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

Weitai Kang[1], Haifeng Huang[2], Yuzhang Shang[3] , Mubarak Shah[3], Yan Yan[1]
[1]University of Illinois Chicago, [2]Zhejiang University, [3]University of Central Florida

ICCV HONOLULU HAWAII OCT 19-23, 2025

Paper Code

Motivation

Background: Aims to build a 3D Large Language Models to serve as a general-purpose assistants in the 3D real world.

Related works: In contrast to the versatile image-text pairs employed for training 2D MLLMs, collecting 3D instruction-following data for 3DLLM remains a significant challenge. Existing works have made progress in generating more instruction data by ChatGPT, or rule-based methods.

Bottleneck: **However,** their data still lack robustness in two aspects: 1) Most of the existing instruction data consist of positive pairs, lacking adversarial or negative samples. 2) Current instruction data lack diversity in language styles.

Contributions: In this paper, we introduce Robin3D, a powerful 3DLLM trained on large-scale instruction-following data generated by our novel data engine, Robust Instruction Generation (RIG) engine.

Method

Scene-Level, Category-Base:

Hybrid Object Probing Evaluation

Can you find any bookshelf; telephone; couch?

Yes. <OBJ001> ❶ and <OBJ0002> ❷ ; No; Yes. <OBJ003>. ❸

Scene-Level, Expression-Base:

Partial Factual 3D Visual Grounding

If you can, please share the ID of the object that fits the description "the kitchen counter that is closer to the shelf".

No.

If you can, please share the ID of the object that fits the description "the pillow under the couch".

It is "lying on". <OBJ004>.❹

If you can, please share the ID of the object that fits the description "the mouse on the desk".

Yes. <OBJ005>.❺

Object-Level, Category-Base:

Hybrid Referring Object Classification

Is it correct that <OBJ006>❻ is a whiteboard; <OBJ007>❼ is a table; <OBJ008>❽ is a clock?

Yes; No, <OBJ007>❼ is a office chair; Yes.

Scene-Level, Expression-Base:

Faithful 3D Question Answering

What is small and blue and sits underneath the desk near the tall bookshelves in the room? If you can, answer the question using a word or a phrase. And provide all the IDs for objects related to the question and answer.

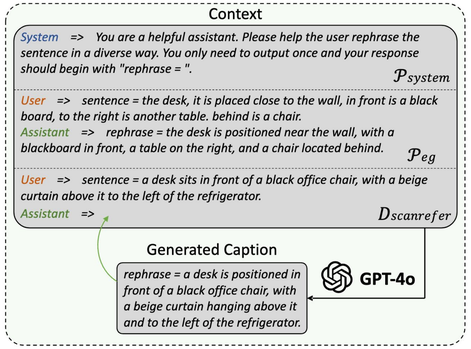
Trash can. <OBJ009>❾, <OBJ001>❶, <OBJ002>❷, and <OBJ010>.❿

What is on top of the bed? If you can, answer the question using a word or a phrase. And provide all the IDs for objects related to the question and answer.

No.

Design Overview:

- **Data Engine**:
  - Adversarial instruction-following data
  - Diverse instruction-following data
- **Model Improvement**:
  - Relation-Augmented Projector
  - Sequence Organization



Design Details:

[Data] Adversarial instruction-following data:

Mix negative and positive data samples with grounded instance to enhance the model's discriminative understanding

[Data] Diverse instruction-following data:

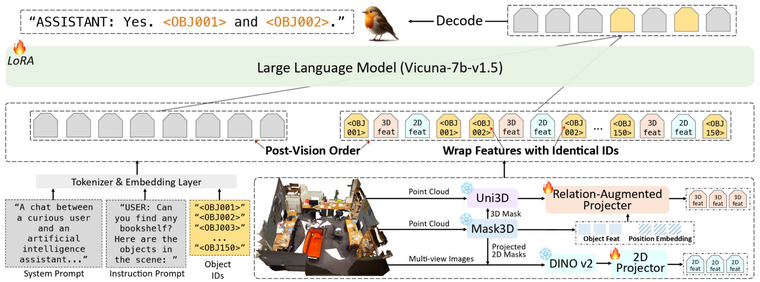
Use ChatGPT and its in-context learning ability to diversify the instruction styles to enhance model's generalization.

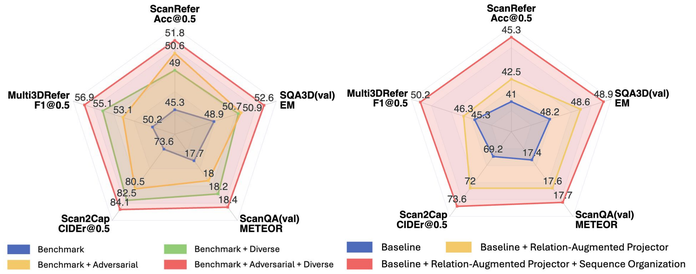
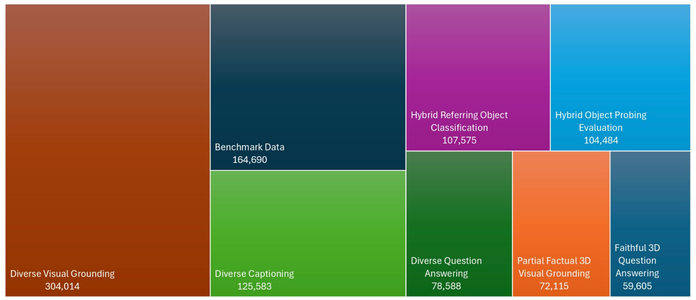
[Model] Relation-Augmented Projector:

Fuse scene-level, position embeddings, and unified object features to obtain relation-aware object-centric features.

[Model]Relation-Augmented Projector:

Use dual object-id and post-vision order to enhance ID-Feature connection.



Experiment

Model	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA(val)		SQA3D(val)		SQA3D(test)	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	M	R	EM	EM-R	EM	EM-R
Task-Specific Training												
ScanRefer	37.3	24.3	-	-	-	-	-	-	-	-	-	-
EDA	53.8	41.7	-	-	-	-	-	-	-	-	-	-
Concretenet	50.6	46.5	-	-	-	-	-	-	-	-	-	-
M3DRef-CLIP	51.9	44.7	42.8	38.4	-	-	-	-	-	-	-	-
Scan2Cap	-	-	-	-	23.3	39.1	-	-	-	-	-	-
Vote2Cap-DETR++	-	-	-	-	37.1	67.6	-	-	-	-	-	-
ScanQA	-	-	-	-	-	-	13.1	33.3	-	-	-	-
SQA3D	-	-	-	-	-	-	-	-	-	-	46.6	-
Joint Training												
D3Net	-	37.9	-	32.2	35.7	62.6	-	-	-	-	-	-
3DJCG	49.6	37.3	-	26.6	31.0	49.5	-	-	-	-	-	-
3D-VLP	51.4	39.5	-	-	32.3	54.9	-	-	-	-	-	-
3D-VisTA	50.6	45.8	-	-	34.0	66.9	13.9	35.7	-	-	48.5	-
PQ3D	-	51.2	-	50.1	36.0	80.3	-	-	-	-	47.1	-
3DLLM	30.3	-	-	-	-	-	14.5	35.7	-	-	-	-
Oryx	-	-	-	-	-	-	15.0	37.3	-	-	-	-
LL3DA	-	-	-	-	36.8	65.2	15.9	37.3	-	-	-	-
LEO	-	-	-	-	38.2	72.4	20.0	49.2	-	-	50.0	52.4
Scene-LLM	-	-	-	-	-	-	16.6	40.0	-	-	54.2	-
Chat-3D v2	35.9	30.4	-	-	15.5	28.1	16.1	40.1	-	-	-	-
Grounded-3DLLM	47.9	44.1	45.2	40.6	35.5	70.6	15.2	37.1	-	-	-	-
Chat-Scene	55.5	50.2	57.1	52.4	36.3	77.1	18.0	41.6	53.2	56.1	54.6	57.5
Robin3D (Ours)	**60.8**	**55.1**	**64.9**	**59.7**	**38.4**	**87.2**	**19.2**	**44.0**	**56.0**	**58.6**	**56.9**	**59.8**