

SCENE-R1: SUPPLEMENTARY MATERIAL AND CODE GUIDE

Anonymous authors

Paper under double-blind review

OVERVIEW

Currently, utilizing large language models to understand the 3D world is becoming popular. Yet existing 3D-aware LLMs act as black boxes: they output bounding boxes or textual answers without revealing how those decisions are made, and they still rely on pre-trained 3D detectors to supply object proposals.

We introduce Scene-R1, a video-grounded framework that learns to reason about 3D scenes without any point-wise 3D instance supervision by pairing reinforcement-learning-driven reasoning with a two-stage grounding pipeline.

In the temporal grounding stage, we explicitly reason about the video and select the video snippets most relevant to an open-ended query. In the subsequent image grounding stage, we analyze the image and predict the 2D bounding box. After that, we track the object using SAM2 to produce pixel-accurate masks in RGB frames, and project them back into 3D, thereby eliminating the need for 3D detector-based proposals while capturing fine geometry and material cues.

Scene-R1 can also adapt to the 3D visual question answering task to answer free-form questions directly from video. Our training pipeline only needs task-level 2D boxes or textual labels without dense 3D point-wise labels. Scene-R1 surpasses existing open-vocabulary baselines on multiple datasets, while delivering transparent, step-by-step rationales. These results show that reinforcement-learning-based reasoning combined with RGB-D video alone offers a practical, annotation-efficient route to trustworthy 3D scene understanding.

This document provides an overview of Scene-R1, including setup instructions, the training process, and evaluation guidelines.

SETUP

```
conda create -n vlm python=3.11
conda env create -f environment.yml
conda activate vlm
```

TRAINING

Scene-R1 training involves the following steps:

1. Data Preprocessing:

Download the dataset. Before training, you need to preprocess the video data.

```
1 bash preprocess_scannet.sh
```

2. GRPO Training:

```
1 cd scripts
2 # First stage
3 bash run_scanrefer.sh
4 # Second stage
5 bash run_scanrefer_image.sh
6
```

EVALUATION

After training, evaluate your model's performance:

```
1 bash scripts/evaluate.sh # Use evaluate.sh for evaluation.
```