Appendix for: Pseudo-Q Generating Pseudo Language Queries for Visual Grounding

A. Statistics of RefCOCO Dataset

In Figure 1, we show the statistics of the training set of RefCOCO [5] dataset to demonstrate spatial relationship is one of the dominant components in language queries. As we can see, spatial relationships exists in almost 60% of queries. Furthermore, the most common spatial relationships in RefCOCO are *left* and *right*. In addition, other spatial relationships, *i.e., middle*, *front*, *top*, and *bottom*, are also frequently found in language queries.

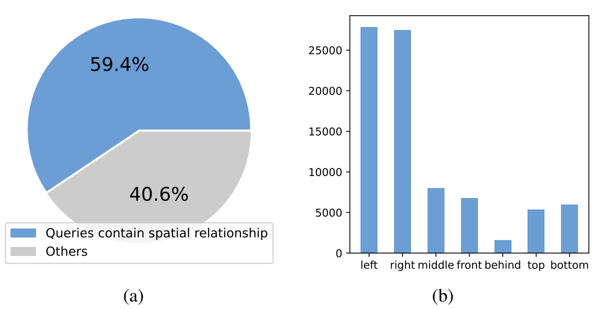

Figure 1. **Statistics of the training set of RefCOCO [5] dataset.** (a): The percent of language queries that contain spatial relationships. (b): The number of different spatial relationships.

B. Pseudo-Query Templates

Our pseuod-queries are generated following the templates shown in Table. 1. All the possible templates is considered in our method for the purpose of obtaining as many candidate pseudo-samples as possible. Honestly, this strategy will inevitably produce some ungrammatical pseudo-samples. Our approach is similar to all the pseudo-label based methods, such as semi-supervised learning, which can't guarantee every single pseudo-query is correct. Overall, these pseudo-queries provide valuable supervision signals and eventually benefit the training of the model.

C. Visual-Language Model

In this section, we provide more details about the architecture of the visual encoder and the language encoder.

In the visual encoder, a CNN backbone and a transformer-based network are stacked sequentially for image feature extraction. The CNN backbone is a ResNet-50 model [4] pre-trained on ImageNet [2], and the transformer-based network is the encoder part of DETR network [1] which consists of six transformer layers. Moreover, the pretrained weights of DETR are utilized for initialization. The output feature maps of the ResNet-50 are fed into a 1×1 convolutional layer for dimension reduction. Then, they are flattened into 1D vectors for the transformer network.

In the language encoder, a token embedding layer and a linguistic transformer are employed to extract textual features. Specifically, the token embedding layer is leveraged to convert the discrete words into continuous language vectors. Since BERT [3] has been successfully applied for text feature extraction, the BERT architecture which has 12 transformer layers is adopted as the linguistic transformer.

Table 1. Pseudo-query templates. *Attr* and *Rela* represents attribute and relationship, respectively.

Pseudo Query Template	Example
{*Noun*}	"man", "building" etc.
{*Noun*} {*Attr*}	"man standing" etc.
{*Attr*} {*Noun*}	"talk man", "wooden building" etc.
{*Noun*} {*Rela*}	"man on the right" etc.
{*Rela*} {*Noun*}	"center man", "left building" etc.
{*Noun*} {*Attr*} {*Rela*}	"man standing on the right" etc.
{*Noun*} {*Rela*} {*Attr*}	"man right standing" etc.
{*Attr*} {*Noun*} {*Rela*}	"standing man on the right" etc.
{*Attr*} {*Rela*} {*Noun*}	"standing right man" etc.
{*Rela*} {*Noun*} {*Attr*}	"right man standing" etc.
{*Rela*} {*Attr*} {*Noun*}	"right standing man" etc.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional

transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1