3D-LLaVA: Towards Generalist 3D LMMs with Omni Superpoint Transformer

Supplementary Material

6. Supplementary Implementation Details

Sparse 3D U-Net. The 3D scene encoder of our 3D-LLaVA, *i.e.*, Sparse 3D U-Net, follows the practice of [36, 37, 54]. Specifically, it is a 5-layer U-Net. The channel size of the first encoder stage is set to 32. Every following stage of the encoder layer will enlarge the channel size of the previous stage by 32, and each stage of the decoder layer will reduce the channel size of the previous stage by 32, *i.e.*, the channel size of the whole 3D scene encoder set as “32-64-96-128-160-128-96-64-32”. A linear layer is applied to the output of Sparse 3D U-Net to convert dimension into 256, which will then be leveraged as the input of the Omni Superpoint Transformer.

Omni Superpoint Transformer (OST). The proposed OST is composed of 3 encoder layers. The channel sizes of the hidden embedding and the feed-forward embedding are set to 256 and 1024, respectively. The head number of the distance-adaptive self-attention layer is set to 8. The distance bias factor σ is produced by applying a linear layer to the query feature. The classification head, mask head, and alignment head are all composed of only a linear layer. The output channel of the classification head is set to 199 (198 for instance categories and 1 for background). The output of the alignment head and mask head is set to 1024.

LoRA Parameters. LoRA [24] is applied to every linear layer int the LLM, *i.e.*, Vicuna-1.5-7B, except for the logits head. The lora_r is set to 64 and lora_alpha is set to 128.

7. Supplementary Experimental Analysis

Comparison on Nr3D. In the manuscript, we compare our 3D-LLaVA on referring segmentation benchmarks including ScanRefer [7] and Multi3DRefer [66]. In this supplementary material, we further include the comparison on Nr3D [1]. The performance is reported in Table 6. Our 3D-LLaVA consistently achieves the best performance among the competitors, further demonstrating the effectiveness of the proposed paradigm to re-use the OST for grounding the description on masks.

Effect of Frozen OST as The Mask Decoder. In this experiment, we analyze the effect of leveraging the frozen Omni Superpoint Transformer (OST) as the mask decoder for grounding open-ended language description to 3D masks. The results are presented in Table 7. The models shown in the table have the same architecture and are trained with the same instruction tuning strategy. For the model with an external decoder, the parameters of the mask decoder are randomly initialized and are all optimized in instruction tuning. It achieves 42.1% mIoU. In contrast, our default setting with frozen OST only trains a projection MLP for the [SEG] hidden state while boosting the mIoU to 43.3%. Besides, leveraging the frozen OST as the referring mask decoder will not add an extra module to the universal framework, which turns out to be more elegant. This comparison further demonstrates the advantages of the proposed 3D-LLaVA that makes use of the perception prior in OST.

Table 6. **Performance comparison on Nr3D [1].** Please note that we report the performance of 3D-LLaVA without finetuning on this dataset, while all of the other competitors are finetuned on it. BUTD-DETR* and EDA* are re-implemented by [21].

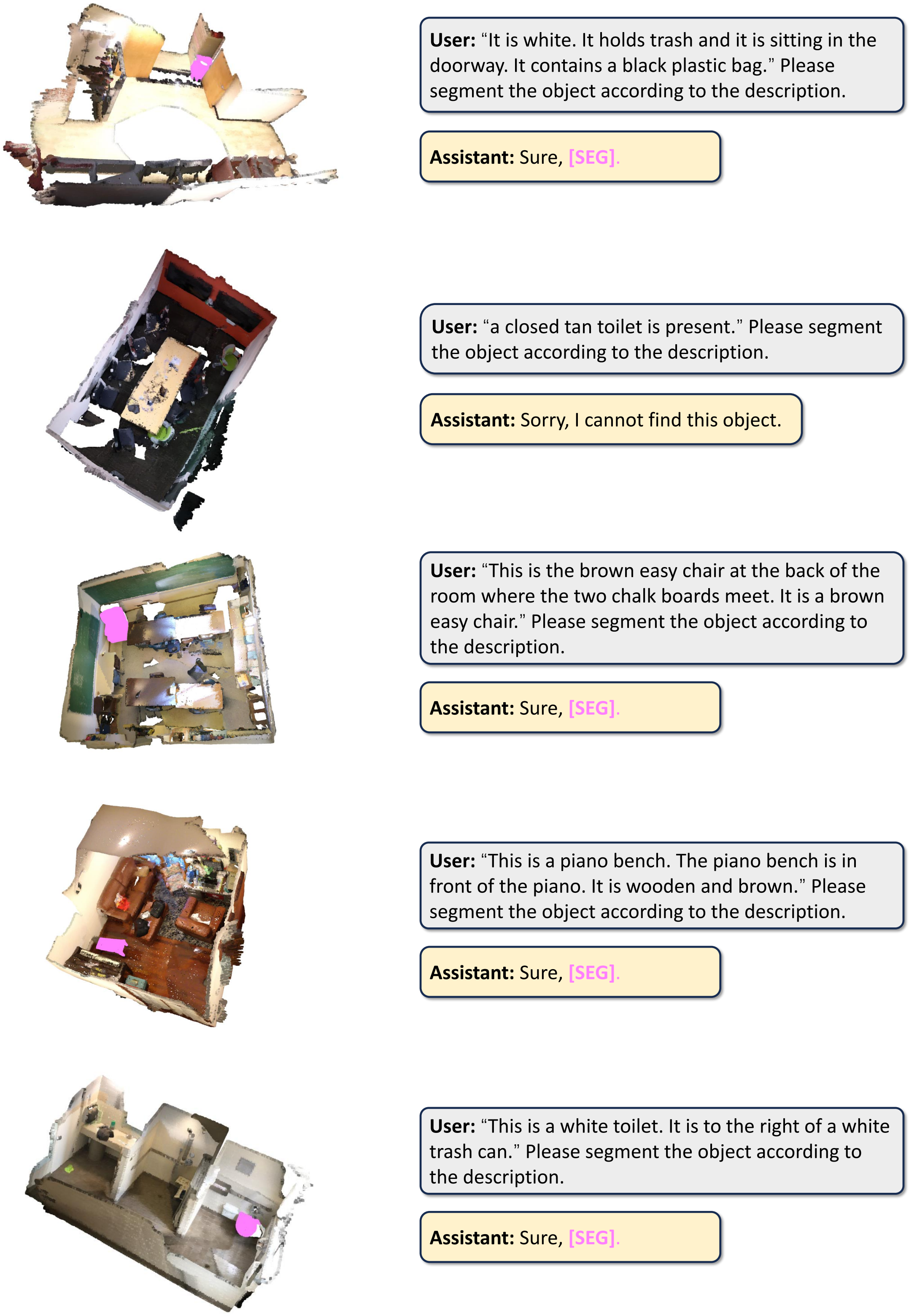
Mask Decoder	Nr3D (test) - mIoU
M3DRef-CLIP [66]	27.0
BUTD-DETR* [30]	27.5
EDA* [60]	29.3
SegPoint [21]	32.2
3D-LLaVA (ours)	**32.6**

Table 7. **Ablation Study on the Mask Decoder.** Our default setting is highlighted with light blue.

Mask Decoder	ScanRefer (val) - mIoU
External Decoder	42.1
Frozen OST	**43.3**

8. Supplementary Qualitative Results

This supplementary material provides additional qualitative results, showcasing the performance of our model across various room types, question categories, and target objects. Specifically, Figures 6, 7, 8, and 9 illustrate the qualitative results for 3D referring segmentation, 3D visual question answering, 3D situated question answering, and dense captioning, respectively. These examples demonstrate that 3D-LLaVA consistently delivers accurate responses across a diverse range of tasks.



User: "It is white. It holds trash and it is sitting in the doorway. It contains a black plastic bag." Please segment the object according to the description.

Assistant: Sure, [SEG].

User: "a closed tan toilet is present." Please segment the object according to the description.

Assistant: Sorry, I cannot find this object.

User: "This is the brown easy chair at the back of the room where the two chalk boards meet. It is a brown easy chair." Please segment the object according to the description.

Assistant: Sure, [SEG].

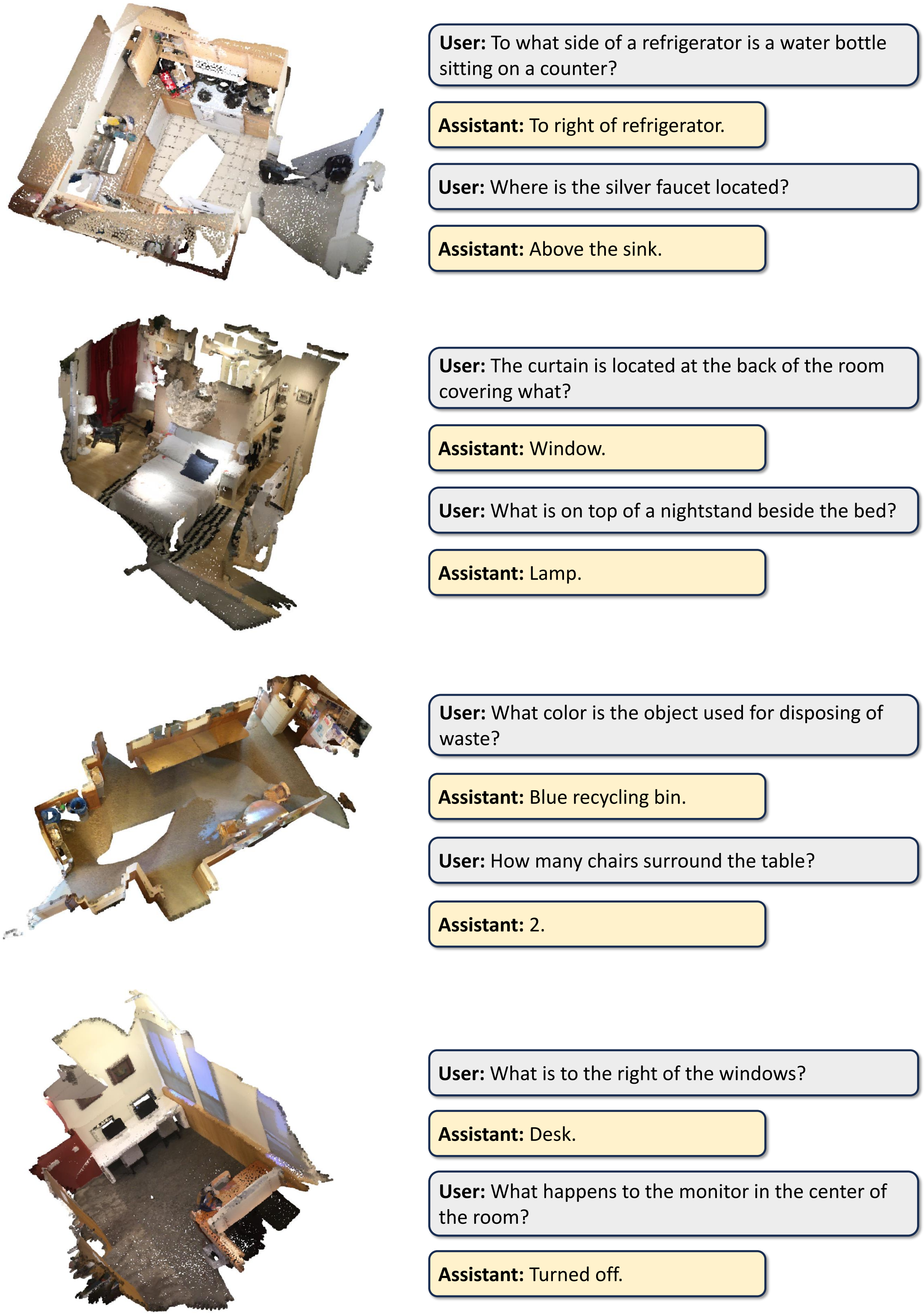
User: "This is a piano bench. The piano bench is in front of the piano. It is wooden and brown." Please segment the object according to the description.

Assistant: Sure, [SEG].

User: "This is a white toilet. It is to the right of a white trash can." Please segment the object according to the description.

Assistant: Sure, [SEG].

Figure 6. Qualitative result of referring segmentation. We visualize different scenes including the kitchen, office, bedroom, and toilet. Within these examples, the second scene does not include the referred object. Our 3D-LLaVA manages to judge whether the object exists.



User: To what side of a refrigerator is a water bottle sitting on a counter?

Assistant: To right of refrigerator.

User: Where is the silver faucet located?

Assistant: Above the sink.

User: The curtain is located at the back of the room covering what?

Assistant: Window.

User: What is on top of a nightstand beside the bed?

Assistant: Lamp.

User: What color is the object used for disposing of waste?

Assistant: Blue recycling bin.

User: How many chairs surround the table?

Assistant: 2.

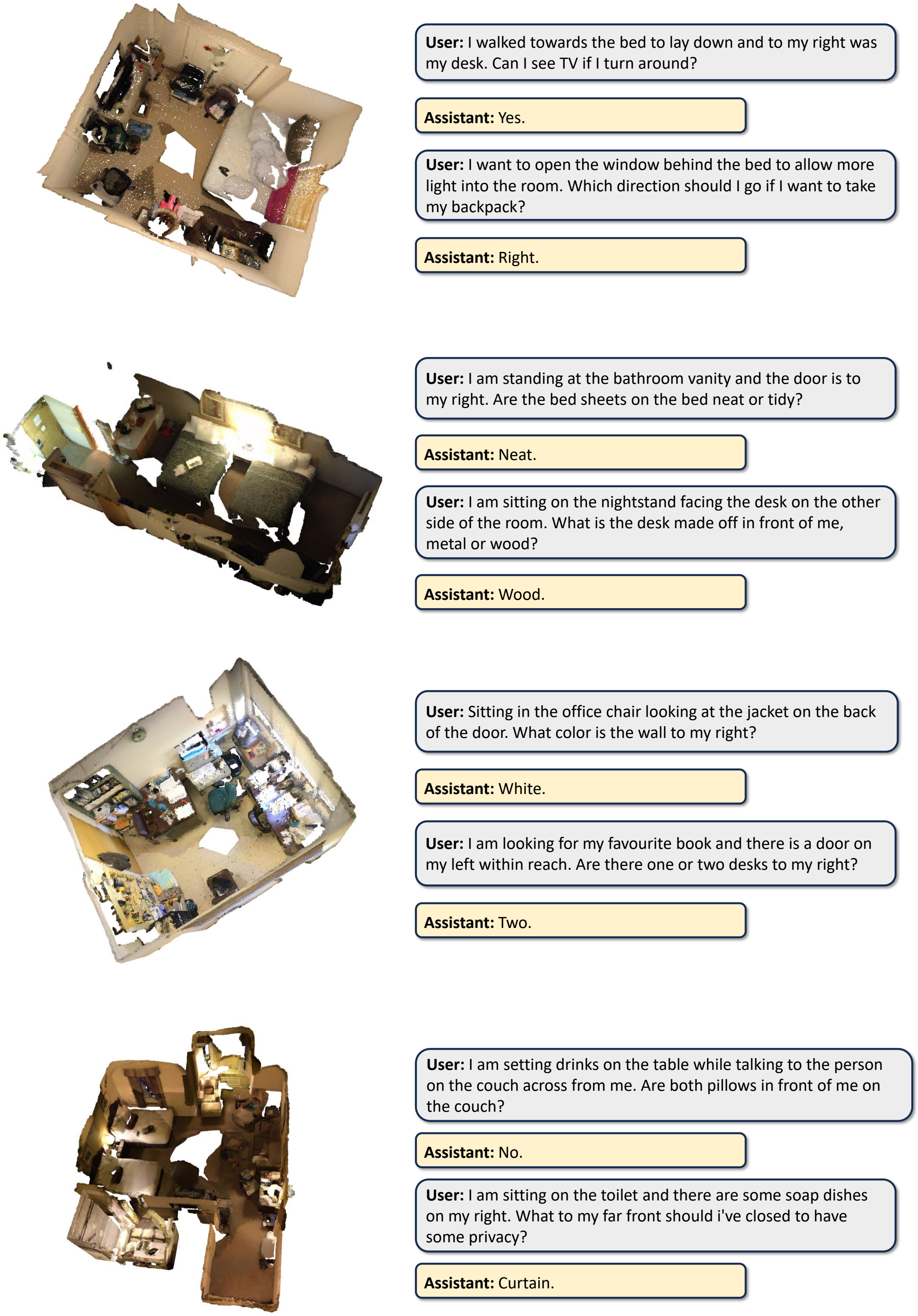
User: What is to the right of the windows?

Assistant: Desk.

User: What happens to the monitor in the center of the room?

Assistant: Turned off.

Figure 7. Qualitative result of visual question answering. We visualize different questions asking for objects' relationship, category, counting, color, and status.



User: I walked towards the bed to lay down and to my right was my desk. Can I see TV if I turn around?

Assistant: Yes.

User: I want to open the window behind the bed to allow more light into the room. Which direction should I go if I want to take my backpack?

Assistant: Right.

User: I am standing at the bathroom vanity and the door is to my right. Are the bed sheets on the bed neat or tidy?

Assistant: Neat.

User: I am sitting on the nightstand facing the desk on the other side of the room. What is the desk made off in front of me, metal or wood?

Assistant: Wood.

User: Sitting in the office chair looking at the jacket on the back of the door. What color is the wall to my right?

Assistant: White.

User: I am looking for my favourite book and there is a door on my left within reach. Are there one or two desks to my right?

Assistant: Two.

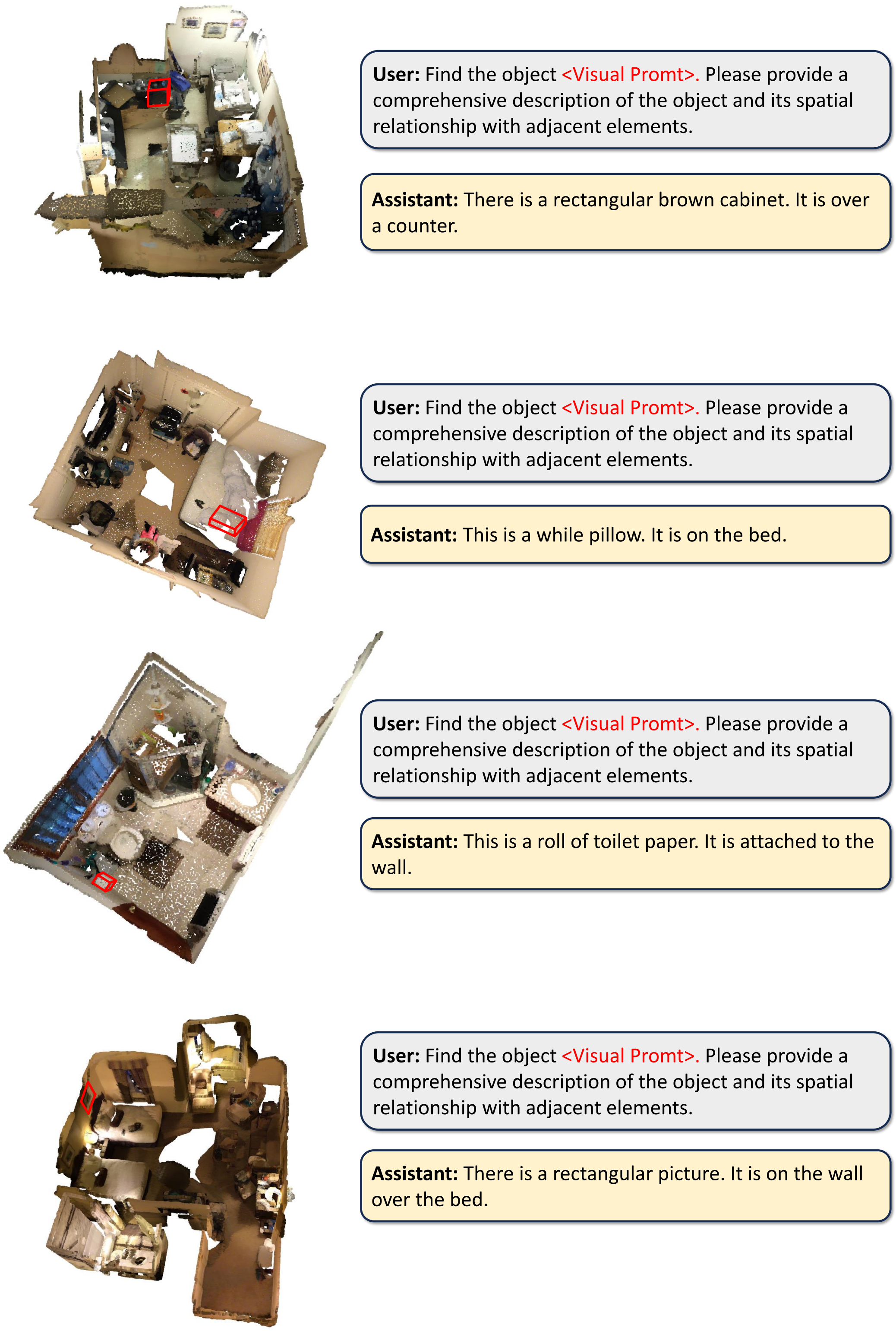
User: I am setting drinks on the table while talking to the person on the couch across from me. Are both pillows in front of me on the couch?

Assistant: No.

User: I am sitting on the toilet and there are some soap dishes on my right. What to my far front should i've closed to have some privacy?

Assistant: Curtain.

Figure 8. Qualitative result of situated question answering. We visualize different question types including those asking for direction, attribute, counting, and judgment.



User: Find the object <Visual Promt>. Please provide a comprehensive description of the object and its spatial relationship with adjacent elements.

Assistant: There is a rectangular brown cabinet. It is over a counter.

User: Find the object <Visual Promt>. Please provide a comprehensive description of the object and its spatial relationship with adjacent elements.

Assistant: This is a while pillow. It is on the bed.

User: Find the object <Visual Promt>. Please provide a comprehensive description of the object and its spatial relationship with adjacent elements.

Assistant: This is a roll of toilet paper. It is attached to the wall.

User: Find the object <Visual Promt>. Please provide a comprehensive description of the object and its spatial relationship with adjacent elements.

Assistant: There is a rectangular picture. It is on the wall over the bed.

Figure 9. Qualitative result of dense captioning. We visualize four different types of objects that are positioned in different rooms.