Supplementary Materials: Context-Aware Indoor Point Cloud Object Generation through User Instructions

1 DATASET TRANSFORMATION

In this section, we provide more details about the data transformation from ReferIt3D to Nr3D-SA and Sr3D-SA. As depicted in Fig. 2, The data pipeline can be summarized as follows. (a) Paraphrase the descriptive texts into generative instructions based on LLM prompt engineering. (b) Filter out erroneous results by rules. (c) Re–paraphrase the incorrect ones using GPT-3.5 or GPT-4 according to the perplexity of sentences. (d) Proofread the revised sentences manually. The steps (b) and (c) are repeated until the rule-based filters detect no errors. After that, the step (d) is performed.

1.1 Prompting Templates

To generate diverse instructions without breaking changes in the semantics of original texts, we use dynamic prompting templates with different manually designed imperative verbs. We leverage the ChatGPT [2] API to rewrite each original descriptive text. The prompts for calling the ChatGPT API are shown in Fig. 1. The verbs are selected randomly and inserted into the corresponding slots during each call to the API. A weight is also assigned to each verb to ensure that the language is more natural. The verbs are listed as follows: **add** (10%), **put** (10%), **place** (10%), **set** (10%), **create** (10%), **generate** (10%), **insert** (10%), **produce** (10%), **lay** (5%), **deposit** (5%), **position** (5%), and **situate** (5%). To improve the diversity of the generated instructions (e.g., passive sentences and clauses), we reduce the likelihood of producing imperative sentences to 0.5.

1.2 Rule-based Filtering

Although LLMs have tremendous power, errors still occur when the original sentences are too complex, particularly for the Nr3D dataset. To detect errors in generated instructions, we employ rule-based filtering methods to identify obvious errors. The following are descriptions of our filtering rules:

(a) Locating words that are not transformed properly should be considered erroneous. The word blacklist covers: *find*, *pick*, *choose*, *select*, *locate*, *identify*, *search*, *seek*, *spot*, *gaze*, etc.
(b) Sentences without any generative verbs in 1.1 should be considered incorrect.
(c) Missing negative words and antonyms indicate high risks of changing the semantics of original sentences, such as *no*, *not*, *nowhere*, and *nothing*.

1.3 Mixed Correction

As a means of revising the error-prone sentences, we propose a mixed correction process involving both GPT and human labor. We first repeat the paraphrasing process on the incorrect sentences. We observe that Sr3D generates much better quality sentences than Nr3D due to its concise grammar structure. Since the proportion of incorrect sentences in Nr3D is smaller than that of the entire dataset, we perform manual proofreading on paraphrased sentences only from Nr3D as the final step. By the end of the process, only 335 sentences out of 41K sentences from Nr3D are required to be manually revised by two workers.

Table 1: The results of utilizing the generated data as augmented data for visual grounding.

Metrics \ Dataset	Nr3D w/o Aug.	Nr3D w/ Aug.
Easy(%, ↑)	35.2±0.3	**42.5**±0.3
Hard(%, ↑)	24.5±0.3	**30.5**±0.4
V-Dep(%, ↑)	28.4±0.2	**35.1**±0.4
V-Indep(%, ↑)	30.4±0.3	**37.0**±0.3
Among-True(%, ↑)	47.1±0.2	**51.7**±0.3
Overall(%, ↑)	29.7±0.2	**36.4**±0.4

Prompts:
You are a helpful chatbot.
Following sentences locate ONLY ONE object in a scene.
Transform the sentence to create this object.
Include generative verbs such as '`{I-VERB}`' to create it.
Change 'the' to 'a' or 'an' properly.
Imperative sentences are prefered.
Declarative sentences such as 'there is' are disallowed.
Avoid multiple imperative sentences.

`{TEXT}`

Figure 1: Dynamic prompting templates with slots. Imperative verbs `{I-VERB}` are selected randomly from a manually designed list with weights. The likelihood of preference for *imperative sentences* are set to 0.5. The original texts are placed to `{TEXT}` slot.

2 DETAILS OF METHODOLOGY

2.1 Losses

The object classification loss, denoted as ℓobj, represents a cross-entropy loss formulated as:

$$\ell_{obj} = -\sum_{c=1}^{C} \text{logits}_c \cdot \log(target_c) \tag{1}$$

Here, logits_c refers to the language logits produced by the point cloud encoding model, while $target_c$ represents the actual class of the object.

Similarly, the language loss, denoted as $\ell lang$, also follows a cross-entropy formulation:

$$\ell_{lang} = -\sum_{c=1}^{C} \text{logits}_c \cdot \log(target_c) \tag{2}$$

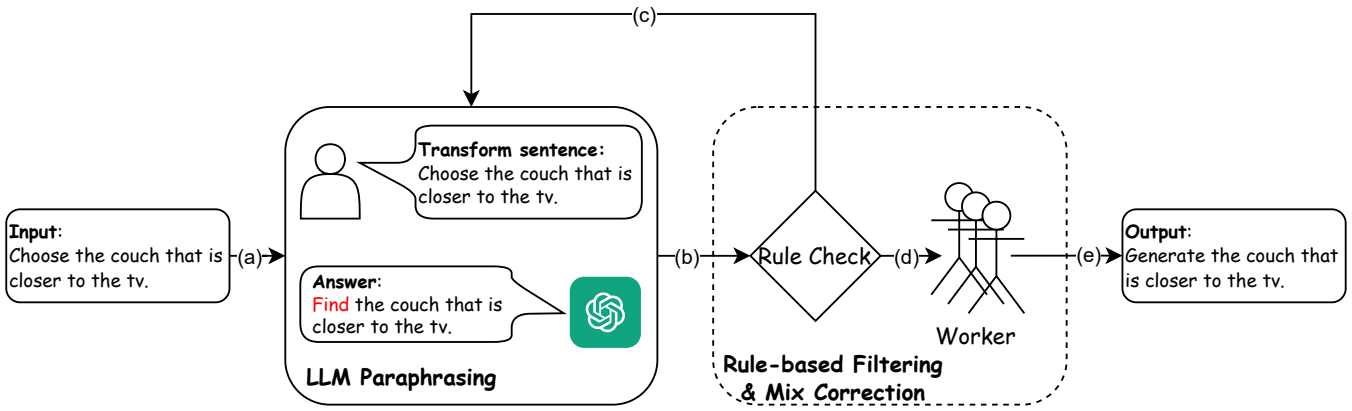


Figure 2: Data pipeline. Input texts are first processed through steps (a) and (b). If the generated texts are considered incorrect, step (c) would be taken to re-run the paraphrasing process until no error is found. After that, manual proofreading and correction are applied as step (d) to output the final results.

In this context, logits_c pertains to the language logits generated by the underlying BERT model, and $target_c$ signifies the desired class of the text.

The Point-E Loss function can be seen as an Mean Squared Error (MSE, or L2), expressed by:

$$\ell_{point-e} = \frac{1}{N}\sum_{i=1}^{N}(\text{denoised}_i - \text{target}_i)^2 \tag{3}$$

Here, N denotes the total number of elements. denoised_i represents the i-th denoised output of the model, while target_i represents the i-th corresponding original data point.

3 ADDITIONAL EXPERIMENT RESULTS

3.1 Quantitative Results

Table 3 shows the complete results of EMDs and classification accuracy in response to the Experiments Section. The table is sorted according to the proportion of objects within the entire dataset for ease of comparison. It is noteworthy that the classification accuracy is higher for object classes with more data, whereas the performance drops drastically for object classes with less data.

3.2 Qualitative Results

We present additional augmented scenes created by our method to enhance the qualitative analysis. Figure 4 presents 3×3 examples generated by our method. Both the generated and reference objects are annotated to assess the performance of our method. While some of the samples may not perfectly match the ground-truths, the generated objects align well with the given instructions and context surroundings.

However, our methods also have limitations due to the lack of sufficient training data. Figure 3 shows some typical failure cases in our proposed method. Due to the difficulty in accurately determining the correct location, the generated objects may occasionally have errors in their positions (Fig. 3a) or deviate slightly from the actual ground-truth position (Fig. 3b). Additionally, the diffusion process necessitates a substantial volume of data to reconstruct an object, making it more challenging to reconstruct objects from low-quality categories like doors and curtains in the ScanNet dataset (Fig.3c & Tab. 3). Further advancements in 3D modeling could help alleviate these issues by reducing the shortage of 3D data.

4 APPLICATION

We select visual grounding tasks to demonstrate our proposed method. We use the MVT model [1] and Nr3D dataset for training in visual grounding. The initial Nr3D dataset is evenly split into two parts referred to as part I and part II. For training without augmentation, we utilize part I to train the MVT dataset and assess performance on the test dataset. To train with augmented data created from our generated objects, we initially generate objects using part II from the Nr3D-SA dataset and then merge part I from Nr3D dataset with generated objects to form the training dataset.

The evaluation is conducted using the official evaluation script from MVT repository[1]. The "easy" and "hard" splits depend on whether the scene contains more than two distractors as the same category as the reference object. The "view-dep." and "view indep." splits depend on whether the referring expression is dependent on the speaker's view or not. Table 1 illustrates that our approach improves the performance of subsequent tasks, thus showing the effectiveness of the method we have suggested.

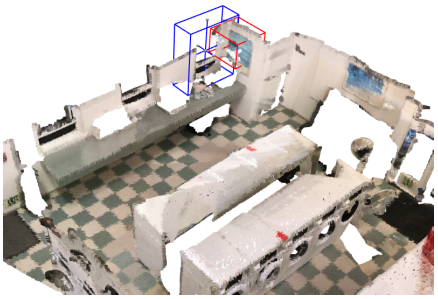
5 OVERHEAD

We train and test our model on 4 RTX 4090 GPUs. During training, we utilize a batch size of 16 across various datasets and sampling points. During the inference phase, we use a batch size of 1 for the diffusion process. The model training and inference process overheads for each configuration are displayed in table 2.

[1] https://github.com/sega-hsj/MVT-3DVG

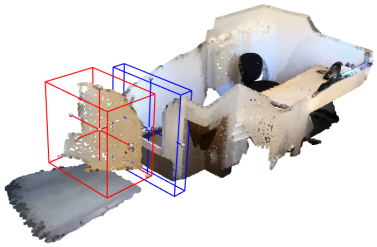
Set up a cabinet that is closer to the refrigerator.

(a) Incorrect location



Put a window on the far right when facing the three windows.

(b) Fine-grained deviation



Produce a door that is on the left side of the chair.

(c) Low-quality object

Figure 3: Typical failure cases involved in our method. Each augmented scene is accompanied by an instruction, in which a red and blue bounding box represents the generated and reference objects, respectively.

Table 2: Model training and inference overhead.

Method	# of Points	Train. VRAM(GiB)	Train. Step Latency(ms)	Infer. VRAM(GiB)	Infer. Latency(ms)
Nr3D-SA	1024	46.85	309.67	2.41	1764.94
Nr3D-SA + Sr3D-SA	1024	46.85	301.20	2.41	1754.98
Nr3D-SA + Sr3D-SA	2048	79.19	546.45	3.21	4000.53

REFERENCES

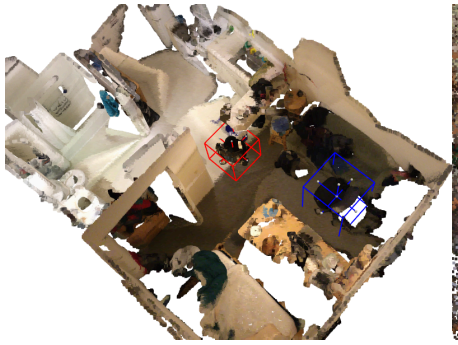
[1] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. 2022. Multi-view transformer for 3d visual grounding. In *CVPR*. 15524–15533.

[2] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

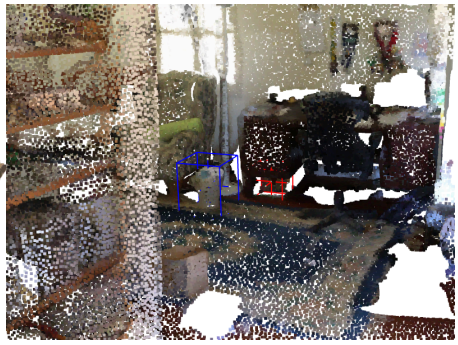
Object Class	Ours						Point-E Only					
	MMD	COV	1-NNA	JSD	Acc@1	Acc@5	MMD	COV	1-NNA	JSD	Acc@1	Acc@5
chair(7.58%)	11.52	33.10	98.80	2.474	80.59	95.91	11.15	21.63	99.63	2.78	62.39	91.56
door(6.72%)	7.66	29.43	99.78	2.918	0.09	5.14	7.57	21.39	99.93	2.936	2.16	13.89
trash can(4.78%)	10.38	36.47	99.13	2.983	31.66	56.38	11.62	19.37	99.69	3.7	6.77	30.6
window(4.76%)	9.56	30.08	99.17	3.416	31.12	62.55	9.69	27.53	99.77	3.309	39.4	64.27
table(4.70%)	12.59	30.61	98.88	3.988	47.34	80.55	13.2	20.93	99.37	4.614	21.73	45.58
cabinet(3.71%)	12.22	36.38	98.57	2.960	16.95	64.00	10.82	27.64	99.5	2.974	9.21	38.84
picture(3.53%)	6.44	36.25	99.53	3.146	19.06	50.00	6.98	31.24	99.58	3.309	35.01	67.09
shelf(3.42%)	9.28	35.90	99.65	2.912	35.55	76.18	8.79	23.99	99.69	2.91	24.71	48.07
lamp(3.17%)	13.59	30.17	99.17	4.069	50.41	71.90	13.58	24.1	99.86	4.49	12.53	23.28
desk(3.16%)	14.59	35.11	99.56	3.660	11.56	44.67	13.96	22.85	99.85	3.839	1.9	15.33
pillow(2.36%)	10.96	40.94	98.82	3.033	51.18	69.69	10.71	37.24	99.24	3.18	26.58	50.23
backpack(2.34%)	12.02	36.79	97.65	2.758	33.07	66.34	11.54	27.52	99.21	2.414	27.94	68.31
sink(2.33%)	14.32	31.71	99.59	4.461	60.16	84.15	12.19	25.95	98.96	3.915	41.18	70.59
towel(2.23%)	10.01	35.63	99.23	3.197	8.43	45.21	9.37	29.08	99.66	3.214	7.06	22.35
monitor(2.17%)	8.90	37.85	99.09	3.205	81.38	91.09	8.91	31.31	99.7	3.241	68.21	86.79
box(2.09%)	13.58	40.36	98.32	3.551	35.87	67.04	13.85	28.53	98.87	3.814	7.06	31.98
nightstand(1.77%)	15.13	26.32	98.68	4.392	67.11	86.84	11.86	32.28	99.61	3.868	5.51	50.39
couch(1.77%)	11.71	35.68	98.68	3.382	18.94	50.66	10.08	36.5	99.9	3.417	0.78	6.02
kitchen cabinets(1.61%)	9.99	32.72	99.48	3.590	10.47	40.84	9.2	30.62	99.82	3.247	1.81	26.45
curtain(1.54%)	11.61	33.83	98.12	3.925	2.26	4.51	9.27	24.93	99.46	3.688	0.54	5.96
bookshelf(1.47%)	12.62	28.26	99.46	3.478	24.46	70.11	11.91	27.17	99.73	3.533	0.54	12.23
office chair(1.40%)	13.98	33.69	96.79	3.288	12.30	80.21	10.94	33.73	100	2.857	1	56.49
bed(1.39%)	8.60	36.73	98.98	3.440	10.88	35.03	10.2	29.76	99.39	3.604	1.56	5.71
stool(1.37%)	16.27	30.88	99.26	4.114	11.76	32.35	14.85	22.35	100	3.855	12.35	27.65
keyboard(1.34%)	7.37	32.14	99.82	3.099	18.21	54.29	7.82	36.3	99.67	3.27	11.52	22.83
file cabinet(1.29%)	14.85	32.43	97.57	4.202	35.68	64.32	17.31	18.35	99.61	4.267	5.68	40.83
plant(1.26%)	14.60	28.57	98.51	3.102	7.14	23.81	11.89	25.17	99.49	2.8	1.7	11.56
dresser(1.21%)	16.01	32.35	99.71	4.916	14.12	33.53	11.82	19.67	99.48	4.18	6.9	44.56
mirror(1.21%)	13.24	31.16	98.91	5.045	41.30	63.77	12.74	27.37	99.08	4.829	16.32	40.79
coffee table(1.16%)	14.65	42.31	99.36	5.082	15.38	61.54	17.95	18.98	99.54	5.75	0.46	14.81
kitchen cabinet(1.03%)	11.40	42.61	97.39	4.211	17.39	46.09	10.48	33.54	99.07	3.696	15.53	42.86
whiteboard(0.95%)	7.54	36.02	98.45	3.721	2.48	14.91	7.65	31.08	99.4	3.197	2.79	13.55
shoes(0.90%)	8.88	37.14	98.00	3.073	21.14	44.57	8.67	25.3	99.51	2.712	62.77	80.54
book(0.89%)	10.44	38.07	99.08	4.513	8.72	21.56	10.21	37.31	99.74	4.431	0.52	5.44
computer tower(0.89%)	16.41	44.44	98.46	4.249	32.10	72.84	16.99	27.44	99.09	4.4	74.39	91.46
radiator(0.84%)	21.52	14.29	100.00	5.829	64.29	85.71	12.25	30.26	98.68	4.306	5.26	17.11
bag(0.83%)	15.11	38.41	97.83	3.212	3.62	28.99	13.93	34.62	97.95	2.777	0.26	7.69
toilet paper(0.82%)	16.72	36.50	99.27	5.129	32.85	57.66	14.11	34.22	98.41	4.571	8.22	21.49
armchair(0.75%)	10.42	34.86	98.78	2.723	4.89	50.76	11.07	26.88	98.27	2.754	0.41	16.5
laptop(0.71%)	13.40	50.00	92.86	5.011	71.43	78.57	11.99	32.58	97.75	3.3	9.55	43.82
toilet(0.68%)	15.71	39.29	97.32	3.882	73.21	78.57	11.03	27.75	98.99	3.022	26.59	55.78
books(0.68%)	21.02	21.59	99.43	5.030	22.73	80.68	17.56	37.5	91.67	5.814	0	41.67
kitchen counter(0.68%)	12.98	29.41	98.53	5.147	52.94	70.59	12.71	31.4	99.71	4.469	32.56	61.05
telephone(0.67%)	15.12	46.43	97.62	5.341	28.57	76.19	13.83	36.08	99.05	3.999	0	3.8
cup(0.66%)	18.65	26.14	100.00	4.969	3.27	13.73	15.46	37.69	99.62	4.939	7.69	21.54
suitcase(0.65%)	17.81	23.53	98.53	5.521	32.35	58.82	13.12	24.41	99.61	3.972	2.62	52.23
microwave(0.65%)	17.97	39.29	100.00	5.309	14.29	39.29	12.85	40.63	99.22	3.791	16.41	62.5
recycling bin(0.59%)	12.59	34.78	99.28	5.246	10.14	31.88	16.19	19.19	99.75	3.904	34.85	69.7
bottle(0.52%)	16.89	30.49	97.56	4.702	1.22	28.05	12.67	30.05	100	4.545	0.55	9.29
ottoman(0.48%)	24.26	40.00	100.00	6.081	33.33	73.33	16.94	21.35	98.88	4.941	0	1.69
light(0.45%)	17.70	48.48	96.97	4.974	6.06	42.42	16.69	23.6	99.44	5.459	1.12	6.74
end table(0.43%)	15.94	35.19	98.15	4.795	5.56	27.78	16.13	28.21	97.01	4.759	1.71	26.5
printer(0.42%)	17.69	33.33	98.68	3.697	1.75	27.19	16.09	39.09	99.09	4.201	4.55	52.73
sofa chair(0.37%)	12.23	36.11	100.00	5.060	0.00	0.00	13.2	31.61	97.99	3.489	0	9.2
board(0.35%)	18.76	32.14	96.43	5.207	3.57	10.71	12.13	42	99	4.034	0	2
laundry hamper(0.34%)	13.85	31.37	100.00	6.181	37.25	58.82	27.29	18.42	100	5.54	0	2.63
coffee maker(0.33%)	20.93	42.86	96.43	5.455	28.57	28.57	12.77	38.46	98.08	5.048	0	0
blanket(0.31%)	7.79	42.50	100.00	4.693	0.00	0.00	14.37	31.06	100	3.726	3.79	11.36
mouse(0.31%)	19.35	50.00	94.44	5.246	0.00	5.56	13.8	47.37	98.25	6.24	1.75	3.51
paper towel dispenser(0.31%)	19.98	41.67	96.88	4.282	0.00	0.00	16.07	36.36	99.43	4.07	6.82	15.91
bathroom stall door(0.30%)	20.53	29.17	97.92	5.230	2.08	29.17	6.6	21.55	99.57	3.771	0	0
person(0.26%)	5.81	38.00	100.00	3.947	4.00	26.00	17.79	22.22	97.22	5.141	0	0
bathroom stall(0.26%)	23.72	30.30	100.00	6.124	0.00	21.21	19.66	24.49	99.49	5.524	0	0
cabinets(0.20%)	7.49	28.57	100.00	3.893	0.00	0.00	14.2	27.66	98.4	5.083	1.06	11.7
bar(0.20%)	17.38	51.43	95.71	4.977	0.00	0.00	6.78	31.08	99.32	4.403	5.41	16.22
bench(0.19%)	8.23	35.00	100.00	4.002	0.00	2.50	24.26	30.23	98.84	5.925	0	0
wardrobe closet(0.18%)	12.43	36.76	97.06	5.618	2.94	4.41	20.66	33.33	83.33	5.908	0	0
doors(0.16%)	13.57	37.93	100.00	5.928	0.00	0.00	6.92	20.93	99.42	3.354	0	0
storage bin(0.16%)	8.24	31.82	100.00	4.015	10.61	27.78	15.94	29.03	98.39	4.775	0	1.08
blackboard(0.15%)	12.78	43.40	97.64	3.426	0.00	1.89	14.98	20.51	100	4.479	0	0
soap dish(0.14%)	24.56	35.71	96.43	6.213	0.00	0.00	14.65	30.72	99.4	5.64	0	1.81
sign(0.13%)	18.46	23.68	100.00	6.281	0.00	2.63	12.04	44.19	100	5.732	0	2.33
rail(0.12%)	7.57	29.96	100.00	3.941	3.24	23.89	7.24	30.47	100	4.345	2.34	12.89
cart(0.08%)	12.52	32.00	98.67	3.140	0.00	0.00	15.3	21.25	98.75	3.831	0	0
oven(0.07%)	20.80	27.78	100.00	5.574	0.00	0.00	19.65	27.78	97.22	5.498	0	5.56
pipe(0.05%)	19.96	31.71	100.00	5.653	0.00	0.00	19.06	29.55	98.86	6.03	0	2.27

Table 3: Complete experiment results for 32,000 objects randomly drawn. MMD is multiplied by 10^2 and JSD is multiplied by 10^1.

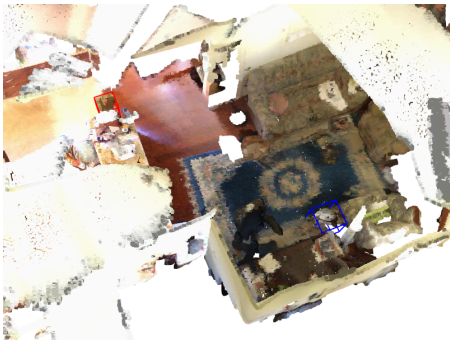
Situate a window near the armchair.



Generate a chair that is between the couch and the suitcase.

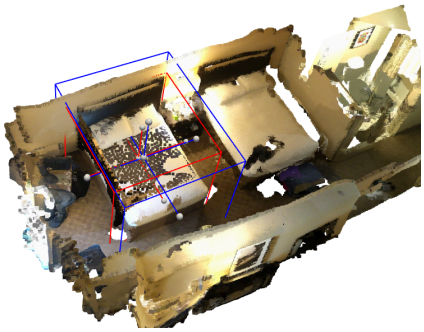


Add a trash can that is far from the cabinet.



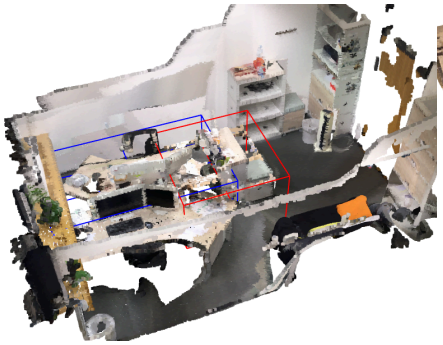
Position a trash can in the middle of the table and the desk.

Insert a pillow that is far from the curtain.



Place a bed that is far away from the toilet paper.

Add a table that is farthest from the curtain.



Produce a desk that is close to a backpack.

Position the suitcase near the chair.

Figure 4: Additional qualitative results. Each augmented scene is accompanied by an instruction, in which a red and blue bounding box represents the generated and reference objects, respectively.