

# How RL Enhances Spatial Understanding?

Zeyu Zhang



Homepage



3D-R1

Talk @ 3DCVer, Oct 9, 2025

# Some Quotes From The Bitter Lesson

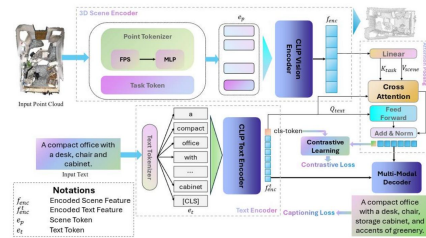
*“One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are **search and learning**.”*

*— Rich Sutton*

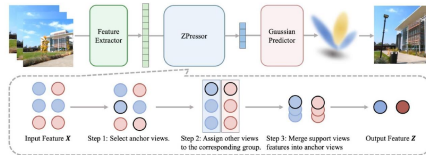


# From Specialized Methods to General Methods

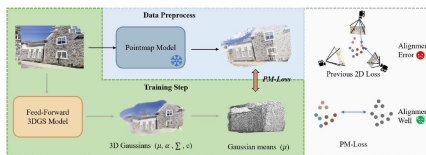
3D Dense Captioning



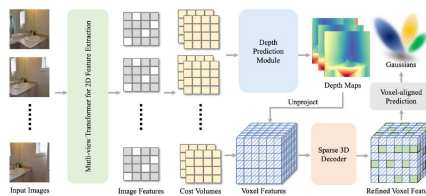
3D CoCa (2025)



ZPressor (NeurIPS 2025)



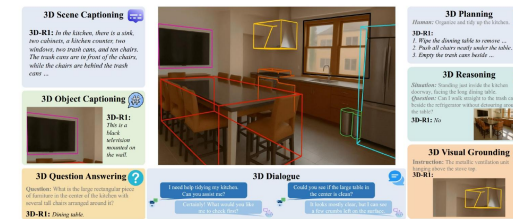
PM-Loss (2025)



VolSplat (2025)

Evolve

3D Foundation Model



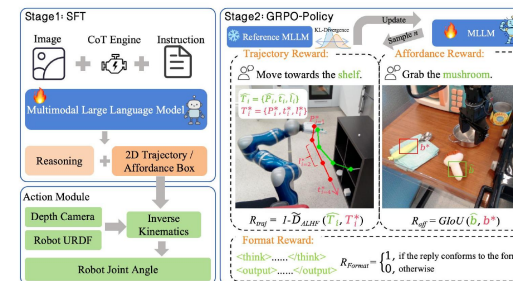
3D-R1 (2025)

Navigation Foundation Model



Nav-R1 (2025)

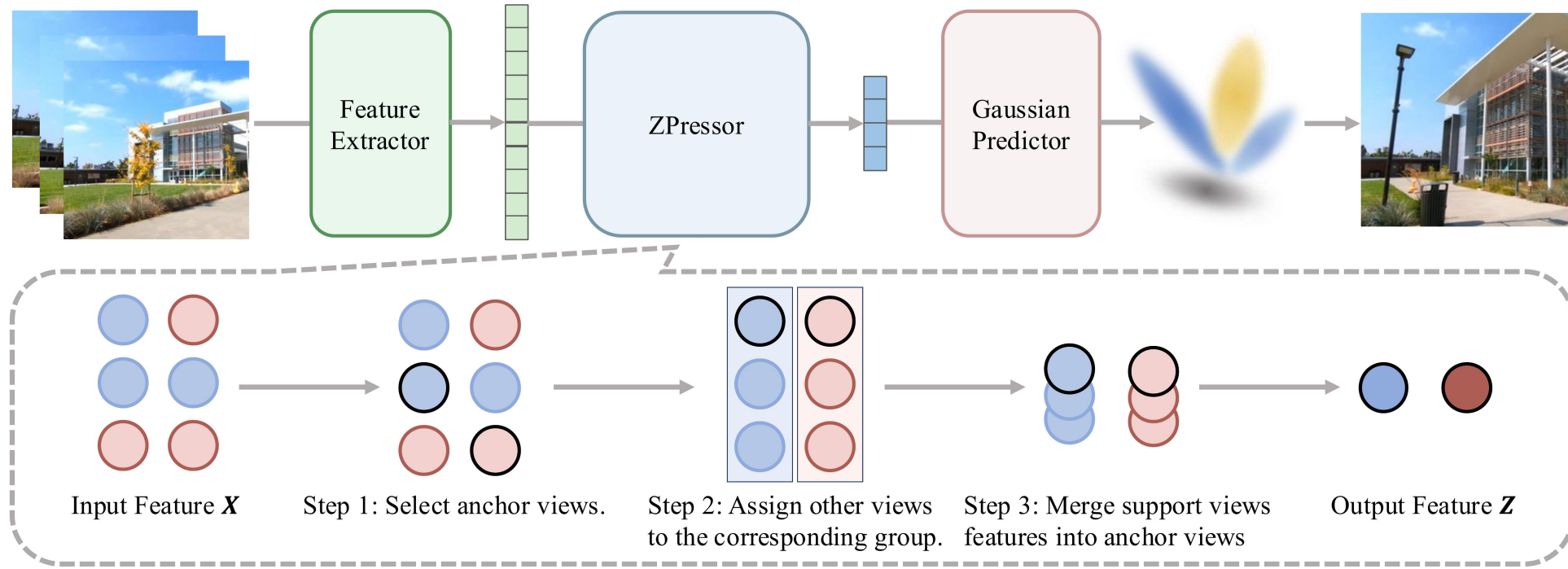
Manipulation Foundation Model



VLA-R1 (2025)

3D models represent a significant shift from specialized methods, which are designed for specific tasks, to foundation models that can handle a wide range of tasks.

# Efficient 3D Reconstruction: ZPressor (2025)



**ZPressor** is an efficient feed-forward 3D scene reconstruction model with bottleneck-aware compression.

Weijie Wang, Yuedong Chen, Zeyu Zhang et al. *ZPressor: Bottleneck-Aware Compression for Scalable Feed-Forward 3DGS* (NeurIPS 2025)

# Results of ZPressor (2025)

## Visualization on DL3DV (36 Input Views)



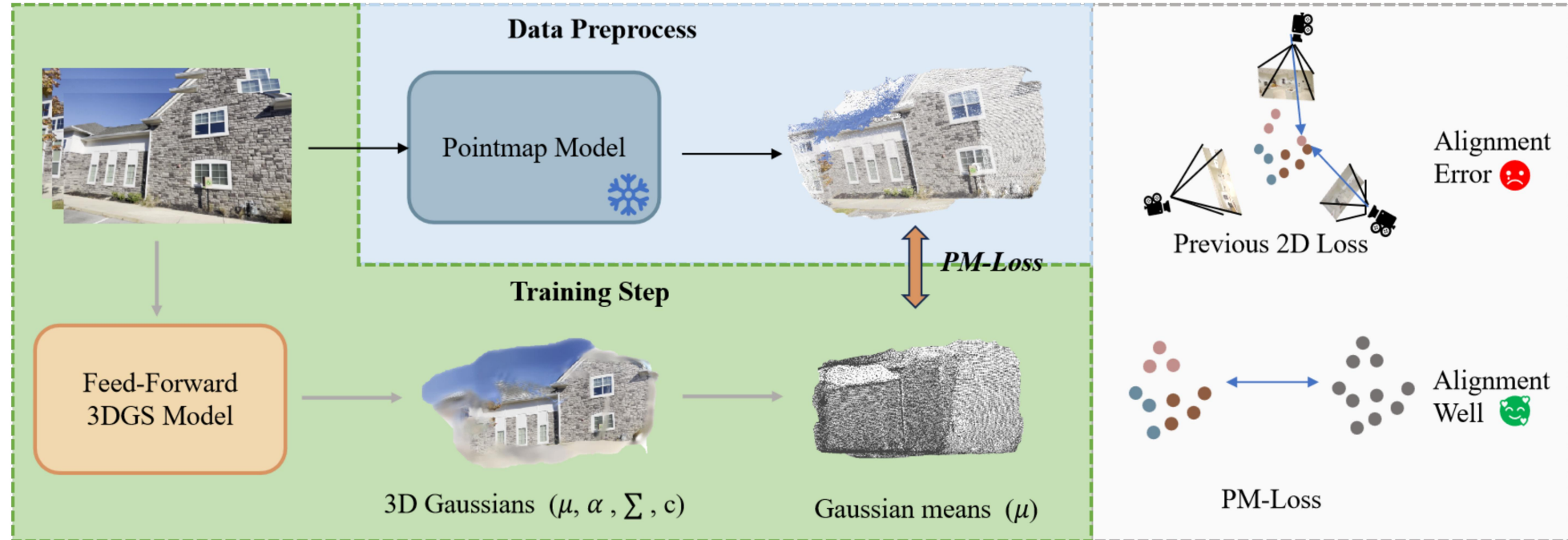
a62c330f5403e2e41a82a74c4e865b705c5706843b992fae2fe2e538b122d984



63798f5c6fbfcb4eb686268248b8ecbc8d87d920b2bcce967eeaedfd3b3b6d82



# Geometric Prior Matters: PM-Loss (2025)

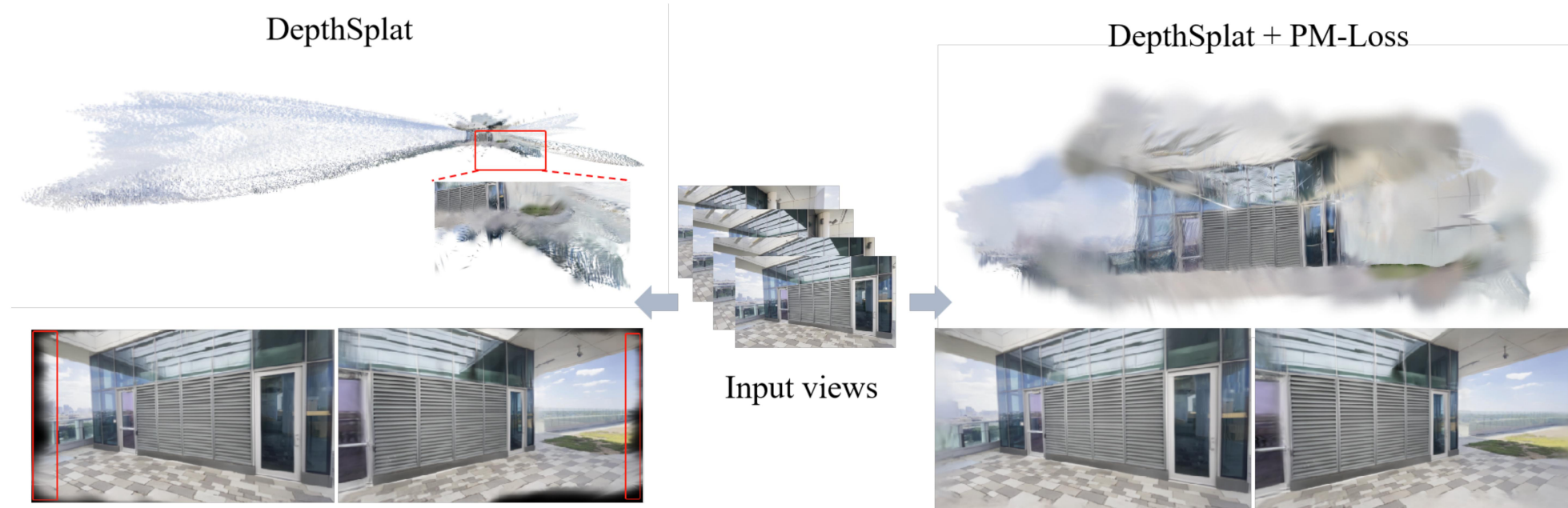


**PM-Loss** is a novel regularization loss based on a learned point map for feed-forward 3DGS, leading to smoother 3D geometry and better rendering.

Duochao Shi, Weijie Wang, Yuedong Chen, Zeyu Zhang et al. *Revisiting Depth Representations for Feed-Forward 3D Gaussian Splatting* (2025)



# Results of PM-Loss (2025)



However, depth discontinuities at object boundaries often lead to fragmented or sparse point clouds, degrading rendering quality—a well-known limitation of depth-based representations. To tackle this issue, we introduce **PM-Loss**, a novel regularization loss based on a pointmap predicted by a pre-trained transformer (Chamfer distance). Although the pointmap itself may be less accurate than the depth map, it effectively enforces geometric smoothness, especially around object boundaries.

# Results of PM-Loss (2025)

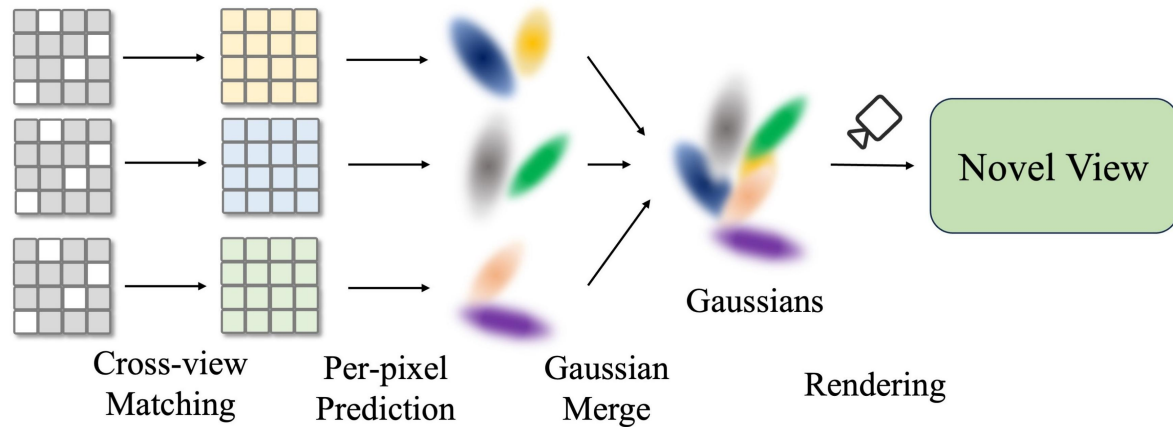
Depthsplat



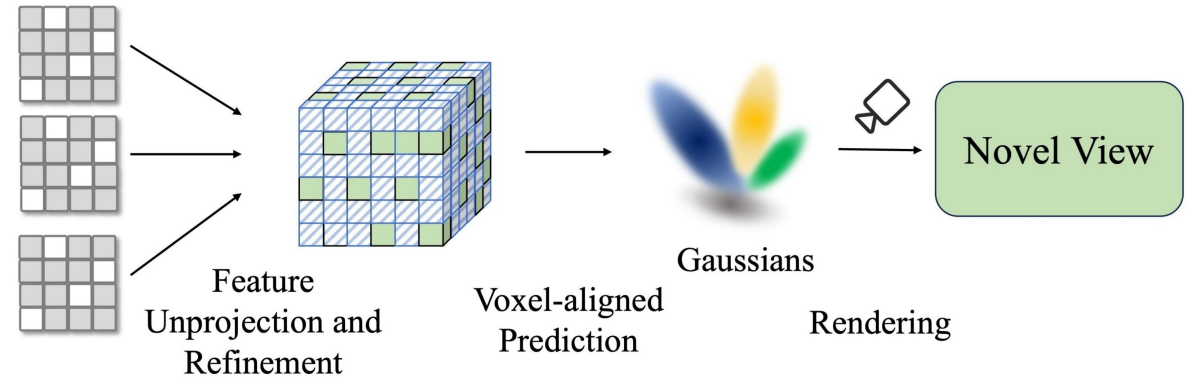
Depthsplat + PM-Loss



# Voxel-Aligned Matters: VolSplat (2025)



(a) Pixel-aligned Feed-forward 3DGS

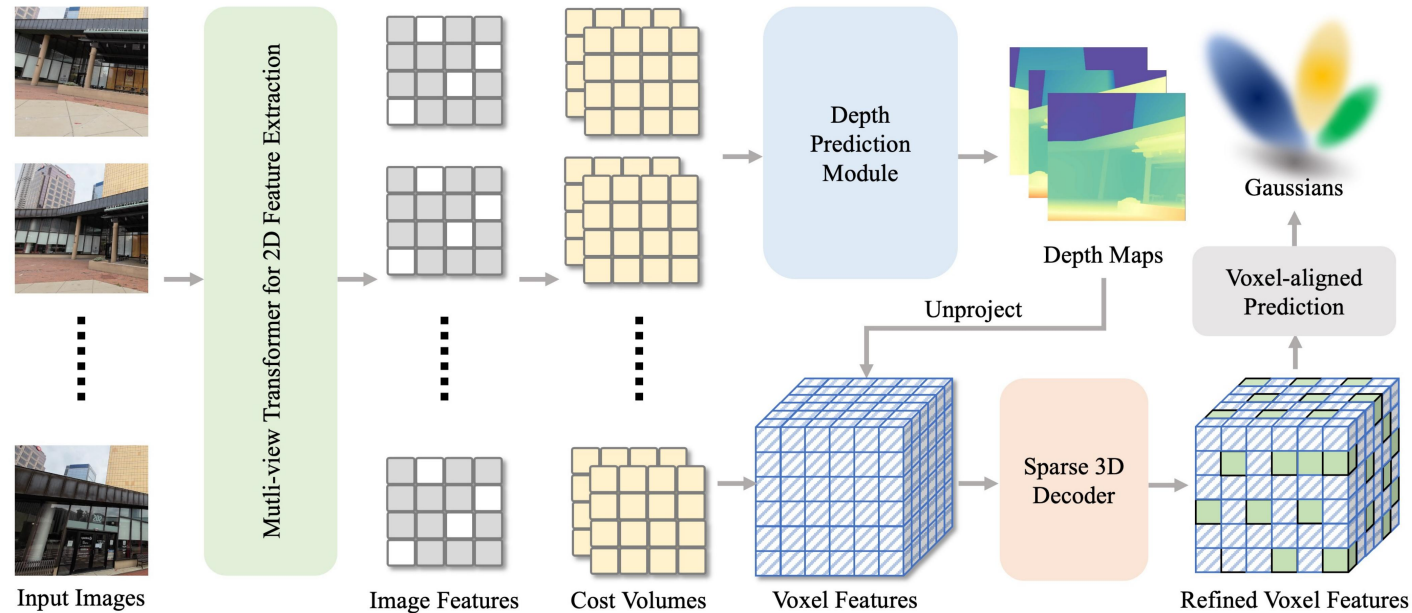


(b) Voxel-aligned Feed-forward 3DGS (Ours)

Pixel-aligned feed-forward 3DGS methods suffer from two primary limitations: 1) 2D feature matching struggles to effectively resolve the multi-view alignment problem, and 2) the Gaussian density is constrained and cannot be adaptively controlled according to scene complexity.

Weijie Wang, Yeqing Chen, Zeyu Zhang et al. *VolSplat: Rethinking Feed-Forward 3D Gaussian Splatting with Voxel-Aligned Prediction* (2025)

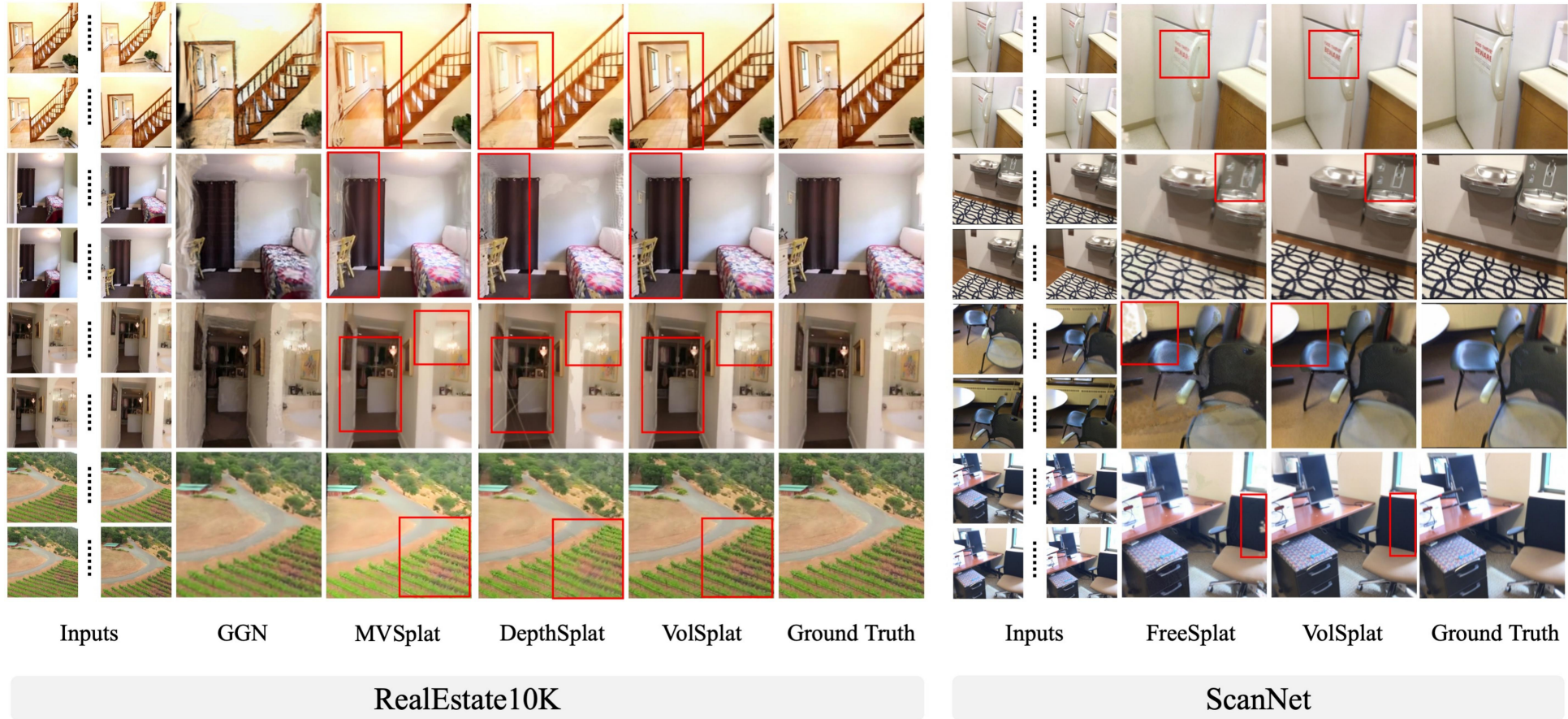
# Voxel-Aligned Matters: VolSplat (2025)



Given multi-view images as input, we first extract 2D features for each image using a Transformer-based network and construct per-view cost volumes with plane sweeping. Depth Prediction Module then estimates a depth map for each view, which is used to unproject the 2D features into 3D space to form a voxel feature grid. Subsequently, we employ a sparse 3D decoder to refine these features in 3D space and predict the parameters of a 3D Gaussian for each occupied voxel. Finally, novel views are rendered from the predicted 3D Gaussians.

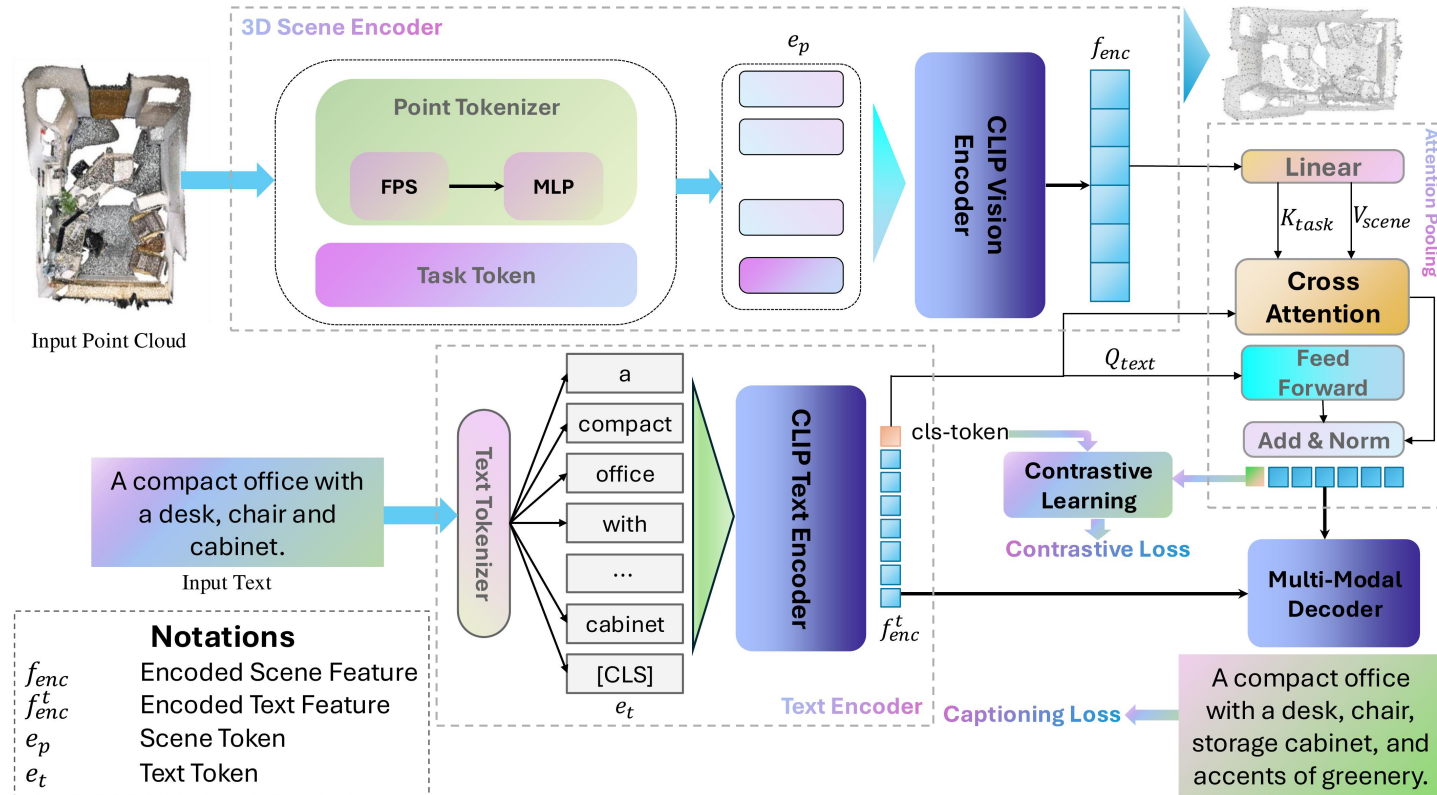


# Results: VolSplat (2025)



The results on the left are from RealEstate10K, and the results on the right are from ScanNet.

# 3D Representation Learning: 3D CoCa (2025)



**3D CoCa** leverages 3D multimodal representation learning to tackle scene understanding through large-scale contrastive pretraining.

Ting Huang, Zeyu Zhang et al. *3D CoCa: Contrastive Learners are 3D Captioners* (2025)



# Results of 3D CoCa (2025)



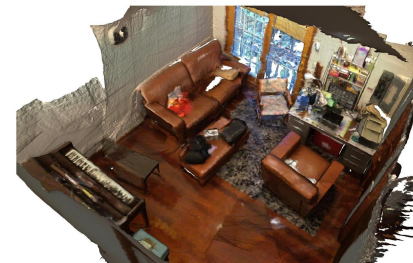
**Vote2Cap-DETR++:** A room with a large wooden dining table and multiple chairs.



**Vote2Cap-DETR++:** A room with several rectangular tables and various items on them.



**Vote2Cap-DETR++:** A room with a few tables, cluttered items on top, and several chairs nearby.



**Vote2Cap-DETR++:** A living room with two sofas and a small side table.

**Ours:** A spacious dining area featuring a long wooden table surrounded by several chairs, with a painting on the wall.

**Ours:** An open space designed for work or study, with multiple tables and chairs arranged to form a collective workspace, and ample floor space around them.

**Ours:** A messy workspace, with various documents or tools scattered on the tables and a few chairs and electronic devices placed around.

**Ours:** A cozy lounge area featuring two brown sofas and a coffee table, with a rug on the floor and some decorative items nearby.

**GT:** In a bright dining room, a long wooden table is flanked by neatly arranged chairs. Light filters in through the window, and a decorative painting adorns the wall.

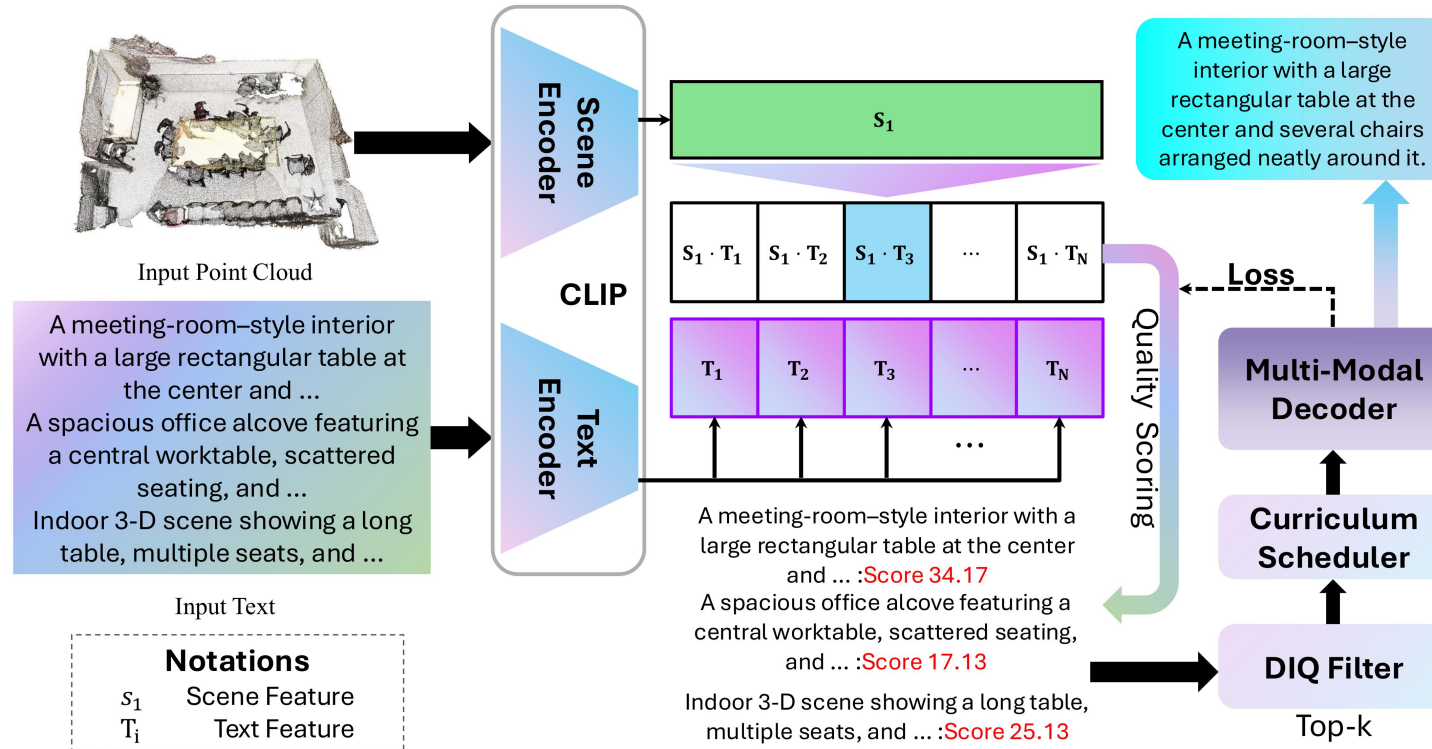
**GT:** A spacious indoor setting with several parallel tables and chairs, offering walking and working areas on all sides. The layout resembles a classroom.

**GT:** An office area, where tabletops are covered with multiple items and documents. Chairs and computer accessories are set around the room.

**GT:** A comfortable living room setup with two leather sofas, a small coffee table, and a rug on the floor. The corner have a musical instrument and ornaments.

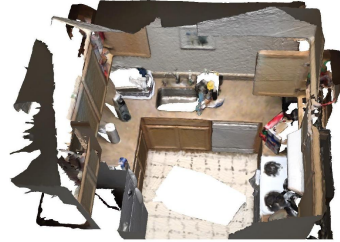
A visual comparison on the ScanRefer dataset showcasing indoor scenes described by Vote2Cap-DETR++, 3D CoCa (Ours), and the ground truth (GT).

# 3D Data-Centric Learning: DC-Scene (2025)



Point clouds and captions are encoded, scored with 3D CLIP, and filtered by the Dual-Indicator Quality (DIQ) module to select top- $k$  candidates. A Curriculum Scheduler trains the Multi-Modal Decoder, while a feedback loop updates CLIP scores with caption loss, forming a data-centric learning cycle.

# Results of DC-Scene (2025)



**Baseline(full data):** a small kitchen with cabinets, a sink, and a white appliance on the right.

**DC-Scene(Top-75%):** a kitchen where wooden cabinets frame a metal sink beneath a wall picture, while a white washer-dryer sits to the right of the light-tiled floor that opens into a carpeted hallway.

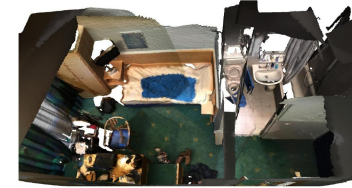
**GT:** A compact galley kitchen with wooden upper and lower cabinets, a stainless-steel sink centered along the back work-top, and a white washer-dryer unit standing on the right side of the tiled floor.



**Baseline(full data):** a bedroom with two beds, a desk and some clutter on the floor.

**DC-Scene(Top-75%):** a cramped dormitory room with parallel single beds, a back-wall desk piled high with textbooks, and clothing and an open teal suitcase strewn over the dark-blue carpet.

**GT:** A student dorm room containing two single beds along opposite walls, a wooden study desk cluttered with books and electronics, and clothes plus an open turquoise suitcase scattered across the blue carpet.



**Baseline(full data):** a small bedroom with a bed, green carpet, and a bathroom to the right.

**DC-Scene(Top-75%):** a bedroom featuring a bed topped with a bright blue blanket, a wicker hamper near the footboard, and an ensuite bathroom on the right where a basin and shower are visible through the open doorway.

**GT:** A small bedroom with a light-wood single bed covered by a blue throw, green carpet flooring, a wicker laundry basket at the foot of the bed, and an adjoining bathroom on the right showing a white sink and shower stall.

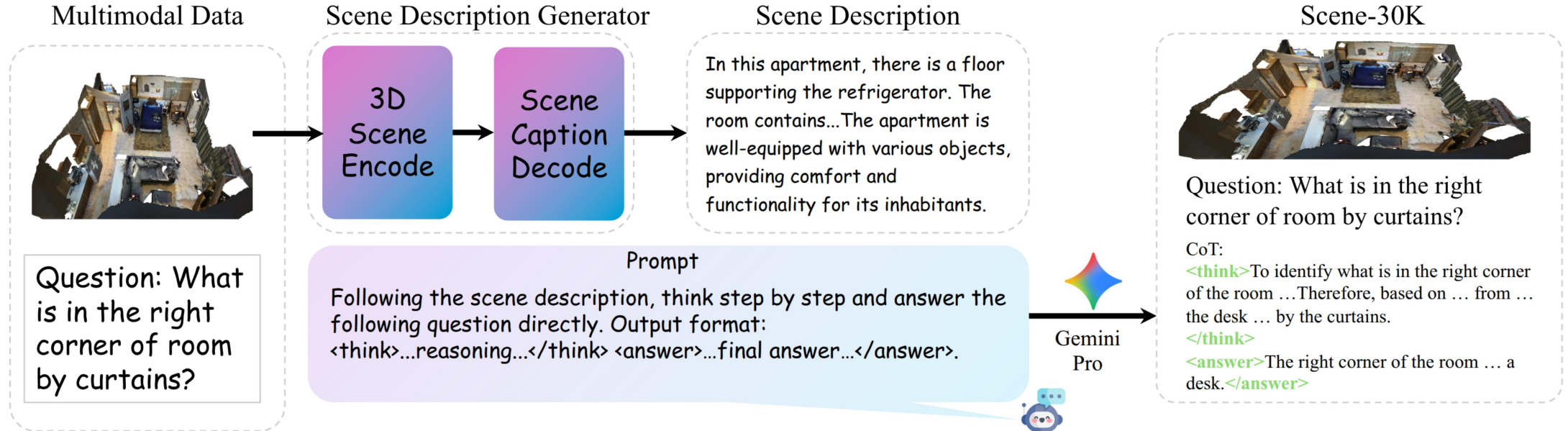
For three validation scenes from the ScanRefer dataset, we present the rendered point cloud mesh (top row), followed by captions generated by three sources: the full-data baseline model (in pink), our **DC-Scene** model trained on the top-75% DIQ samples (in red), and the human-annotated ground truth (in green).



# What's next for 3D foundation models?

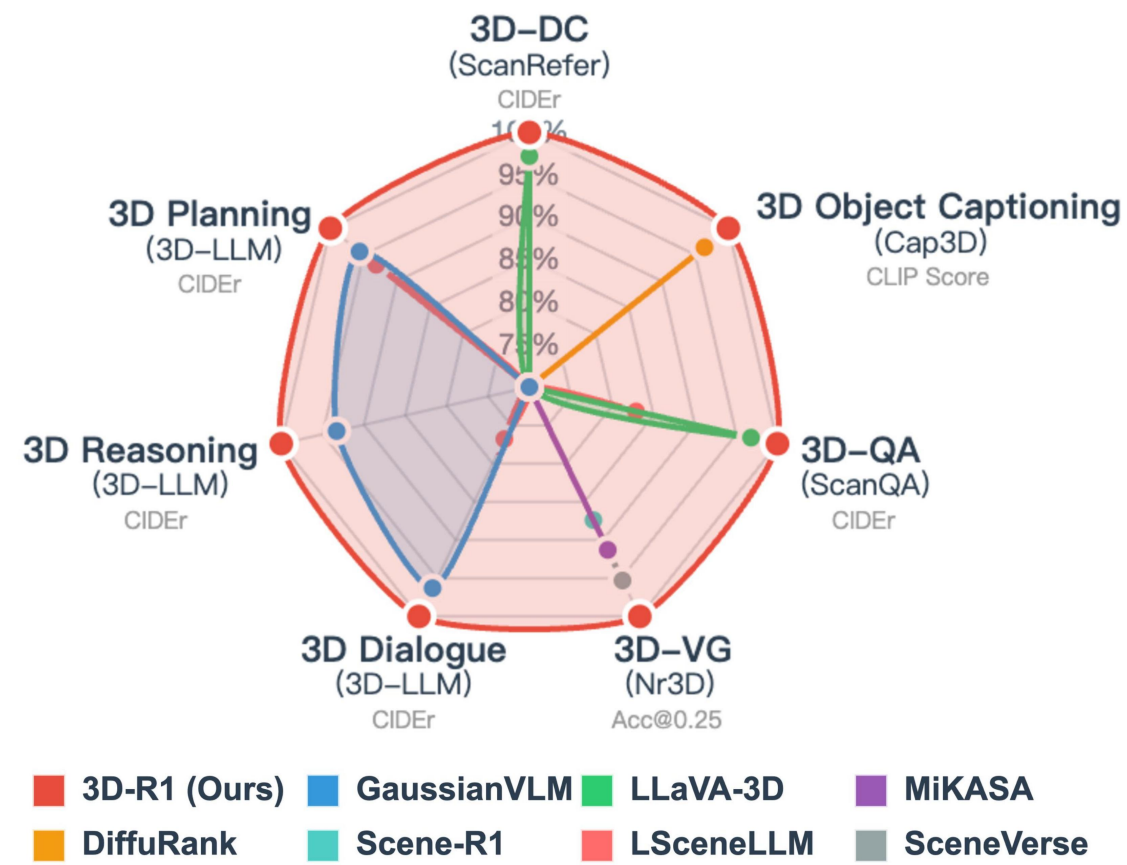
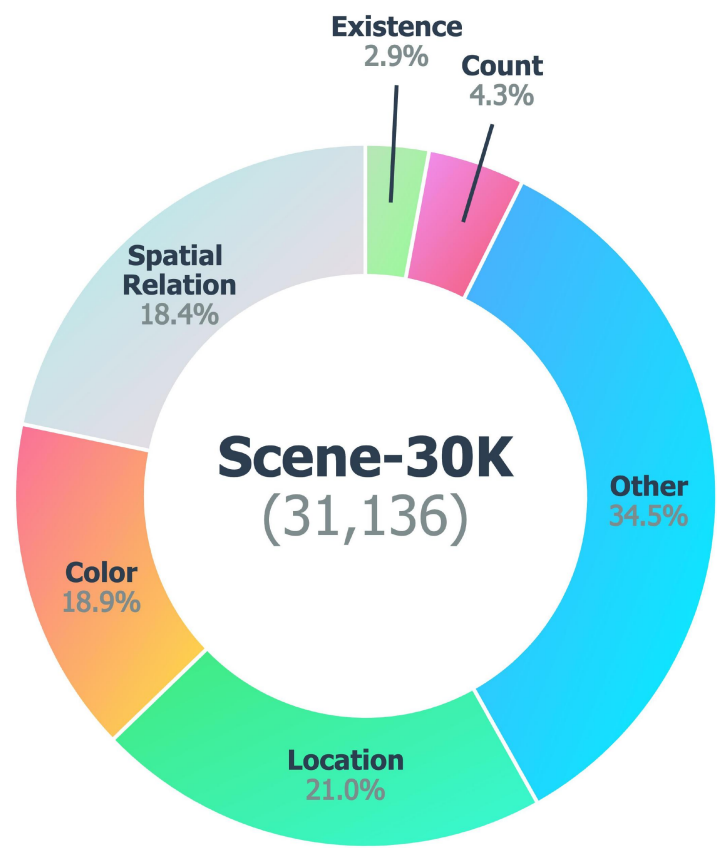
- How can we achieve zero-shot generalizability across different tasks given the domain knowledge gap between them?
- How can we adjust a foundation model after conventional supervised post-training when the outcomes are unsatisfactory on specific tasks?
- And most importantly, how can we ground our foundation model in the physical world?

# Synthetic Data Helps Enhance Generalizability: 3D-R1



**CoT Data Engine.** The point cloud of a scene is first sent to scene description generator to get a description of the scene. Then based on the description, we apply Gemini-Pro to synthetic CoT data.

# Scene-30K in 3D-R1



**Scene-30K** contains diverse scene categories and question types. **3D-R1** demonstrates strong performance across various tasks.

# Generalizability: 3D-R1

## 3D Scene Captioning

**3D-R1:** In the kitchen, there is a sink, two cabinets, a kitchen counter, two windows, two trash cans, and ten chairs. The trash cans are in front of the chairs, while the chairs are behind the trash cans ...

## 3D Object Captioning

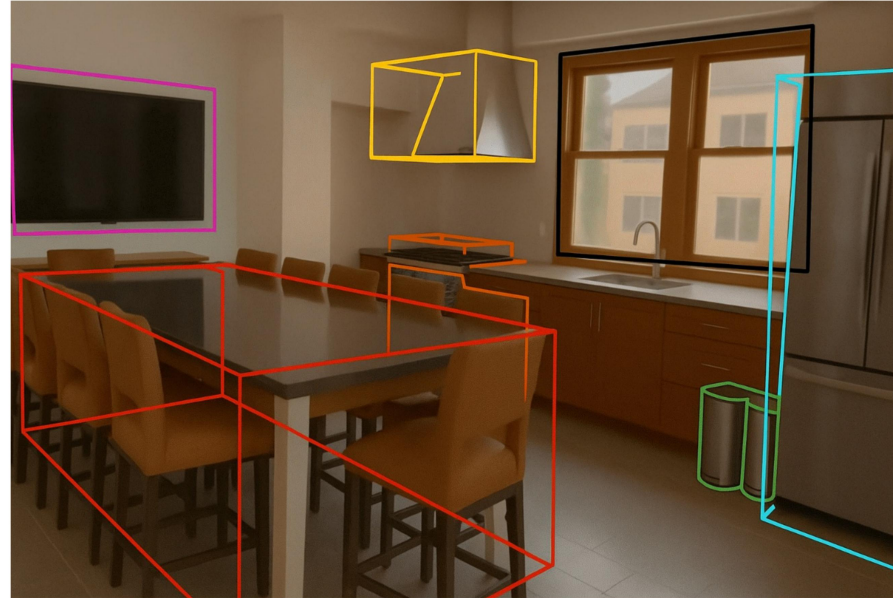


**3D-R1:**  
This is a black television mounted on the wall.

## 3D Question Answering

**Question:** What is the large rectangular piece of furniture in the center of the kitchen with several tall chairs arranged around it?

**3D-R1:** Dining table.



## 3D Dialogue

I need help tidying my kitchen.  
Can you assist me?

Certainly! What would you like me to check first?

Could you see if the large table in the center is clean?

It looks mostly clear, but I can see a few crumbs left on the surface.

## 3D Planning

**Human:** Organize and tidy up the kitchen.

**3D-R1:**

1. Wipe the dinning table to remove ...
2. Push all chairs neatly under the table...
3. Empty the trash cans beside ...

## 3D Reasoning

**Situation:** Standing just inside the kitchen doorway, facing the long dining table.

**Question:** Can I walk straight to the trash cans beside the refrigerator without detouring around the table?

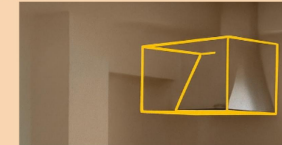
**3D-R1:** No



## 3D Visual Grounding

**Instruction:** The metallic ventilation unit hanging above the stove top.

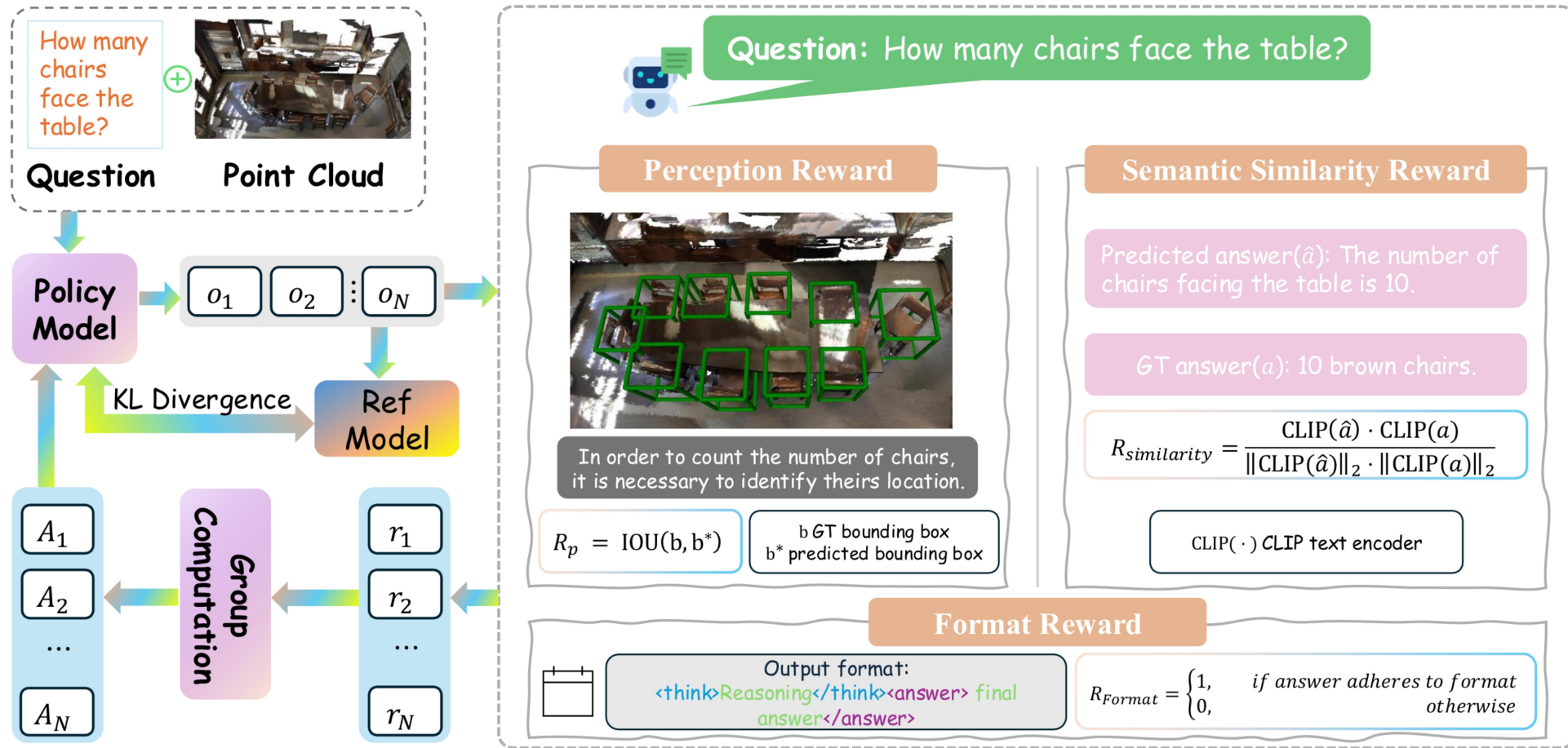
**3D-R1:**



**3D-R1** is a generalist model capable of handling various downstream tasks and applications in a zero-shot manner with incredible generalizability, significantly reducing the need for expensive adaptation.

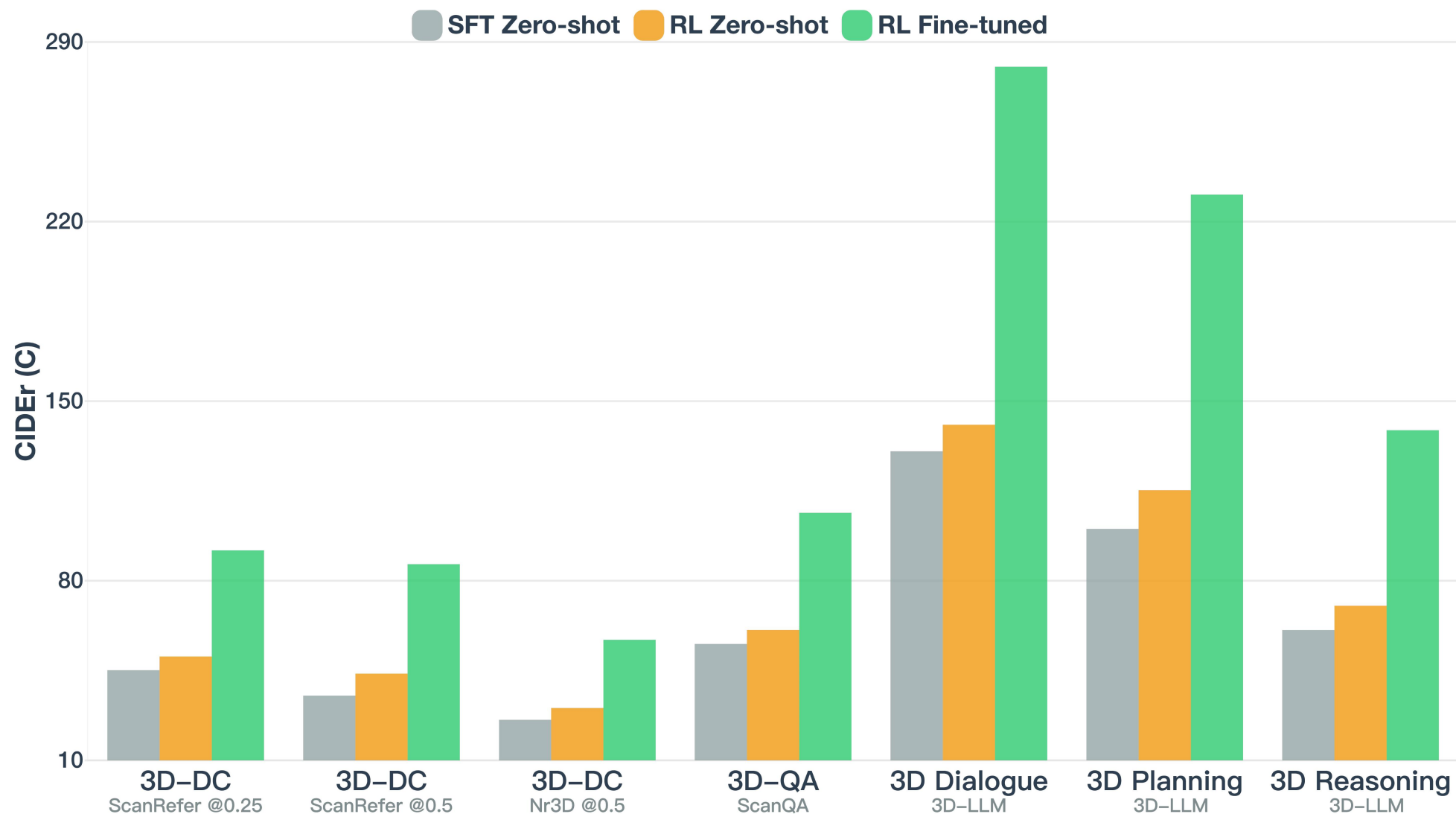


# Adjust Output: Reinforcement Learning with Verifiable Rewards (RLVR)



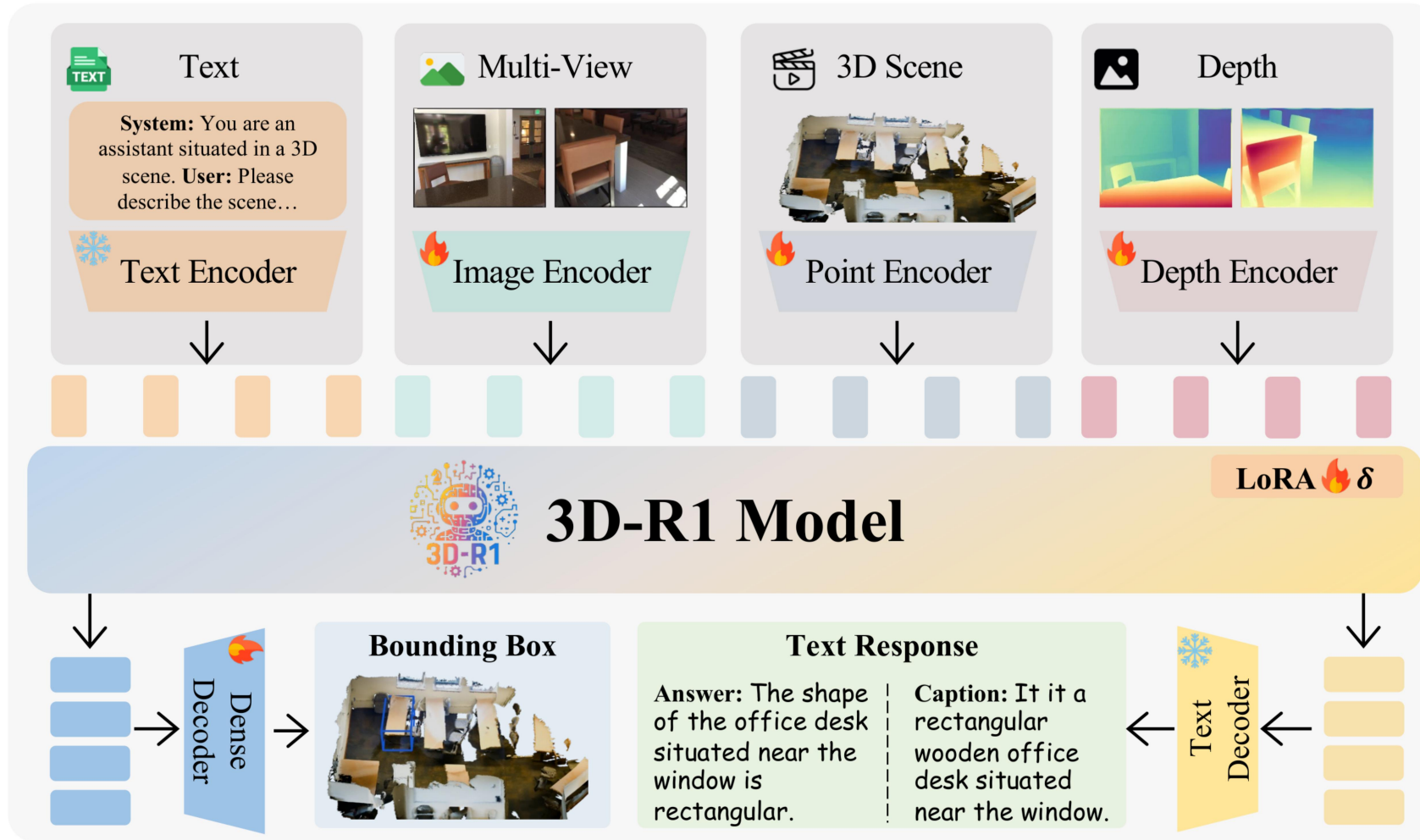
The policy model generates  $N$  outputs from a point cloud and question. Then perception IoU, semantic CLIP-similarity, and format-adherence rewards are computed, grouped, and combined with a KL term to a frozen reference model to update the policy.

# Enhanced Reasoning: 3D-R1



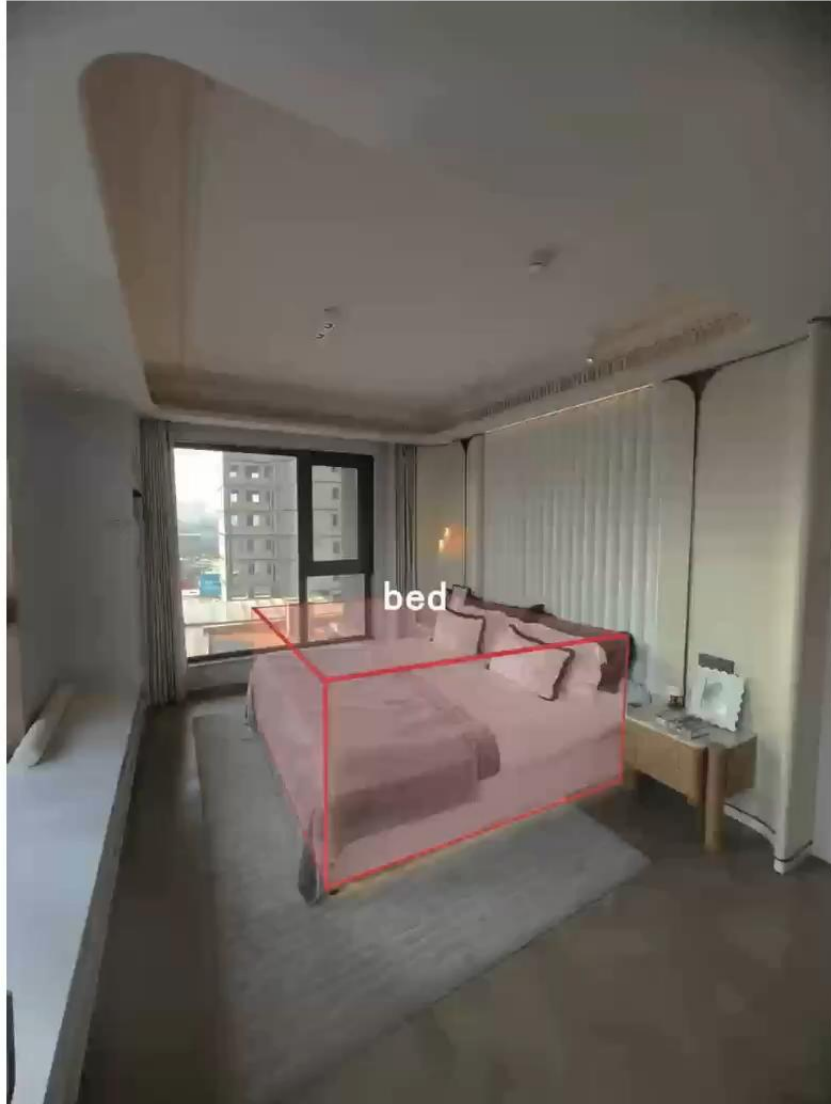
**3D-R1** exhibits remarkable generalizability with enhanced reasoning capabilities.

# Foundation Model: 3D-R1



**3D-R1** is an open-source generalist model that enhances the reasoning of 3D VLMs for unified scene understanding.

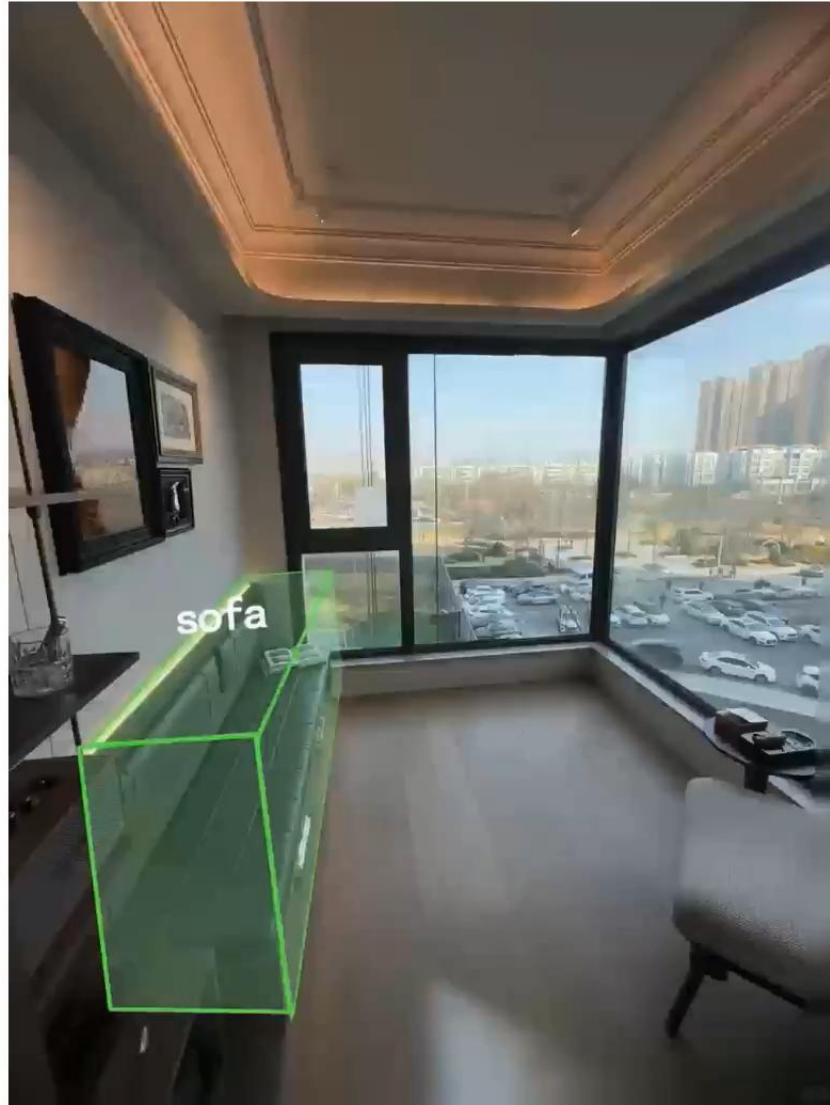
# 3D Scene Dense Captioning (3D-DC)



**3D-DC**



# 3D Object Captioning



**3D Object Captioning**

# 3D Visual Grounding (3D-VG)



**3D-VG**

# 3D Question Answering (3D-QA)



**3D-QA**

# 3D Dialogue



**3D Dialogue**



# 3D Reasoning



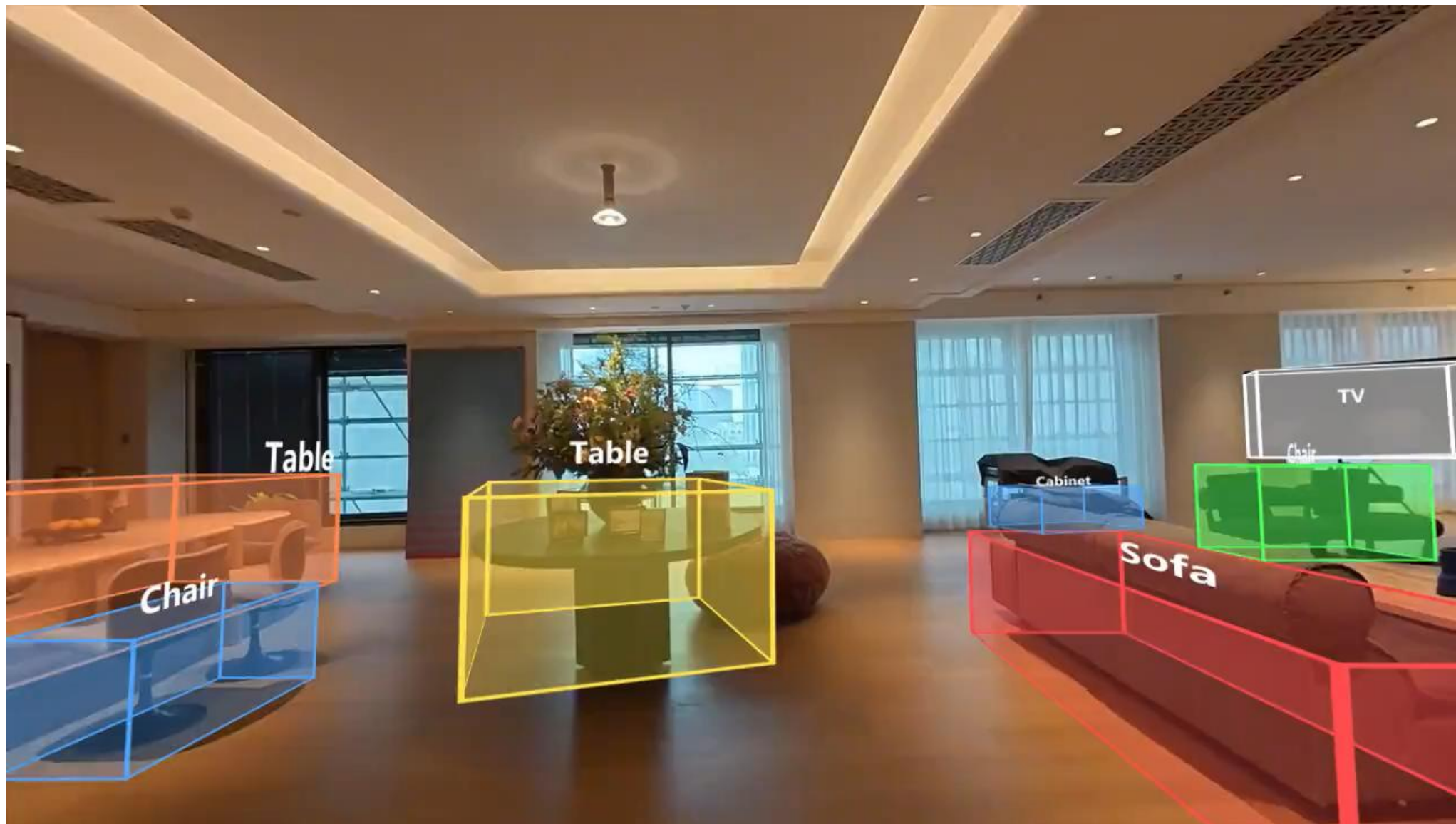
**3D Reasoning**

# 3D Planning



**3D Planning**

# Zero-Shot Results



# System and Memory: Nav-R1

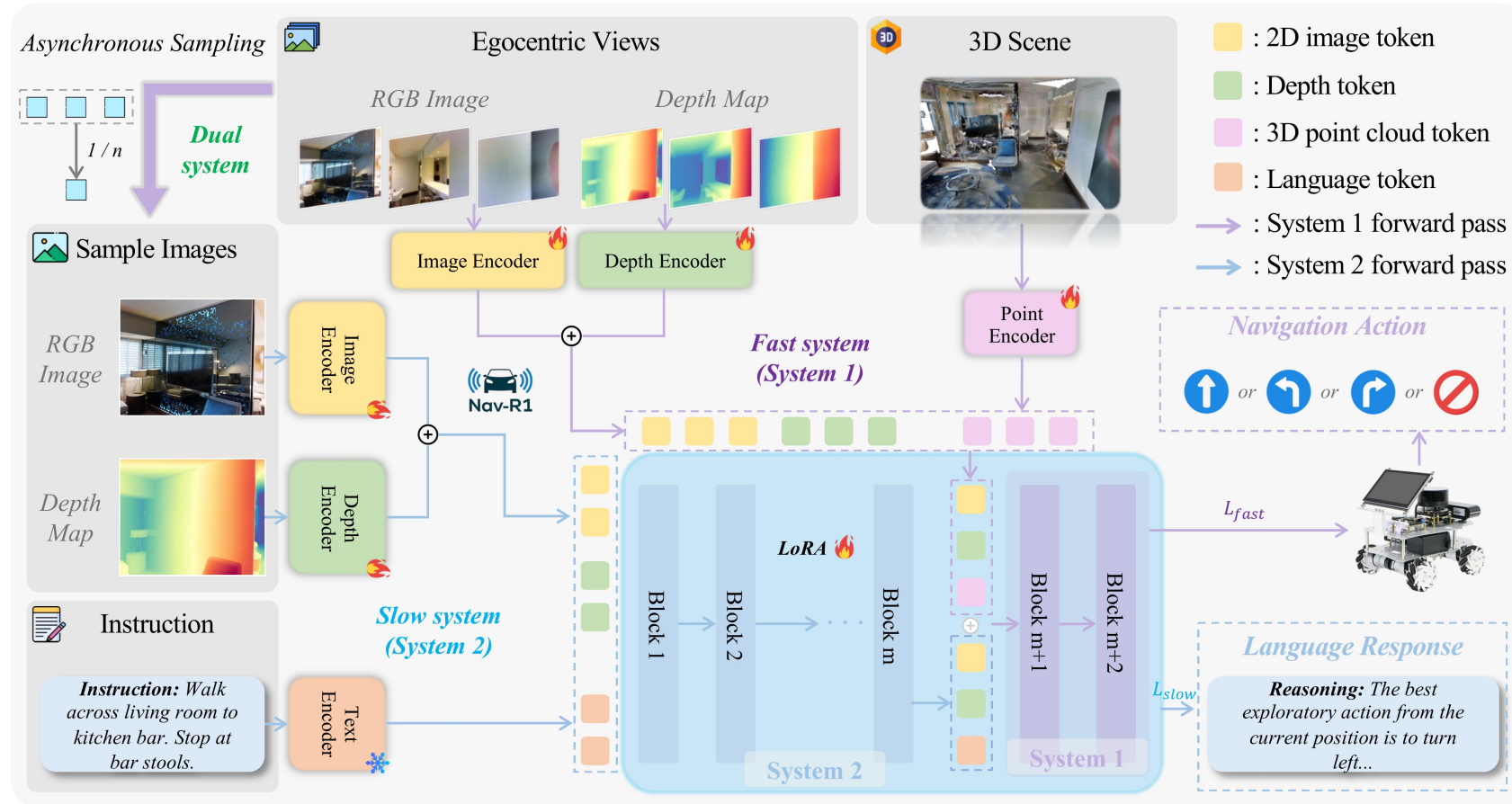
What if we ground a 3D foundation model in embodied scenes? How can its reasoning approach human-level intelligence? This is inspired by psychology.

*“The division of labor between System 1 (fast) and System 2 (slow) is highly efficient: it minimizes effort and optimizes performance.”*

*— Daniel Kahneman (Nobel Prize in Economics)*

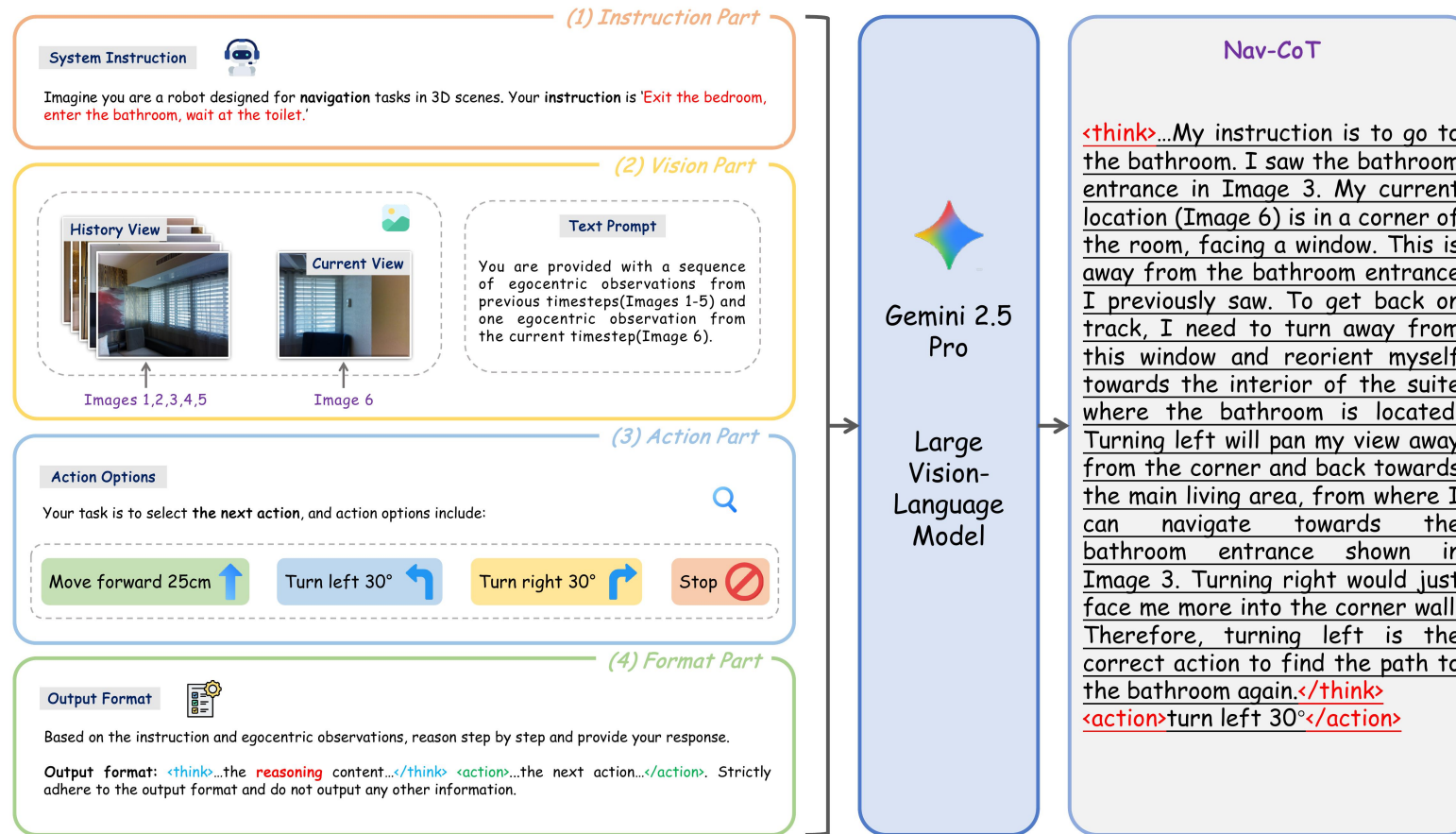


# Fast-in-Slow: Nav-R1



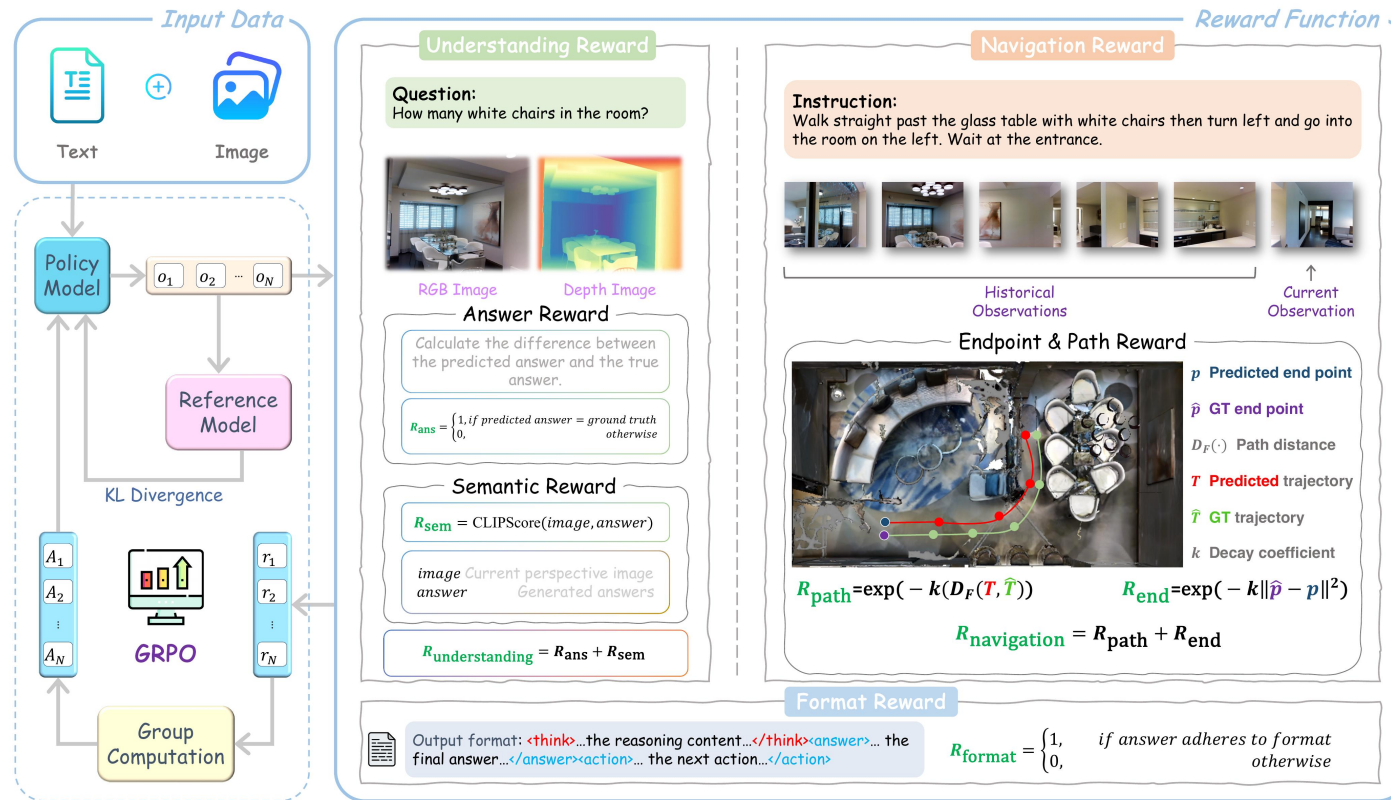
**Nav-R1** features a Fast-in-Slow design that ensures rapid decision-making within long-horizon planning..

# Synthetic Data: Nav-CoT-110K



We construct the **Nav-CoT-110K** dataset by defining navigation instructions, integrating egocentric visual inputs, providing action options and specifying the output format. These components are fed into Gemini 2.5 Pro, which generates step-by-step reasoning and action decisions aligned with navigation goals.

# Adjust Output: RLVR



**The pipeline of RL Policy.** The policy model generates  $N$  outputs from text-image input. Then understanding reward (answer correctness and semantic alignment), navigation reward (path fidelity and endpoint accuracy), and format reward (structure adherence) are computed, grouped, and combined with a KL term to a frozen reference model to update the policy.



# Navigation Foundation Model: Nav-R1



**Nav-R1** is an embodied foundation model that integrates dialogue, reasoning, planning, and navigation capabilities to enable intelligent interaction and task execution in 3D environments.

# Results: Nav-R1 but Reduces Forgetting

TABLE IV  
EMBODIED DIALOGUE AND PLANNING RESULTS ON 3D-LLM [18]. EMBODIED REASONING RESULTS ON SQA3D [33].

| Method               | Embodied dialogue |              |              |              | Embodied planning |              |              |              | Embodied reasoning |              |              |              |
|----------------------|-------------------|--------------|--------------|--------------|-------------------|--------------|--------------|--------------|--------------------|--------------|--------------|--------------|
|                      | C↑                | B-4↑         | M↑           | R↑           | C↑                | B-4↑         | M↑           | R↑           | C↑                 | B-4↑         | M↑           | R↑           |
| LL3DA [10]           | 190.01            | 23.95        | 23.50        | 40.61        | 128.80            | 12.95        | 17.05        | 39.25        | -                  | -            | -            | -            |
| Spatial 3D-LLM [46]  | -                 | -            | -            | -            | 195.92            | 14.65        | 18.95        | 36.93        | -                  | -            | -            | -            |
| LSceneLLM [56]       | 104.98            | -            | 21.26        | 36.00        | 214.63            | -            | 21.05        | 47.05        | -                  | -            | -            | -            |
| LEO [22]             | -                 | -            | -            | -            | -                 | -            | -            | -            | 124.70             | 9.40         | 25.50        | 48.40        |
| 3D-R1 [23]           | 280.34            | <b>39.45</b> | 66.89        | <b>55.34</b> | 230.50            | 25.45        | <b>48.34</b> | 55.67        | 138.67             | <b>23.56</b> | 35.45        | <b>60.02</b> |
| <b>Nav-R1 (Ours)</b> | <b>281.20</b>     | 39.34        | <b>67.53</b> | 55.12        | <b>230.52</b>     | <b>25.98</b> | 47.11        | <b>56.23</b> | <b>139.98</b>      | 23.20        | <b>36.15</b> | 59.50        |

The results show that our parameter-efficient tuning during hierarchical post-training effectively reduces forgetting when adapting the foundation model from understanding to navigation, while the model still maintains comparable performance on embodied scene understanding.



# Results: Nav-R1

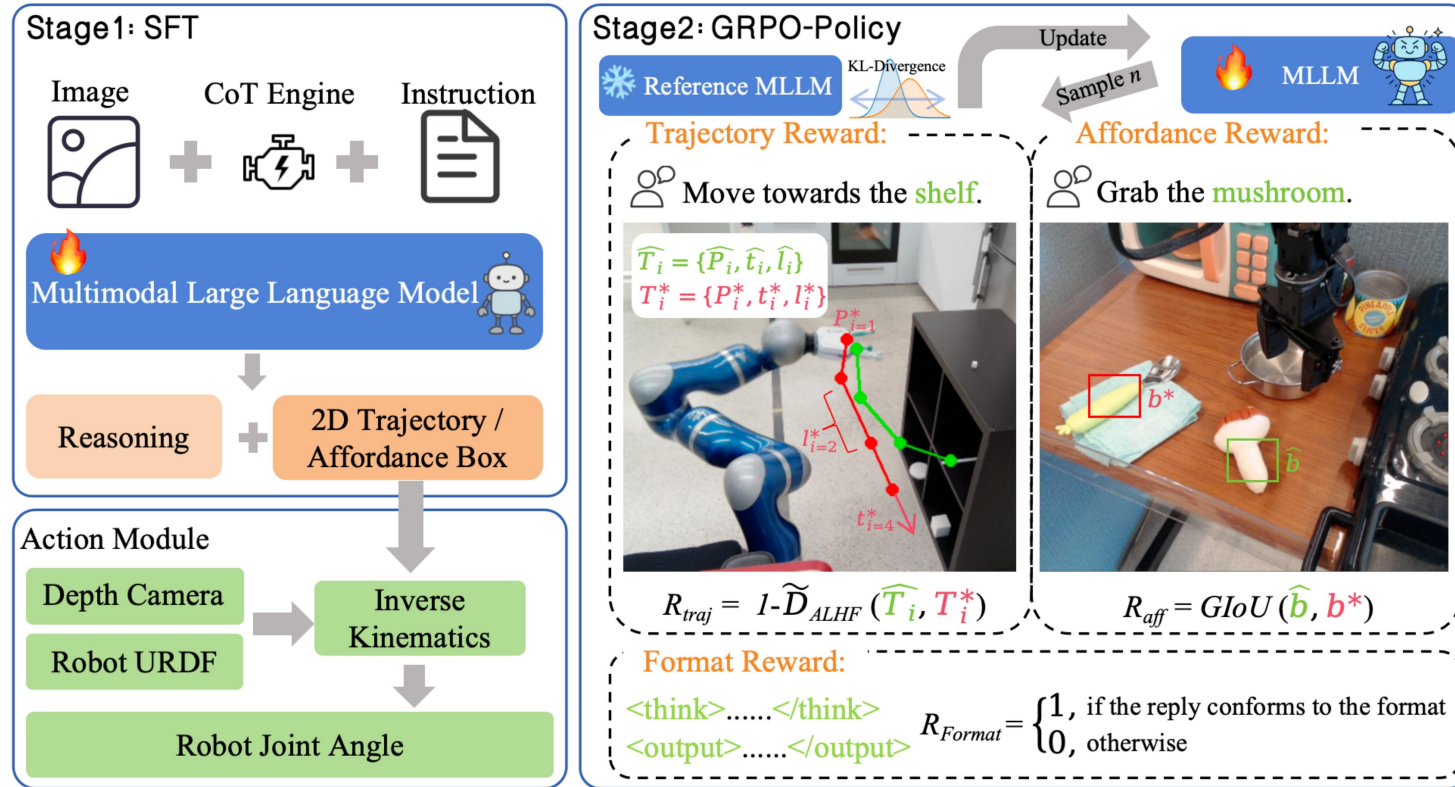


## Nav-R1: Reasoning and Navigation in Embodied Scenes

Qingxiang Liu, Ting Huang, Zeyu Zhang, Hao Tang



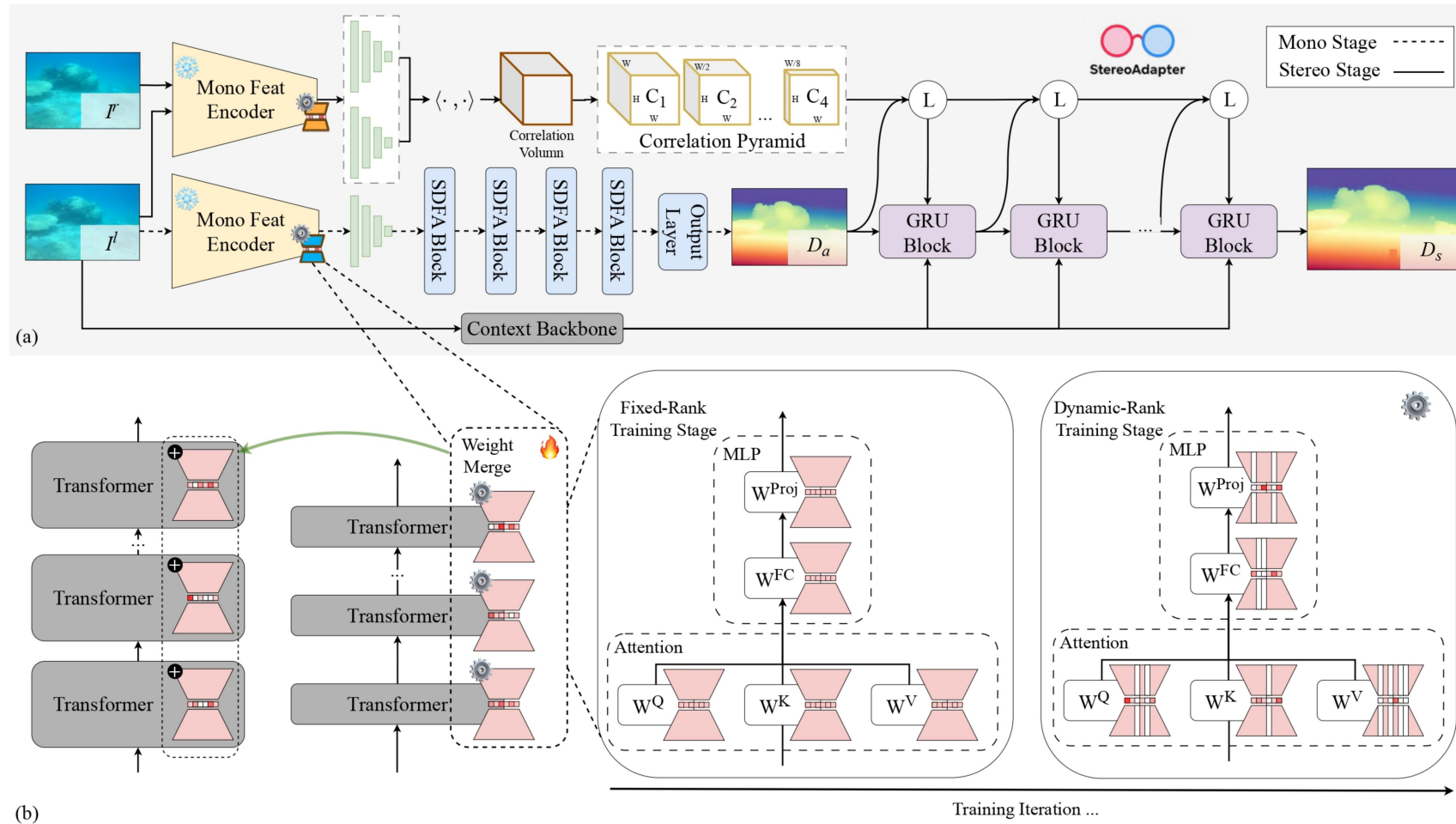
# Similar Idea for Arm Robots: VLA-R1



Training has two stages: **Stage 1** uses SFT with CoT supervision to learn reasoning over images and instructions; **Stage 2** refines reasoning and actions via RL with verifiable rewards (GRPO). **During inference**, a control stack converts outputs into joint-level robot commands.

## VLA-R1: Enhancing Reasoning in Vision-Language-Action Models

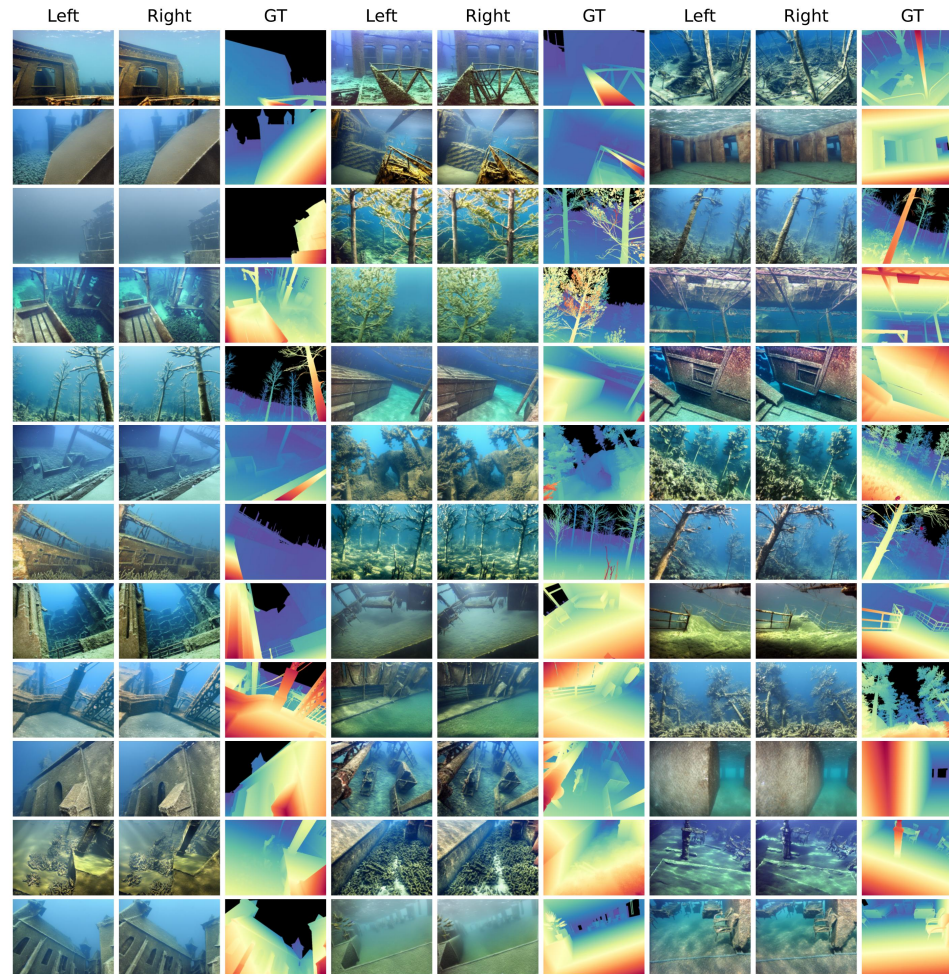
# Bridging the Domain Gap in Post-Training: StereoAdapter



**StereoAdapter** is a self-supervised adaptive model that allows robust underwater depth estimation.



# Synthetic Data: UW-StereoDepth-40K



**Data synthesis.** Unreal Engine 5 rendering for UW-StereoDepth-40K dataset.

# Results: StereoAdapter

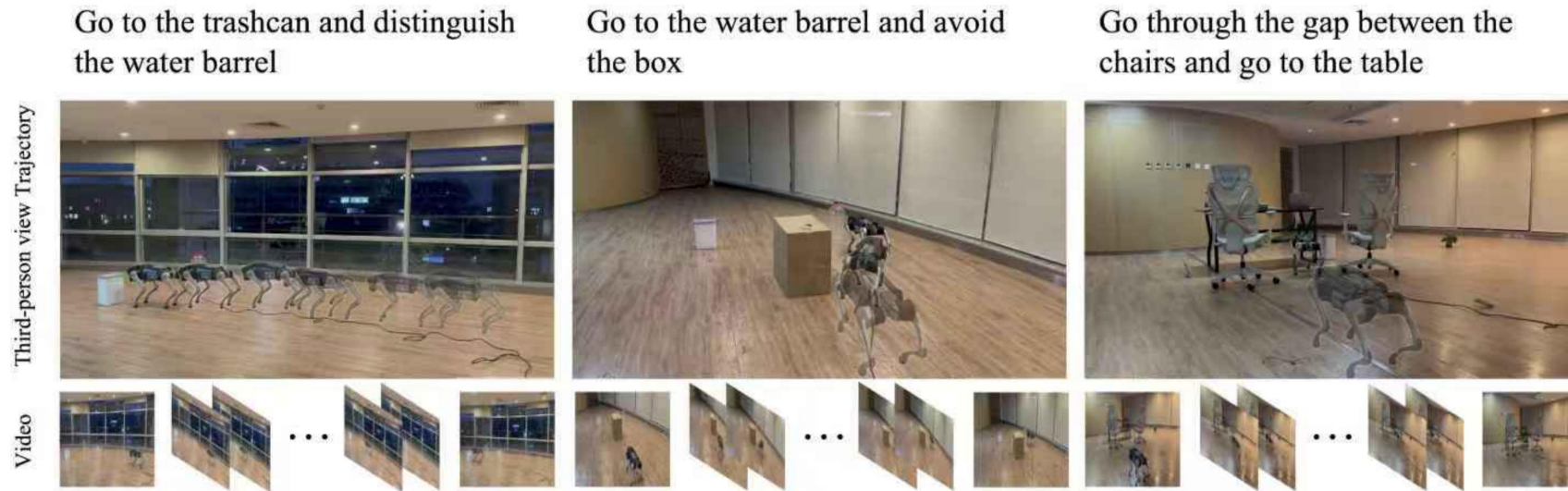


## StereoAdapter: Adapting Stereo Depth Estimation to Underwater Scenes

Zhengri Wu, Yiran Wang, Yu Wen, Zeyu Zhang, Biao Wu, Hao Tang

# Works in Progress

- Vision–Language–Action models for mobile robots such as robot dogs, UAVs, and humanoid robots.
- Real-time 3D Reconstruction for Mobile Robots
- Video World Models



Our mobile robot's VLA model follows user instructions to perform scene understanding, navigation, and action.

# Takeaways

- Do not abuse reinforcement learning for post-training; use RL only to adjust the foundation model's output.
- Synthetic data and data-driven methods are the key to achieving scalability and generalizability.
- Work on unimodal LLMs that perform next-token prediction will not achieve advanced machine intelligence. If you are interested in human-level intelligence, do not rely solely on LLMs; instead, enhance spatial awareness in visual foundation models.



End

Thank you.



Homepage



3D-R1