Spatiality-guided Transformer for 3D Dense Captioning on Point Clouds (Supplementary Materials)

Heng Wang, **Chaoyi Zhang**, **Jianhui Yu** and **Weidong Cai**
School of Computer Science, University of Sydney, Australia
{heng.wang, chaoyi.zhang, jianhui.yu, tom.cai}@sydney.edu.au

In this supplementary for SpaCap3D [Wang *et al.*, 2022], we provide more details of the learnable positional encoding in Section 1. We visualize the attention mechanism used in our SpaCap3D framework in Section 2 and provide more qualitative results of our method in Section 3.

1 Learnable Positional Encoding

Figure 1 illustrates the three different learnable positional encoding approaches for tokens to the encoder. To generate the C-dim positional encoding vector for each token input to the encoder, the random one, as used in 2D detection Transformer [Carion *et al.*, 2020], randomly learns weight parameters during training and such learnt weights are used as positional encoding for M proposals/tokens during inference, which are fixed for different scene inputs. To make positional encoding object-variant, [Liu *et al.*, 2021] further proposed to generate positional encoding based on predicted box parameters, box center and box size optionally, as shown as the green bounding box in Figure 1. We implement such approach by directly using the predicted bounding box center and size from detection backbone for each proposal. The vote center-based way is similar to the box center-based one but vote centers are the M centers after grouping in proposal module from the detection backbone. Red dots in Figure 1 refer to the votes after voting module from which vote centers are generated using farthest point sampling technique.

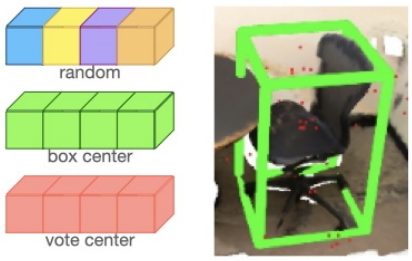


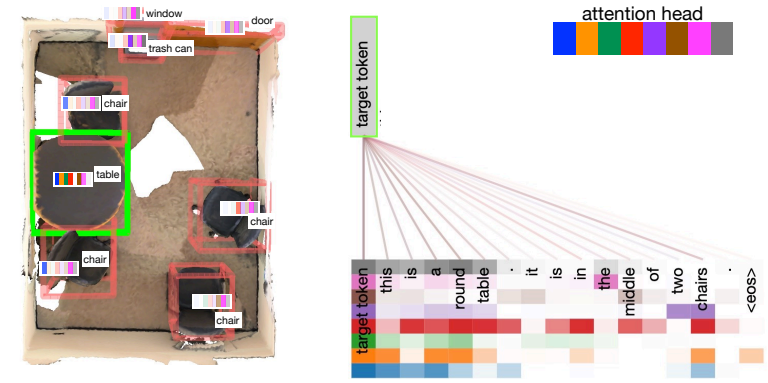
Figure 1: Illustration of different ways of positional encoding. *random* refers to the randomly learnable weights, while *box center* and *vote center* use features originated from detected box center (and its size optionally) and vote cluster center, respectively. *vote center* achieves the best results in Table 3 of our main paper.

2 Attention Visualization

We provide three examples of how the attention works in our proposed method in Figure 2. We follow [Vig, 2019] to visualize the attentions which are extracted from the last block in both encoder and decoder and values from different heads (eight in total) are marked with different colors. We use opacity to represent the magnitude. The more transparent the color, the smaller the value. In each example, the target object to be described is highlighted in green and the surrounding objects are marked in red. The left figure in each example shows how the surrounding objects contribute to the target object representation learning. As the spatial relations between the target object and its neighbors are different, attentions learnt for different surrounding objects are different. We also present how the spatiality-enhanced target vision token contributes to the generation of each predicted word in the right figure of each example. Taking Figure 2(a) for instance, we observe that our target vision token contributes differently to the predicted words. It especially emphasizes the words describing the target object itself (*"round table"*), the words expressing the spatial relation (*"the middle of"*), and the words about the neighboring objects (*"two chairs"*), which demonstrates the successful incorporation of relative 3D spatiality in the representation learning phase through our proposed spatiality-guided encoder.

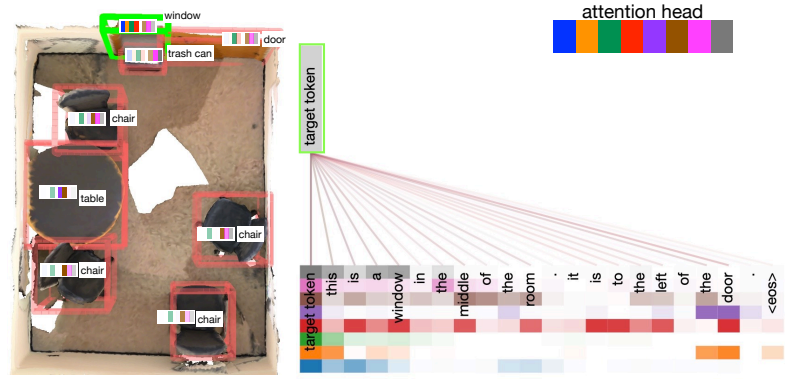
3 More Qualitative Results

We display more detect-and-describe results from models trained with and without our proposed token-to-token (T2T) spatial relation guidance in Figure 3. In Figure 3(a), our method with T2T guidance captures two spatial relations for the target object file cabinet, one is with the chair on the right and the other is with the desk on the top. Without T2T guidance, the predicted relation *"to the left of a desk"* is incorrect. Figure 3(b) shows the case when T2T guidance boosts more precise description generation - not merely *"at a table"* but *"at the far end of the table"*. Figure 3(c) and 3(d) present more cases when T2T guidance improves relation variety - relations between whiteboard with wall and table, and between table with the whole room and the surrounding chairs, respectively. Figure 3(e) highlights the case when the lack of T2T guidance could lead to the wrong prediction of the target object itself.



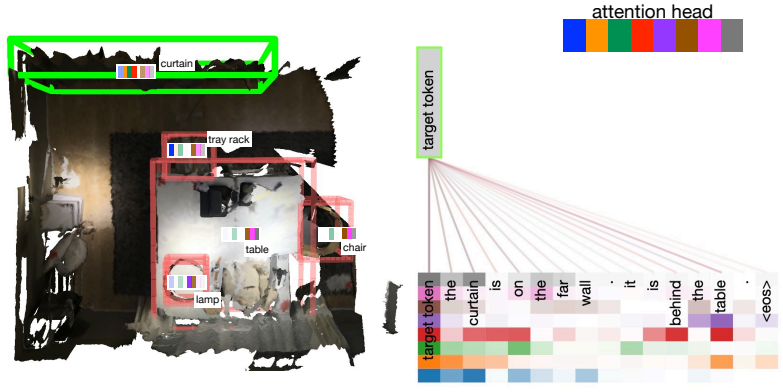
Generated caption for the table object:
this is a round table. it is in the middle of two chairs.

(a)



Generated caption for the window object:
this is a window in the middle of the room. it is to the left of the door.

(b)



Generated caption for the curtain object:
the curtain is on the far wall. it is behind the table.

(c)

Figure 2: Examples of our encoder and decoder attention for a target object marked in green in a 3D scene. In each eight-color vector, different colors represent different attention heads. The more transparent the color is, the smaller the attention value is. The colorful vector shown on each object represents the eight-head attention values between the target object marked in green and its surrounding objects in red. The decoder attention between the target vision token and the generated caption words is shown as the eight-color vector underlying each generated word.

Ours w/o T2T: this is a white cabinet. it is to the left of a desk.

Ours: the file cabinet is under the desk. it is to the left of the chair.

GT: a brown cabinet under the table. it is to the right of the door.

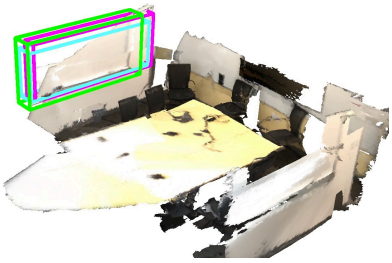
(a)

Ours w/o T2T: this is a wooden chair. it is at a table.

Ours: this is a brown chair. it is at the far end of the table.

GT: this is a brown chair. it is turned toward the end of the table.

(b)

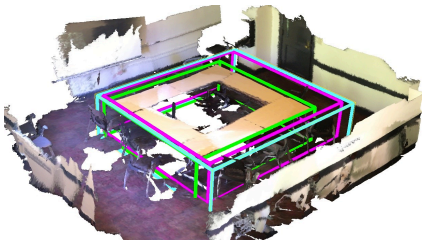


Ours w/o T2T: this is a large whiteboard. it is to the left of the table.

Ours: the whiteboard is on the wall. it is to the left of the table.

GT: this whiteboard is on the left side surface. the white board is attached to wall.

(c)

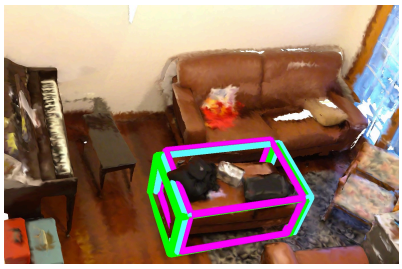


Ours w/o T2T: this is a large table. it is in the center of the room.

Ours: this is a large table in the middle of the room. it is surrounded by chairs.

GT: the table is cream color and is in the center of the room. there are chairs around the table.

(d)



Ours w/o T2T: this is a black tv stand. it is in front of a couch.

Ours: this is a brown ottoman. it is in front of a couch.

GT: this is a brown ottoman in front of a brown sofa.

(e)

Figure 3: More qualitative results from our methods with and without token-to-token (T2T) spatial relation guidance. Caption boxes share the same color with detection bounding boxes for ground truth (green), ours with T2T (blue), and ours without T2T (pink). Imprecise parts of sentences produced by ours without T2T are marked in red, and correctly expressed descriptions predicted by T2T-guided method are highlighted using underscores.

References

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[Liu *et al.*, 2021] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3D object detection via transformers. In *ICCV*, pages 2949–2958, 2021.

[Vig, 2019] Jesse Vig. A multiscale visualization of attention in the transformer model. In *ACL: System Demonstrations*, pages 37–42, 2019.

[Wang *et al.*, 2022] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3D dense captioning on point clouds. In *IJCAI*, 2022.