Spatial Intelligence

From Virtual to Real Worlds

Zeyu Zhang

Talk @ Yahaha, Sep 18, 2025

Some Quotes

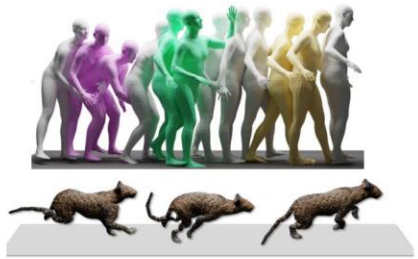
"There are several characteristics of intelligent behavior. For example, the capacity to understand the physical world, the ability to remember and retrieve information, the ability to reason, and the ability to plan. These are four essential characteristics of intelligent systems or entities."

— *Yann LeCun*

From Specialist to Generalist

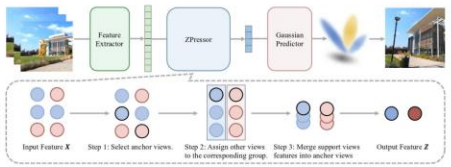
Motion Anything (2025)

Motion Avatar (2024)



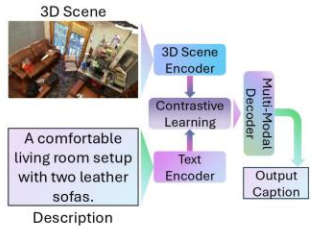
Animal & Object Generation

ZPressor (2024)



Scene Reconstruction & Generation

3D CoCa (2025)

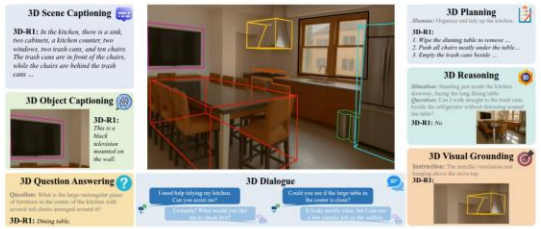


Scene Understanding

Evolve

UniMo (2025)

Motion Foundation Model



3D-R1 (2025)

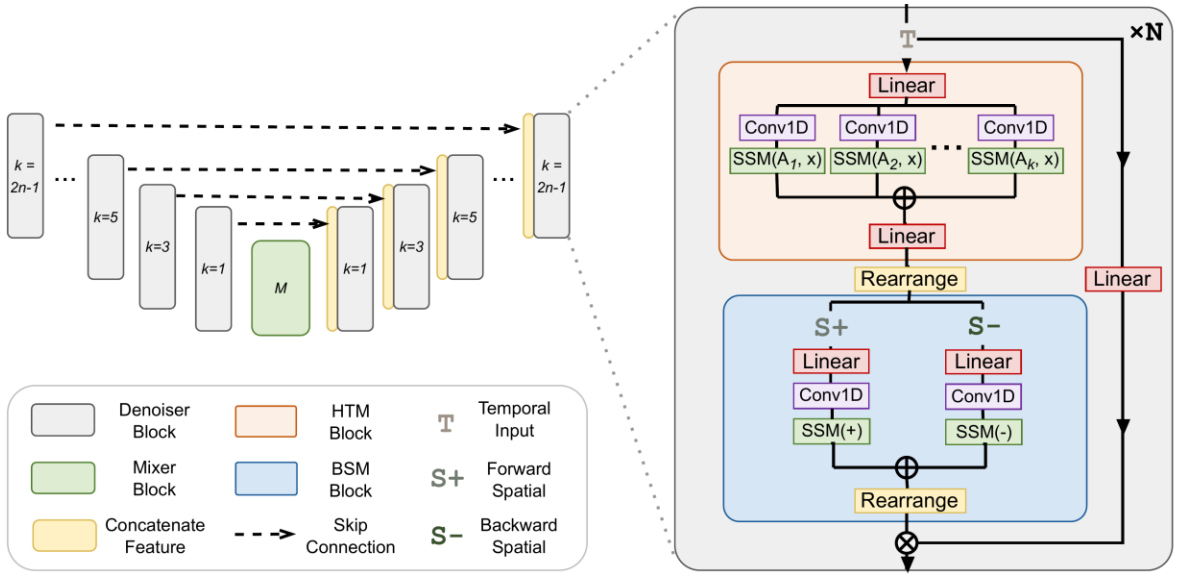
3D Foundation Model

Nav-R1 (2025)

Embodied Foundation Model

3D and 4D models represent a significant shift from specialist models, which are designed for specific tasks, to foundation models that can handle a wide range of tasks.

Specialist Model: Motion Mamba (2024)



Motion Mamba is an efficient text-to-motion model with linear complexity.

Zeyu Zhang et al. *Motion Mamba: Efficient and Long Sequence Motion Generation* (ECCV 2024)

Specialist Model: Motion Mamba (2024)

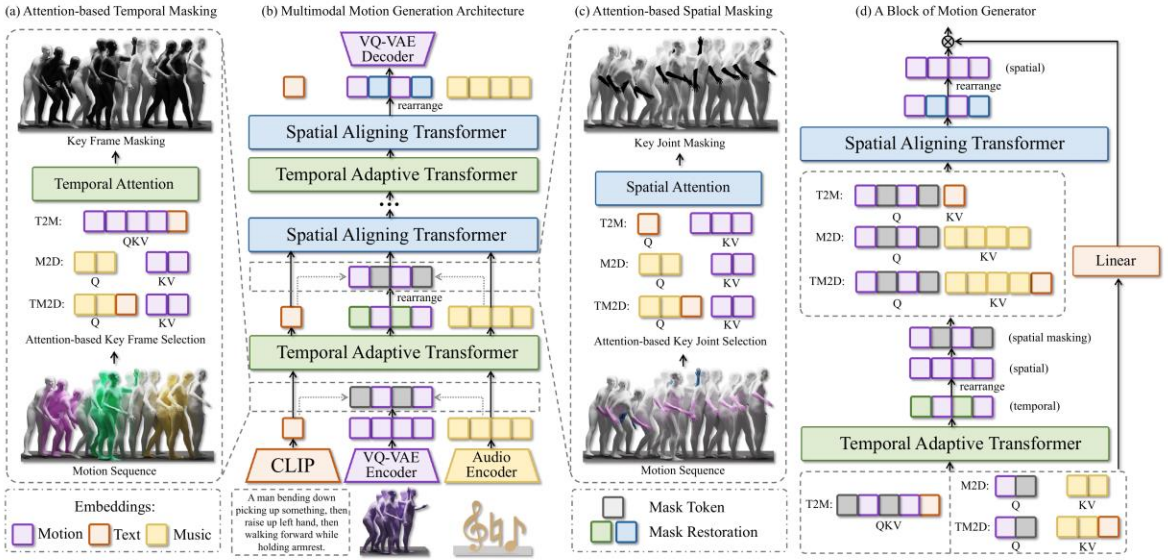
MotionDiffuse

MDM

MLD

Motion Mamba (Ours)

Specialist Model: Motion Anything (2025)

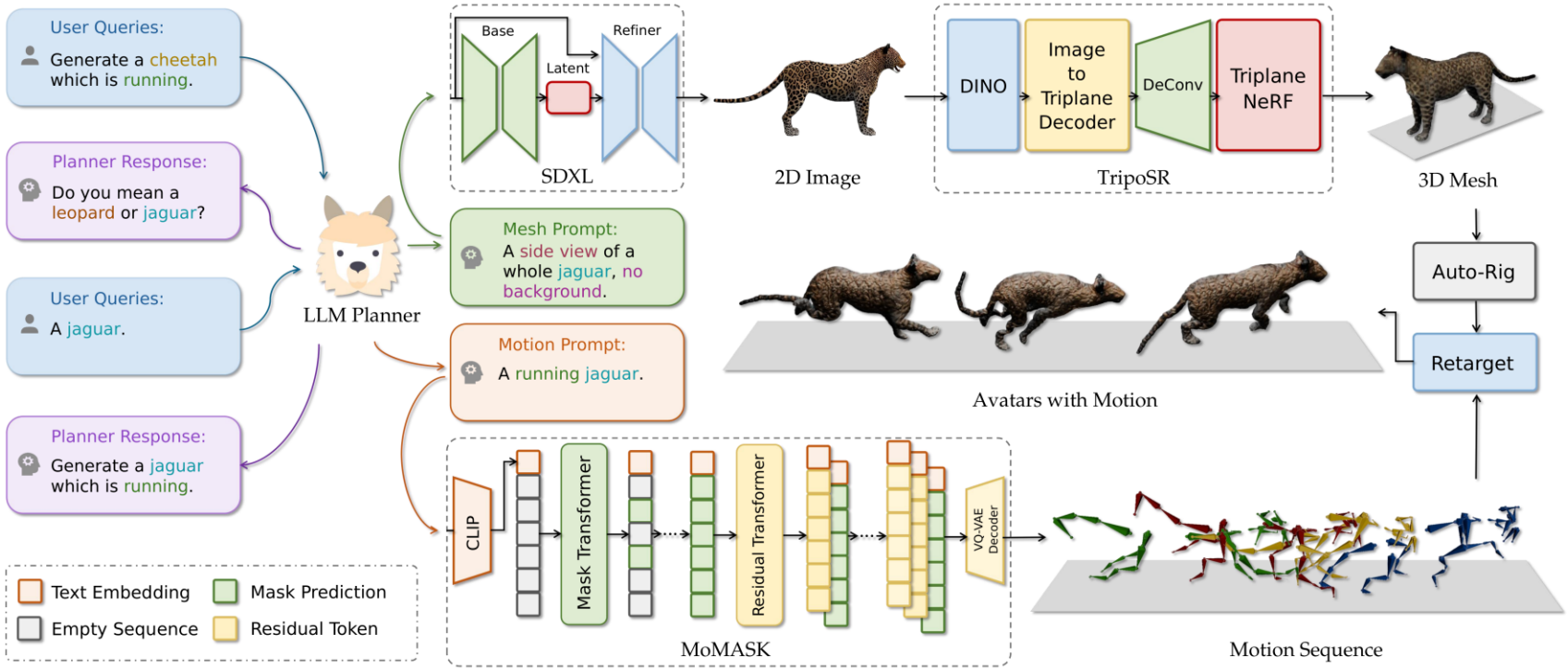


Motion Anything is an multimodal-conditioned motion generation model.

Zeyu Zhang et al. *Motion Anything: Any to Motion Generation*

Specialist Model: Motion Anything (2025)

Specialist Model: Motion Avatar (2024)



Motion Avatar is a feed-forward 4D generative model that can generate animatable meshes of both humans and animals, conditioned on text and/or images.

Zeyu Zhang et al. *Motion Avatar: Generate Human and Animal Avatars with Arbitrary Motion* (BMVC 2024)

Specialist Model: Motion Avatar (2024)

A red robot is boxing.

A red robot is doing hip pop dancing.

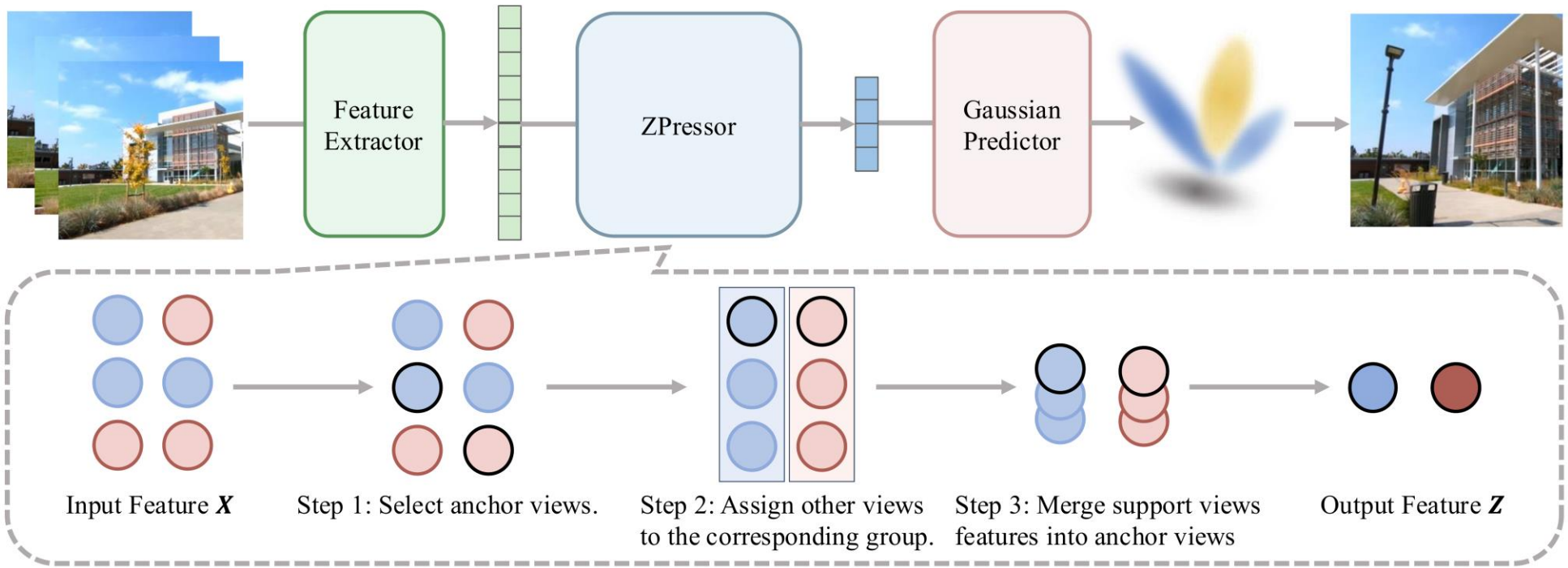
A red robot is saluting.

A bear runs then stands then runs again.

A bear walks then stands.

A bear runs then walks hesitantly.

Specialist Model: 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* (2025)

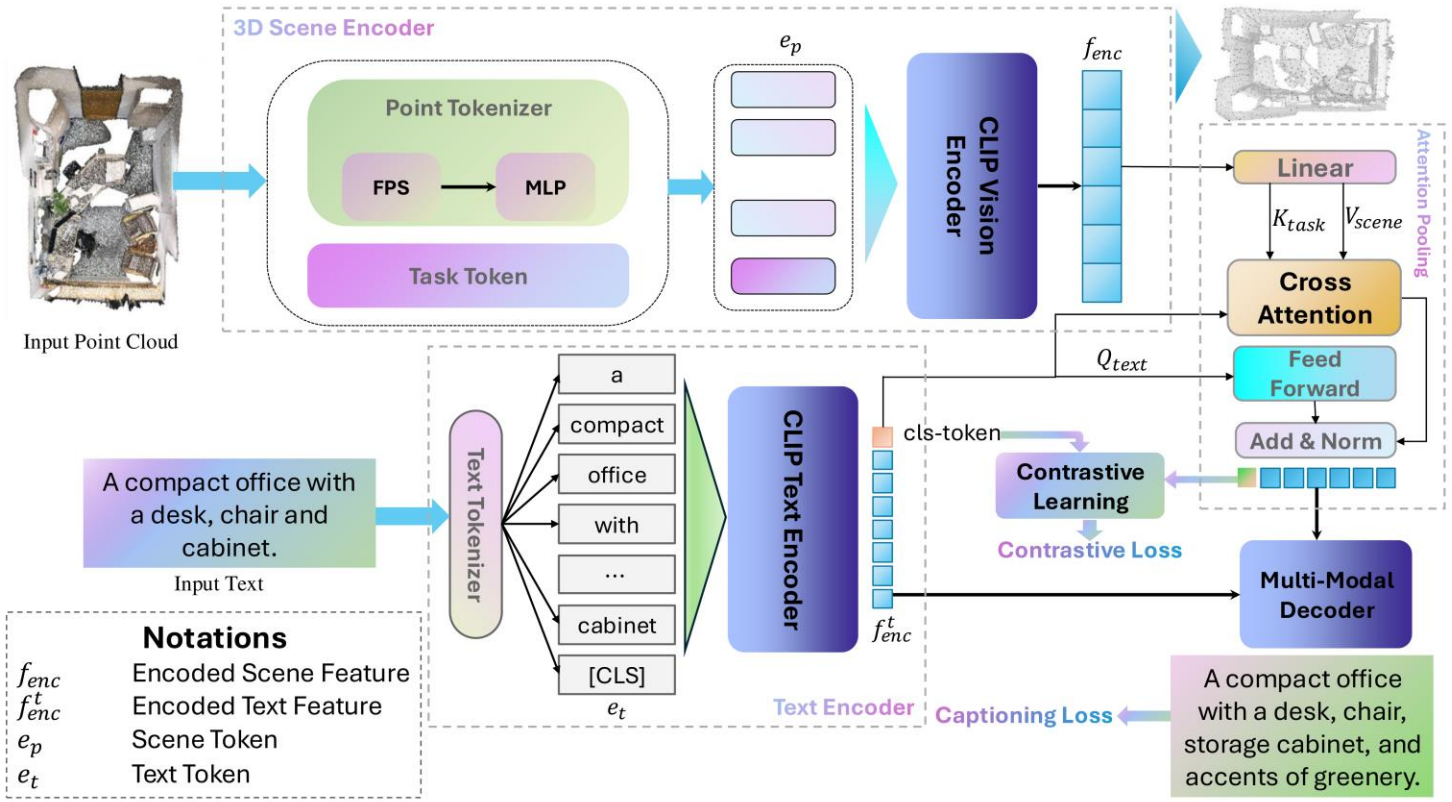
Specialist Model: ZPressor (2025)

Visualization on DL3DV (36 Input Views)

a62c330f5403e2e41a82a74c4e865b705c5706843b992fae2fe2e538b122d984

63798f5c6fbfcb4eb686268248b8ecbc8d87d920b2bcce967eeaedfd3b3b6d82

Specialist Model: 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)

Specialist Model: 3D CoCa (2025)

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

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

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.

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

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.

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

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

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

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

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

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

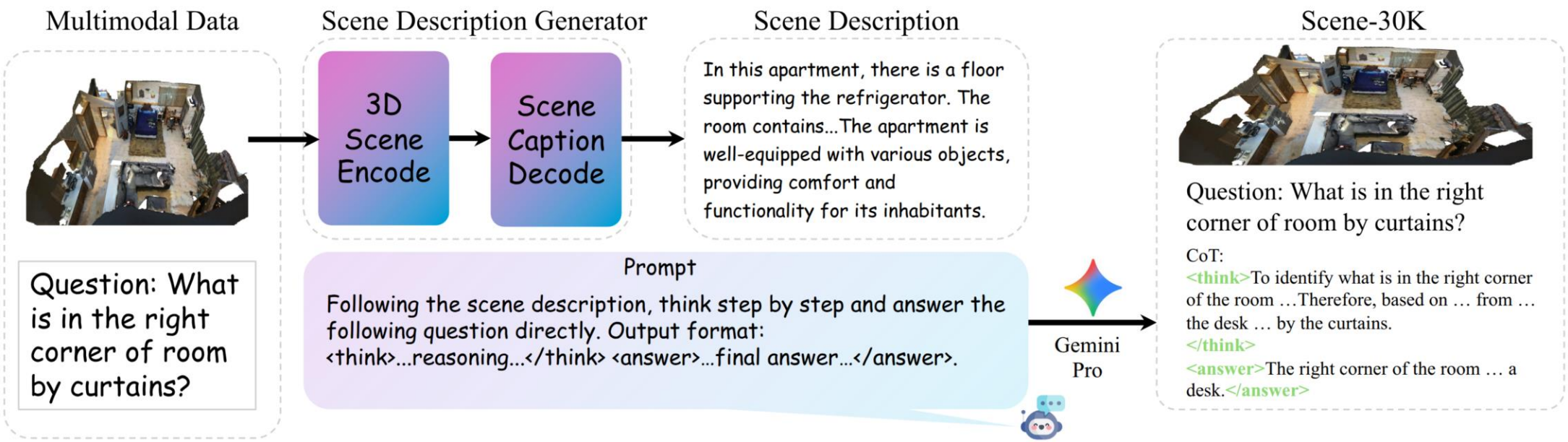
GT: A comfortable living room setup with two leather sofas, a small coffee table, and a rug on the floor. The corner have a musical instrument and ornaments.

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

What's next for 3D foundation models?

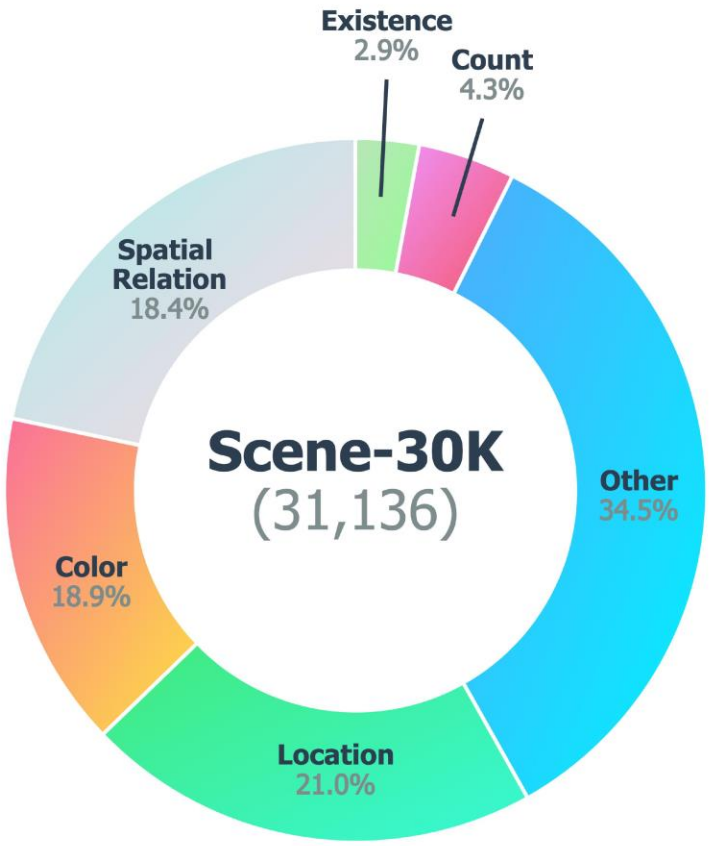
- How can we achieve zero-shot generalizability across different tasks given the domain knowledge gap between them?
- How can we adjust a foundation model after conventional supervised post-training when the outcomes are unsatisfactory on specific tasks?
- And most importantly, how can we ground our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

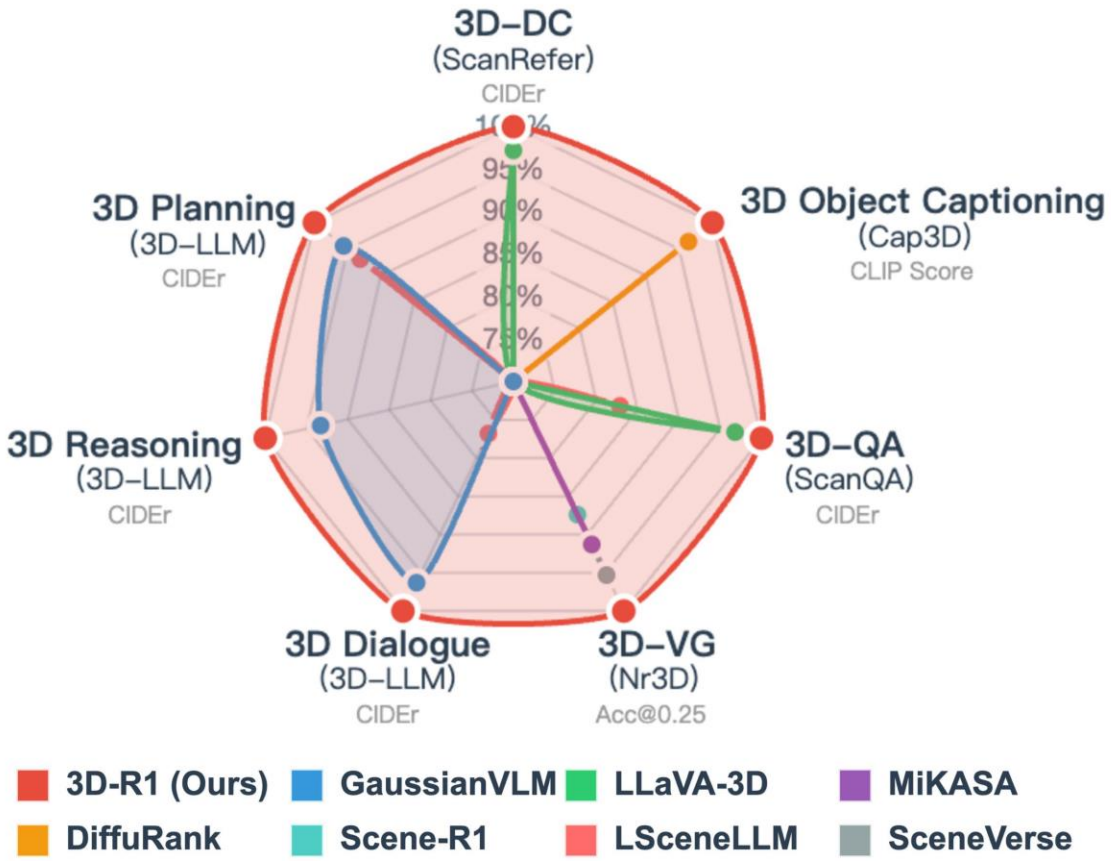
Synthetic Data Helps Enhance Generalizability: 3D-R1



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

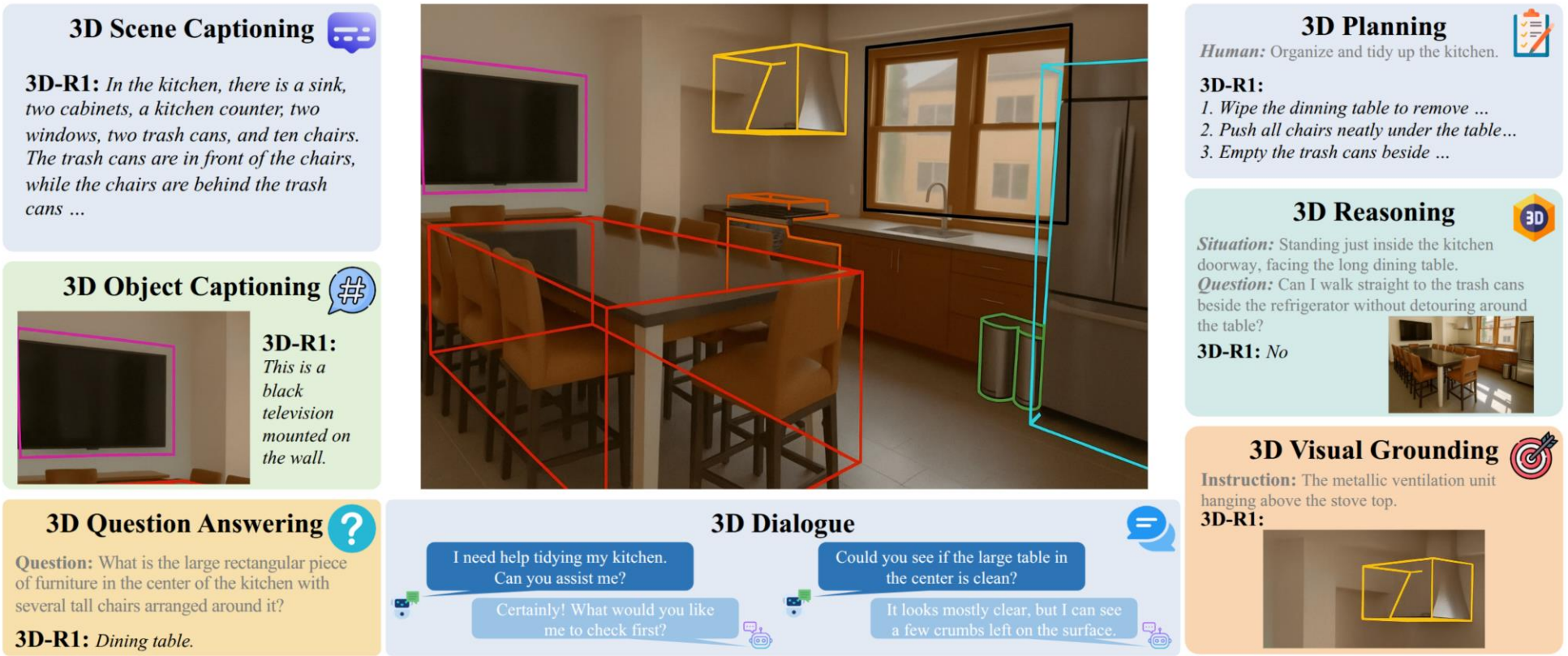
Scene-30K in 3D-R1





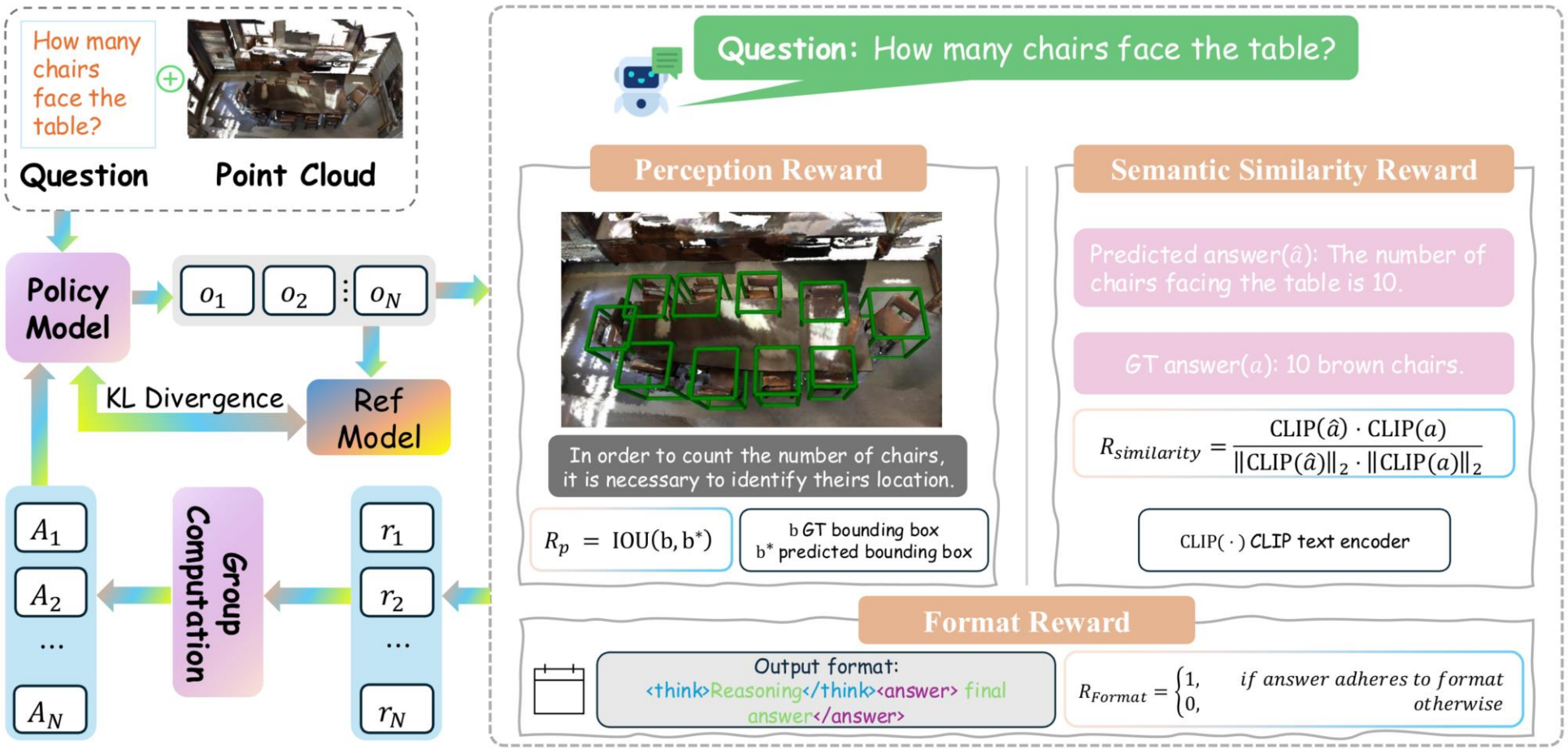
Scene-30K contains diverse scene categories and question types. **3D-R1** demonstrates strong performance across various tasks.

Generalizability: 3D-R1



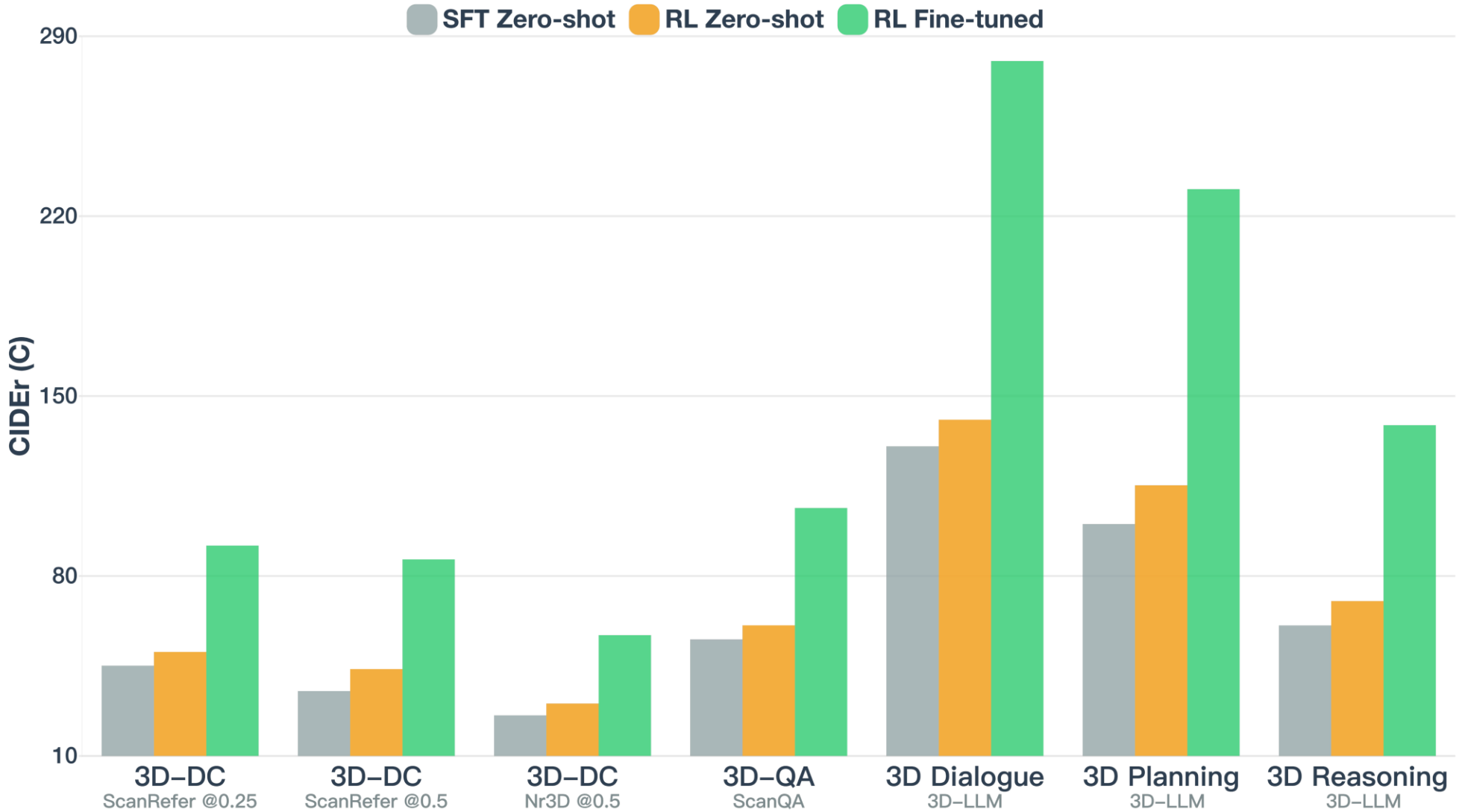
3D-R1 is a generalist model capable of handling various downstream tasks and applications in a zero-shot manner with incredible generalizability, significantly reducing the need for expensive adaptation.

Adjust Output: Reinforcement Learning with Verifiable Rewards (RLVR)



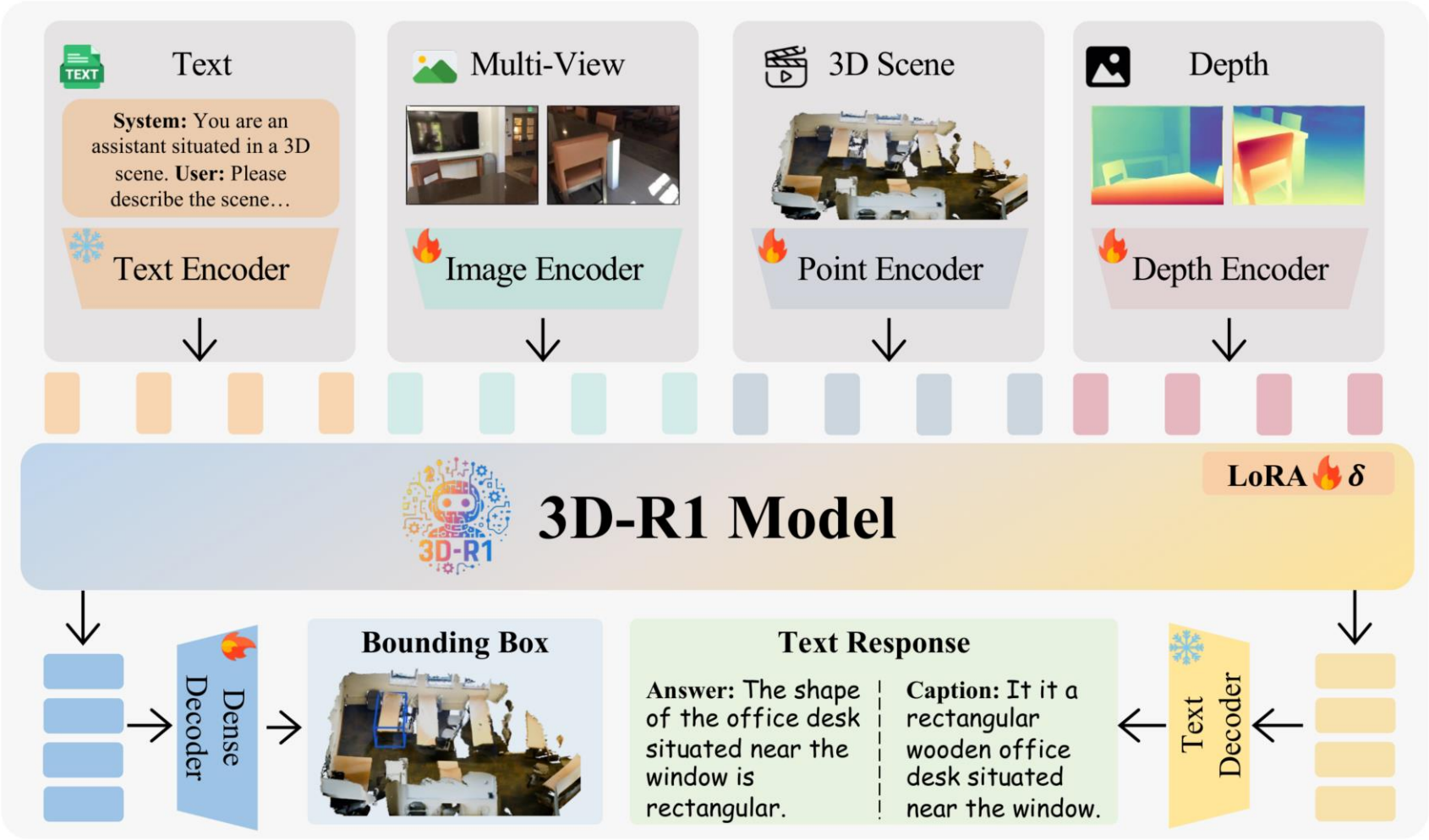
The policy model generates N outputs from a point cloud and question. Then perception IoU, semantic CLIP-similarity, and format-adherence rewards are computed, grouped, and combined with a KL term to a frozen reference model to update the policy.

Enhanced Reasoning: 3D-R1



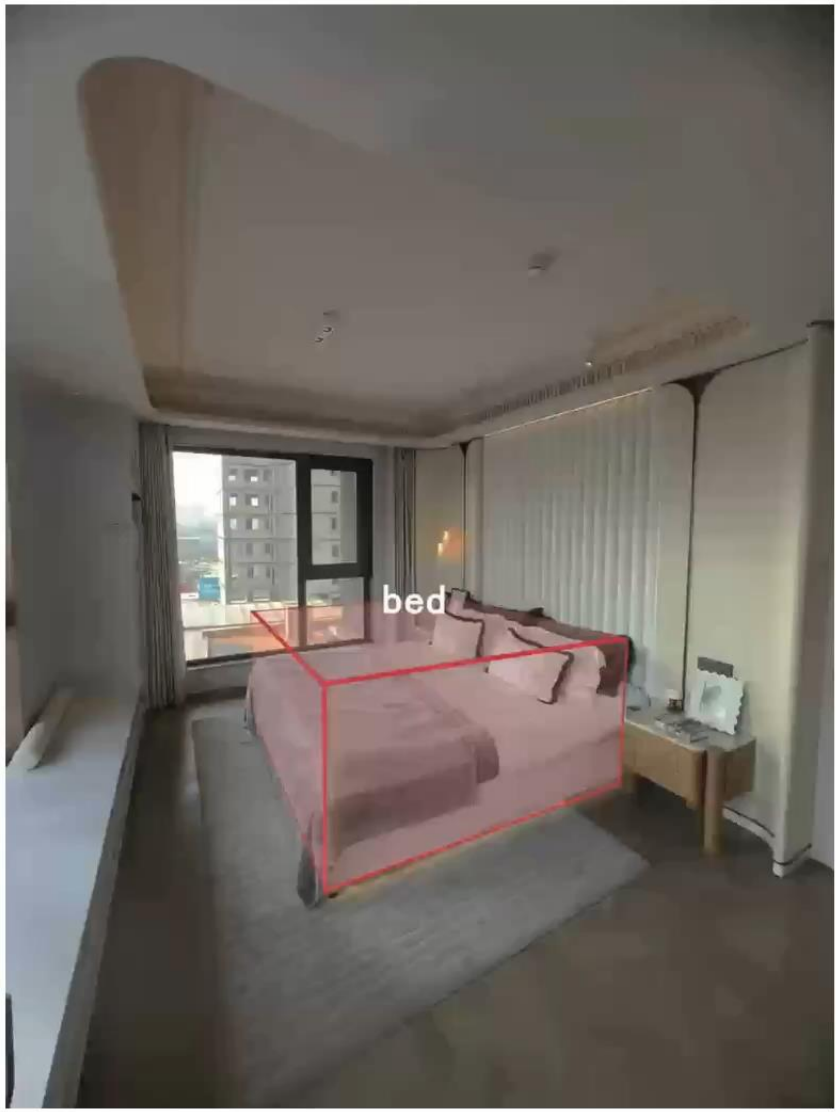
3D-R1 exhibits remarkable generalizability with enhanced reasoning capabilities.

Foundation Model: 3D-R1



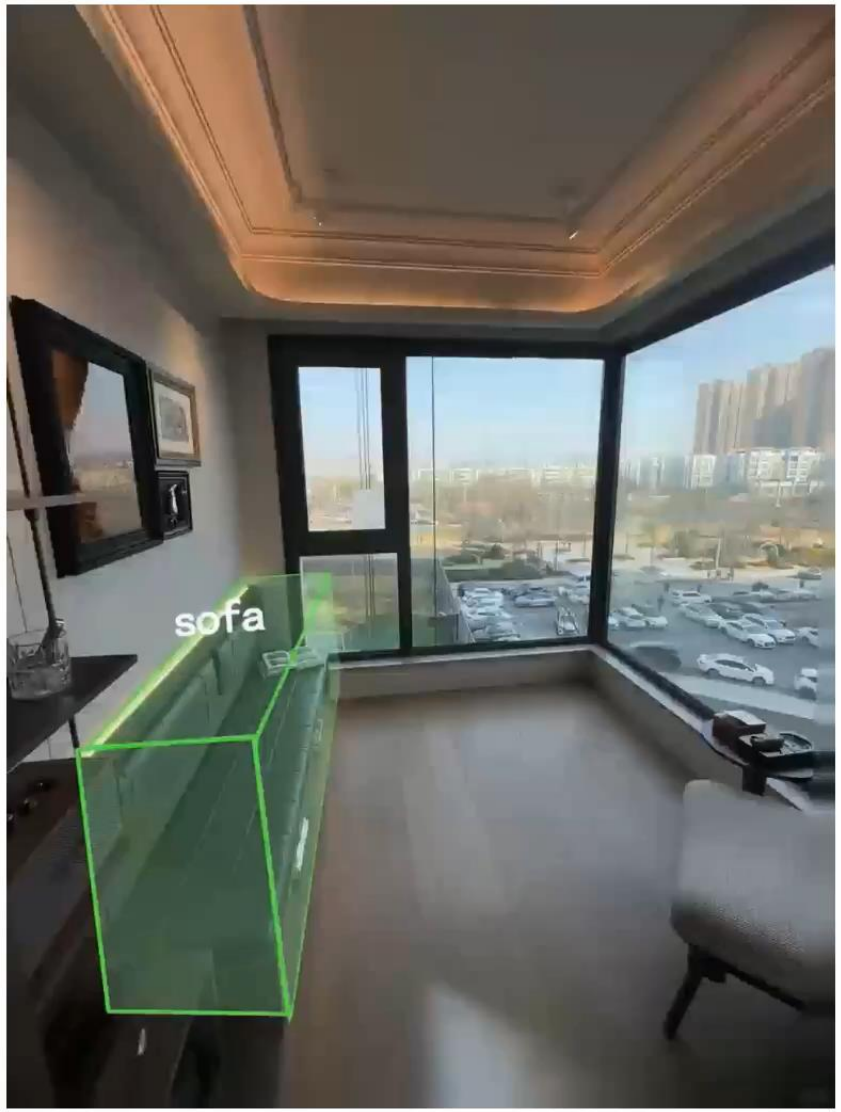
3D-R1 is an open-source generalist model that enhances the reasoning of 3D VLMs for unified scene understanding.

3D Scene Dense Captioning (3D-DC)



3D-DC

3D Object Captioning

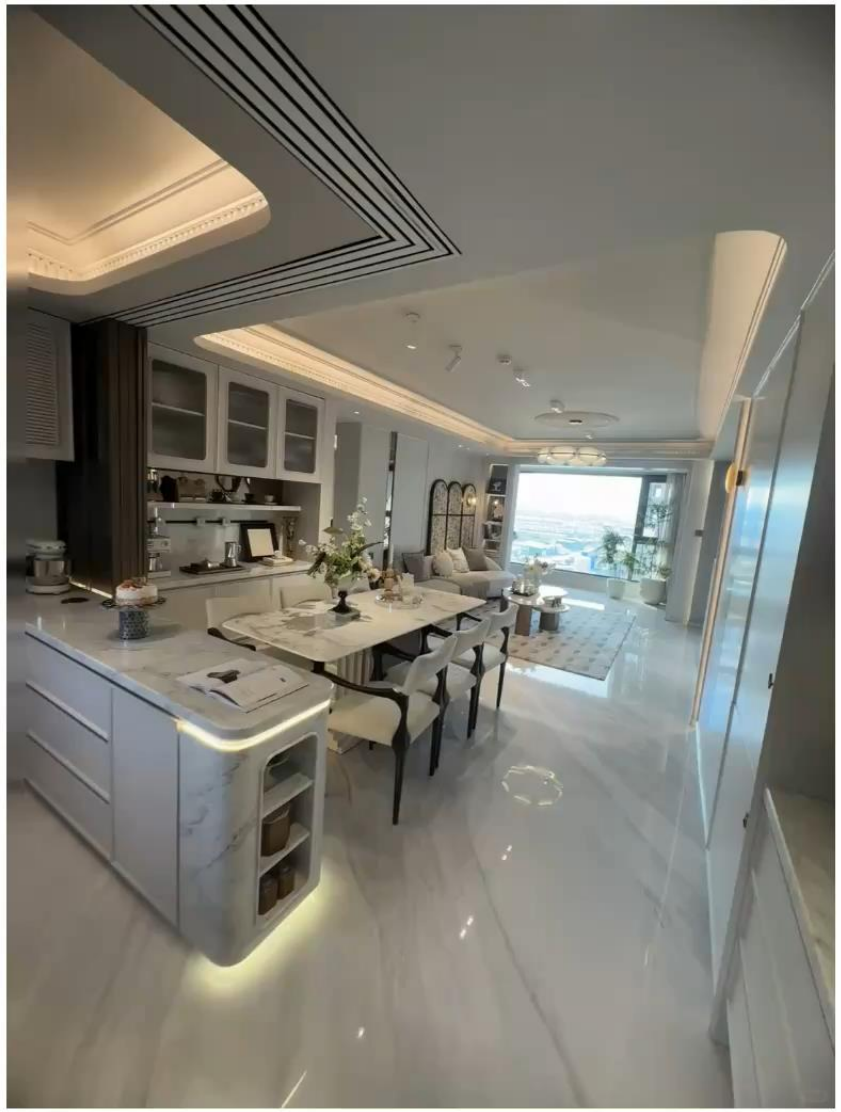


3D Object Captioning

3D Visual Grounding (3D-VG)

3D-VG

3D Question Answering (3D-QA)



3D-QA

3D Dialogue

3D Dialogue

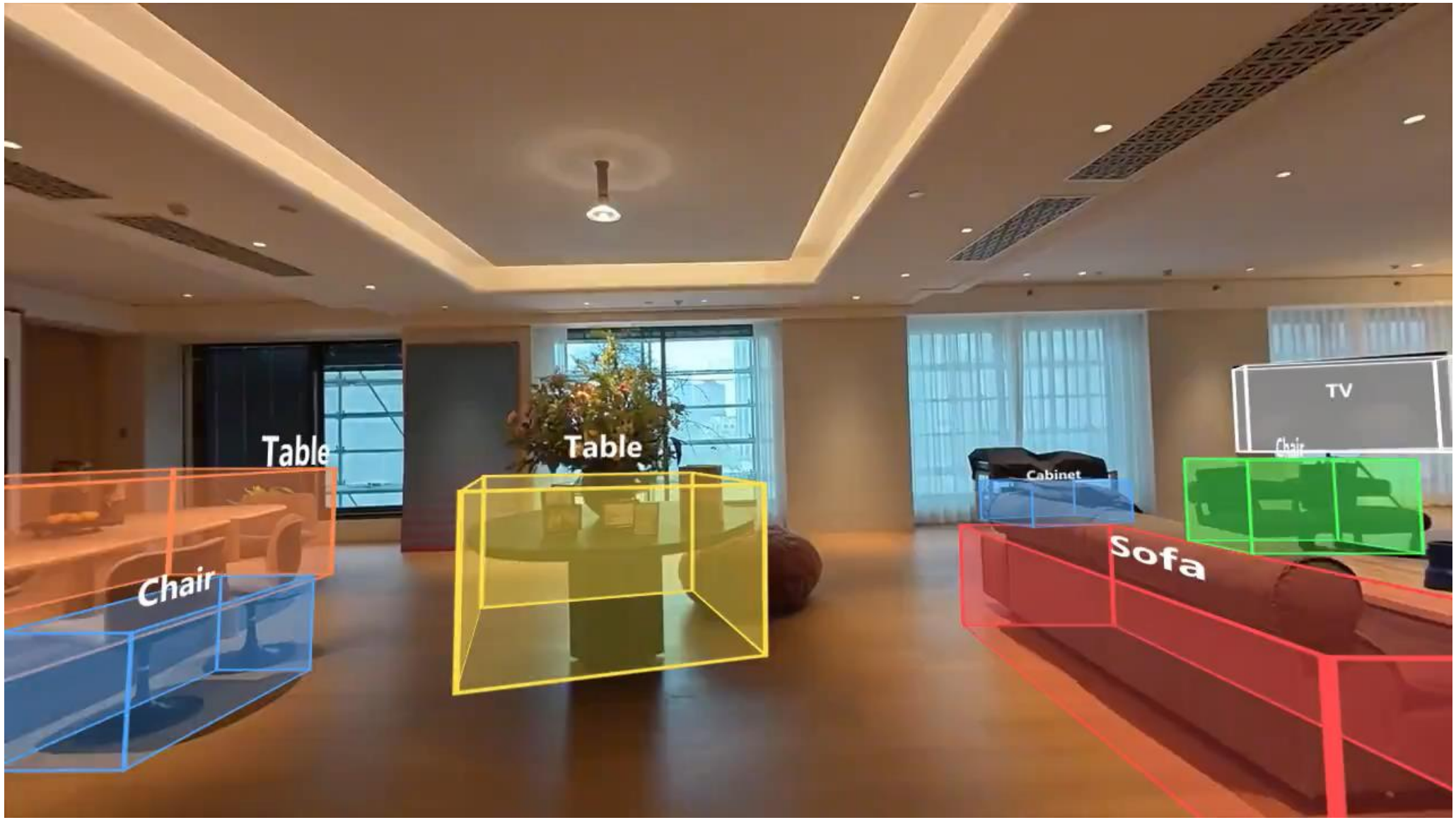
3D Reasoning

3D Reasoning

3D Planning

3D Planning

Zero-Shot Results



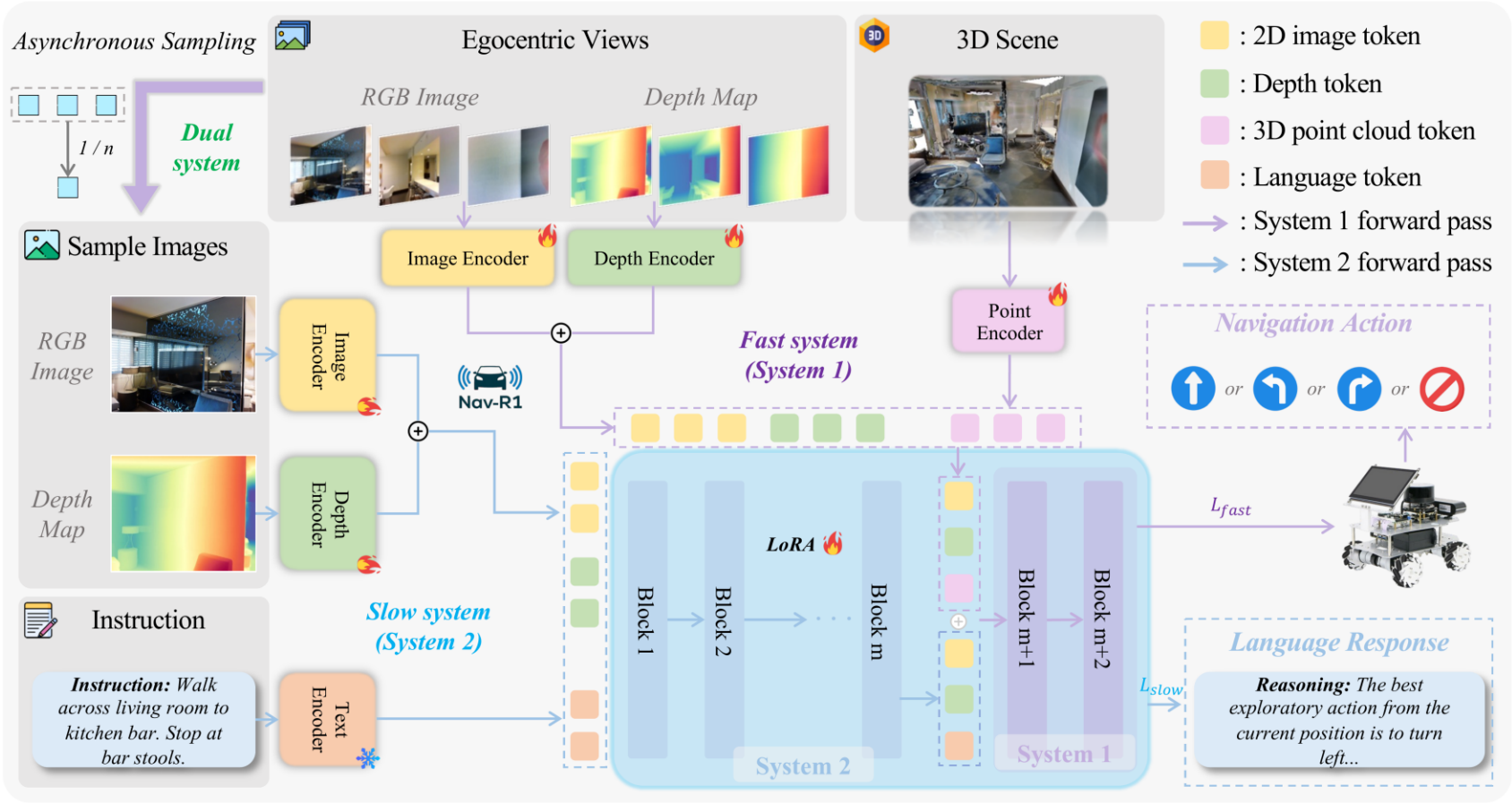
System and Memory: Nav-R1

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..

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

Nav-R1

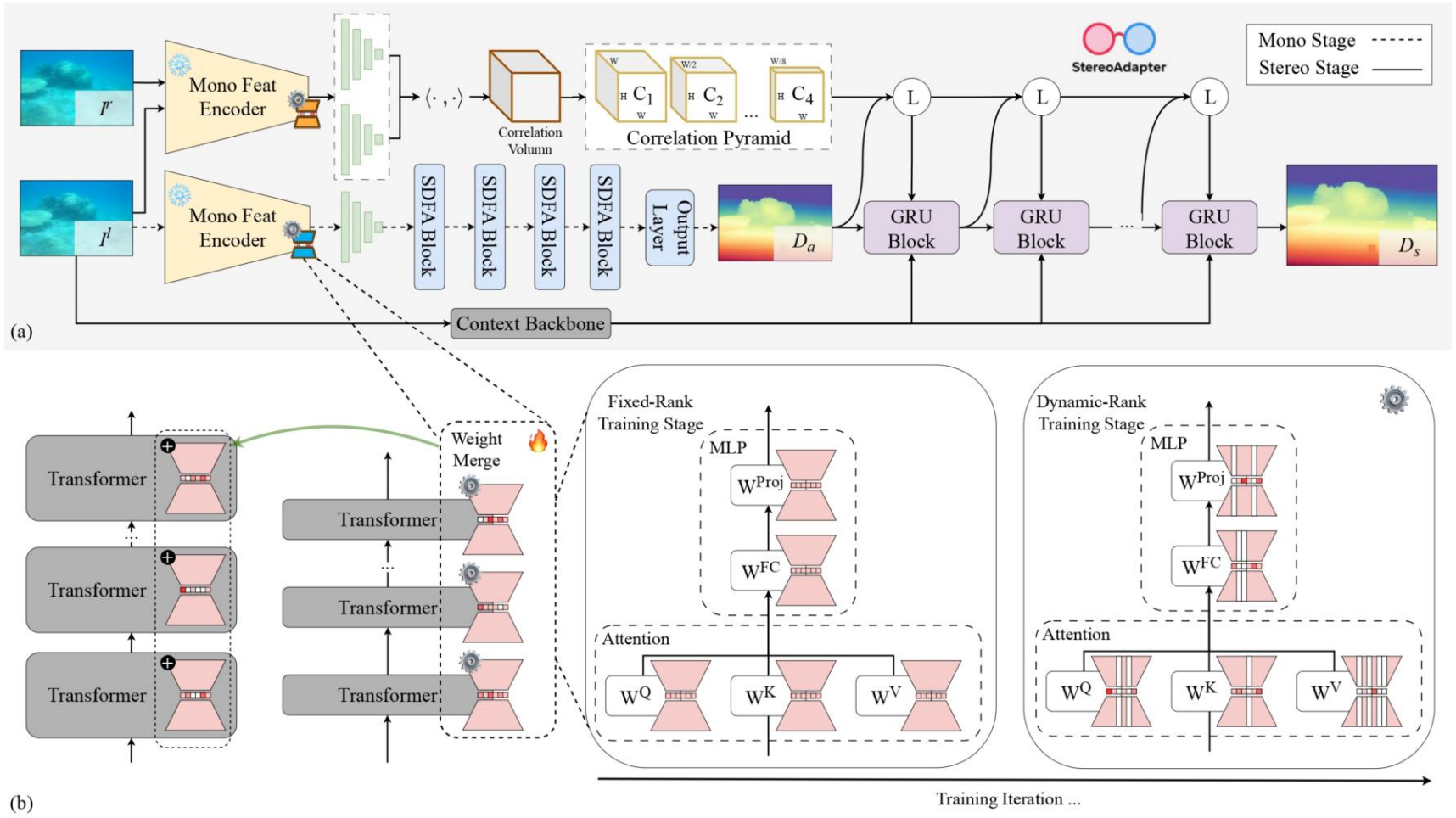
北京大学 PEKING UNIVERSITY

Nav-R1: Reasoning and Navigation in Embodied Scenes

Qingxiang Liu, Ting Huang, Zeyu Zhang, Hao Tang

Nav-R1

Bridging the Domain Gap in Post-Training: StereoAdapter



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

Results: StereoAdapter

StereoAdapter

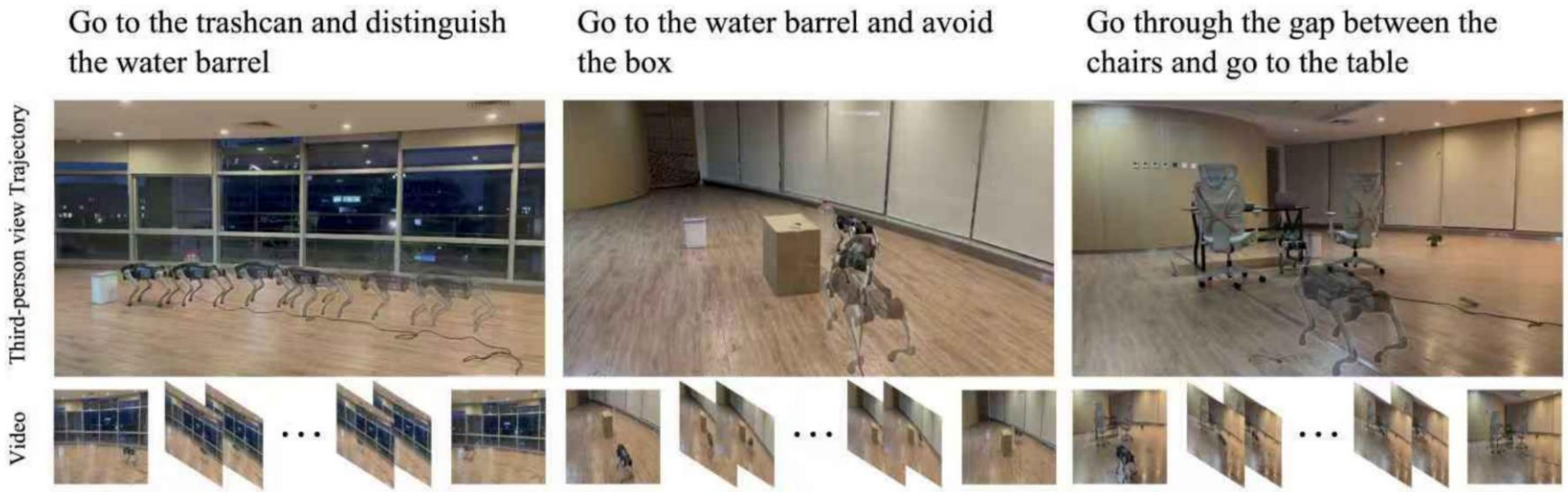
北京大学 PEKING UNIVERSITY

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.
- Video World Models
- 4D Generative 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.