rely on the object-level semantic anchor even when the modified caption is not fully coherent from a commonsense perspective.

Moreover, the first row of Figure 7 exemplifies the behavior that would be expected if anisotropy and approximation were strongly correlated: the model correctly localizes the object for the original caption, fails to reproduce the annotated bounding box under the lowest-similarity replacement, and recovers the same bounding box in the highest-similarity case.

Visual inspection of Figure 6 and Figure 7 indicates that the predicted bounding boxes correspond to existing objects in the image, regardless of the caption semantics. A broader qualitative examination over additional samples reveals that this behavior holds in the majority of cases. This consistency suggests that both models may rely on an internal notion of *objectness*, conceptually analogous to the objectness score in Faster R-CNN [29], which steers predictions toward dominant object regions even under counterfactual inputs.

6. Conclusions

In this work, we investigated the role of language embedding geometry in approximation behavior observed in visual grounding models under counterfactual conditions. By introducing a similarity-controlled counterfactual caption generation protocol, we were able to systematically control perturbations in embedding space while keeping the input image fixed and analyze model behavior as a function of semantic proximity by cosine similarity.

From an explainability standpoint, we highlight that counterfactual robustness is crucial to understanding why visual grounding models fail to revise their predictions

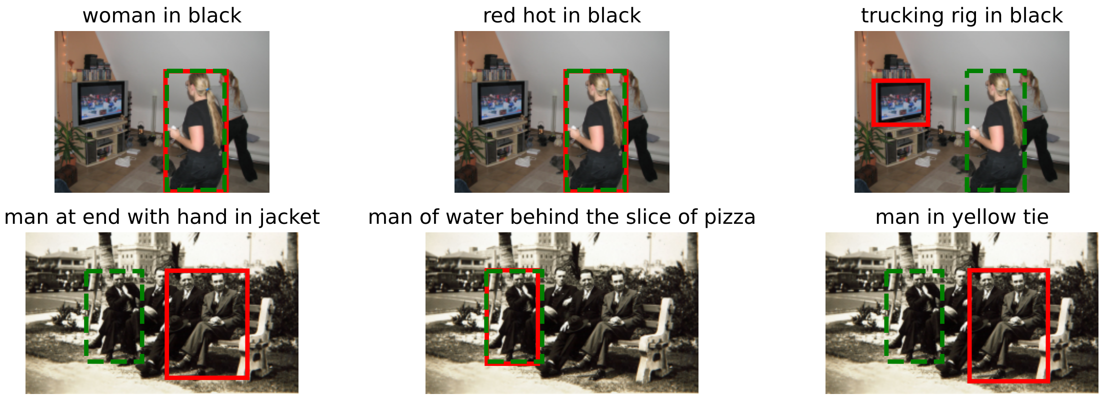


Figure 6. Qualitative results produced by TransVG on the replacement datasets, showing approximation behavior. The top row corresponds to object replacement, while the bottom row shows context replacement. From left to right, the columns display results for the original caption, the edited caption sampled from the lowest similarity bin, and the edited caption sampled from the highest similarity bin.

Figure 7. Qualitative results produced by SwimVG on the replacement datasets, showing approximation behavior. The top row corresponds to object replacement, while the bottom row shows context replacement. From left to right, the columns display results for the original caption, the edited caption sampled from the lowest similarity bin, and the edited caption sampled from the highest similarity bin.

when semantically relevant components are altered.

Across models relying on language encoders with different geometric structures of the embedding space, our results reveal only a weak positive correlation between embedding anisotropy and approximation behavior. These findings suggest that such phenomenon cannot be solely attributed to global anisotropy of the embedding space. While our study rules out this factor as a primary cause, further investigation is needed to determine whether grounding failures are better explained by other geometric properties of the language space beyond global anisotropy or instead arise from complex multimodal interactions and the reliance on shortcuts tied to object-level semantic anchors.

Qualitative evidence suggests that both visual grounding models may learn an internal notion of objectness. Nevertheless, confirming this hypothesis requires a systematic and quantitative investigation, which is deferred to future studies.

While our analysis focuses on two representative Transformer-based grounding models with distinct embedding geometries, the observed behavior is likely to extend to other models relying on object-centric representations. Future work should validate these findings on more recent open-vocabulary and cross-encoder models.

Future work may also focus on extending the presented experiments by taking into account other aspects of the embedding space geometry, such as relative distances and cluster density. Changes in the attention dynamics may contribute to the approximation behavior as well, and should be the object of subsequent studies, ultimately aiming to design visual grounding models that are semantically faithful under counterfactual perturbations.

References

[1] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Billion-scale pretraining with vision transformers for multi-task visual representations. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 1431–1440, Waikoloa, HI, USA, 2022. IEEE. 1

[2] Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, and Gen Luo. Weakmcn: Multi-task collaborative network for weakly supervised referring expression comprehension and segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9175–9185, 2025. ISSN: 2575-7075. 2

[3] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13636–13652, 2023. 2, 3, 5

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, page 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 1, 3

[5] Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. On isotropy calibration of transformer models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, page 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. 1, 3, 6

[6] Kawin 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)*, page 55–65, Hong Kong, China, 2019. Association for Computational Linguistics. 3

[7] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6371–6381, 2019. ISSN: 2575-7075. 3, 4

[8] Alejandro Fuster Baggetto and Victor Fresno. Is anisotropy really the cause of bert embeddings not being semantic? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 4271–4281, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1, 3, 6

[9] Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 35–48, St. Julian's, Malta, 2024. Association for Computational Linguistics. 3

[10] Lixia Ji, Yunlong Du, Yiping Dang, Wenzhao Gao, and Han Zhang. A survey of methods for addressing the challenges of referring image segmentation. *Neurocomputing*, 583:127599, 2024. 2

[11] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 01–10, 2023. ISSN: 2575-7075. 2

[12] Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, page 285–299, Dublin, Ireland, 2022. Association for Computational Linguistics. 3

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 3992–4003, 2023. ISSN: 2380-7504. 3

[14] Hongbing Li, Xinran Wang, Linyi Yang, Qi Li, and Bo Xiao. Sala: Semantic alignment and localization alignment for visual grounding. *Neurocomputing*, 654:131265, 2025. 2

[15] Menghao Li, Chunlei Wang, Wenquan Feng, Shuchang Lyu, Guangliang Cheng, Xiangtai Li, Binghao Liu, and Qi Zhao. Iterative robust visual grounding with masked reference based centerpoint supervision. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, page 4653–4658, 2023. ISSN: 2473-9944. 3

[16] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. (arXiv:2501.02189), 2025. arXiv:2501.02189 [cs]. 1

[17] 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 *Computer Vision – ECCV 2014*, page 740–755, Cham, 2014. Springer International Publishing. 2, 4

[18] Jingyu Liu, Liang Wang, and Ming Hsuan Yang. Referring expression generation and comprehension via attributes: 16th ieee international conference on computer vision, iccv 2017. page 4866–4874, 2017. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, page 38–55, Cham, 2025. Springer Nature Switzerland. 3

[20] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11–20, 2016. 1, 2

[21] George A. Miller. Wordnet: A lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. 4

[22] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2873–2878, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 3, 6

[23] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2):30:1–30:40, 2023. 1

[24] Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. explosion/spacy: v3.7.2: Fixes for apis and requirements, 2023. 4

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 1

[26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. (arXiv:2306.14824), 2023. arXiv:2306.14824 [cs]. 3

[27] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, page 2641–2649, 2015. 1, 2

[28] 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*, page 8748–8763. PMLR, 2021. 3

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. (arXiv:1506.01497), 2016. arXiv:1506.01497 [cs]. 7

[30] Liangtao Shi, Ting Liu, Xiantao Hu, Yue Hu, Quanjun Yin, and Richang Hong. Swimvg: Step-wise multimodal fusion and adaption for visual grounding. *IEEE Transactions on Multimedia*, page 1–12, 2025. 2, 3, 5

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. (arXiv:1706.03762), 2023. arXiv:1706.03762 [cs]. 2

[32] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, and Liu Ren. Use: Universal segment embeddings for open-vocabulary image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4187–4196, 2024. ISSN: 2575-7075. 3

[33] Robert Wolfe and Aylin 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)*, page 3050–3061, Dublin, Ireland, 2022. Association for Computational Linguistics. 6

[34] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4818–4829, 2024. ISSN: 2575-7075. 3

[35] Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. Reme: A data-centric framework for training-free open-vocabulary segmentation. (arXiv:2506.21233), 2025. arXiv:2506.21233 [cs]. 3

[36] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15481–15491, 2022. 2

[37] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[38] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, page 69–85, Cham, 2016. Springer International Publishing. 1, 2, 3

[39] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 1307–1315, 2018. ISSN: 2575-7075. 2

[40] Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13438–13448, 2024. ISSN: 2575-7075. 3, 4