Ablating Semantic Underspecification: A Comparative Study of CLIP-based Vision Language Models

Anupta Argo
Princeton University
anupta@princeton.edu

Aidan Ward
Princeton University
aw3592@princeton.edu

Eshaan Govil
Princeton University
eshaangovil@princeton.edu

Abstract

Semantic underspecification–intentional vagueness in natural language–creates challenges for language models, particularly in contexts where humans rely on visual or temporal context that language models do not have access to. Building on Pezzelle (2023) who showed that ViT-B/32 CLIP struggles with image-caption similarity scoring when captions are underspecified, we extend this analysis across a broader set of CLIP-based models including CLIP-quickgelu, SigLIP, SigLIP2, and NVIDIA Labs AM-RADIO. We also introduce a new complementary dataset containing incorrect captions that mimick the linguistic alterations of the underspecification dataset introduced by Pezzelle (2023). Our findings show that while all models display performance decreases with underspecified captions, the magnitude of decline varies across architecture. Furthermore, they also show that this performance decrease is independent of model scale. Additionally, for the incorrect caption dataset, the models exhibit nearly the same decreases in performance and further comparison between underspecified and incorrect captioning suggests that these models struggle to differentiate between the two caption types.

1 Introduction

Semantic underspecification, or the intentional vagueness of parts of speech in natural language, is commonly utilized and easily understood by humans (Pezzelle, 2023). The presence of external visual or temporal context allows humans to exploit the efficiency of underspecified language in everyday communication. However, this same property poses a challenge for language models, which lack access to such external context.

This challenge extends to multimodal models like OpenAI's CLIP, which learn joint image-text embeddings through contrastive training on paired visual and textual inputs. Although these models have demonstrated strong performance on a range of vision-language tasks (Pezzelle, 2023), recent work by Pezzelle (2023) found that ViT-B/32 CLIP performs poorly when captions are semantically underspecified.

This raises questions about whether these limitations are specific to a single model or indicative of a more general weakness across CLIP-based architectures. In this work, we extend the existing analysis to a wider set of models, including CLIP-quickgelu, SigLIP, SigLIP2, and NVIDIA's AM-RADIO. We also introduce a complementary dataset of referentially incorrect captions that closely resemble the linguistic structure of underspecified captions used in previous work. Our goal is to compare model performance across both underspecification and incorrectness and to evaluate whether robustness varies with architecture or scale.

2 Background

CLIP is a multimodal architecture introduced by OpenAI that learns to align vision-language by training on image-caption pairs (Radford et al., 2021). Using separate transformer-based encoders for images and text, CLIP maps both encodings into a shared embedding space (Radford et al., 2021). During training, a contrastive objective is applied to encourage correct image-text pairs to have higher cosine similarity than mismatched ones, and this enables CLIP to perform tasks such as zero-shot classification and image-text retrieval by comparing embeddings directly across modalities (Radford et al., 2021).

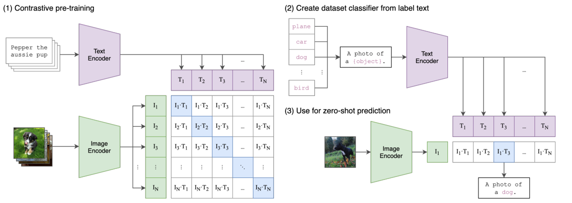

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
T1 T2 T3 ... TN
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1: Original overarching OpenAI CLIP architecture design (Radford et al., 2021)

However, recent benchmarking by Pezzelle (2023), applying the ViT-B/32 CLIP model to their novel underspecification dataset, demonstrates that this model is vulnerable to underspecified captioning and naturally suggests that other CLIP models may also suffer the same vulnerability. The dataset introduces systematic alterations removing context from original image captions, targeting specific linguistic elements to produce six underspecification types: Quantity, Gender, Gender and Number, Location, Object, and Full. For example, the original caption "The woman is standing above the two packed suitcases" is modified into underspecified variants like "The person is standing above the two packed suitcases" (Gender), "They are standing above the two packed suitcases" (Gender+Number), or "They are doing something here" (Full) where context is almost completely missing.

Concretely, Pezzelle (2023) measures the performance of CLIP using scaled cosine similarity between the image embedding and the caption embedding, and with this scoring metric Pezzelle performs two proofs of concept for evaluation:

- The first proof of concept that Pezzelle (2023) performs is calculating the similarity score of each image's original caption as well as the six previously mentioned underspecified captions for the ViT-B/32 CLIP model. He displayed the results of these scores through a violin plot for each type of underspecified caption in comparison to the original.

- The second proof of concept that Pezzelle (2023) performs is a method of comparing the performance of fully underspecified captions to incorrect captions. These incorrect captions were chosen by selecting ten random captions in the dataset excluding the actual caption for a given image, and Pezzelle (2023) calculated the similarity score of all of the captions and determined that "for 82 images out of 100, at least one random caption achieves a higher CLIPScore than `Full`," or the fully underspecified caption.

Pezzelle (2023) found that across all underspecification types, similarity scores decreased compared to the original captions, and this suggests that ViT-B/32 CLIP fails to recover the intended meaning from vague captions, even when the visual input clearly depicts the correct scene.

3 Methodology

3.1 Existing Result Replication

During the course of adding ablations to the original work, we first recreated the first proof of concept violin plot showing the similarity scores of the original and underspecified captions for the ViT-B/32 CLIP model (see Figure 20). However, we chose not to exactly reproduce the second proof of concept due to the analysis of it being fairly limited, and we instead chose to thoroughly expand on the analysis of incorrect captions in comparison to underspecified captions, as will be described in the next section.

3.2 Ablations and Chosen Models

To evaluate whether Pezzelle's results generalize to other CLIP models, we perform the existing proof of concepts across a broader range of CLIP-based models that vary in training objectives, scale, and architectural design. We divide this analysis into two components: (1) breadth analysis, which compares performance across different model families at a fixed scale and resolution, and (2) depth analysis, which investigates changes in performance as scale within each model family changes.

For breadth analysis, to control confounding factors, we fix the model scale and image resolution. Specifically, we evaluate five models: the original CLIP, CLIP with QuickGELU activation, SigLIP, which replaces softmax loss with a sigmoid-based one (Zhai et al., 2023), SigLIP2, which introduces additional objectives on top of SigLIP (Tschannen et al., 2025), and RADIOv2.5-B, a model that adaptively retrieves models for each input image at runtime (Heinrich et al., 2025). We use the ViT-B/16 backbone for each model, and all inputs are resized to 224 × 224 pixels, except for RADIO, which does not operate at a fixed resolution due to its architecture. This specific scale and resolution were chosen simply because all the selected model variants are available in this configuration via the OpenCLIP (Cherti et al., 2023) and NVIDIA Labs (Heinrich et al., 2025) repositories, ensuring a consistent comparison baseline.

For depth analysis, we hold resolution constant within each model family–again with the exception of RADIO–and vary the model scale. This allows us to assess whether increasing model scale improves robustness. The selected configurations are: CLIP and CLIP-quickgelu at ViT-B/16 and ViT-L/14 scales, SigLIP at ViT-B/16, ViT-L/16, and a 400M-14 variant (400 million parameters, patch size 14), SigLIP2 at ViT-B/16, ViT-L/16, and 400M-16, and RADIO at ViT-B/16, ViT-L/16, ViT-H/16, and ViT-H/14. While CLIP and CLIP-quickgelu variants are evaluated at 224 × 224 resolution, SigLIP and SigLIP2 variants are evaluated at 384 × 384. This choice was simply because these resolutions had the most model scales available for each model family in OpenCLIP.

Next, while Pezzelle's second proof of concept result is interesting, there are a few areas where we believed it could be expanded upon. Most notably, Pezzelle (2023) comments that "the model could be 'dazzled' by the presence of words that have a grounded referent in the image ... that could lead it to assign some similarity even when the sentence is completely out of place," meaning that the captions are chosen randomly from the dataset, and thus there is a chance that the random caption could somewhat describe the image and thus get a better score than a truly incorrect caption. Due to this element of randomness, the result that most images had similarity scores where at least one incorrect caption outperformed the underspecified caption does not necessarily imply that a truly incorrect caption is better than an underspecified one.

With these limitations in mind, we wanted to analyze the performance of incorrect captions against the underspecified captions more thoroughly. The revised second proof of concept that we performed followed the same structure as the first proof of concept. Similarly to how the first proof of concept altered captions to be underspecified in certain areas in order to determine how much underspecification impacted the performance of the models, the second proof of concept would now alter the captions in the same categories of Quantity, Gender, Gender and Number, Location, Object, and a fully incorrect caption in order to determine how much incorrectness impacted the performance of the models under the same metric. Furthermore, this would allow us to compare incorrectness results results to those of the underspecified captions. The similarity scores for the altered captions would also be determined for all of the models added in the ablations to the first proof of concept for a full parallel comparison of the two types of captions.

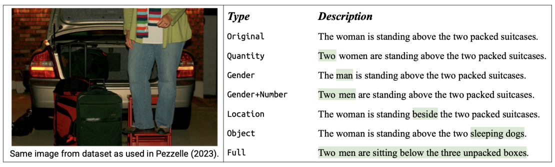
3.3 Dataset Generation

To generate our evaluation dataset for incorrect captioning, we built directly upon the same set of images and original captions used in the underspecified caption dataset introduced by Pezzelle (2023) as previously mentioned. This ensured that any changes we introduced could be fairly compared across both types of caption alteration: underspecification and incorrectness.

To achieve this, we used OpenAI's ChatGPT Python API to programmatically alter each original caption. We crafted prompt templates tailored to each alteration type. For example, a prompt to induce a full incorrectness transformation would ask ChatGPT to change the gender and number of the subject, alter the location, and replace the object of the sentence, ensuring the caption no longer accurately described the image. We provide an example while prompting, making the prompts one-shot. A sample prompt is as follows:

> *"Given this caption: "The woman is standing above the two packed suitcases." Alter the caption by changing the gender and quantity of the subject, the location, and the object of the sentence. Only output the altered caption and nothing else. For example, for the sentence "Three women are having a picnic at the park.", the sentence might change to "Four men are having a basketball game at the house.""*

After generating the altered captions for each image, we manually reviewed the outputs to ensure that they were indeed incorrect in a meaningful way and remained syntactically coherent. This postprocessing step was crucial to filter out any erroneous completions or edge cases where ChatGPT produced a paraphrased or trivially modified version rather than a semantically incorrect one. Figure 2 shows an example of how each category would be changed for one image.



Same image from dataset as used in Pezzelle (2023).

Type	*Description*
Original	The woman is standing above the two packed suitcases.
Quantity	Two women are standing above the two packed suitcases.
Gender	The man is standing above the two packed suitcases.
Gender+Number	Two men are standing above the two packed suitcases.
Location	The woman is standing beside the two packed suitcases.
Object	The woman is standing above the two sleeping dogs.
Full	Two men are sitting below the three unpacked boxes.

Figure 2: Sample transformation from original to incorrect captions

The resulting dataset includes incorrect variants for each of the original semantic alteration categories and allows for a structured comparison between model responses to underspecified versus incorrect inputs.

3.4 Code

The code for all of this paper's work is found at https://github.com/AnuptaA/COS484-Final_Project. Our files can also be found at https://drive.google.com/drive/folders/1LjhNeeRnQiU3TZ3TMmBqtPOApO4CFa7R?usp=share_link.

4 Results

4.1 Depth Results

The line plots summarizing the depth analyses for both the underspecification proof of concept and incorrectness proof of concept–showing the mean similarity scores for each model variant and scale across caption alterations–are included in Appendix A for conciseness. However, these results will still be described in this section and analyzed in Section 5. The main observation from these results was that, across all model families and both proofs of concept, the scale of a model within an individual family did not change how much the altered captions impacted the similarity scores. However, it will be noted that the extent that the types of altered captions impacted the similarity scores, such as Quantity alterations or Location alterations, was different for different models. Most prominently, the RADIO model family was impacted by the fully underspecified and fully incorrect captions much more than the other models, even if some other models resulted in lower absolute scores.

Furthermore, the violin plots for all models for the underspecification proof of concept and incorrectness proof of concept are shown in Appendix B, including the recreation of Pezzelle's first proof of concept determining the similarity scores of underspecified captions for the ViT-B/32 CLIP Model. Again here we see that the performance was very similar across scales for all model families and both caption types.

4.2 Breadth Results

The results for the breadth analysis described in Section 3.2 of the model families for both proofs of concept are shown in this section. Two kinds of visualizations of the mean similarity scores for the ViT-B/16 models at a 224 × 224 resolution in each family of models for each type of altered caption are provided. This first visualization is a line plot of the mean similarity scores of each family for each proof of concept. The second visualization is a heatmap for the differences of the mean similarity scores.

Underspecification Proof of Concept Results

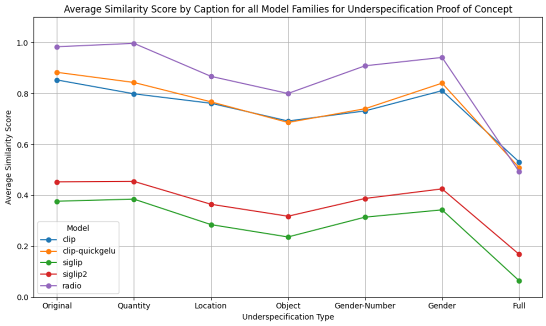


Figure 3: Mean Similarity for All Model Families for Underspecification Proof of Concept

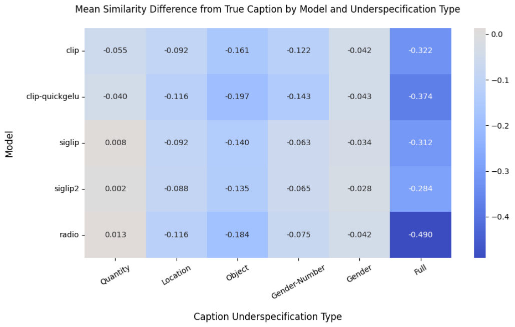


Figure 4: Mean Similarity Difference of Underspecification and Original Caption by Model and Underspecification Type

Figure 3 is the line plot for the underspecification proof of concept. Figure 4 displays the heatmap for the underspecification proof of concept. The difference calculated for this heatmap is between the mean similarity score of the underspecified caption and the original, or true, caption, and this can be stated in a shorthand form as $\text{Sim}_{\text{und}} - \text{Sim}_{\text{true}}$. Therefore, as the plot becomes more negative, the difference between the performance of the underspecified caption and the original caption is greater and the underspecified caption has a particularly larger impact. It can be seen from both the line plot and the heatmap that all model families are impacted by the types of underspecification similarly, though some models are impacted slightly more than others.

Incorrectness Proof of Concept Results

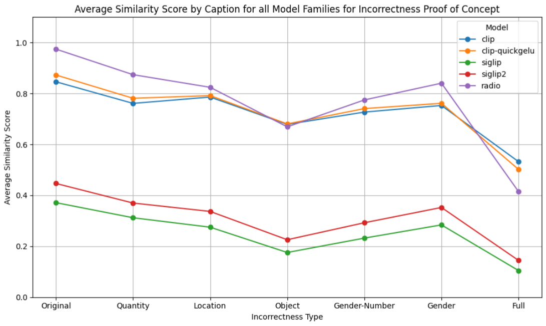


Figure 5: Mean Similarity for All Model Families for Incorrectness Proof of Concept

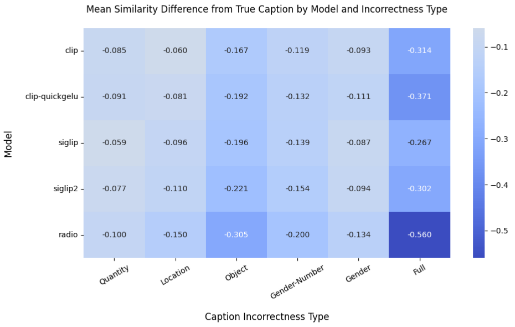


Figure 6: Mean Similarity Difference of Incorrectness and Original Caption by Model and Incorrectness Type

Figure 5 is the line plot for the incorrectness proof of concept. Figure 6 displays the heatmap for the incorrectness proof of concept. The difference calculated for this heatmap is between the mean similarity score of the incorrect caption and the original, or true, caption, and this can be stated in a shorthand form as $\mathrm{Sim}_{\mathrm{inc}} - \mathrm{Sim}_{\mathrm{true}}$. Therefore, as the plot becomes more negative, the difference between the performance of the incorrect caption and the original caption is greater and the incorrect caption has a particularly larger impact. Just as was determined for the underspecification proof of concept, the line plot and heatmap demonstrate both that all model families are impacted similarly by the incorrect captions as well as that this impact also aligns with the impact from underspecified captions.

Comparison Between Underspecification and Incorrectness

To further illustrate the similarities between the impact of underspecified captions and incorrect captions, Figure 7 was created.

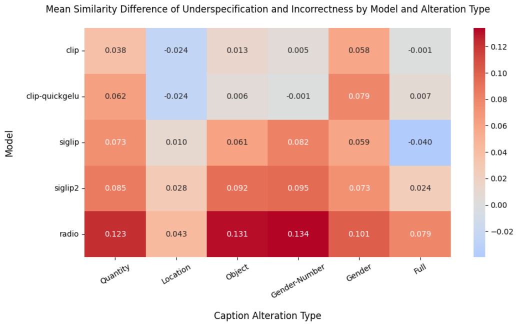


Figure 7: Mean Similarity Difference of Underspecification and Incorrectness by Model and Alteration Type

The difference calculated in Figure 7 is between the mean similarity score of the underspecified caption and the incorrect caption, and this can be stated in a shorthand form as $\mathrm{Sim}_{\mathrm{und}} - \mathrm{Sim}_{\mathrm{inc}}$. Therefore, as the plot becomes more positive, the difference between the performance of the incorrect caption and the underspecified caption is greater and the incorrect caption has a particularly larger impact. It will be noted that, since the similarity scores are on a scale from zero to one, the differences found in this heatmap are much less than the differences in Figure 4 and Figure 6. Furthermore, the differences for the fully underspecified captions and fully incorrect captions are particularly near zero and fluctuate between demonstrating underspecified captions as being worse and incorrect captions as being worse.

5 Analysis

5.1 Understanding Underspecification Across Model Families and Scales

From the results of the depth and breadth analysis, it can be seen that all of the chosen CLIP-based vision-language models were impacted similarly by underspecification in the way that Pezzelle's paper demonstrated with ViT-B/32 CLIP. The underspecification which seemed to perform the worst, excluding the fully underspecified captions, was the underspecification of the object in the caption, and this makes sense as these would be nouns corresponding to the specific scenario in the image that would change the caption the most and therefore require the most context and intuition in our language to understand properly. The model family which handled underspecification the worst was the RADIO models, and this is interesting because they demonstrated themselves to be the best performing models on the original captions.

5.2 Understanding Incorrectness Across Model Families and Scales

Similarly to the analysis of the underspecification proof of concept, the results of the depth and breadth analysis for the incorrectness proof of concept demonstrated that the chosen CLIP-based vision-language models were impacted similarly by incorrectness, with the RADIO model once again being impacted the most and the captions with incorrect objects also performing the worst, excluding the fully incorrect captions, and this can be explained by the same reasoning as the underspecification captions.

5.3 Underspecification as Incorrectness

The most interesting aspect of the results was the expansion on Pezzelle's findings for the second proof of concept. As mentioned previously, Pezzelle was only able to comment that certain incorrect captions had the potential to perform better than fully underspecified ones, likely due to coincidentally specific words greatly improving the similarity score. However, the similar patterns of performance between the impact of underspecification and the impact of incorrectness indicate that the connection between these caption alterations may be more significant than the models preferring certain words in incorrect captions than fully correct underspecified captions: CLIP-based vision-language models may interpret underspecified phrases and alterations in the same way that they interpret incorrect phrases and alterations.

We propose that this is a reasonable conclusion from the results because the incorrect caption dataset being constructed in the same way as Pezzelle's underspecified dataset allows for a fairly direct comparison between the determined similarity scores by limiting other potential variations in the captions which were present in Pezzelle's second proof of concept. Thus, when the similarity score for a type of caption alteration changes a comparable amount for the instance of the caption being made more underspecified and the instance of the caption being made more incorrect, these are the only factors which would have changed, implying that the similar impact for both originates from very practically similar interpretation of both alterations.

6 Conclusion and Future Work

Our study demonstrates a consistent and concerning pattern across CLIP-based vision-language models: they struggle to distinguish between captions that are merely underspecified and those that are factually incorrect. Despite meaningful differences in semantics, the models exhibited similar drops in image-text similarity for both types of captions, suggesting a fundamental limitation in their contextual reasoning capabilities.

This inability to separate ambiguity from outright error has serious real-world implications. Vision-language models are increasingly integrated into critical systems like autonomous vehicles, where misinterpreting vague descriptions as correct or failing to flag incorrect information could lead to catastrophic decision-making.

To address this, we propose two key directions for future work:

- **Exploring Alternative Architectures.** We aim to expand our evaluation to non-vision-transformer based CLIP models, but also to models beyond the CLIP family, including other model architectures—such as cross-attention fusion models, generative captioning models, and retrieval-augmented systems—to assess whether different inductive biases or training paradigms better support semantic underspecification. This would include also examining the effects of model scale, training objectives, and embedding space design.

- **Scaling and Refining the Incorrect Caption Dataset.** A major limitation of our current work is the relatively small size and manual curation of the incorrect caption dataset. Going forward, we plan to automate this pipeline more robustly, improving prompt engineering and validation heuristics to reduce the need for manual oversight. Our goal is to create a large-scale, diverse benchmark of incorrect captions to rigorously test and train models on their ability to flag semantic error—something that will be essential for safe deployment in high-stakes scenarios.

By pursuing these directions, we can hope to advance the reliability and interpretability of vision-language models, moving closer to systems that can reason with nuance and resist failure in real-world settings.

Limitations

While our findings reveal consistent performance drops across caption types, there are several limitations to keep in mind when regarding our conclusions.

First, all experiments were conducted using CLIP models built on vision transformers (ViTs), as opposed to architectures like ResNet-based CLIP. It's possible that the models' inability to differentiate underspecified from incorrect captions is partly a function of their transformer-based structure. Future work should test whether this effect generalizes across different vision backbones.

Second, although we attempted to carefully construct and review the incorrect caption dataset, the generation process is not without noise. Some ChatGPT outputs may have introduced only mild semantic changes, or conversely, altered the caption in multiple dimensions beyond the intended category. This imperfection introduces variability that may have affected model scoring in subtle ways. While we manually checked the outputs, the scale was limited, and further automation with better validation is needed to ensure consistency.

Third, our analysis assumes that similarity score degradation reflects a model's confusion between underspecification and incorrectness. However, we cannot conclusively determine whether models are *confounding* the two phenomena, or whether they simply handle both poorly and produce coincidentally similar outcomes. Disentangling these effects would require a deeper probe into the internal representations and failure modes of the models.

Lastly, while we examined model architecture and scale, we held other variables constant, particularly the resolution and pretrained datasets–which could also influence robustness to caption variation. Additionally, the number of model scales that we examined were also limited. Future work should further explore how performance varies across different resolutions, pretraining datasets, and a broader range of model scales to more thoroughly understand the factors contributing to underspecified caption vulnerability.

Acknowledgements

We would like to acknowledge our COS 484 instructors for their teaching, as well as our project mentor Colin Wang for his support and guidance during our research process.

References

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE.

Greg Heinrich, Mike Ranzinger, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, Pavlo Molchanov, et al. 2025. Radiov2.5: Improved baselines for agglomerative vision foundation models. *arXiv preprint arXiv:2412.07679*.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal NLP. *To appear in the Proceedings of ACL 2023*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

A Appendix: Depth Analyses

In this Appendix, the depth analyses for both the underspecification proof of concept and the incorrectness proof of concept are shown. As previously stated, we believed these results demonstrated that the model scale did not change the impact of underspecification or incorrectness for a given model family due to the similarity of the changes in mean similarities.

A.1 Underspecification Depth Analyses

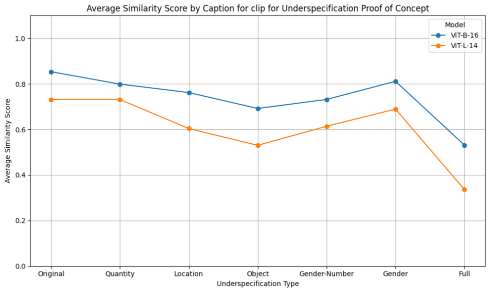


Figure 8: Mean Similarity Score for CLIP Models for Underspecification Proof of Concept

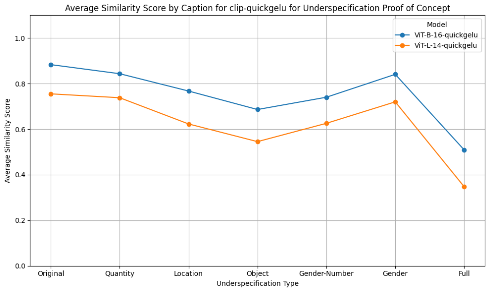


Figure 9: Mean Similarity Score for CLIP with QuickGELU Activation Models for Underspecification Proof of Concept

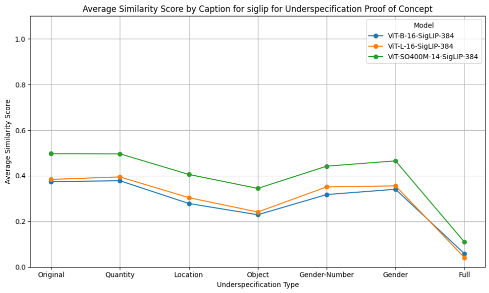


Figure 10: Mean Similarity Score for SigLIP Models for Underspecification Proof of Concept

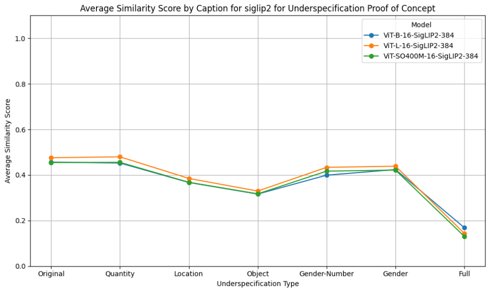


Figure 11: Mean Similarity Score for SigLIP2 Models for Underspecification Proof of Concept

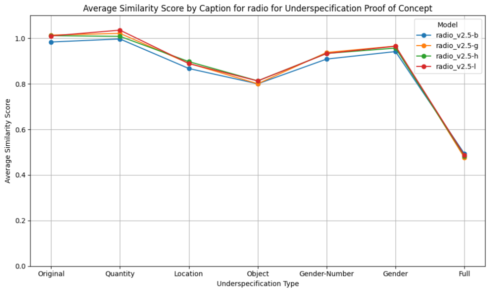


Figure 12: Mean Similarity Score for RADIO Models for Underspecification Proof of Concept

A.2 Incorrectness Depth Analyses

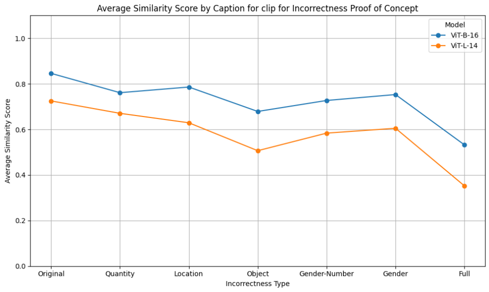


Figure 13: Mean Similarity Score for CLIP Models for Incorrectness Proof of Concept

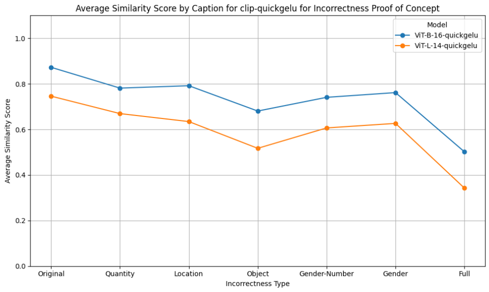


Figure 14: Mean Similarity Score for CLIP with QuickGELU Activation Models for Incorrectness Proof of Concept

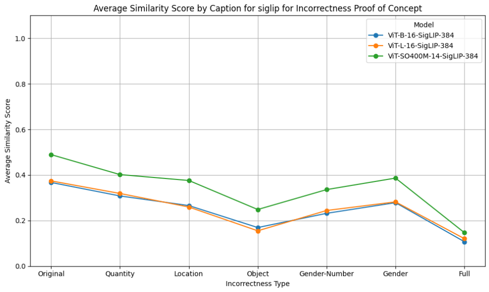


Figure 15: Mean Similarity Score for SigLIP Models for Incorrectness Proof of Concept

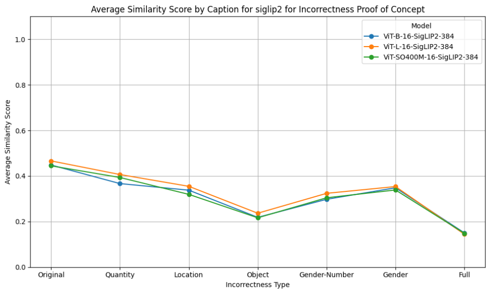


Figure 16: Mean Similarity Score for SigLIP2 Models for Incorrectness Proof of Concept

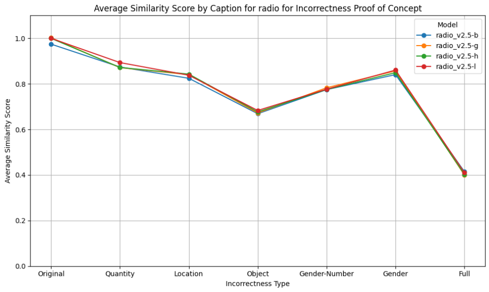


Figure 17: Mean Similarity Score for RADIO Models for Incorrectness Proof of Concept

B Appendix: Violin Plots

This Appendix displays the violin plots made for reproducing Pezzelle's first proof of concept. The violin plot which displays the results of the model that Pezzelle (2023) used, the CLIP ViT-B-32 model, can be see in Figure 20.

B.1 Underspecification Violin Plots

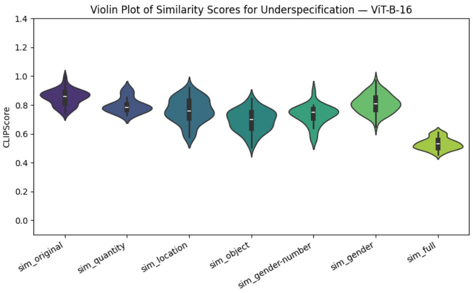


Figure 18: Similarity Score Violin Plot for CLIP ViT-B-16 Model for Underspecification Proof of Concept

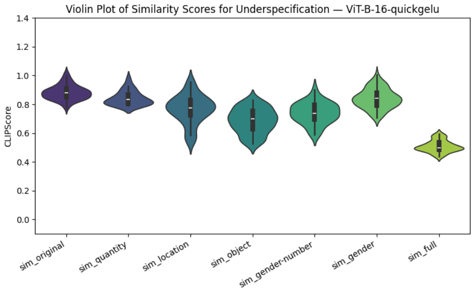


Figure 19: Similarity Score Violin Plot for CLIP ViT-B-16 with QuickGELU Activation Model for Underspecification Proof of Concept

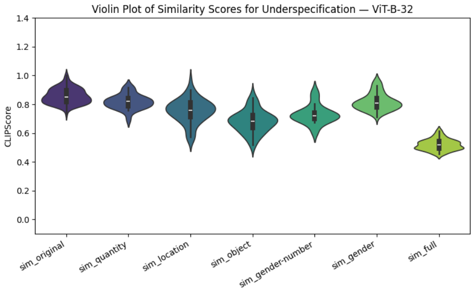


Figure 20: Similarity Score Violin Plot for CLIP ViT-B-32 Model for Underspecification Proof of Concept

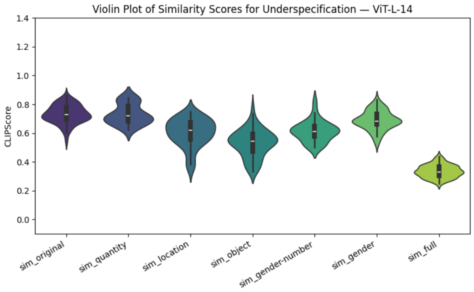


Figure 21: Similarity Score Violin Plot for CLIP ViT-L-14 Model for Underspecification Proof of Concept

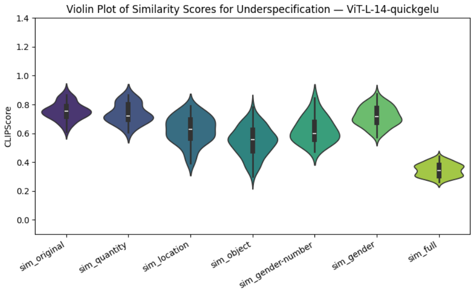


Figure 22: Similarity Score Violin Plot for CLIP ViT-L-14 with QuickGELU Activation Model for Underspecification Proof of Concept

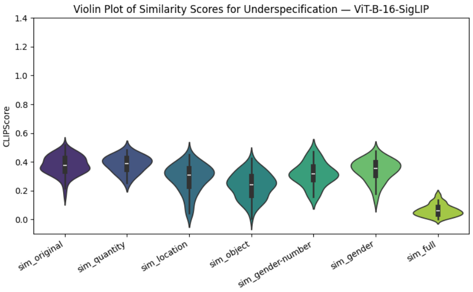


Figure 23: Similarity Score Violin Plot for SigLIP ViT-B-16 Model for Underspecification Proof of Concept

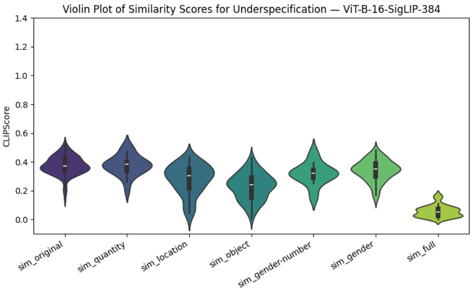


Figure 24: Similarity Score Violin Plot for SigLIP ViT-B-16 with 384x384 Resolution Model for Underspecification Proof of Concept

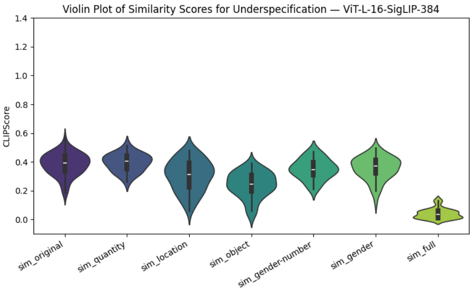


Figure 25: Similarity Score Violin Plot for SigLIP ViT-L-16 with 384x384 Resolution Model for Underspecification Proof of Concept

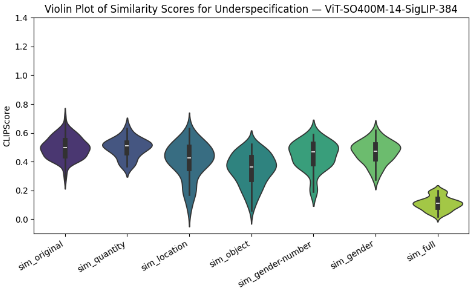


Figure 26: Similarity Score Violin Plot for SigLIP ViT-SO400M-14 with 384x384 Resolution Model for Underspecification Proof of Concept

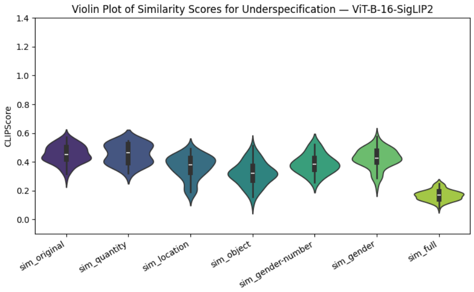


Figure 27: Similarity Score Violin Plot for SigLIP2 ViT-B-16 Model for Underspecification Proof of Concept

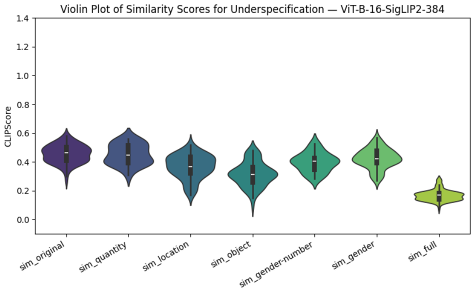


Figure 28: Similarity Score Violin Plot for SigLIP2 ViT-B-16 with 384x384 Resolution Model for Underspecification Proof of Concept

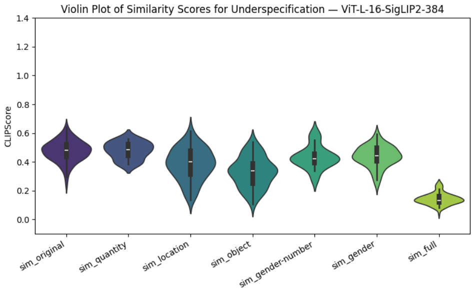


Figure 29: Similarity Score Violin Plot for SigLIP2 ViT-L-16 with 384x384 Resolution Model for Underspecification Proof of Concept

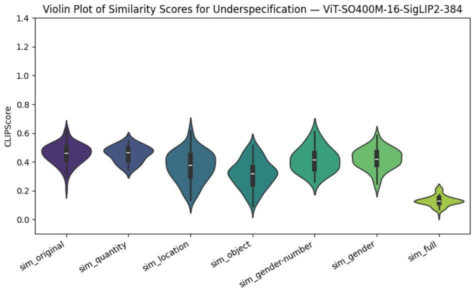


Figure 30: Similarity Score Violin Plot for SigLIP2 ViT-SO400M-16 with 384x384 Resolution Model for Underspecification Proof of Concept

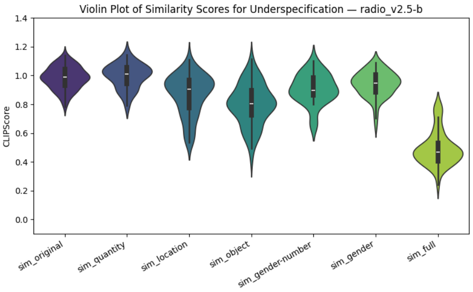


Figure 31: Similarity Score Violin Plot for RADIO-b Model for Underspecification Proof of Concept

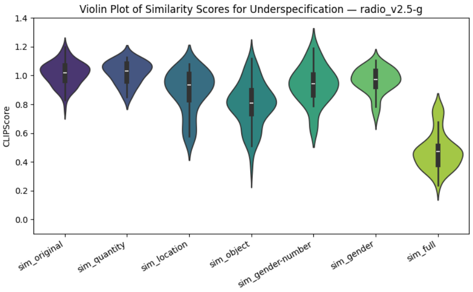


Figure 32: Similarity Score Violin Plot for RADIO-g Model for Underspecification Proof of Concept

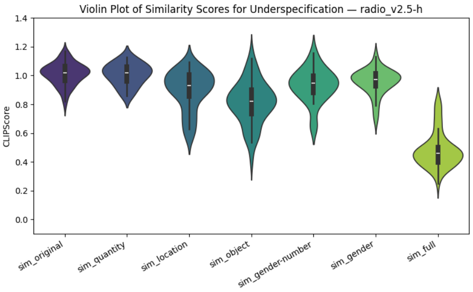


Figure 33: Similarity Score Violin Plot for RADIO-h Model for Underspecification Proof of Concept

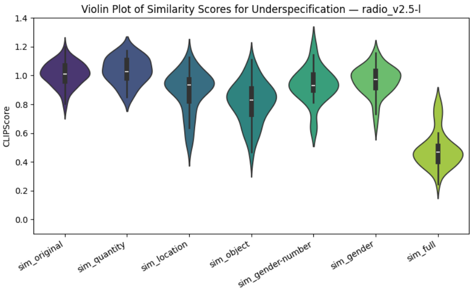


Figure 34: Similarity Score Violin Plot for RADIO-l Model for Underspecification Proof of Concept

B.2 Incorrectness Violin Plots

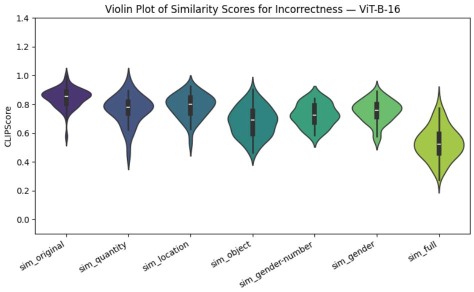


Figure 35: Similarity Score Violin Plot for CLIP ViT-B-16 Model for Incorrectness Proof of Concept

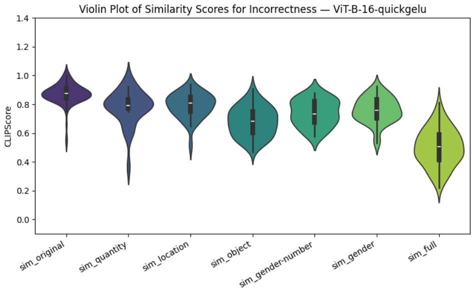


Figure 36: Similarity Score Violin Plot for CLIP ViT-B-16 with QuickGELU Activation Model for Incorrectness Proof of Concept

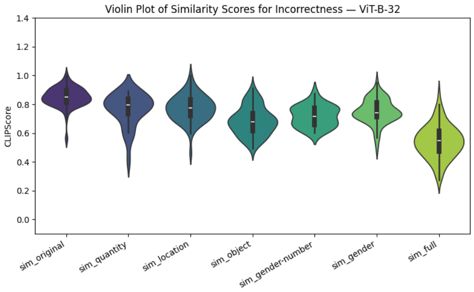


Figure 37: Similarity Score Violin Plot for CLIP ViT-B-32 Model for Incorrectness Proof of Concept

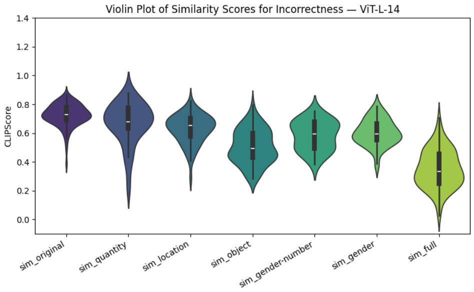


Figure 38: Similarity Score Violin Plot for CLIP ViT-L-14 Model for Incorrectness Proof of Concept

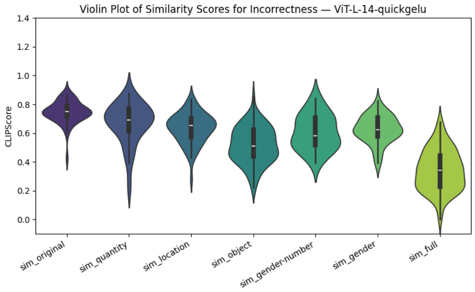


Figure 39: Similarity Score Violin Plot for CLIP ViT-L-14 with QuickGELU Activation Model for Incorrectness Proof of Concept

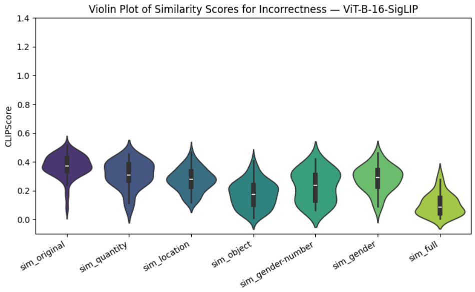


Figure 40: Similarity Score Violin Plot for SigLIP ViT-B-16 Model for Incorrectness Proof of Concept

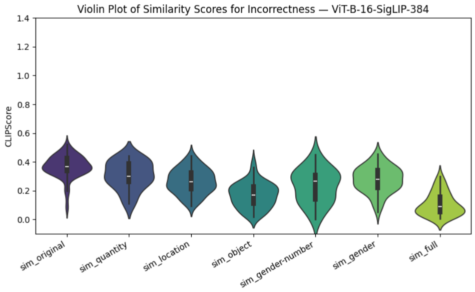


Figure 41: Similarity Score Violin Plot for SigLIP ViT-B-16 with 384x384 Resolution Model for Incorrectness Proof of Concept

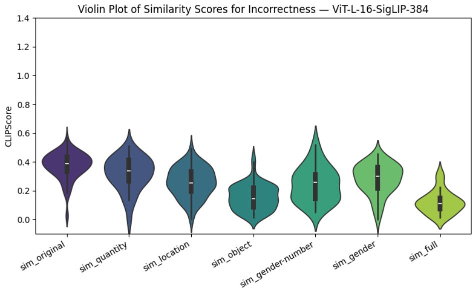


Figure 42: Similarity Score Violin Plot for SigLIP ViT-L-16 with 384x384 Resolution Model for Incorrectness Proof of Concept

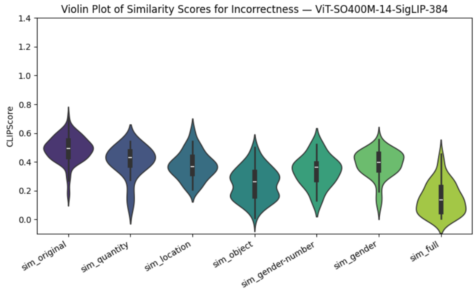


Figure 43: Similarity Score Violin Plot for SigLIP ViT-SO400M-14 with 384x384 Resolution Model for Incorrectness Proof of Concept

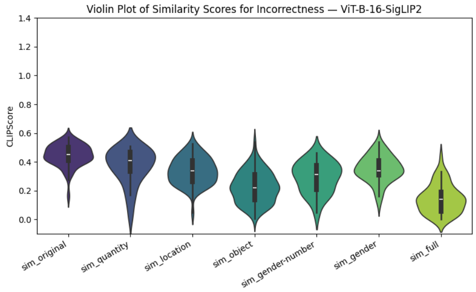


Figure 44: Similarity Score Violin Plot for SigLIP2 ViT-B-16 Model for Incorrectness Proof of Concept

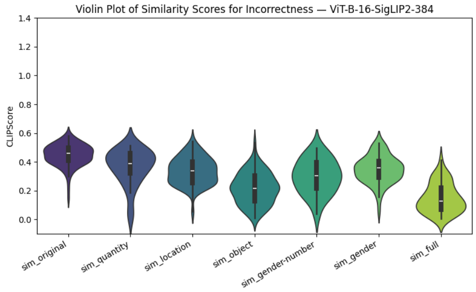


Figure 45: Similarity Score Violin Plot for SigLIP2 ViT-B-16 with 384x384 Resolution Model for Incorrectness Proof of Concept

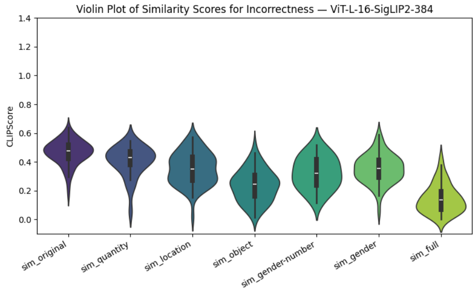


Figure 46: Similarity Score Violin Plot for SigLIP2 ViT-L-16 with 384x384 Resolution Model for Incorrectness Proof of Concept

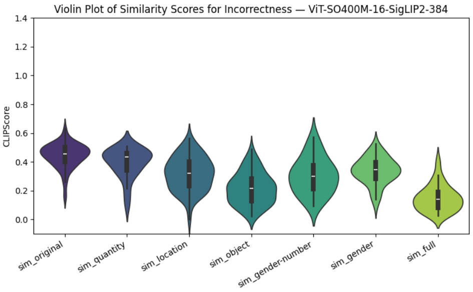


Figure 47: Similarity Score Violin Plot for SigLIP2 ViT-SO400M-16 with 384x384 Resolution Model for Incorrectness Proof of Concept

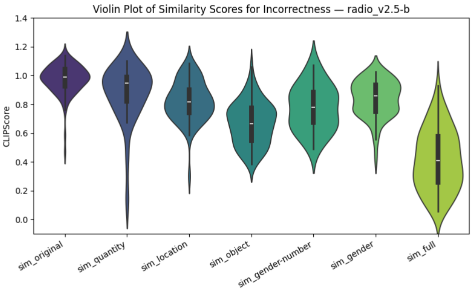


Figure 48: Similarity Score Violin Plot for RADIO-b Model for Incorrectness Proof of Concept

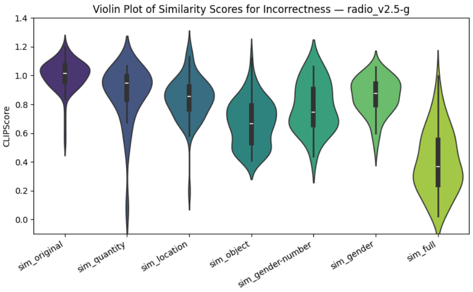


Figure 49: Similarity Score Violin Plot for RADIO-g Model for Incorrectness Proof of Concept

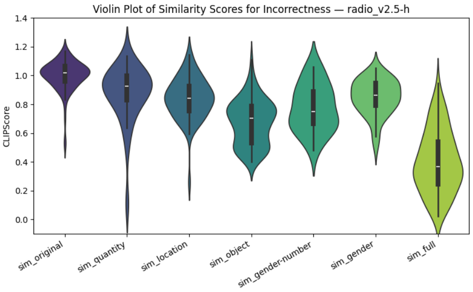


Figure 50: Similarity Score Violin Plot for RADIO-h Model for Incorrectness Proof of Concept

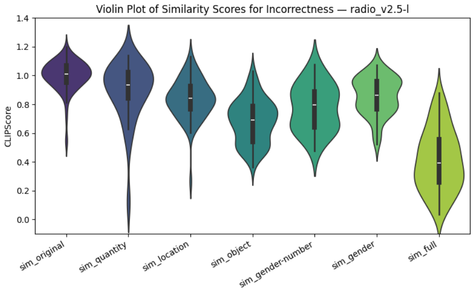


Figure 51: Similarity Score Violin Plot for RADIO-l Model for Incorrectness Proof of Concept