Robin3D: Improving 3D Large Language Model via Robust Instruction Tuning

Supplementary Material

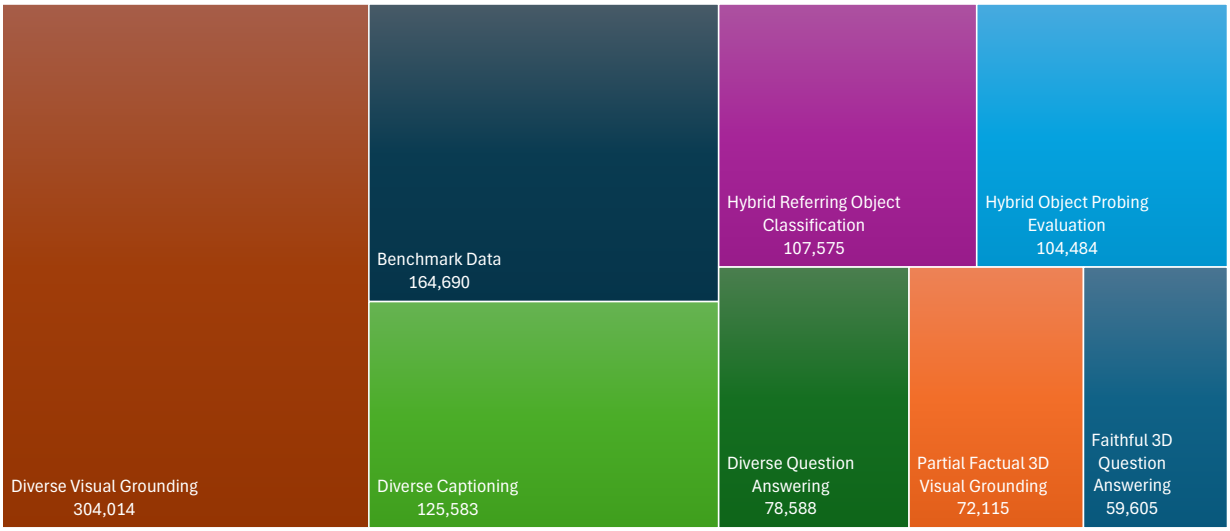

Figure 1. The number of samples for different tasks in our robust data and the visualization of their proportion of the total data.

Figure 2. The word cloud of our adversarial data.

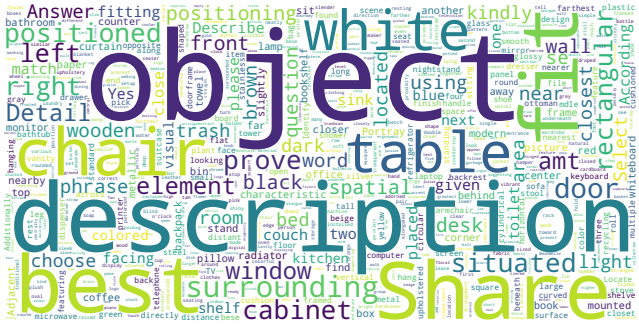

Figure 3. The word cloud of our diverse data.

1. Data Analysis

As shown in Fig. 1, we provide detailed statistics on the number of samples for different tasks in our robust dataset, along with a qualitative result of their respective proportions in the total data. For our Diverse Instruction data, we split it into three parts based on the task categories for statistical purposes, which include Diverse Visual Grounding, Diverse Captioning, and Diverse Question Answering.

We further present the word cloud of our Adversarial Instruction data and Diverse Instruction data in Fig. 2 and Fig. 3, respectively. We exclude the words related to object IDs, as they pertain to the referring and grounding format rather than the actual data content.

In Fig. 4, we provide statistics on the average sentence length for each task in our robust dataset. Here, the sentence length is calculated as the number of words in the question prompt plus the number of words in the answer, excluding the count of object IDs.

2. Detailed Ablation Study of Adversarial data

To further evaluate the effectiveness of each task in our Adversarial Instruction data, we conduct detailed ablation studies on Hybrid Object Probing Evaluation data, Hybrid Referring Object Classification data, Partial Factual 3D Visual Grounding data, and Faithful 3D Question Answering data by adding them to the benchmark data in each experiment. As shown in Tab. 1, all four tasks in the Adversarial data contribute notable improvements compared with solely training on the benchmark data.

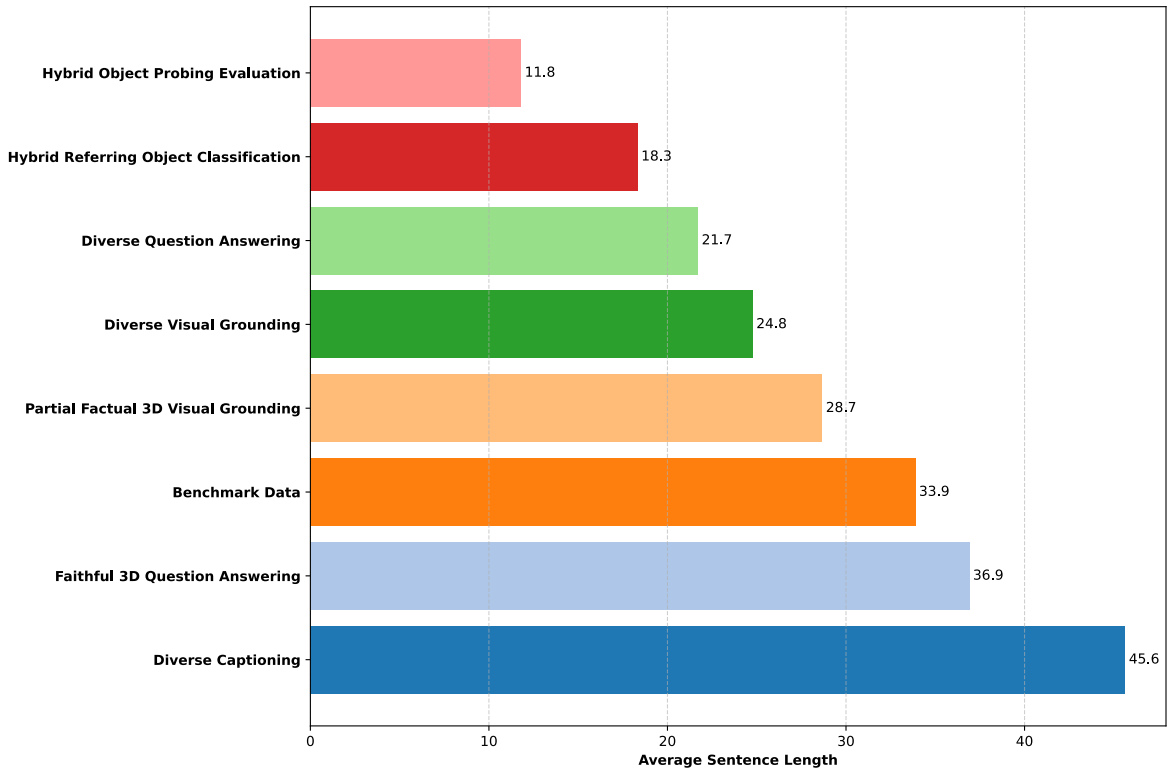

Figure 4. The average sentence length of different tasks in our robust data.

Data	ScanRefer Acc@0.5	Multi3DRefer F1@0.5	Scan2Cap C@0.5	ScanQA(val) M	SQA3D(val) EM
Benchmark	45.3	50.2	73.6	17.7	48.9
+ HOPE	45.8	52.5	76.1	17.8	50.1
+ HROC	47.7	53.0	78.9	18.0	50.3
+ PF-3DVG	45.7	51.0	77.2	17.9	49.6
+ 3DFQA	47.2	52.1	77.5	17.9	50.2

Table 1. **Ablation study on Adversarial Instruction data.** *Benchmark* denotes training on the original training set of the benchmarks. *HOPE* denotes adding the Hybrid Object Probing Evaluation data to the original training set. *HROC* denotes adding the Hybrid Referring Object Classification data to the original training set. *PF-3DVG* denotes adding the Partial Factual 3D Visual Grounding data to the original training set. *3DFQA* denotes adding the Faithful 3D Question Answering data to the original training set.

3. Details of Robin3D

To train a 3D LLM using instruction fine-tuning, we first represent the 3D scene as a sequence of vision tokens, then append it with system and instruction prompts, expressed as sequences of language tokens, to indicate the task. Taking the above tokens as input, a LLM is supervised to output the answer tokens via next token prediction. Specifically, as shown in Fig. 5, given the point cloud of a 3D scene, we use the pre-trained 3D segmenter Mask3D [**?**] to extract object features along with their corresponding 3D masks. Following Chat-Scene [**?**], we further sample each object's point cloud based on the 3D masks, normalize it, and employ the pre-trained Uni3D [**?**] to extract unified object-centric 3D features. Additionally, 2D masks projected from the 3D masks are used to sample and average 2D features, which are extracted by DINO v2 from multi-view images of each object. Our Relation-Augmented Projector fuses the 3D features and position embeddings from Mask3D and Uni3D into our final 3D features. In line with Chat-Scene [**?**], we incorporate special tokens $\{< \texttt{OBJ}_i >\}_{i=1...n}$ as object IDs into the vocabulary. These ID tokens are paired

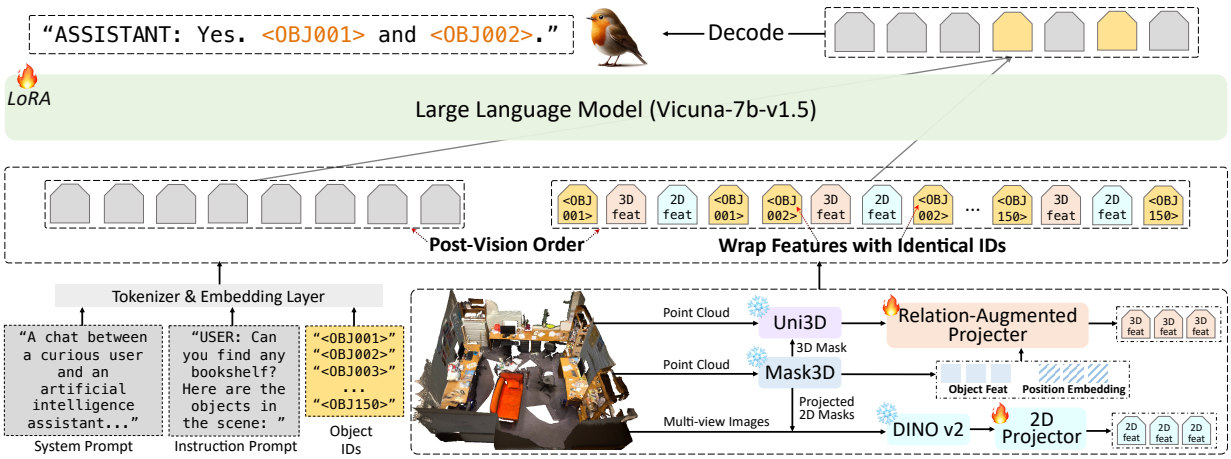

Figure 5. Overview of Robin3D model structure. *Bottom*: Our Relation-Augmented Projecter fuses the features and position embedding from Mask3D and Uni3D to generate final 3D features. 2D features from DINO v2 are projected into the LLM space. We freeze the Mask3D, Uni3D, and DINO v2. *Middle*: We enhance the connection between object IDs and object features by wrapping the features with identical IDs and the Post-Vision order. *Top*: We use LoRA to fine-tune the LLM on our constructed 1 million instruction data.

with 2D and 3D object features to indicate each object, for referring to the object in the input instruction or grounding the object in model's output. We combine each object feature with its corresponding object ID, and appends the system and question prompts at the beginning of the sequence, which are then fed into the LLM.