Supplementary Materials: Enhancing Relation and Semantic Understanding in Multiple Instances for Visual Grounding

Anonymous Authors

In this supplementary material, we provide additional details on our approach and more experimental results. In Section 1, we present more experimental results under the across datasets setting. In Section 2, we provide new ablation experiments regarding semantic-sensitive prior injection. In Section 3, we provide additional implementation details of our method.

1 CROSS-DATASET COMPARASION

In Table 6 of the main text, we provide the performance comparison between our method and TransVG under the cross-dataset settings. In Table 1 of the supplementary materials, we offer further comparisons of cross-dataset performance using different baseline models. From the table, we observe that when models are trained on RefCOCO [6] and tested on RefCOCO+ [6], both baseline methods exhibit significant performance drops. In contrast, our method maintains higher performance while suffering less performance degradation compared to the IID setting (for instance, on the validation set, TransVG decreases by 4.13%, whereas our method only decreases by 0.17%). This demonstrates the generalization ability of our approach.

2 MORE ABLATION STUDIES

Designs of semantic-sensitive visual grounding. Our Semantic-Sensitive Visual Grounding enhances the understanding of fine-grained semantics of target objects by injecting semantic priors generated from the Stable Diffusion model [4]. In this section, we present ablation experiments on using data augmentation similar to our relation-sensitive augmentation to improve the model's understanding of object semantics. Specifically, we utilize the Stable Diffusion to generate images containing the target object based on input query and employ an object detector to detect the bounding box of the target object. We treat the queries, generated images, and detected boxes as new augmented samples, which are mixed with the original dataset to train the model. The performance on the RefCOCO+ dataset is shown in Table 3. It is observed that using data augmentation does not enhance the model's understanding of fine-grained semantics; instead, it leads to a decrease in performance. We attribute this to the fact that the images generated from the query typically contain only a single object and focus on a zoomed-in, object-centric view. In such cases, the model only needs to output the unique object box in the image without needing to understand the fine-grained semantics of the object, which cannot be generalized to the test data. Therefore, the model's performance declines. We also can not use the prompt of class name with a random quantifier to generate images like what we do in relation-sensitive data augmentation. This is because we find that the multiple instances generated by stable diffusion have similar fine-grained attributes, and it is difficult to construct pseudo queries that can distinguish them solely based on fine-grained semantics.

Method	val	testA	testB
Train and test on RefCOCO+			
TransVG[1]	63.50	68.15	55.63
TransVG+ours	**66.13**	**70.95**	**62.06**
VLTVG [5]	73.60	78.37	64.53
VLVTG+ours	**73.95**	**79.53**	**64.88**
Train on RefCOCO, test on RefCOCO+			
TransVG[1]	59.37	65.65	50.14
TransVG+ours	**65.96**	**70.48**	**54.71**
VLTVG [5]	67.69	74.47	58.68
VLVTG+ours	**68.32**	**75.11**	**59.24**

Table 1: Performance of cross-dataset where models are trained on RefCOCO dataset and tested on RefCOCO+ dataset.

Method	val	testA	testB
Baseline(TransVG)	63.50	68.15	55.63
Data Augmentation	62.28	66.50	54.13
Semantic Prior Injection	**66.13**	**70.95**	**62.06**

Table 2: Ablations of the designs of semantic-sensitive visual grounding on RefCOCO+ dataset.

Number of prior images	val	testA	testB
0	63.50	68.15	55.63
1	66.13	70.95	**62.06**
2	**66.17**	70.89	62.03
3	66.09	**71.01**	62.00

Table 3: Ablations on the number of prior images on RefCOCO+ dataset.

Ablation on the number of prior images. In the main text, we generated only one prior image for each query in the Semantic Prior Injection. In Table 3, we attempt to generate more prior images and provide ablation experiments on the number of prior images. It can be observed that compared to the results with zero prior images, using one prior image can significantly improve the performance of the model. This demonstrates that our semantic prior injection module can effectively enhance the model's understanding of the fine-grained semantics of the target object, thus improving performance. However, despite the number of prior images being further increased, there hasn't been a significant improvement in the model's performance. This may be because the stable generative capacity of the Stable Diffusion model ensures that the generated images are highly aligned with the textual queries and thus one prior image provides enough semantic information. Additionally, increasing the number of prior images will increase the model's time consumption. Therefore, we ultimately choose to use only one prior image.

3 IMPLEMENTATION DETAILS

Details of our method based on VLVTG [5]. VLVTG is also a Transformer-based model. It consists of a visual-linguistic verification module to optimize visual features to focus more on the target object referred to by the query, a language-guided context encoder to gather information from visual context to aid in visual grounding, and a multi-stage cross-modal decoder that iteratively fuses visual and textual features from a randomly initialized target query to more accurately retrieve object representations. When applying our approach to VLVTG, we replace the randomly initialized target query in the multi-stage cross-modal decoder with the semantic-aware token from Equation (4) in our main text and keep other modules unchanged. We also use the augmented Relation-Sensitive Training Dataset to train the model to enhance its understanding of spatial relationships.

Details of text-to-image model. In the Relation-Sensitive Data Augmentation and the Semantic Prior Injection, we use the Stable Diffusion (SDXL-turbo) [3] as the text-to-image model.

Details of the pseudo-query generation. In the Relation-Sensitive Data Augmentation, we utilize the method in CPL [2] to generate pseudo-queries for each object. We predefine a set of spatial relationships and determine the spatial relation of each object by comparing the center coordinates and area of each object box output by the detector. Specifically, we have considered seven basic relational terms: left, right, front, behind, middle, top, and bottom. Among them, left, right, middle, top, and bottom will be determined by comparing the central coordinates of the boxes outputted by the detector. The front and behind will be determined based on the size of the box area, based on the assumption that for the objects in the same category, the front one should have a larger object region. In addition to these basic relational terms, we also incorporate ordinal numbers based on the sorting of the box central coordinates (e.g., second left) and consider combinations of these basic relational terms (e.g., left bottom). Finally, we obtain the pseudo queries based on the template '{Rela} {Noun}', such as 'middle orange'.

REFERENCES

[1] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1769–1779.

[2] Yang Liu, Jiahua Zhang, Qingchao Chen, and Yuxin Peng. 2023. Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2828–2838.

[3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[5] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9499–9508.

[6] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, 69–85.