A Better Loss for Visual-Textual Grounding

Supplementary Material

Davide Rigoni
University of Padua
Bruno Kessler Foundation
Padua, Italy
drigoni@fbk.eu

Luciano Serafini
Bruno Kessler Foundation
Povo, Italy
serafini@fbk.eu

Alessandro Sperduti
University of Padua
Padua, Italy
sperduti@unipd.it

ABSTRACT

In this paper, we provide supplementary material for the paper "A Better Loss for Visual-Textual Grounding", which has been accepted to be presented at the 37th ACM/SIGAPP Symposium On Applied Computing.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; **Object detection**.

KEYWORDS

Computer Vision, Visual Textual Grounding, Semantic Loss

ACM Reference Format:
Davide Rigoni, Luciano Serafini, and Alessandro Sperduti. 2022. A Better Loss for Visual-Textual Grounding: Supplementary Material. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477314.3507047

1 BACKGROUND

We use the following notation: lower case symbols for scalars and indexes, e.g. n; italics upper case symbols for sets, e.g. A; upper case symbols for textual sentences, e.g. S; bold lower case symbols for vectors, e.g. $\boldsymbol{a}$; bold upper case symbols for matrices and tensors, e.g. $\boldsymbol{A}$; the position within a tensor or vector is indicated with numeric subscripts, e.g. $\boldsymbol{A}_{ij}$ with $i, j \in \mathbb{N}^+$; calligraphic symbols for domains, e.g. $\mathcal{Q}$.

1.1 Intersection over Union (IoU)

Given a pair of bounding box coordinates $(\boldsymbol{b}_i, \boldsymbol{b}_j)$, the *Intersection over Union (IoU)*, also known as Jaccard index, is an evaluation metric used mainly in object detection tasks, which aims to evaluate how much the two bounding boxes refer to the same content in the image. Specifically, it is defined as:

$$IoU(\boldsymbol{b}_i, \boldsymbol{b}_j) = \frac{|\boldsymbol{b}_i \cap \boldsymbol{b}_j|}{|\boldsymbol{b}_i \cup \boldsymbol{b}_j|}, \tag{1}$$

where $|\boldsymbol{b}_i \cap \boldsymbol{b}_j|$ is the area of the box obtained by the intersection of boxes $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$, while $|\boldsymbol{b}_i \cup \boldsymbol{b}_j|$ is the area of the box obtained by the union of boxes $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$. It is invariant to the bounding boxes sizes, and it returns values that are strictly contained in the interval $[0, 1] \subset \mathbb{R}$, where 1 means that the two bounding boxes refer to the same image area, while a score of 0 means that the two bounding boxes do not overlap at all. The fact that two bounding boxes that do not overlap have *IoU* score equal to 0, is the major issue of this metric: the zero value does not represent how much the two bounding boxes are far from each other. For this reason, in its standard definition, the intersection over union is mainly used as an evaluation metric rather than as a component of a loss function for learning.

1.2 Complete Intersection over Union (CIoU)

In order to solve the issue of *IoU* when considering it as a loss function, several alternative formulations were suggested in the literature, e.g. [12] proposed the *Generalized IoU (GIoU)* loss, [16] proposed the *Distance IoU (DIoU)* loss, while only recently [17] proposed the *Complete IoU (CIoU)* loss, which has shown promising results and faster convergence than *GIoU* and *DIoU*. It is defined as:

$$\mathcal{L}_{CIoU}(\boldsymbol{b}_i, \boldsymbol{b}_j) = S(\boldsymbol{b}_i, \boldsymbol{b}_j) + D(\boldsymbol{b}_i, \boldsymbol{b}_j) + V(\boldsymbol{b}_i, \boldsymbol{b}_j) \tag{2}$$

$$S(\boldsymbol{b}_i, \boldsymbol{b}_j) = 1 - IoU(\boldsymbol{b}_i, \boldsymbol{b}_j); \tag{3}$$

$$D(\boldsymbol{b}_i, \boldsymbol{b}_j) = \frac{\rho(\boldsymbol{p}_i, \boldsymbol{p}_j)^2}{c^2}; \tag{4}$$

$$V(\boldsymbol{b}_i, \boldsymbol{b}_j) = \alpha \frac{4}{\pi^2} \left(\arctan \frac{wt_j}{ht_j} - \arctan \frac{wt_i}{ht_i} \right) \tag{5}$$

where $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ are two bounding boxes, $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are their central points, $IoU(\boldsymbol{b}_i, \boldsymbol{b}_j)$ is the standard *IoU*, ρ is the euclidean distance between the given points, c is the diagonal length of the *convex hull* of the two bounding boxes, α is a trade-off parameter, wt_i and ht_i are the width and the height of the bounding box $\boldsymbol{b}_i$, respectively. Differently from the standard *IoU*, the *Complete IoU* is formulated in such a way to return meaningful values, leveraging the bounding boxes geometric shapes, even when two bounding boxes are not overlapped.

2 MODEL

As outlined in the main paper, our model follows a typical basic architecture for visual-textual grounding tasks. It is based on a two-stage approach in which, initially, a pre-trained object detector is used to extract, from a given image $\boldsymbol{I}$, a set of k bounding box proposals $\mathcal{P}_{\boldsymbol{I}} = \{\boldsymbol{p}_i\}_{i=1}^k$, where $\boldsymbol{p}_i \in \mathbb{R}^4$, jointly with features $H^v = \{\boldsymbol{h}_i^v\}_{i=1}^k$, where $\boldsymbol{h}_i^v \in \mathbb{R}^d$, where d is the number of returned

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507047

features. The features represent the internal object detector activation values before the classification layers and regression layer for bounding boxes. Moreover, our model extracts the spatial features $H^s = \{\boldsymbol{h}_i^s\}_{i=1}^k$, where $\boldsymbol{h}_i^s \in \mathbb{R}^5$ from all the bounding boxes proposals. Specifically, the spatial features for the proposal $\boldsymbol{p}_i$ are defined as:

$$\boldsymbol{h}_i^s = \left[\frac{x1}{wt}, \frac{y1}{ht}, \frac{x2}{wt}, \frac{y2}{ht}, \frac{(x2-x1)\times(y2-y1)}{wt\times ht}\right], \tag{6}$$

where $(x1, y1)$ refers to the top-left bounding box corner, $(x2, y2)$ refers to the bottom-right bounding box corner, wt and ht are the width and height of the image, respectively. We also assume that the object detector returns, for each $\boldsymbol{p}_i$, a probability distribution $Pr_{Cls}(\boldsymbol{p}_i)$ over a set Cls of predefined classes, i.e. the probability for each class $\xi \in Cls$ that the content of the bounding box $\boldsymbol{p}_i$ belongs to ξ.

Regarding the textual features extraction, given a noun phrase $\boldsymbol{q}_j$, initially all its words $W^{\boldsymbol{q}_j} = \{w_i^{\boldsymbol{q}_j}\}_{i=1}^l$ are embedded in a set of vectors $E^{\boldsymbol{q}_j} = \{\boldsymbol{e}_i^{\boldsymbol{q}_j}\}_{i=1}^l$ where $\boldsymbol{e}_i^{\boldsymbol{q}_j} \in \mathbb{R}^w$, where w is the size of the embedding. Then, our model applies a LSTM [4] neural network to generate from the sequence of word embeddings only one new embedding $\boldsymbol{h}_j^\star$ for each phrase $\boldsymbol{q}_j$. This textual features extraction is defined as:

$$\boldsymbol{h}_j^\star = L1\left(LSTM(E^{\boldsymbol{q}_j})\right), \tag{7}$$

where $\boldsymbol{h}_j^\star \in \mathbb{R}^t$ is the LSTM output of the last word in the noun phrase $\boldsymbol{q}_j$, and $L1$ is the L1 normalization function.

Once vector $\boldsymbol{h}_j^\star$ has been generated from the noun phrase $\boldsymbol{q}_j$, the model performs a multi-modal feature fusion operation in order to combine the information contained in $\boldsymbol{h}_j^\star$ with each of the proposal bounding boxes $\boldsymbol{h}_z^v$. For this operation, we have decided to use a simple function that merges the multi-modal features together rather than relying on a more complex operator, such as bilinear-pooling or deep neural network architectures. The multi-modal fusion component we adopted returns the set of new vectorial representations $H^{||} = \{\boldsymbol{h}_{jz}^{||}\}_{j\in[1,\dots,m],z\in[1,\dots,k]}$, where vectors $\boldsymbol{h}_{jz}^{||}$ are defined as:

$$\boldsymbol{h}_{jz}^{||} = LR\left(\boldsymbol{W}^{||}\left(\boldsymbol{h}_j^\star \,||\, \boldsymbol{h}_z^s \,||\, L1(\boldsymbol{h}_z^v)\right) + \boldsymbol{b}^{||}\right), \tag{8}$$

where $||$ indicates the concatenation operator, $\boldsymbol{h}_{jz}^{||} \in \mathbb{R}^c$, LR indicates the leaky-relu activation function, $\boldsymbol{W}^{||} \in \mathbb{R}^{c\times(t+s+v)}$ is a matrix of weights, and $\boldsymbol{b}^{||} \in \mathbb{R}^c$ is a bias vector.

Finally, the model predicts the probability $\boldsymbol{P}_{jz}$ that a given noun phrase $\boldsymbol{q}_j$ is referred to a proposal bounding box $\boldsymbol{p}_z$ as:

$$\boldsymbol{P}_{jz} = \frac{\exp(\boldsymbol{W}^g \times \boldsymbol{h}_{jz}^{||} + b^g)}{\sum_{i=1}^k \exp\left(\boldsymbol{W}^g \times \boldsymbol{h}_{ji}^{||} + b^g\right)}, \tag{9}$$

where $\boldsymbol{W}^g \in \mathbb{R}^{1\times c}$ and $b^g \in \mathbb{R}$ are weights.

Indeed, the representations $\boldsymbol{h}_{jz}^{||}$ of the proposals bounding box features conditioned with the textual features can also be used to refine the proposal bounding box coordinates, that are generated by the object detector independently by the textual features. Specifically, our model does not predicts new bounding box coordinates, but offsets for the coordinates defined as

$$\boldsymbol{o}_{jz} = \boldsymbol{W}^{\mathcal{B}} \times \boldsymbol{h}_{jz}^{||} + \boldsymbol{b}^{\mathcal{B}}, \tag{10}$$

where $\boldsymbol{W}^{\mathcal{B}} \in \mathbb{R}^{4\times c}$ and $\boldsymbol{b}^{\mathcal{B}} \in \mathbb{R}^4$ are a matrix of weights and a bias vector, respectively. The final predicted bounding boxes coordinates are then obtained as the sum of the proposal bounding boxes coordinates with the predicted offsets.

3 DATASETS DETAILS

Flickr30k Entities and ReferIt constitute the two most common datasets used in the literature, although other datasets have been used (e.g., [2, 6, 9, 14]).

The Flickr30k Entities dataset [11, 13] contains 32K images, 275K bounding boxes, 159K sentences, and 360K noun phrases. Each image is associated with five sentences with a variable number of noun phrases, and each noun phrase is associated with a set of bounding boxes ground truth coordinates. Following all works in the literature, if a noun phrase corresponds to multiple ground truth bounding boxes, we merged the boxes and used their union region as its ground-truth. On the contrary, a noun phrase with no associated bounding box was removed from the dataset. We used the standard split for training, validation, and test set as defined in [11], consisting of 30K, 1K, and 1K images, respectively.

The ReferIt [7] dataset contains 20K images, 99K bounding boxes, and 130K noun phrases. This dataset differs from Flickr30k Entities since it does not contain sentences, which means that the noun phrases are mutually independent. For this reason, the state-of-the-art models that depend on a sentence linking all the noun phrases, since they use a feature fusion operator that assumes the presence of the input sentence containing all the noun phrases, cannot be applied to it. We used the same split as in [11] that consists of 9K images of training, 1K images of validation, and 10K images of test.

4 IMPLEMENTATION DETAILS

Our model extracts the words vocabulary using the SpaCy [5] framework for both datasets. Each word embedding is initialized using the GloVe [10] pre-trained weights, which our model does not train, while the remaining weights are initialized according to Xavier [3]. To compare objectively the experimental results with state-of-the-art models, we have used the same object detector adopted in [15], which consists of a Faster R-CNN pre-trained object detector [1] on the Visual Genome [8] dataset that uses ResNet-101 as backbone model[1]. The features associated to each bounding box are extracted from the ResNet-101's layer *pool5_flat*. Following [15], our object detector returns for each bounding box proposal a probability distribution over 1600 classes. We could have applied other object detectors or bounding box proposals which would have lead to further improvements, however this research direction is not related to the aim of this paper. Our model adopts the normalized bounding boxes coordinates with the following representation:

$$\boldsymbol{b} = \left[\frac{x1+x2}{2}, \frac{y1+y2}{2}, bwt, bht\right], \tag{11}$$

where bwt and bht are the width and height of the bounding box, respectively.

Regarding the parameter *alpha* in Eq. 5, we just used the value specified in [17] which is identified by a specific formula.

[1]The ResNet-101 weights were pre-trained on COCO for initialization.

Regarding the application of our losses to the state-of-the-art DDPN [15] model, we have used the authors' official code of the object detector to extract the bounding boxes proposals with their probabilities, and then we have re-implemented their DDPN model in Pytorch. Specifically, we implemented their model following the architecture and the hyper-parameters reported in their article, because the official implementation, as reported in the official repository[2], presents a slightly different architecture that leads to different results. On the re-implemented model, maintaining the same architecture and hyper-parameters, we have implemented our losses that lead to better results as reported in our main article.

5 QUALITATIVE RESULTS

In Figures 1-12, we have reported some qualitative results obtained by our model in both Flickr30k Entities and ReferIt datasets. Figures 1, 2, 3, 4, 5, 6 are examples of the Flickr30k Entities test set images, while Figures 7, 8, 9, 10, 11, 12 are examples of the ReferIt test set. We can see that in both the datasets, very often the predicted bounding boxes that have an intersection over union value under 0.5, are still close to the ground truths bounding boxes. Only in Figure 5, the model predicts a bounding box for the query "one hand" that is located very far from its ground truth.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[2] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020. Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10083–10092. https://doi.org/10.1109/CVPR42600.2020.01010

[3] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010 (JMLR Proceedings, Vol. 9)*, Yee Whye Teh and D. Mike Titterington (Eds.). JMLR.org, 249–256. http://proceedings.mlr.press/v9/glorot10a.html

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[5] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*.

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[7] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferIt Game: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[9] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 11–20. https://doi.org/10.1109/CVPR.2016.9

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. https://doi.org/10.3115/v1/d14-1162

[11] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2641–2649. https://doi.org/10.1109/ICCV.2015.303

[12] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 658–666. https://doi.org/10.1109/CVPR.2019.00075

[13] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics* 2 (2014), 67–78. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229

[14] 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.

[15] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 1114–1120. https://doi.org/10.24963/ijcai.2018/155

[16] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 12993–13000. https://aaai.org/ojs/index.php/AAAI/article/view/6999

[17] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. 2020. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *CoRR* abs/2005.03572 (2020). arXiv:2005.03572 https://arxiv.org/abs/2005.03572

[2] https://github.com/XiangChenchao/DDPN

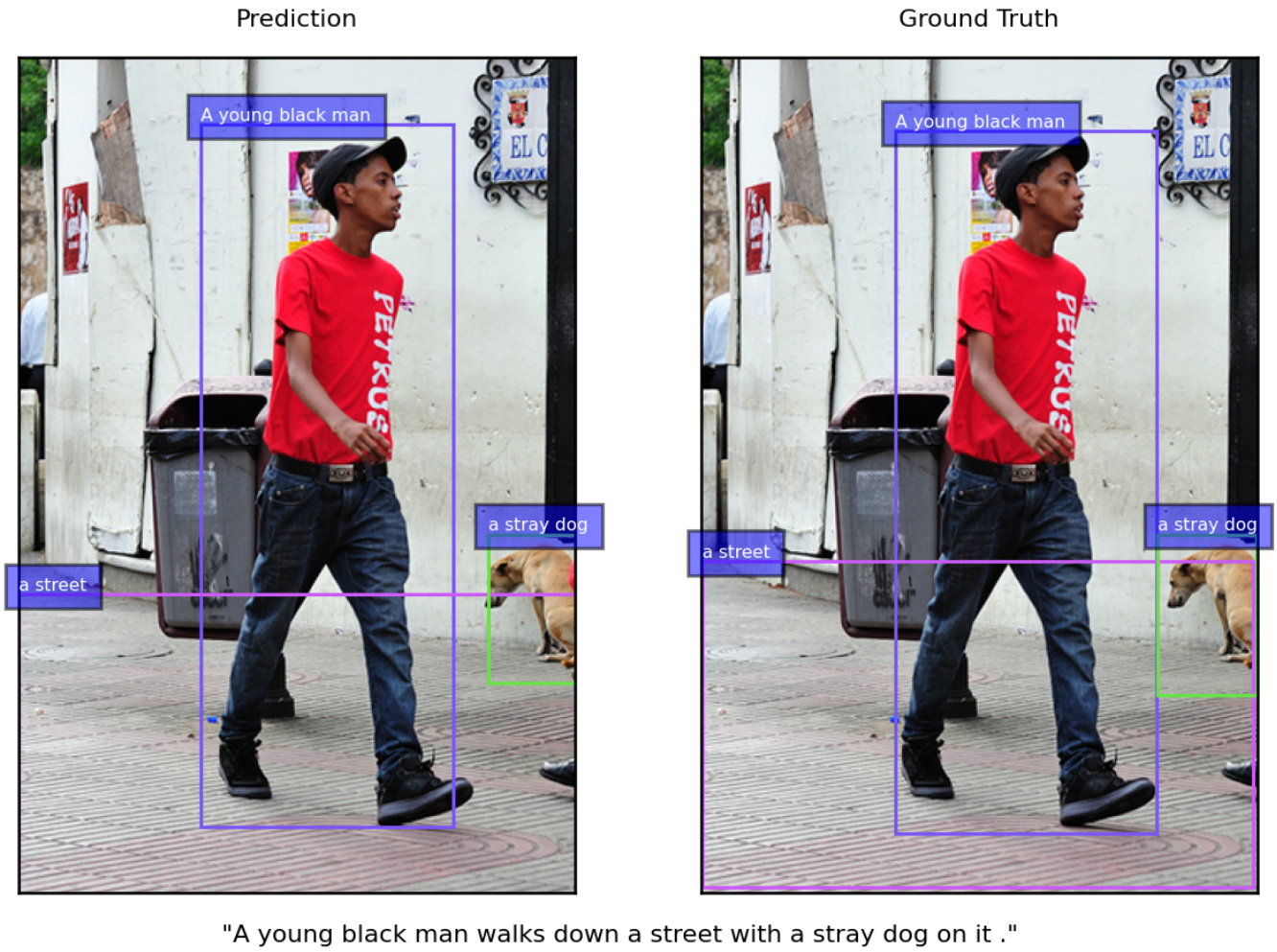

Figure 1: Qualitative result obtained by our model on the Flickr30k Entities test set. All bounding boxes are predicted correctly.

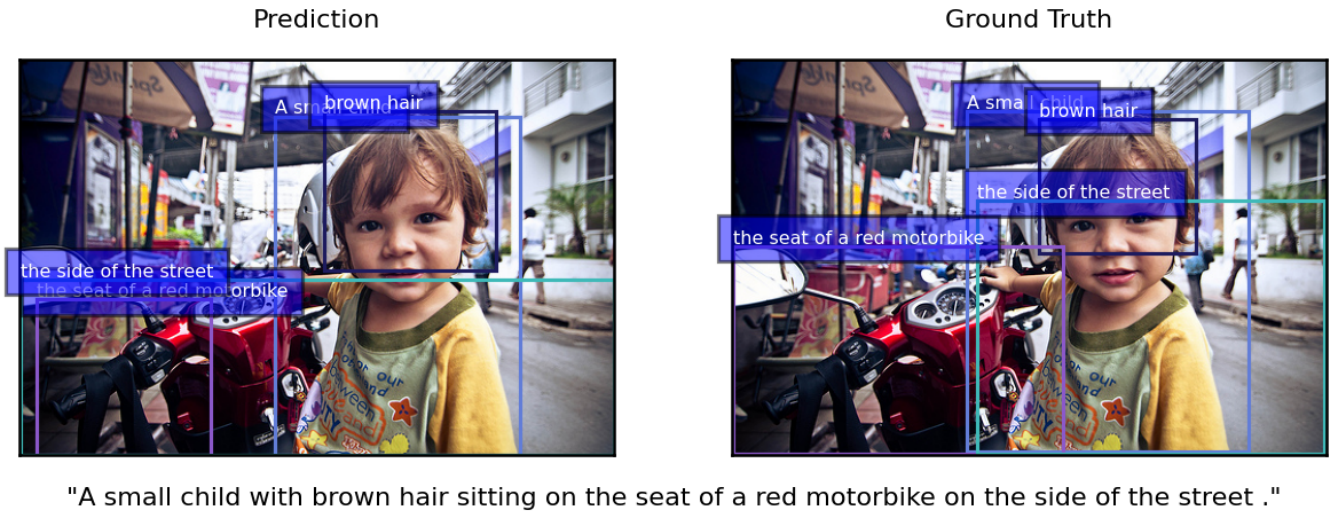

Figure 2: Qualitative result obtained by our model on the Flickr30k Entities test set. The bounding boxes aligned with the queries "the seat of a red motorbike" and "the side of the street" present an intersection over union value with their ground truths that is lower than 0.5.

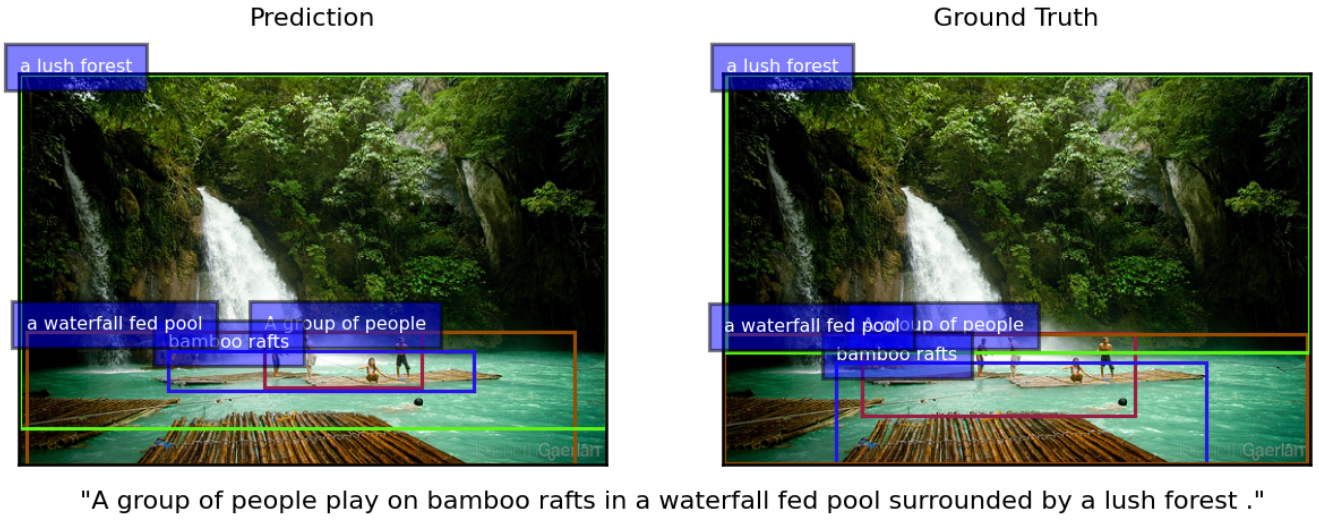

Figure 3: Qualitative result obtained by our model on the Flickr30k Entities test set. The bounding boxes aligned with the queries "A group of people" and "bamboo rafts" present an intersection over union value with their ground truths that are lower than 0.5.

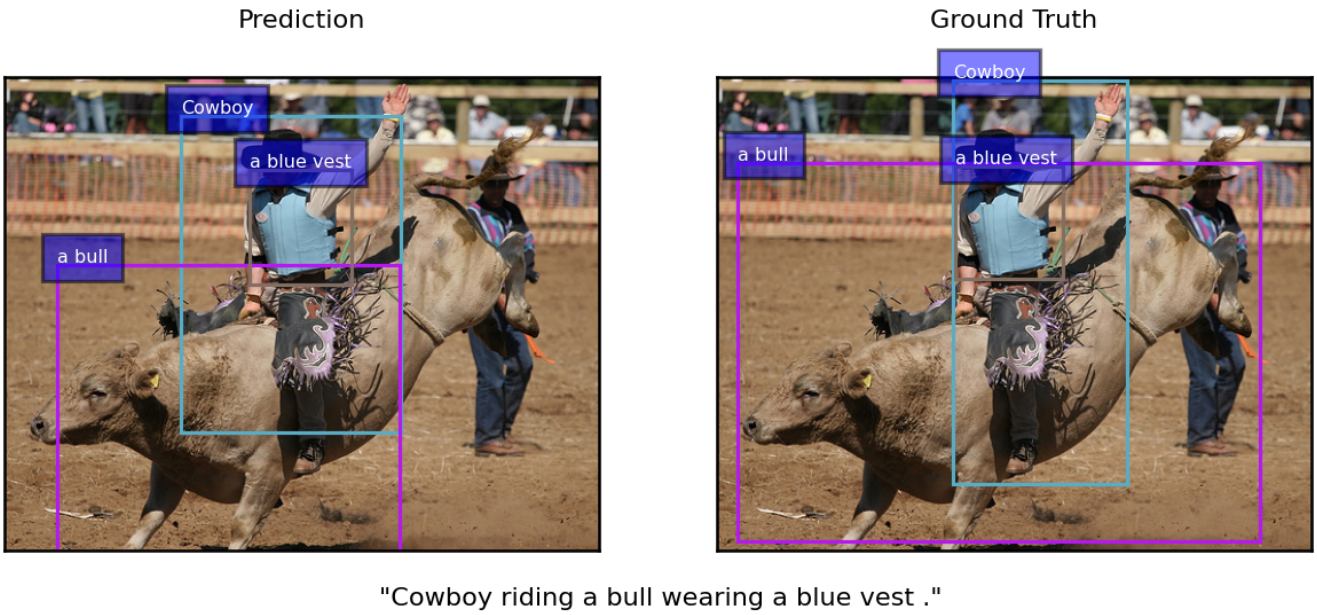

Figure 4: Qualitative result obtained by our model on the Flickr30k Entities test set. The bounding box aligned with the query "a bull" presents an intersection over union value with its ground truth that is lower than 0.5.

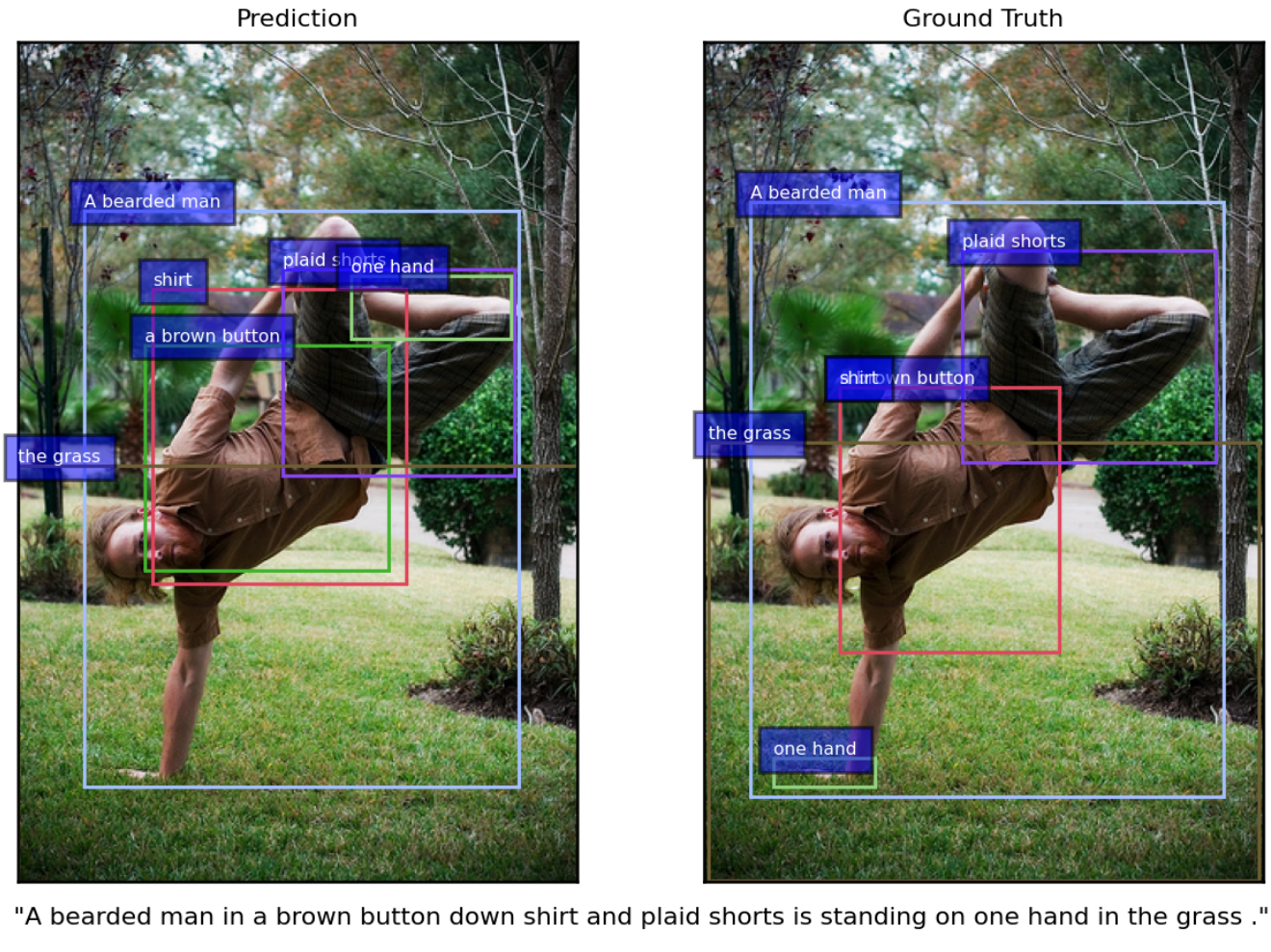

Figure 5: Qualitative result obtained by our model on the Flickr30k Entities test set. The bounding boxes aligned with the queries "shirt" and "one hand" present an intersection over union value with their ground truths that are lower than 0.5.

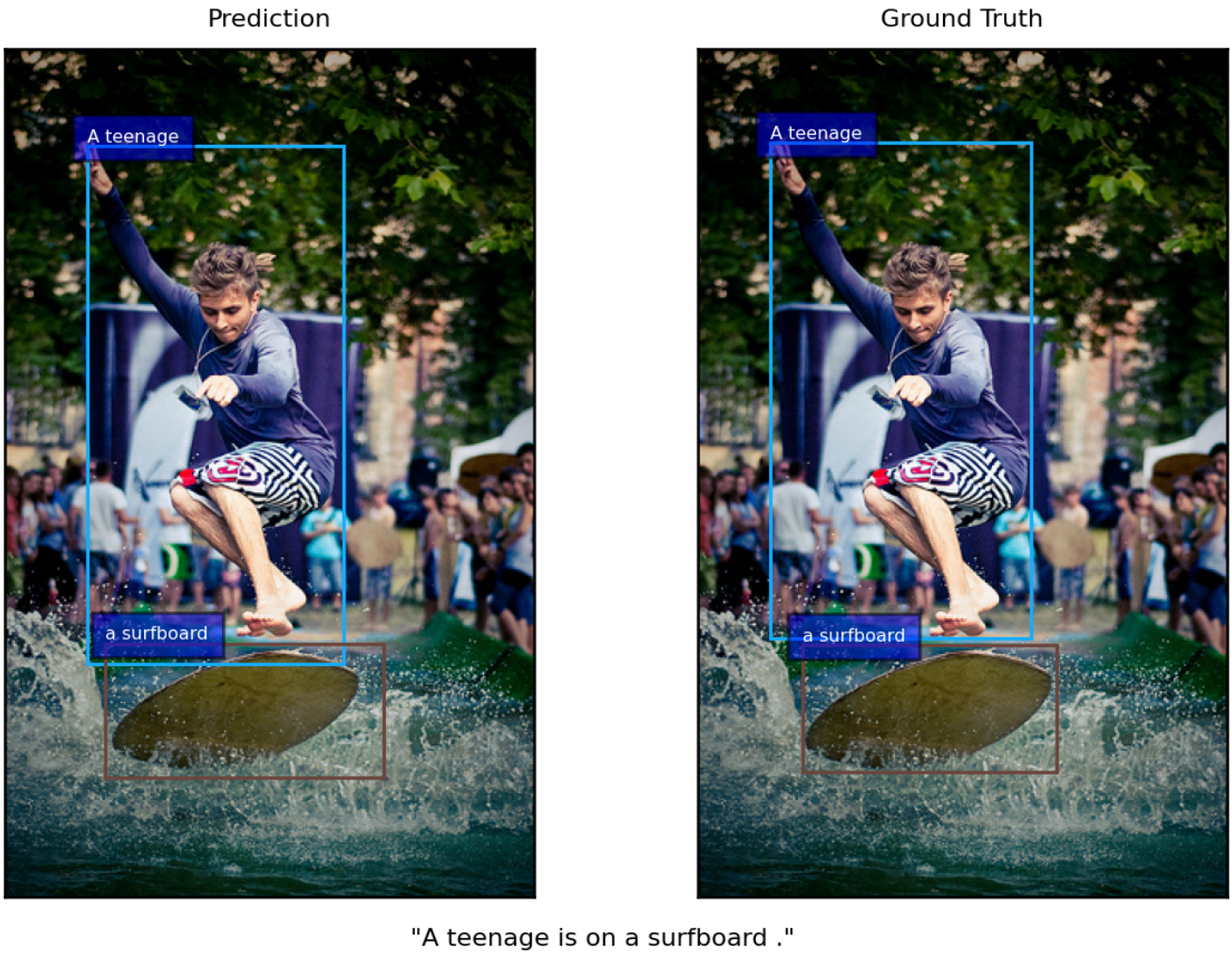

Figure 6: Qualitative result obtained by our model on the Flickr30k Entities test set. All bounding boxes are predicted correctly.

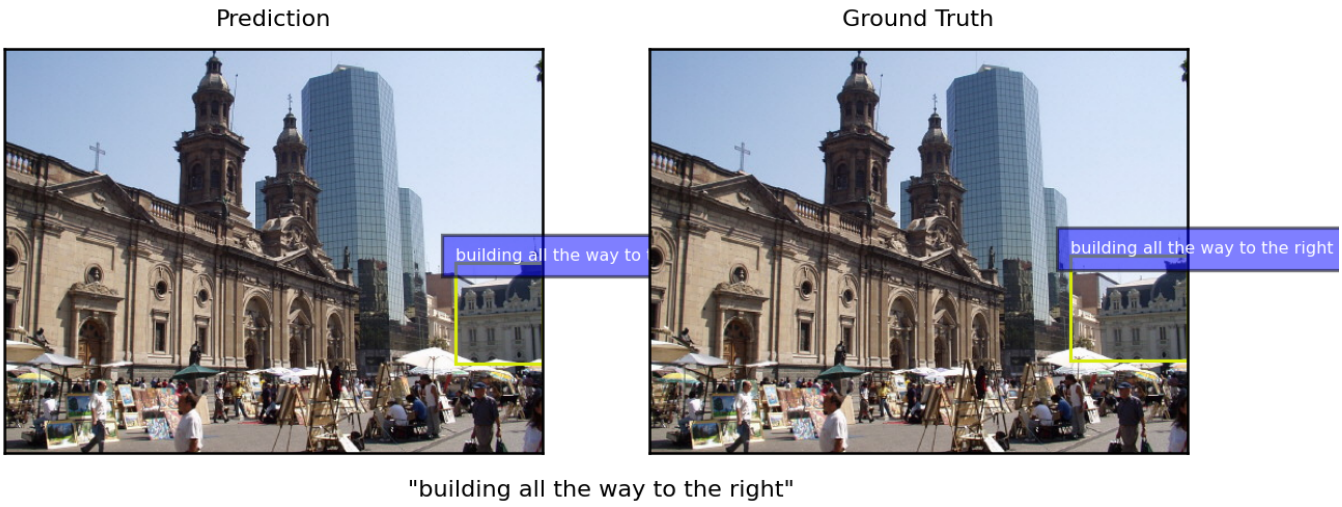

Figure 7: Qualitative result obtained by our model on the ReferIt test set. The bounding box is predicted correctly.

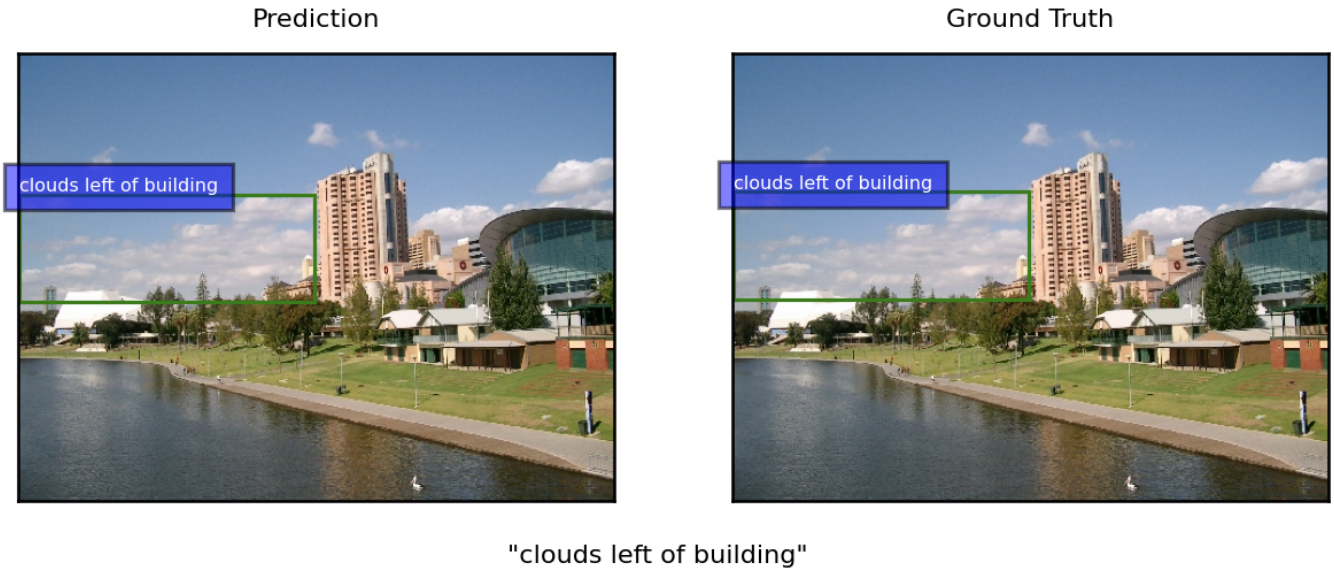

Figure 8: Qualitative result obtained by our model on the ReferIt test set. The bounding box is predicted correctly.

Figure 9: Qualitative result obtained by our model on the ReferIt test set. The bounding box is predicted correctly.

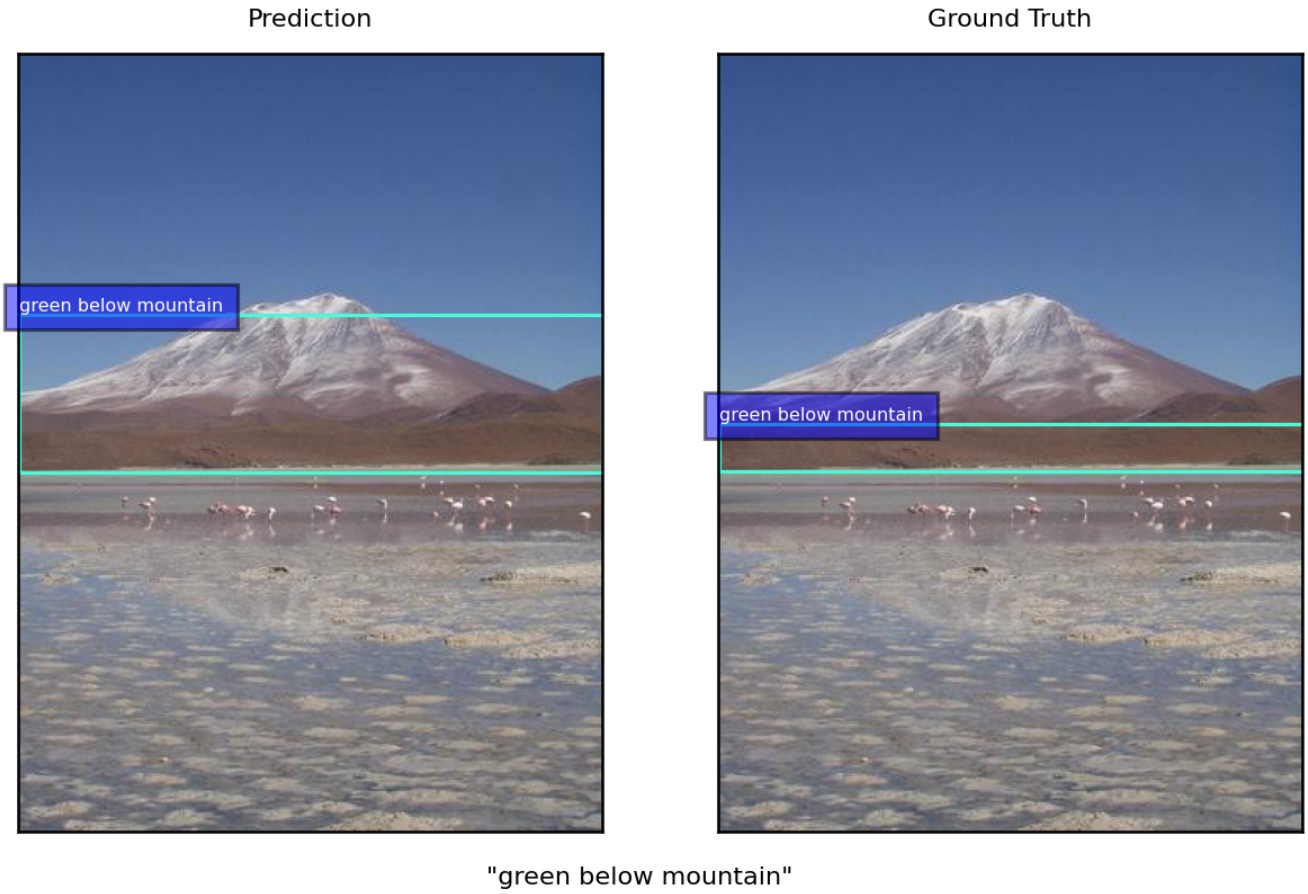

Figure 10: Qualitative result obtained by our model on the ReferIt test set. The predicted bounding box presents an intersection over union value with the ground truth of 0.30.

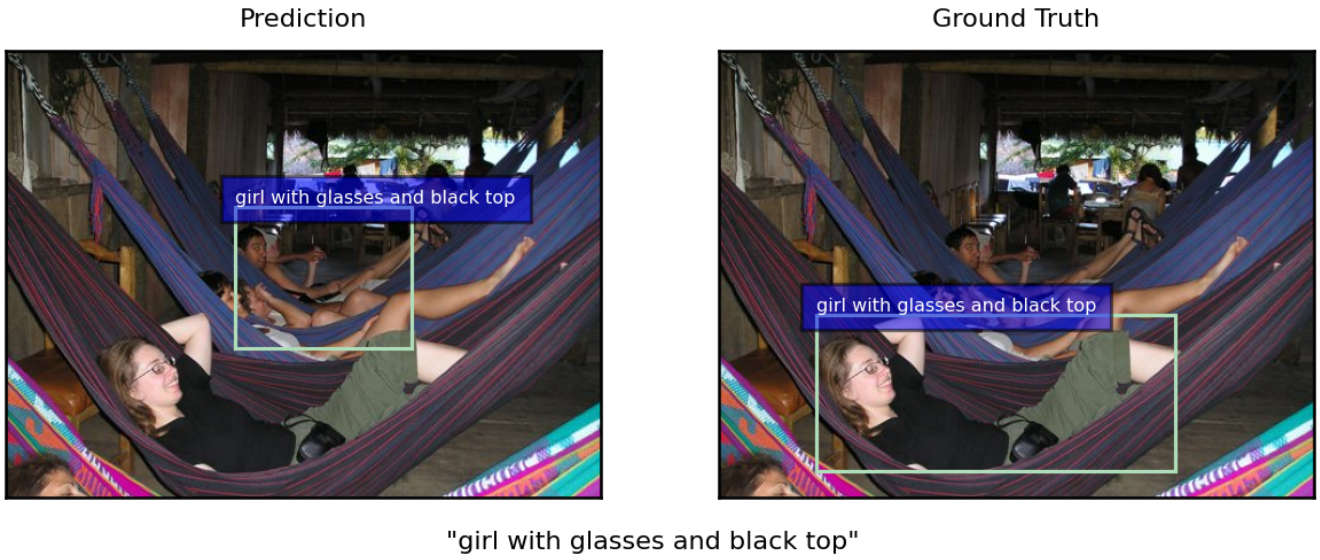

Figure 11: Qualitative result obtained by our model on the ReferIt test set. The predicted bounding box presents an intersection over union value with the ground truth of 0.08.

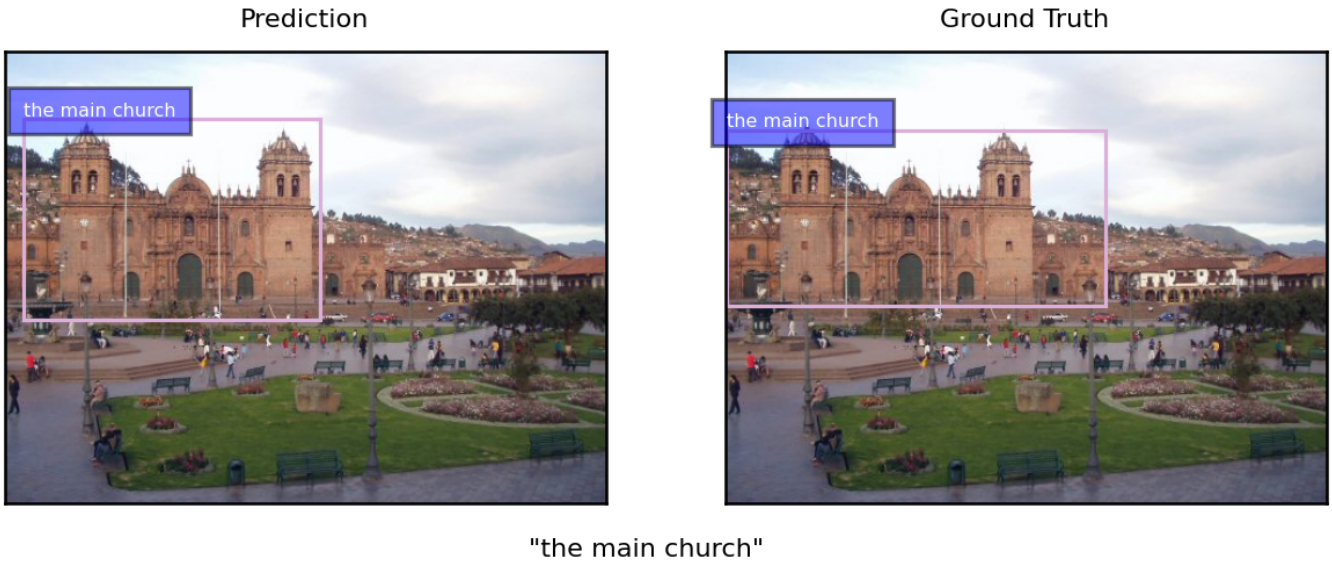

Figure 12: Qualitative result obtained by our model on the ReferIt test set. The bounding box is predicted correctly.