Appendix

A. Implementation Details

A.1. Downstream Tasks

ScanRefer [8]: The ScanRefer dataset contains 51,583 sentences written by humans to describe 800 scenes in ScanNet. We used the official split and allocated 36,665 and 9,508 samples for training and validation, respectively. The dataset is categorized into unique and multiple subsets based on whether the target object is a unique class in the scene. In this task, we need to find the target object described by a sentence. The evaluation metric for this task is accuracy under intersection over union (IoU) 0.25 and 0.5.

Nr3D/Sr3D [1]: The Sr3D dataset comprises of 83,572 utterances that are automatically generated using a template that focuses on the target-anchor spatial relationship. The Nr3D contains 45,503 human utterances. Both Sr3D and Nr3D are split by "Easy"/"Hard" and "ViewDep"/"ViewIndep". Hard samples are the ones with two or more distractors in a scene. The view-dependent samples contain language descriptions that rely on viewing directions. These two datasets are also used for visual grounding like ScanRefer. But grounding accuracy with ground truth object proposal is evaluated in this setting.

ScanQA [3]: ScanQA is a dataset for 3D question answering with 41,363 questions and 58,191 answers. Different from 2D QA, ScanQA focuses more on spatial relations. We follow [3] to use exact matches EM@1 and EM@10 as the evaluation metric. EM@K means the percentage of top K answers from the model matches one of the ground-truth answers. Also, we include text similarity metrics to evaluate answers, including BLEU-4, ROUGE, METEOR, and CIDEr.

SQA3D [36]: SQA3D is a benchmark for scene understanding of embodied agents with 6.8k unique situations, 20.4k descriptions, and 33.4k diverse reasoning questions. Given a situation, an embodied agent must understand embodied activities, navigation instructions, and common sense, and perform multi-hop reasoning. The evaluation metric is answer accuracy under different types of questions.

A.2. Model Architecture

For the scene encoder, we use a three-layer Pointnet++ [39] with radius 0.2, 0.4, and sample all points to aggregate a 768-dimension feature. For all text and object tokens, the dimension is 768 in the following multi-modal fusion layers. In the unified encoder, the number of attention heads is set to 12 and the dimension of feedforward layers is set to 2048. For the visual grounding head, we use a two-layer MLP with a hidden dimension of 384. For the question-answering head and the situated reasoning head, we use a two-layer MLP with input dimensions 512 (from the attention flat layer) and 768.

A.3. Training settings

The settings of pre-training including mask ratio, and optimization hyperparameters are introduced in the main paper. In this part, we elaborate on the fine-tuning details.

3D Visual Grounding: We only use a cross-entropy loss for fine-tuning 3D-VisTA on ScanRefer, Nr3D, and Sr3D. For all these grounding tasks, we set the batch size to 64, and the learning rate to 1e-4, We multiply the learning rate of the text encoder by 0.1 to stabilize the training process. We fine-tune the pre-trained 3D-VisTA for 100, 100, and 50 epochs for ScanRefer, Nr3D, and Sr3D, respectively. AdamW with $\beta_1 = 0.9, \beta_2 = 0.98$ is chosen as the optimizer. We use a warmup of 5,000 steps and a cosine annealing learning rate schedule.

3D Question Answering: We use a cross-entropy answer classification loss and a visual grounding loss for ScanQA. The batch size is 64 and the learning rate is 1e-4. 3D-VisTA is fine-tuned for 30 epochs with 2000 warmup steps for this task. Other optimization parameters are the same as the visual grounding task.

3D Situated Reasoning: Answer classification loss is used for SQA3D. We fine-tune 3D-VisTA for 50 epochs. Other optimization parameters are the same as the 3D question-answering task.

A.4. ScanScribe

In the main paper, we introduce our method of generating new scene-text pairs from scene graphs and large language models. More examples and cases are provided in this section. We support 40 relations and the mapping of relations to descriptions for the template-based generation is shown in Table A1.

With these relations, we can use templates like "This is a `object`, a `neighbor` is `relation` to `object`" and utilize GPT-3 to increase text diversity. During pre-training, to balance the proportion of template and GPT-3 generated texts in the 3R-Scan dataset, we duplicate texts from GPT-3 to 15 times for pre-training. Examples from both template-based generation and GPT-3 are presented in Fig. A1. We can observe that given entities and relations in a scene, GPT-3 can summarize them into a fluent and natural sentence.

B. Additional Results

We provide ablation studies on the use the template-generated text and GPT-3-generated text. As shown in Table A2, GPT-3-generated text improves Sr3D and Nr3D by 1.0% and 1.5%, while having little impact on ScanRefer and ScanQA. More qualitative results including failure cases are provided in Fig. A2.

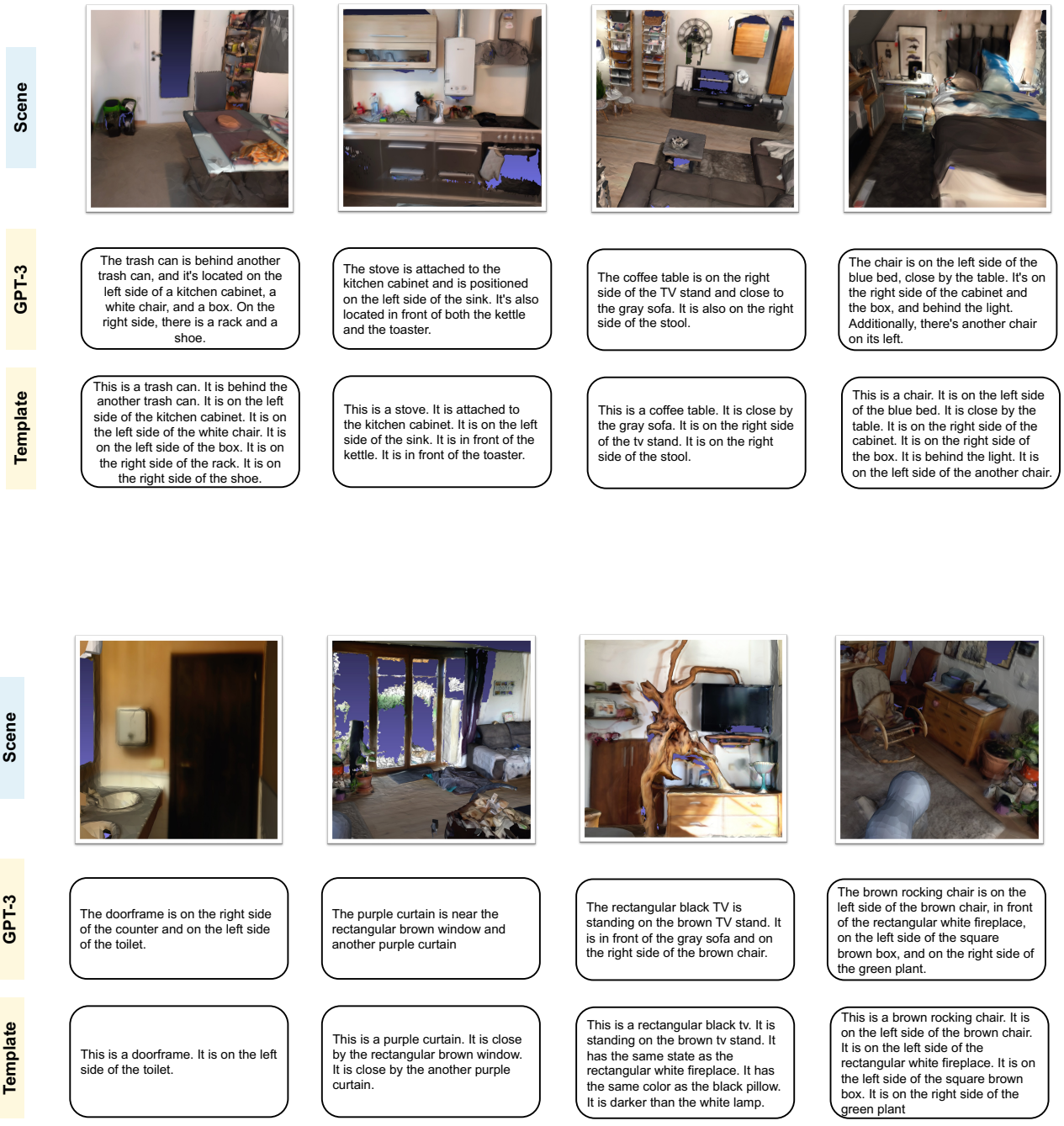

Figure A1: Examples of both template and GPT-3 generated text in ScanScribe dataset. GPT-3 generated text is more natural than template-generated text.

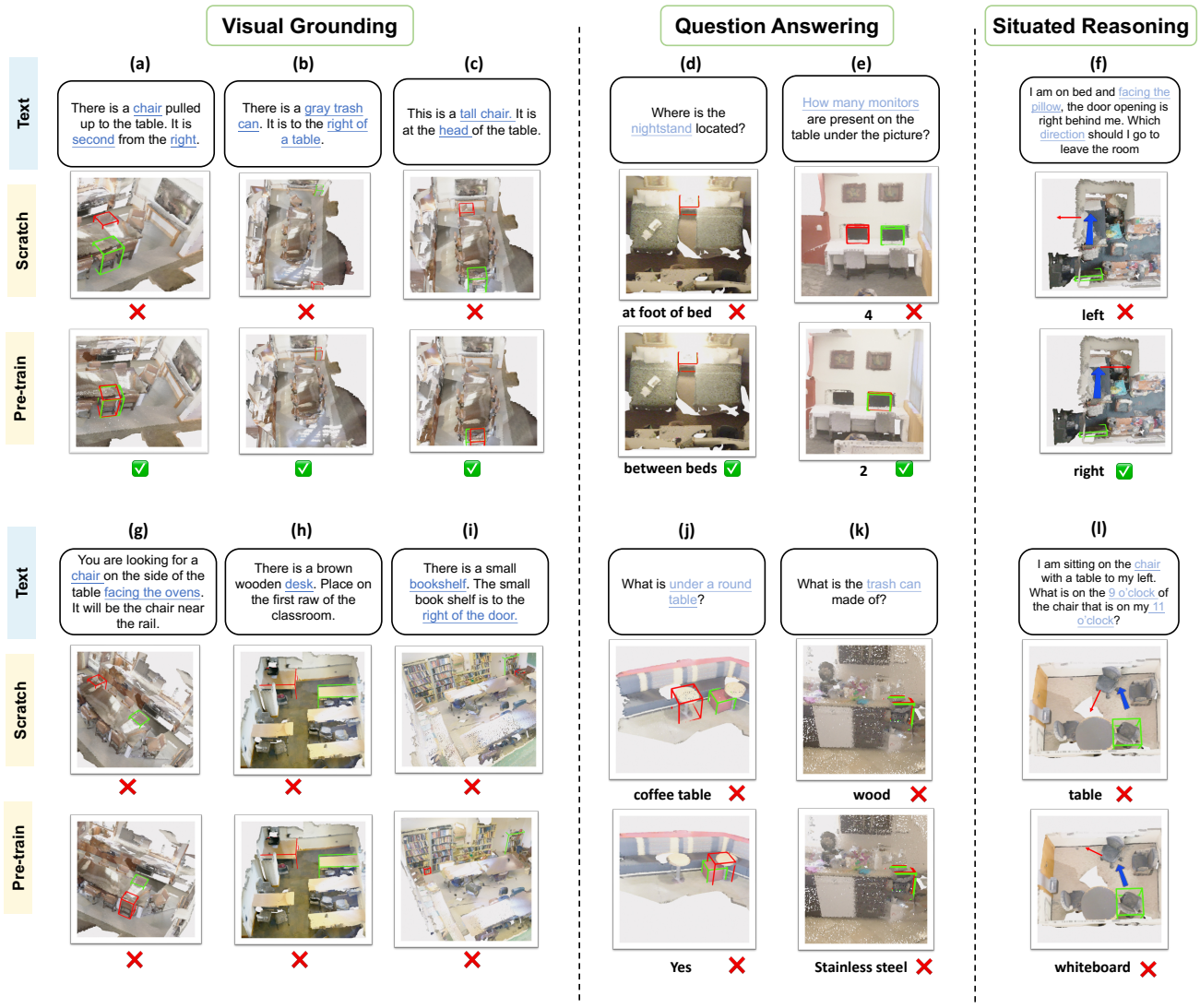

Figure A2: Qualitative results on ScanRefer, ScanQA, and SQA3D. Green and red denote the ground-truth and predicted object boxes, respectively. As shown in (a,b,c,d,e,f), the pre-trained 3D-VisTA shows advantages in spatial reasoning, concept grounding, and situation understanding. In spite of these advantages, (g, h, j) indicate that for some complicated cases with spatial relations, the pre-trained model still cannot understand them. (i, k) show that our model is still limited by the semantic information extracted by point clouds, which fail to locate the right object or understand texture. From (l), we can observe that our model may fail in the case requiring complex multi-hop reasoning.

Table A1: The mapping of relations to descriptions.

Relation	**Description**
supported by	is supported by the
left	is on the left side of the
right	is on the right side of the
front	is in front of the
behind	is behind the
close by	is close by the
inside	is inside the
bigger than	is bigger than the
smaller than	is smaller than the
higher than	is higher than the
lower than	is lower than the
same symmetry as	has the same symmetry as the
same as	is the same as the
attached to	is attached to the
standing on	is standing on the
lying on	is lying on the
hanging on	is hanging on the
connected to	is connected to the
leaning against	is leaning against the
part of	is part of the
belonging to	is belonging to the
built in	is built in the
standing in	is standing in the
covers	covers the
lying in	is lying in the
hanging in	is hanging in the
same color	has the same color as the
same material	has the same material as the
same texture	has the same texture as the
same shape	has the same shape as the
same state	has the same state as the
same object type	has the same object type as the
messier than	is messier than the
cleaner than	is cleaner than the
fuller than	is fuller than the
more closed	is more closed to the
more open	is more open than the
brighter than	is brighter than the
darker than	is darker than the
more comfortable than	is more comfortable than the

Table A2: Ablation studies on the template and GPT-3 generated text from 3R-Scan. We report the results on ScanRefer, Sr3D, Nr3D and ScanQA.

Template	GPT-3	ScanRefer	Sr3D	Nr3D	ScanQA
✓	×	57.4	75.4	62.7	23.7
✓	✓	57.4	76.4	64.2	23.8