3D Change Localization and Captioning from Dynamic Scans of Indoor Scenes (Supplementary Material)

This supplementary material provides additional details on the proposed DyS2Change dataset, including the caption templates and the relationship definition. We also provide additional experimental results on DyS2Change in the last part of this material.

A. Additional Details on DyS2Change Dataset

Caption Templates. We provide all 30 caption templates used in the generation of DyS2Change and SUN-RGBD2Change datasets in Table 5. We record scene change information, including changed object types and the spatial relationships of all objects during the scene change generation process. The change captions are then automatically generated from the recorded change information, object relationships, and caption templates.

Relationship Definitions. The spatial location of objects is useful to specify an object in a room, especially while there are multiple objects from the same class. Here, as shown in Figure 7, we introduce two types of relationships to refer to the change object in the DyS2Change dataset, including object-room relationships indicating the spatial location of an object in the room and object-object relationships for referring an object from another. During the caption generation, the <rel> is determined based on randomly selecting an existing relationship of the changed object with the room or other objects. For example, when the relationship between a changed object "microwave" and an anchor object "countertop" is supported, we can generate a sentence as "The microwave that is on the countertop...".

B. Additional Experimental Results on DyS2Change Dataset

We show additional experimental results on the proposed DyS2Change dataset in Figure 8. Compared to Baseline3D, both two proposed DenseChangeCaps exhibited higher performance on both change detection and captioning in those three examples in Figure 8. We also found that there is still room for improvement in the accuracy of the detection of small objects and the description of detailed relationships in DenseChangeCaps.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020.

Table 5. Caption templates used in DyS2Change dataset. <target> refers to the object that has been changed. <anchor>, <anchor1>, and <anchor2> refer to a nearby object that have a specified spatial relationship to the changed object <target>. <rel>, <rel1>, and <rel2> represent a specified spatial relationship between <target> and <anchor>. The word [that] [is] is randomly included or not included in the sentence (we also randomly use the synonym "which" to replace "that").

Change type	Caption templates
Add	A <target> [that] [is] <rel> <anchor> has been added. A <target> [that] [is] <rel> <anchor> shows up. There is a new <target> [that] [is] <rel> <anchor>. A new <target> [that] [is] <rel> <anchor> is visible. Someone added a <target> [that] [is] <rel> <anchor>.
Delete	The <target> [that] [was] <rel> <anchor> has disappeared. The <target> [that] [was] <rel> <anchor> is no longer there. The <target> [that] [was] <rel> <anchor> is missing. There is no longer a <target> [that] [was] <rel> <anchor>. Someone removed the <target> [that] [was] <rel> <anchor>.
Open	The <target> [that] [is] <rel> <anchor> has been opened. Someone opened the <target> [that] [is] <rel> <anchor>. The <target> [that] [is] <rel> <anchor> has been turned on. Someone turned on the <target> [that] [is] <rel> <anchor>. Someone turned the <target> [that] [is] <rel> <anchor> on.
Close	The <target> [that] [is] <rel> <anchor> has been closed. Someone closed the <target> [that] [is] <rel> <anchor>. The <target> [that] [is] <rel> <anchor> has been turned off. Someone turned off the <target> [that] [is] <rel> <anchor>. Someone turned the <target> [that] [is] <rel> <anchor> off.
Move	The <target> [that] [was] <rel> <anchor> changed its location. The <target> [that] [was] <rel> <anchor> is in a different location. The <target> [that] [was] <rel> <anchor> was moved from its original location. The <target> [that] [was] <rel> <anchor> has been moved. Someone changed the location of the <target> [that] [was] <rel> <anchor>. The <target> [that] [was] <rel1> <anchor1> changed its location to <rel2> <anchor2>. The <target> [that] [was] <rel1> <anchor1> is <rel2> <anchor2> now. The <target> [that] [was] <rel1> <anchor1> has been moved to <rel2> <anchor2>. Someone changed the location of <target> [that] [was] <rel1> <anchor1> to <rel2> <anchor2>. The <target> [that] [was] <rel1> <anchor1> was moved from its original location to <rel2> <anchor2>.

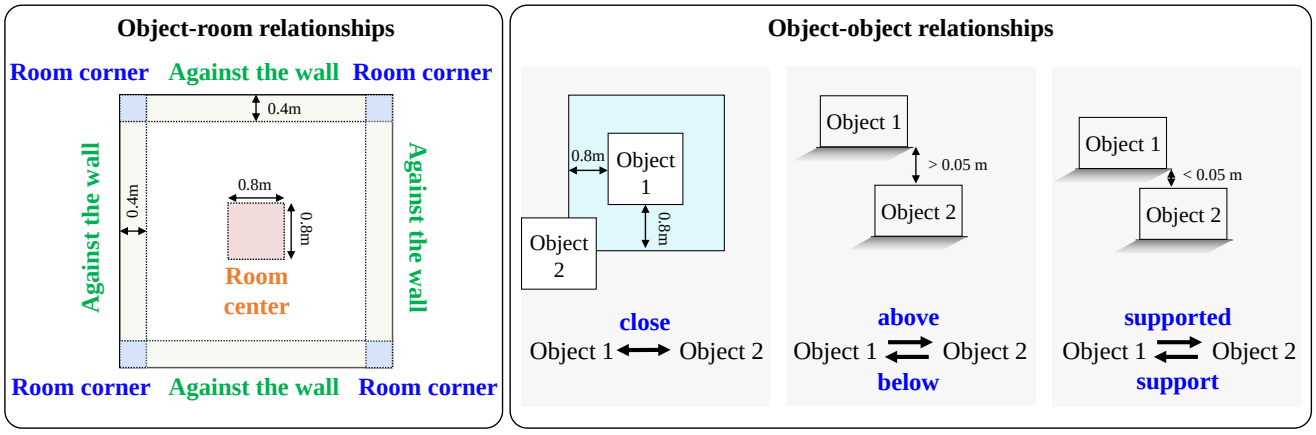

Figure 7. Illustration of relationships used in the DyS2Change dataset. We consider three object-room relationships where objects are located in the room corner, against the wall, or room center. Similar to [1], we consider three object-object relationships. The "close" relationship is determined based on the horizontal distance of objects. If two objects have close relationships or have intersections horizontally, the above and below relationships (with close relationships), supported and support relationships (with horizontal intersections) are determined based on vertical relationships.

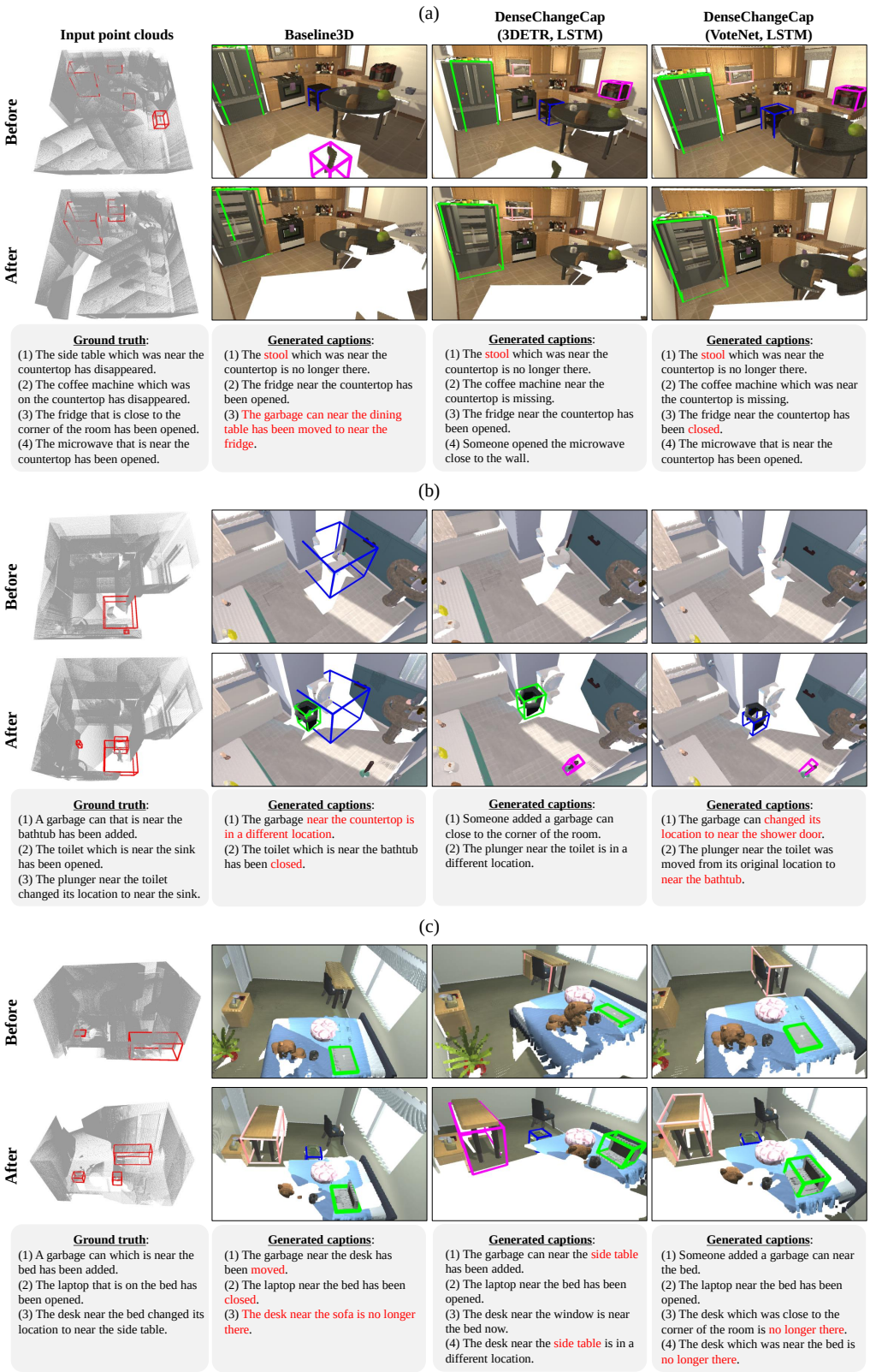

Figure 8. Additional experimental results on DyS2Change dataset. The first column shows the input point clouds, the second, third, and fourth columns show the results of the Baseline3D and two DenseChangeCaps, respectively. The ground truth bounding boxes are highlighted in red and the predicted bounding boxes are highlighted in other colors. Incorrect change captions are highlighted in red.