Supplementary Material of MORE

Yang Jiao[1,2⋆], Shaoxiang Chen[3⋆], Zequn Jie[3], Jingjing Chen[1,2⋆⋆], Lin Ma[3⋆⋆], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] Meituan Inc.

A Additional Method Details

A.1 Training and Inference

The final loss function is a weighted combination of a detection loss $\mathcal{L}_{det}$ as introduced in [4] and a conventional cross-entropy captioning loss $\mathcal{L}_{cap}$ as done in [5, 3, 2]:

$$\mathcal{L} = \rho_{det}\mathcal{L}_{det} + \rho_{cap}\mathcal{L}_{cap}, \tag{1}$$

where $\rho_{det} = 10, \rho_{cap} = 0.1$ are the coefficients empirically set to balance the loss terms. During training, the captioning loss is only computed between the ground-truth and the caption whose predicted bounding box has the largest IoU with the ground-truth box. During inference, non-maximum suppression is leveraged to filter redundant object proposals.

A.2 Relational Words Dictionary

By analyzing the corpus of the validation sets of ScanRefer dataset [1], we manually extract relational words and classify them into two groups, *"simple"* and *"complex"*, according to whether more than two objects should be jointly considered to infer the relation. The specific vocabularies are listed as below:

- ***"simple"*:** *on, in, left, right, top, bottom, above, under, below, front, behind, besides, next to, center, corner, against, opposite, head of, end of*
- ***"complex"*:** *between, rightmost, leftmost, surrounded by, closest, furthest, farthest, nearest, first, second, third, fourth, last*

The union of the *"simple"* and *"complex"* vocabularies are the whole relational word dictionary.

⋆ Equal contribution.
⋆⋆ Corresponding authors.

Table 1. Ablation studies of our method with different point features as input. "xyz","rgb" and "mul" represent the coordinates, point color and multiview features in respectively.

	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5IoU
MORE (xyz)	55.23	32.63	25.48	53.17	35.85	19.72	20.61	41.97	28.69
MORE (xyz+rgb)	55.82	34.21	25.72	54.68	36.99	20.64	20.81	43.08	29.18
MORE (xyz+rgb+normal)	58.89	35.41	26.36	55.41	38.98	**23.01**	21.65	44.33	31.93
MORE (xyz+mul)	60.60	36.04	26.44	55.63	38.48	21.60	21.03	42.87	33.53
MORE (xyz+mul+normal)	**62.91**	**36.25**	**26.75**	**56.33**	**40.94**	22.73	**21.66**	**44.42**	**33.75**

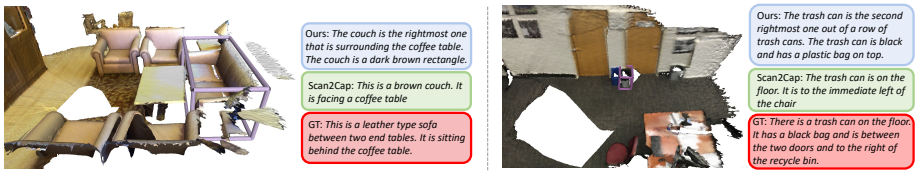

Fig. 1. The illustration of comparison between the captions generated by our method and the baseline method (Scan2Cap). To focus on the target object and its nearby context, we crop and show a part of the scene.

B More Ablation Studies

B.1 Influence of different input features

Generally, available input point features include coordinates (xyz), colors (rgb), normals and multiview features in the ScanRefer dataset [1]. In the main paper, we only report the results of using the input features of "xyz+normal+rgb" and "xyz+normal+multiview" for comparison with the previous work [2]. In this section, to inspect the different input point features' influence on the dense captioning performance, we further conduct ablation studies as shown in the Table.1.

From the results, we have the following observations. First, the point normal, as a critical attribute of the 3D point, can benefit the model's performances. This is because that normals increase the geometric information which can be helpful in recognizing the spatial positions of objects. Moreover, better visual representation can greatly promote caption generation. Appending either the color feature (rgb) or the multiview feature (mul) to the point coordinates can bring significant performance gains, which can be proved by comparing the results of the second row (*"xyz+rgb"*) and fourth row (*"xyz+mul"*) with the results of the first row (*"xyz"*) of the Table.1.

C More Qualitative Results

We further compare some captions generated by our method and the baseline (i.e., Scan2Cap) in Fig.1, where we crop and show a part of the scene with the target object enclosed by a purple bounding box. As the two samples show,

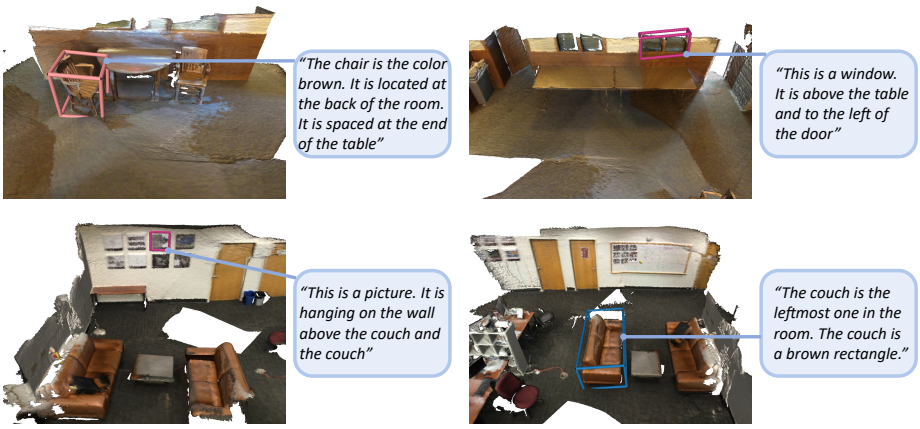

Fig. 2. More qualitative results of our proposed MORE. For clarity, we indicate the target objects with bounding boxes and show the corresponding captions generated by our proposed MORE.

our method can generally attend to more surrounding objects and explore their relations with the target object to assist the caption generation. And compared to the baseline, our generated captions contain more complicated spatial relations and are more accurate when using the ground-truth as references.

We also give more qualitative results of captions generated by our MORE for objects within 3D scenes as shown in the Fig.2. From the results, we can observe that our method can capture abundant contextual cues and describe target objects with rich relational words, which demonstrates the superiority of our MORE in modeling inter-object relations.

References

1. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European Conference on Computer Vision. pp. 202–221. Springer (2020)
2. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3193–3203 (2021)
3. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
4. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
5. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)