6. Supplementary

In this section we provide more details for experiments on RefCOCO+ [40] and ReferIt [14], and provide additional qualitative examples.

Method	RefCOCO+‡		RefCOCO+§	
	test A	test B	test A	test B
InfoGround [11]	39.80	41.11	40.10	40.62
VMRM [10]	58.87	50.32	60.29	50.39
ALBEF [17]	69.37	53.77	69.40	54.04
AMC	80.34	64.55	80.33	65.02

Table 7. We show pointing accuracy results on the RefCOCO+ validation and testing sets with original and clean images. ‡ indicates original image splits and § indicates splits with clean images.

RefCOCO+ Clean As discussed in Anderson et al [2], there are around 51K images from Visual Genome (VG) [16] that are also present in the COCO dataset [21]. Moreover, images in the RefCOCO+ validation/testing sets come from the COCO dataset as well. While there is no overlap in the training, validation and testing sets for RefCOCO+, methods that use VG to pretrain object detectors might use some overlapping data which would make object detectors on some part of the validation and testing sets artificially accurate. In order to fully investigate whether this issue affects the generalization of previous methods, we further explore a more restricted version of the validation and test sets for RefCOCO+ so that no overlap exists with VG and re-run previous methods along with our method on this subset. After cross-referencing images in the VG training set from Anderson et al [2] and images from the RefCOCO+ validation/testing sets, we find 574 and 569 overlapping images in the RefCOCO+ validation/testing sets. In order to correct this, we also evaluate and compare our method with previous methods on a *clean* version of the RefCOCO+ validation and testing sets with 926 and 931 images respectively.

Table 7 shows that in fact this overlap did not have much of an effect on previous methods – and our method also performs at a high accuracy. Our method still outperforms VMRM [10] and InfoGround [11] by a large margin. We also report results for ALBEF [17] and compare it with InfoGround and VMRM which uses bounding boxes for object detectors during training. Even though ALBEF does not use any box information, it still achieves good performance on the RefCOCO+ dataset. Our method, which uses box information, can further improve the pointing accuracy results under both settings.

Sample Spatial Prompts In our main paper, we discuss how to construct textual descriptions using bounding boxes and attributes. In Figure 4, we show several examples of such constructed data. In total, we generate $924,807$ text descriptions using attributes and $168,442$ descriptions with spatial references.

Qualitative Results Figure 5 shows additional qualitative results on the test set of RefCOCO+ and Figure 6 shows similar qualitative results for ReferIt. We show heatmaps generated by our method given images and text phrases. Our model can successfully localize the target object even though there exist other similar objects in the same image.

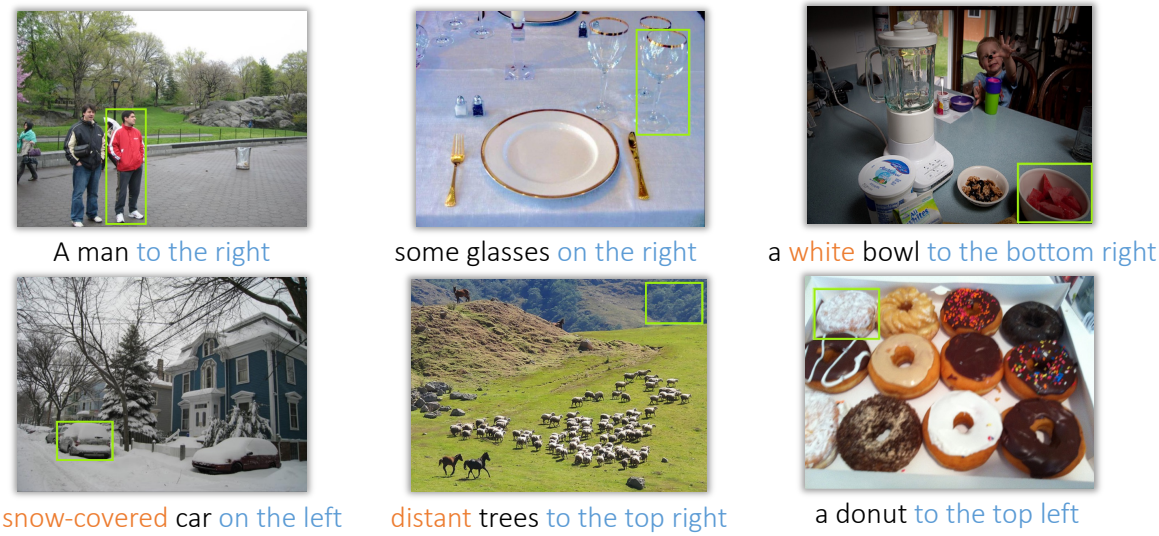
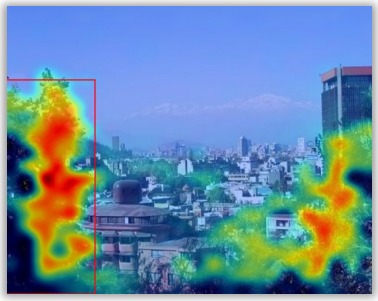
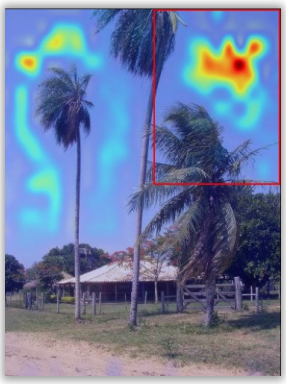


Figure 4. We show some constructed textual descriptions with colored attributes and spatial references.

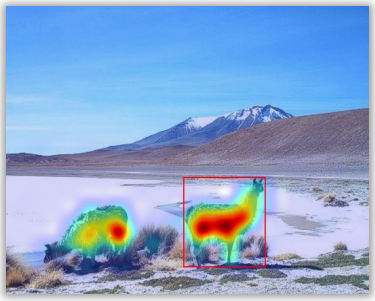
Figure 5. We show more qualitative examples for the RefCOCO+ testing set. Ground truth boxes are marked as red boxes. Below each image we provide with one input phrase.



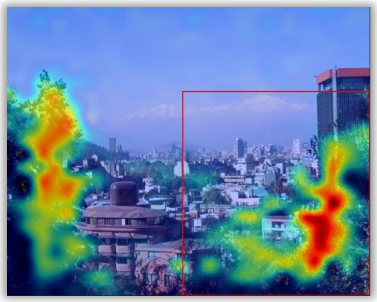
tree on left



sky upper right

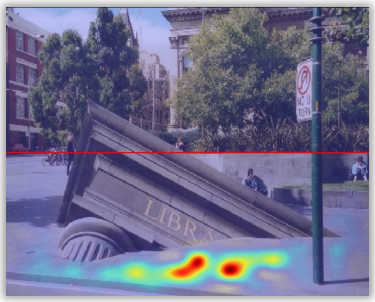


rightmost animal

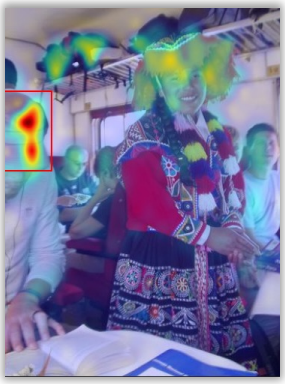


trees right

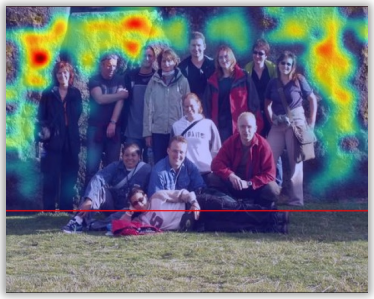
dark brown rock wall
right of building



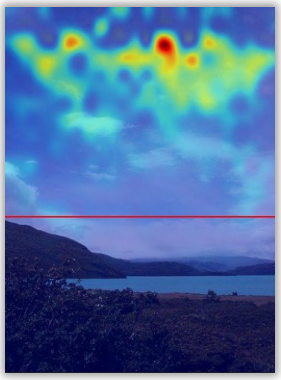
bottom concrete



hat on left



all the top rock



top sky

Figure 6. We show more qualitative examples for the ReferIt testing set. Ground truth boxes are marked as red boxes. Below each image we provide with one input phrase.