Latest Advances in Embodied Reasoning

Zeyu Zhang

Homepage

3D-R1

Talk @ CVLife, Mar 3, 2026

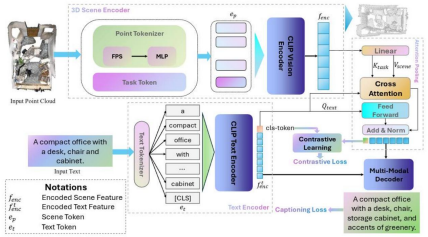
Some Quotes From Paving the Path to Generalizable Robotics

"I'm really excited to see whether we can build generalizable robotic systems with a single large neural network that can quickly learn new skills, similar to what we have already seen in computer vision and natural language processing... Once both components (logical network or foundation model, unsupervised RL for robot-specific learning component) are trained, this logical network, if well optimized, would be highly prepared to acquire new skills very quickly. These skills could be learned through **imitation learning** *from human demonstrations,* **reinforcement learning with a specified reward function**, *or* **reinforcement learning with a human** *in the loop who provides examples to guide the learning process in the right direction. This helps avoid undesirable outcomes that can occur when the reward function is misspecified."*

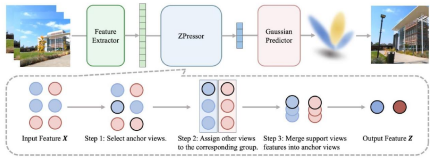
— *Pieter Abbeel*

From Specialized Methods to General Methods

3D Dense Captioning

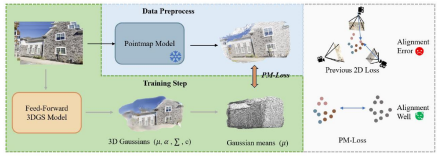


3D CoCa (2025)

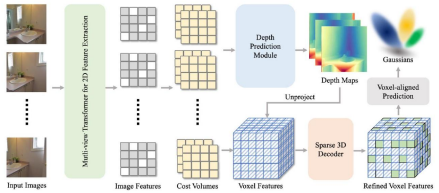


ZPressor (NeurIPS 2025)

3D Reconstruction

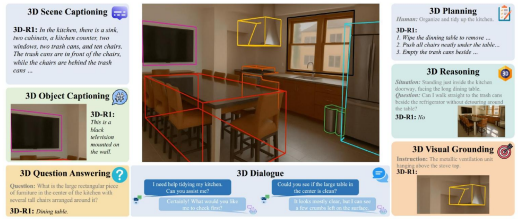


PM-Loss (2025)



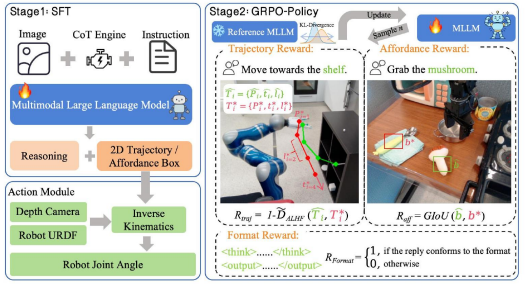
VolSplat (2025)

Evolve



3D-R1 (2025)

Nav-R1 (2025)



VLA-R1 (2025)

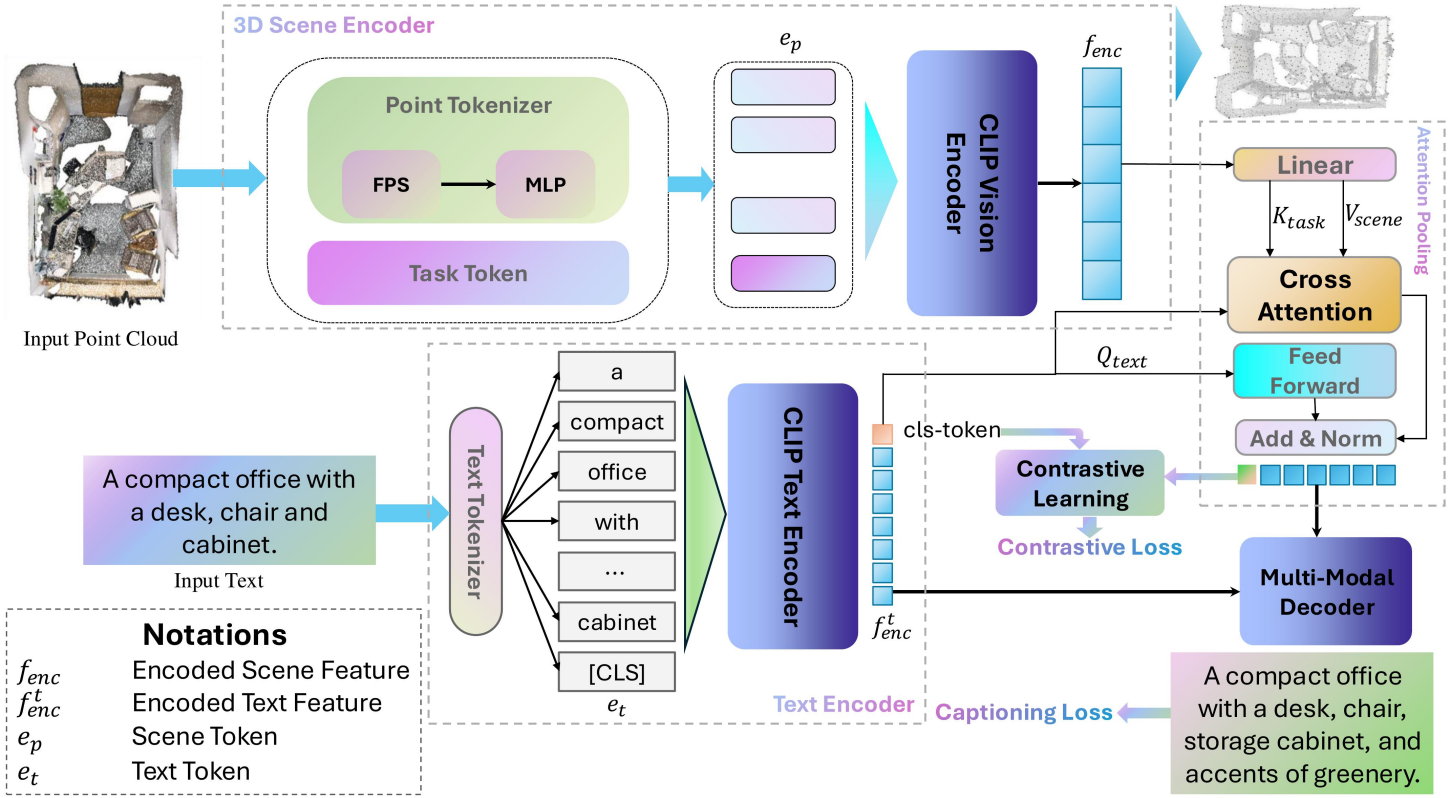
3D Foundation Model

Navigation Foundation Model

Manipulation Foundation Model

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.

3D Representation Learning: 3D CoCa (3DV 2026)



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

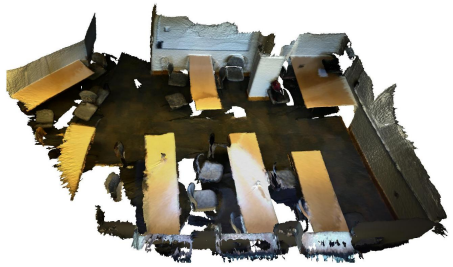
Ting Huang, Zeyu Zhang et al. *3D CoCa: Contrastive Learners are 3D Captioners* (3DV 2026)

Results of 3D CoCa (3DV 2026)

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.

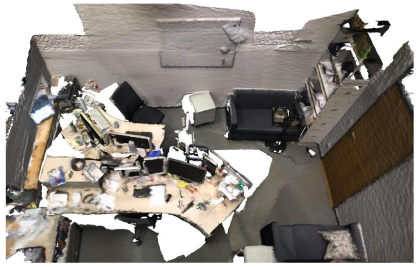
GT: In a bright dining room, a long wooden table is flanked by neatly arranged chairs. Light filters in through the window, and a decorative painting adorns the wall.



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.

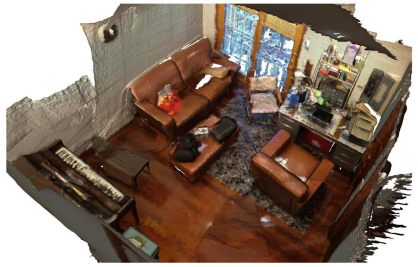
GT: A spacious indoor setting with several parallel tables and chairs, offering walking and working areas on all sides. The layout resembles a classroom.



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.



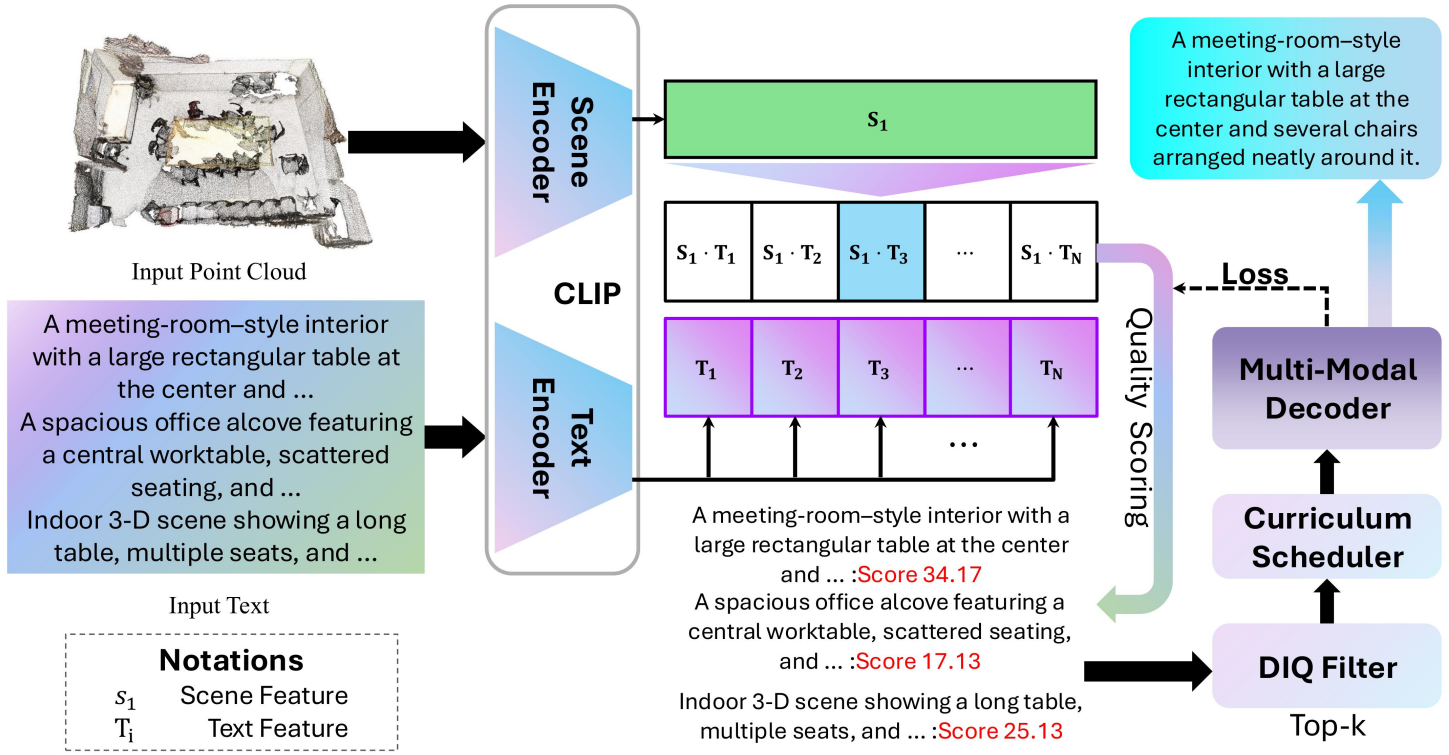
Vote2Cap-DETR++: A living room with two sofas and a small side table.

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).

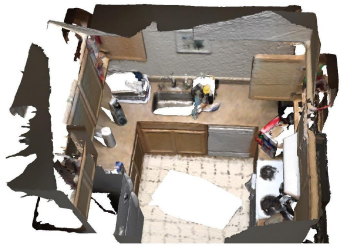
3D Data-Centric Learning: DC-Scene (3DV 2026 EE)



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.

Ting Huang, Zeyu Zhang et al. *DC-Scene: Data-Centric Learning for 3D Scene Understanding* (3DV 2026 Exploration Edge Track)

Results of DC-Scene (3DV 2026 EE)



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

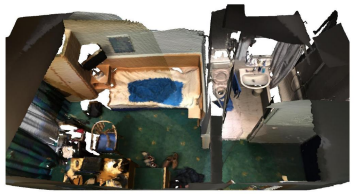
DC-Scene(Top-75%): a kitchen where wooden cabinets frame a metal sink beneath a wall picture, while a white washer-dryer sits to the right of the light-tiled floor that opens into a carpeted hallway.

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.



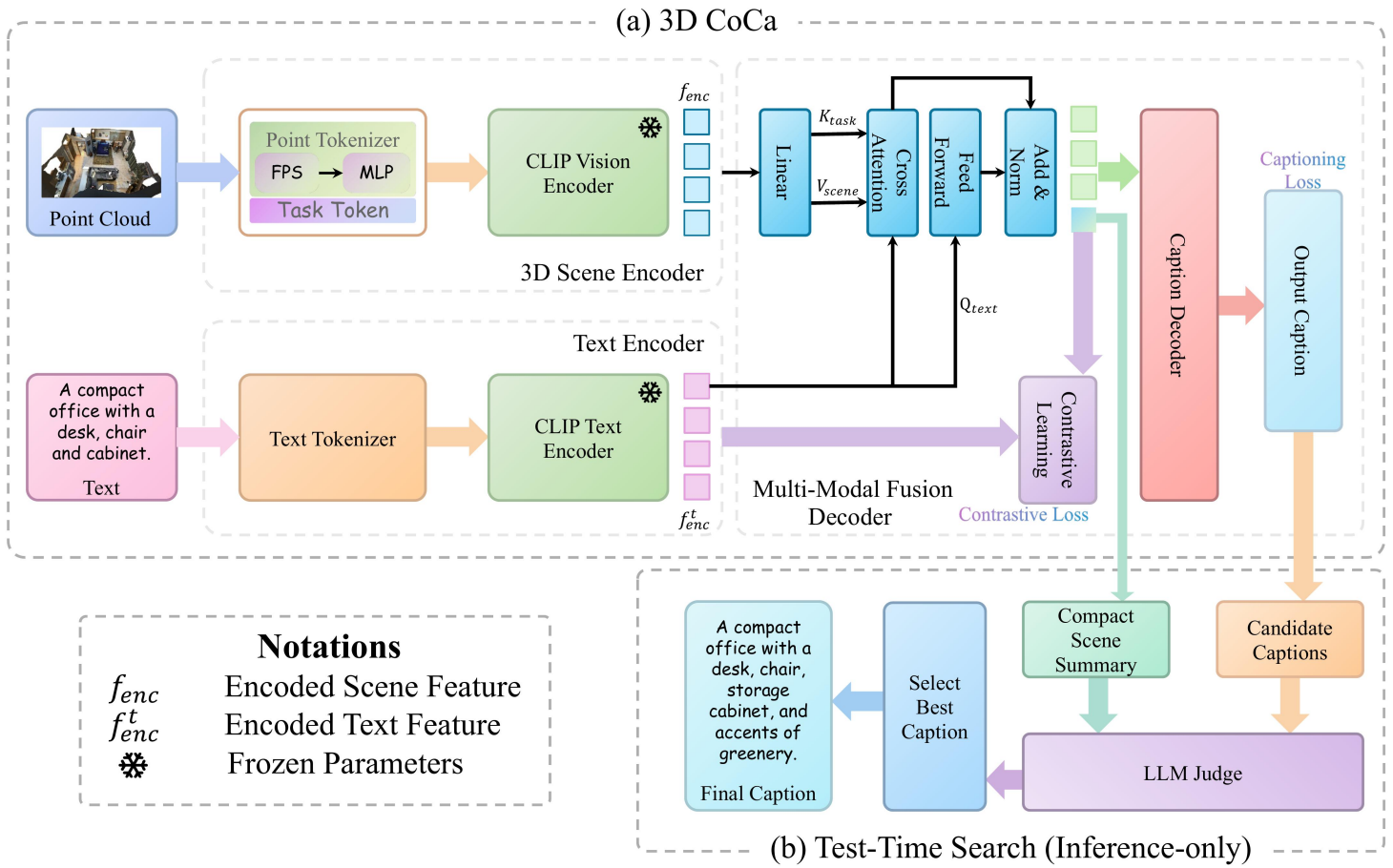
Baseline(full data): a small bedroom with a bed, green carpet, and a bathroom to the right.

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).

3D CoCa v2: Enhancing Generalization with Test-Time Search (2026)



Test-Time Search (inference-only) improves robustness without any parameter updates by generating best-of-N candidate captions from the backbone, conditioning an external LLM judge on a compact scene summary, and selecting the highest-scoring candidate as the final caption.

3D CoCa v2: Generalizable Results

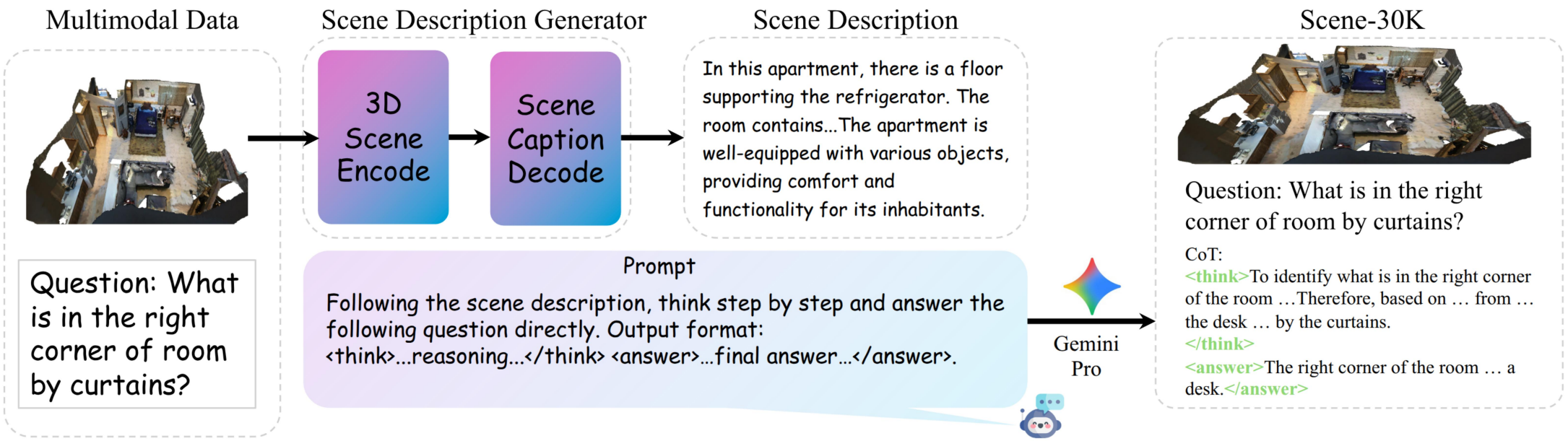
CAM_FRONT	CAM_BACK	CAM_FRONT	CAM_BACK	CAM_FRONT	CAM_BACK	CAM_FRONT	CAM_BACK
							
Vote2Cap-DETR++: A room with a walls and a floor.		**Vote2Cap-DETR++:** A room with a simple layout.		**Vote2Cap-DETR++:** A corridor with two walls and open floor.		**Vote2Cap-DETR++:** A room with open floor.	
3D CoCa: An indoor corridor with walls and a floor.		**3D CoCa:** An indoor room with a simple layout.		**3D CoCa:** An indoor corridor with walls.		**3D CoCa:** An indoor room with a open floor and a simple layout.	
3D CoCa v2: A campus **roadside** with a large signboard, trees, and modern **buildings** beside a paved walkway.		**3D CoCa v2:** A grassy slope under an open sky with a paved path, showing a quiet **outdoor** roadside environment.		**3D CoCa v2:** A straight urban road with lane markings, **buildings** and trees on the sides; a vehicle is visible in the rear view.		**3D CoCa v2:** A rainy urban **street** intersection with wet pavement, lane markings, **traffic lights**, and surrounding buildings.	
GT: An outdoor roadside environment with a signboard, trees, and buildings along a paved pedestrian path.		**GT:** A grassy outdoor area with a sloped hill and a paved path, captured from front and rear cameras.		**GT:** A driving-scene street view with lane markings and surrounding trees; a vehicle is visible in the rear camera view.		**GT:** A rain-soaked city roadway at a signalized intersection, the scene shows lane markings, traffic lights, and buildings.	

Qualitative results on TOD^{3}Cap (OOD, zero-shot). We compare captions generated by the indoor-trained Vote2Cap-DETR++, 3D CoCa and 3D CoCa v2 on outdoor scenes with paired front and back views. Vote2Cap-DETR++ and 3D CoCa often exhibit a strong indoor bias, producing generic indoor descriptions, whereas 3D CoCa v2 generates more scene-consistent outdoor captions that better reflect key semantics. Ground-truth (GT) captions are shown for reference. Red words highlight informative details captured by 3D CoCa v2 but missing in the baseline.

What's next for 3D and embodied foundation models?

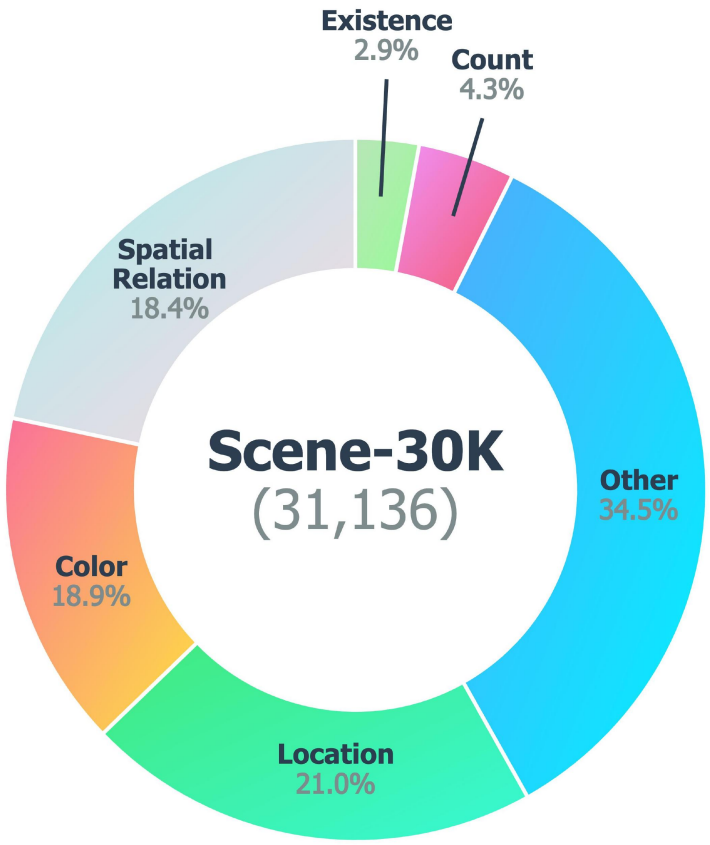
- How can we achieve zero-shot generalizability across different tasks given the domain knowledge gap between them?
- How can we adjust a foundation model after conventional supervised post-training when the outcomes are unsatisfactory on specific tasks?
- How can we ground our foundation model in the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

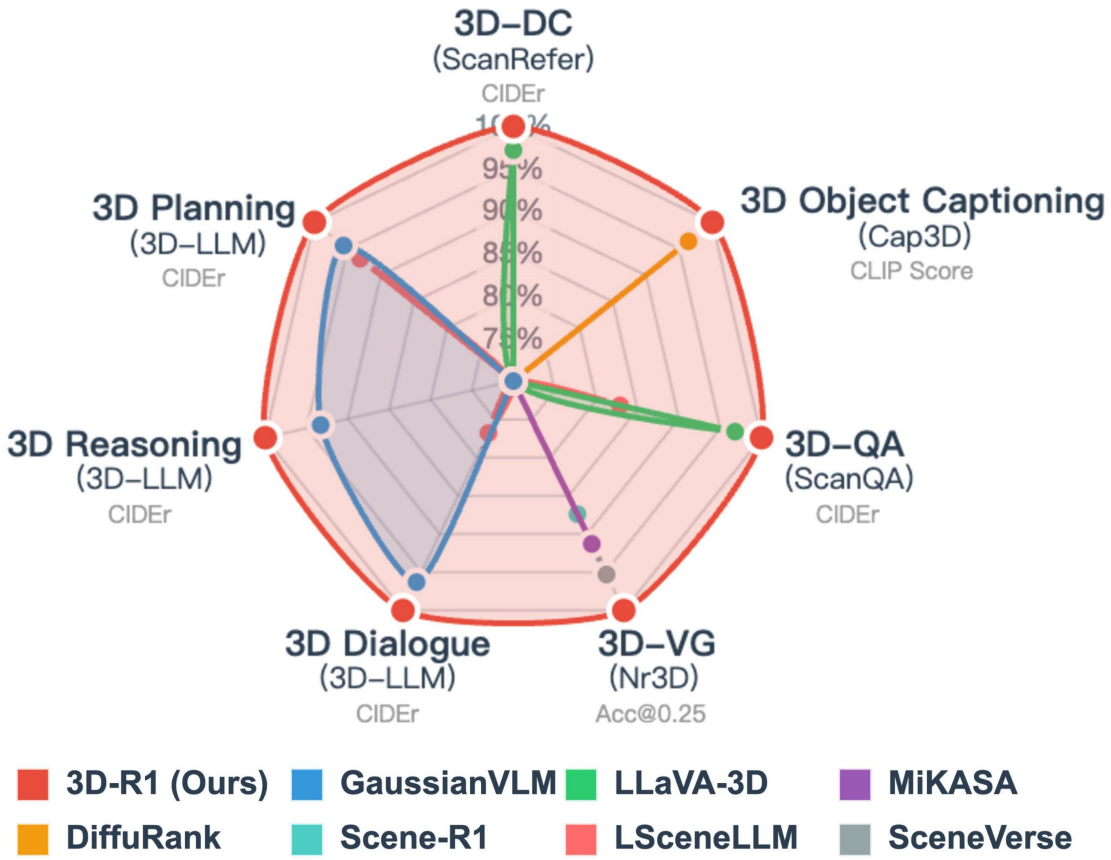
Synthetic Data Helps Enhance Generalizability: 3D-R1



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.

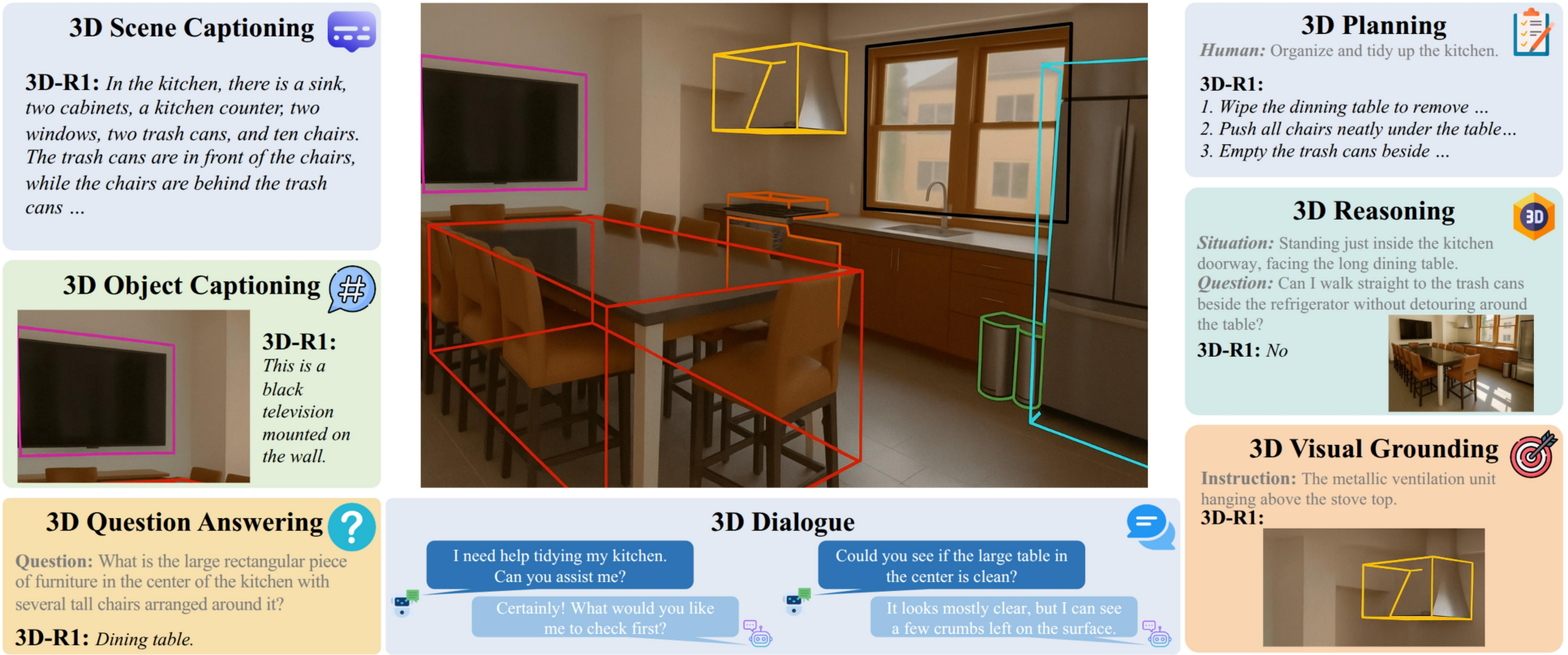
Scene-30K in 3D-R1





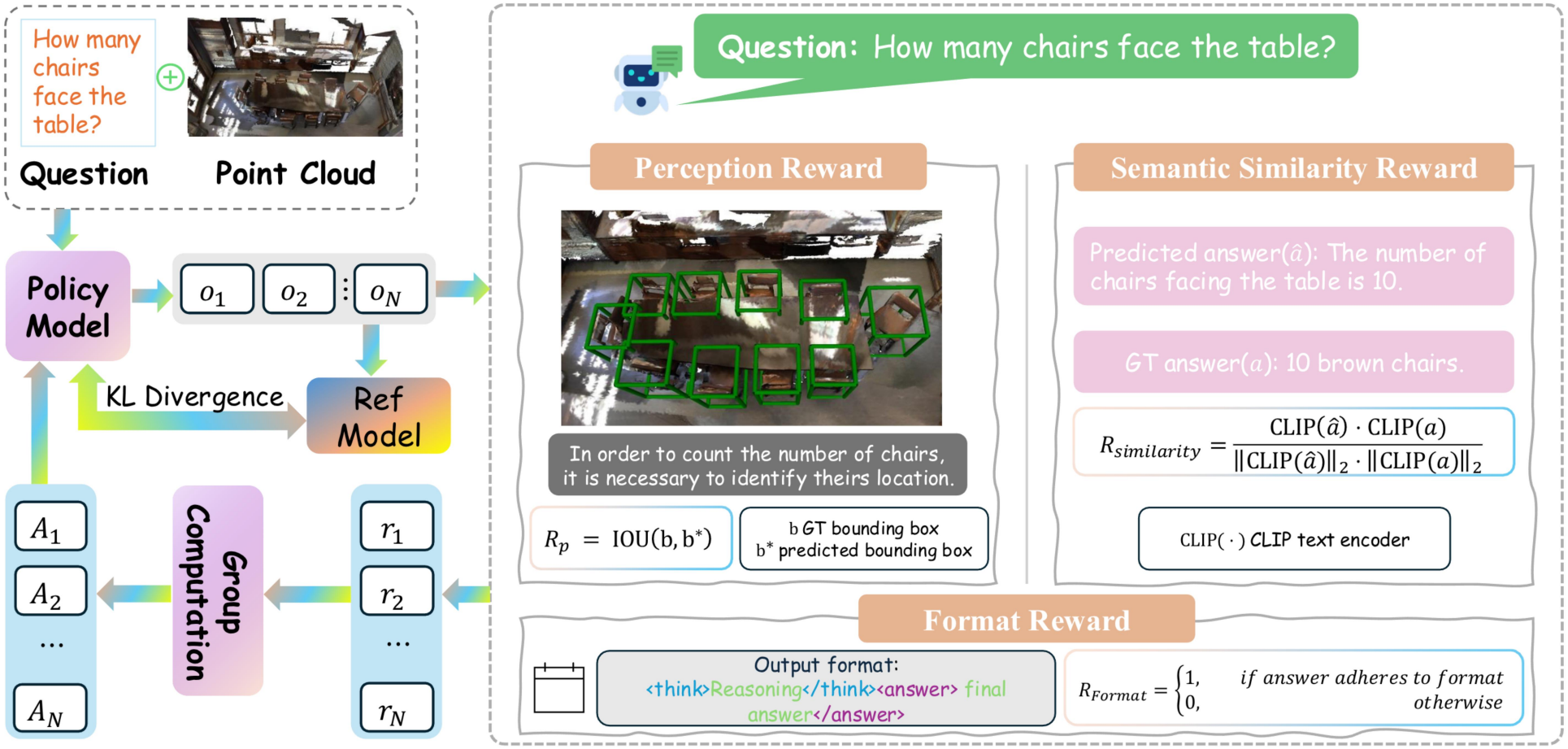
Scene-30K contains diverse scene categories and question types. **3D-R1** demonstrates strong performance across various tasks.

Generalizability: 3D-R1



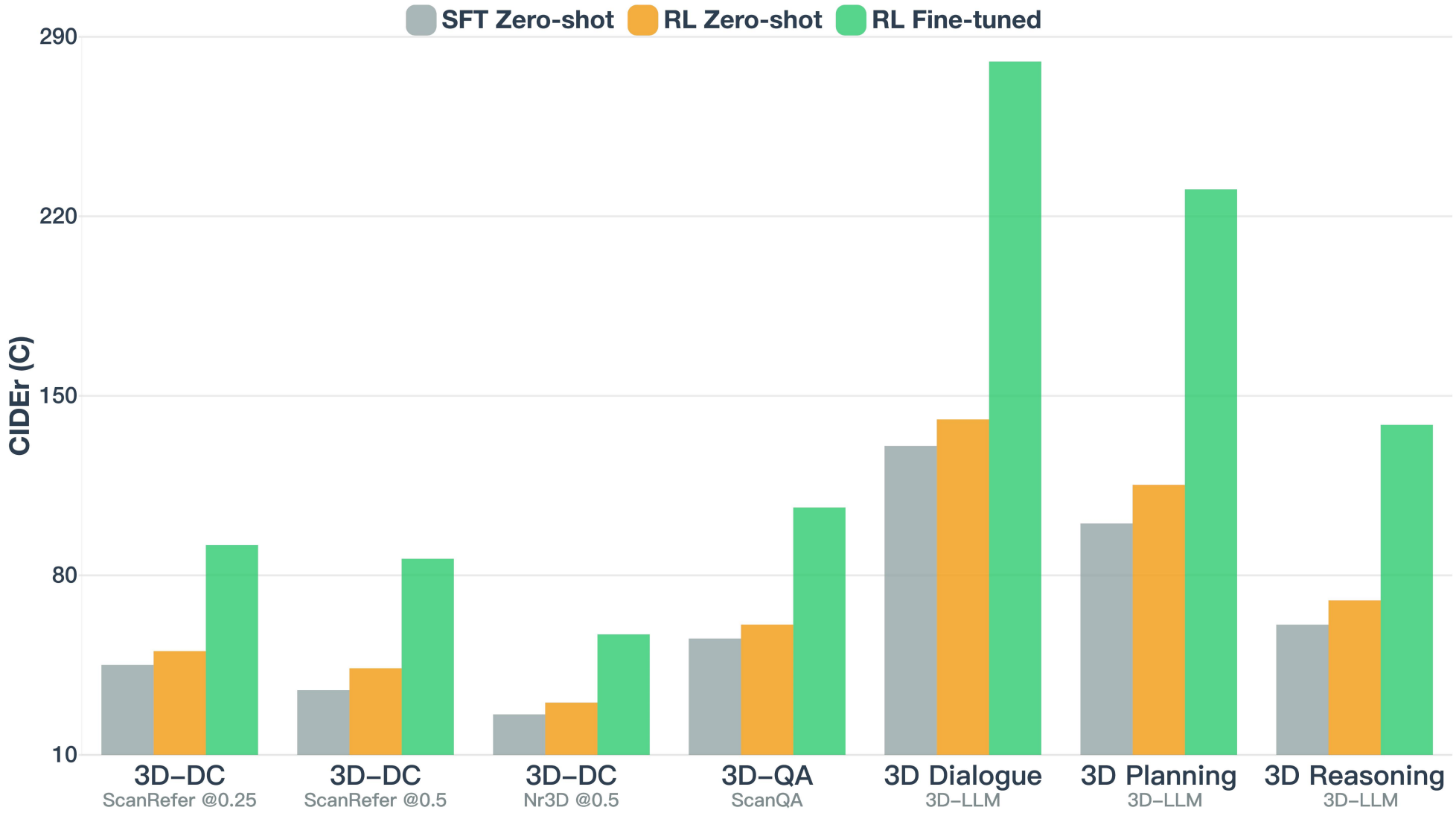
3D-R1 is a generalist model capable of handling various downstream tasks and applications in a zero-shot manner with incredible generalizability, significantly reducing the need for expensive adaptation.

Adjust Output: Reinforcement Learning with Verifiable Rewards (RLVR)



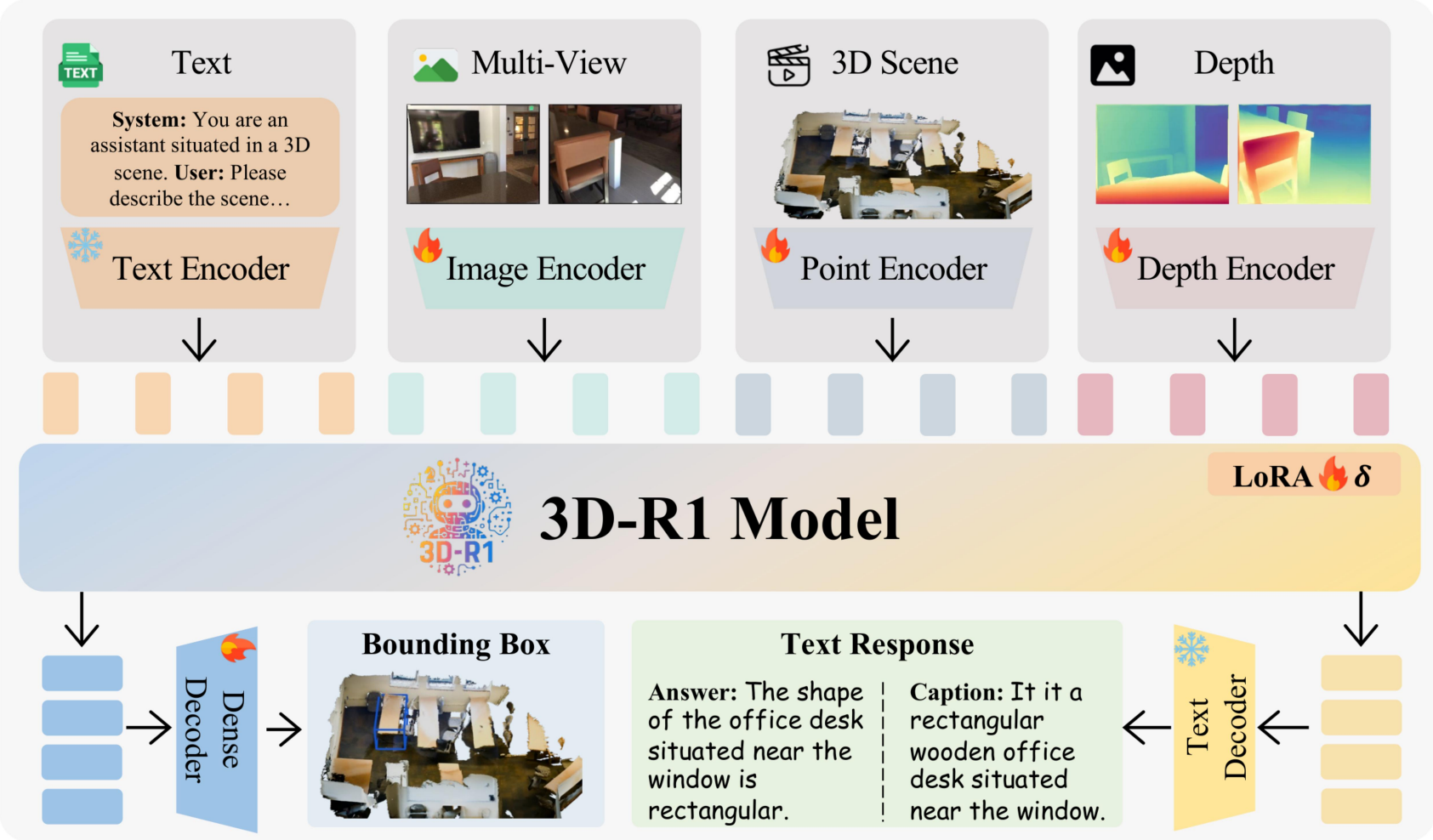
The policy model generates N outputs from a point cloud and question. Then perception IoU, semantic CLIP-similarity, and format-adherence rewards are computed, grouped, and combined with a KL term to a frozen reference model to update the policy.

Enhanced Reasoning: 3D-R1



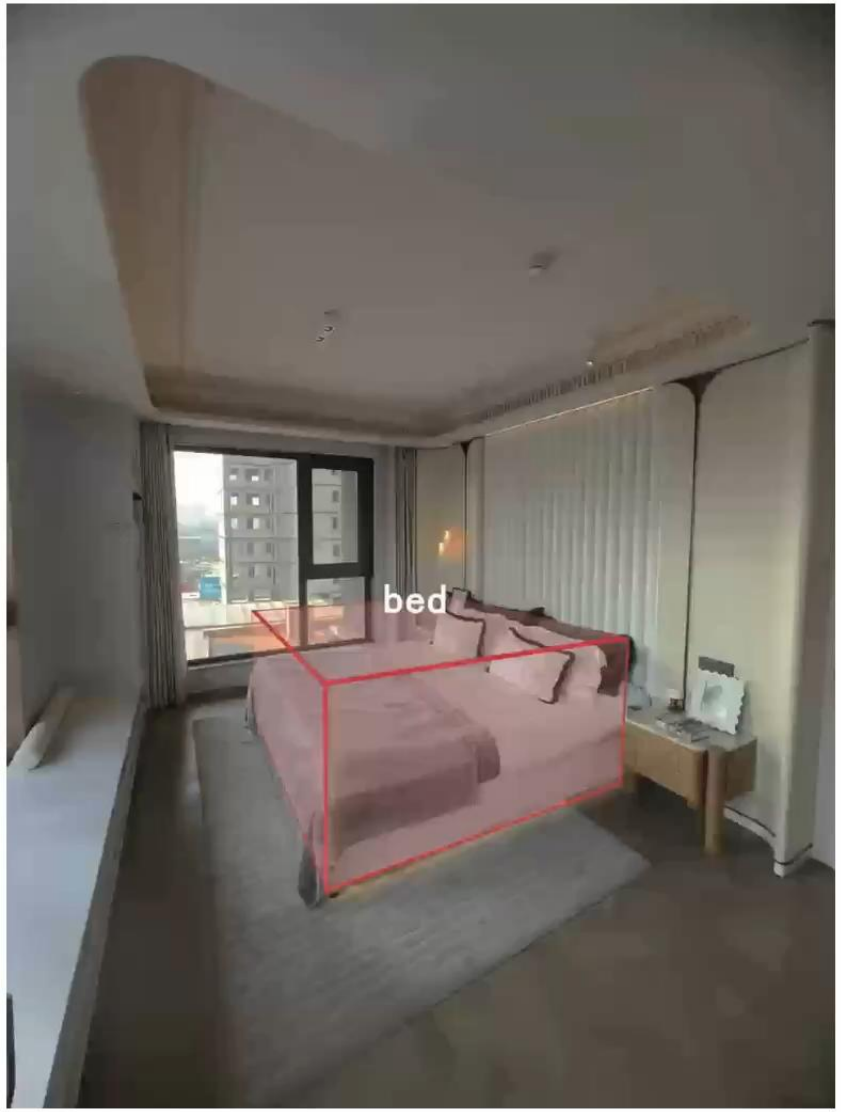
3D-R1 exhibits remarkable generalizability with enhanced reasoning capabilities.

Foundation Model: 3D-R1



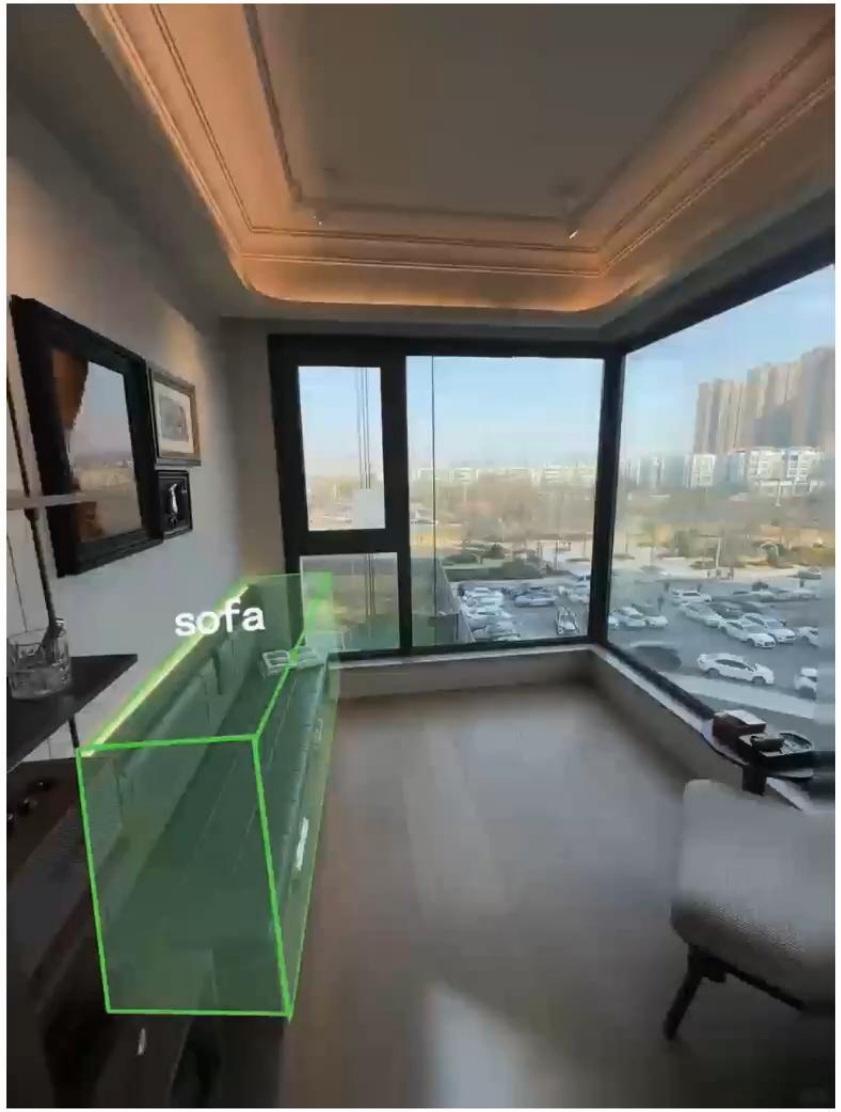
3D-R1 is an open-source generalist model that enhances the reasoning of 3D VLMs for unified scene understanding.

3D Scene Dense Captioning (3D-DC)



3D-DC

3D Object Captioning

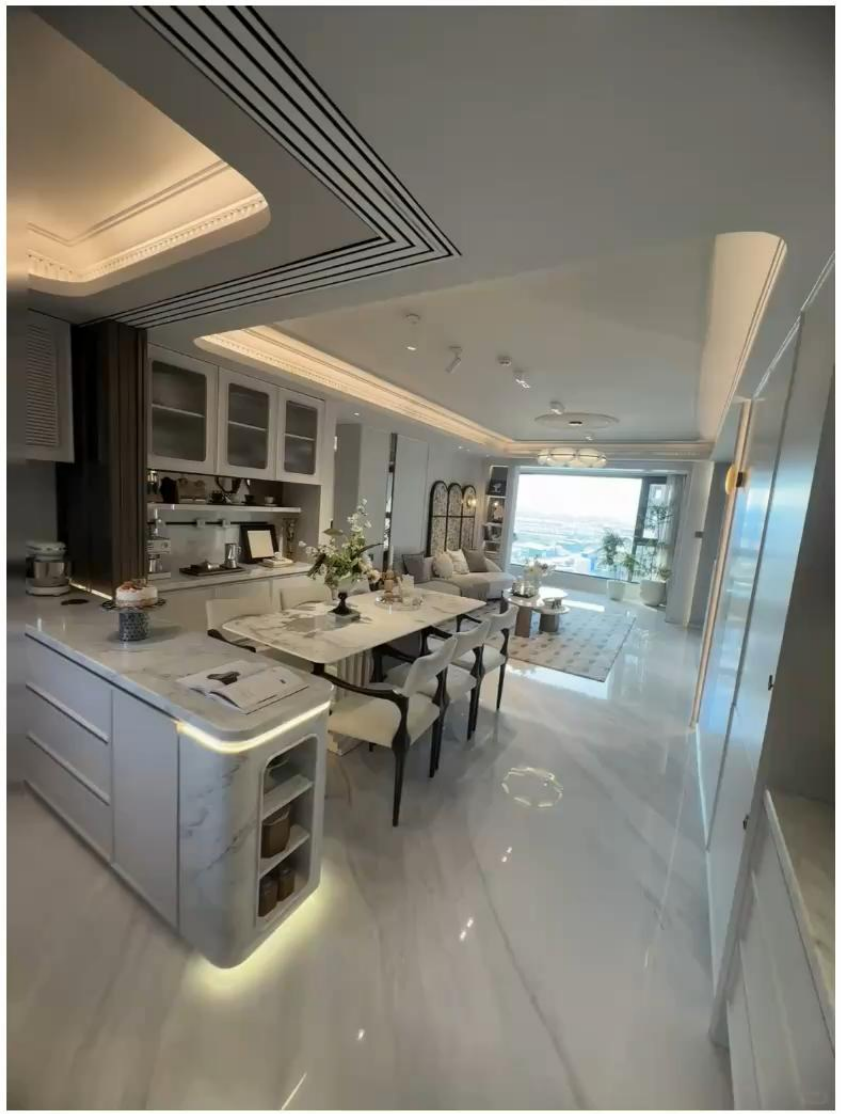


3D Object Captioning

3D Visual Grounding (3D-VG)

3D-VG

3D Question Answering (3D-QA)

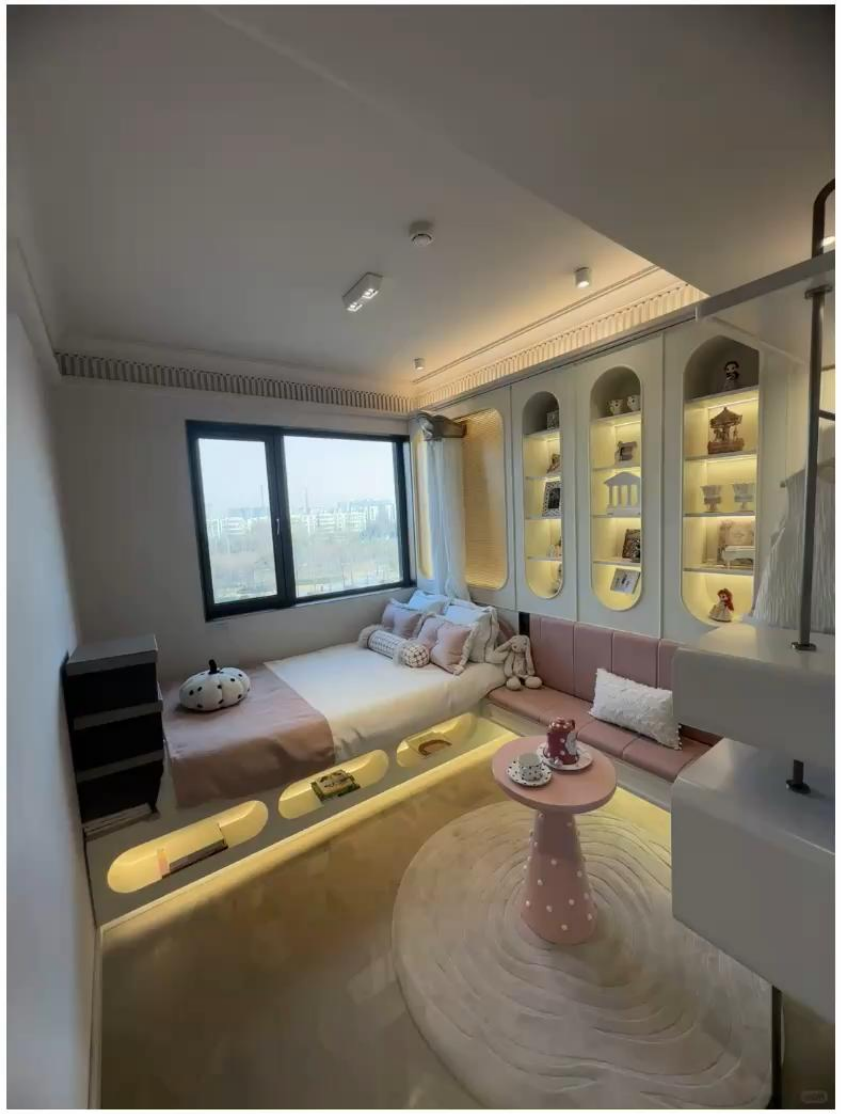


3D-QA

3D Dialogue

3D Dialogue

3D Reasoning

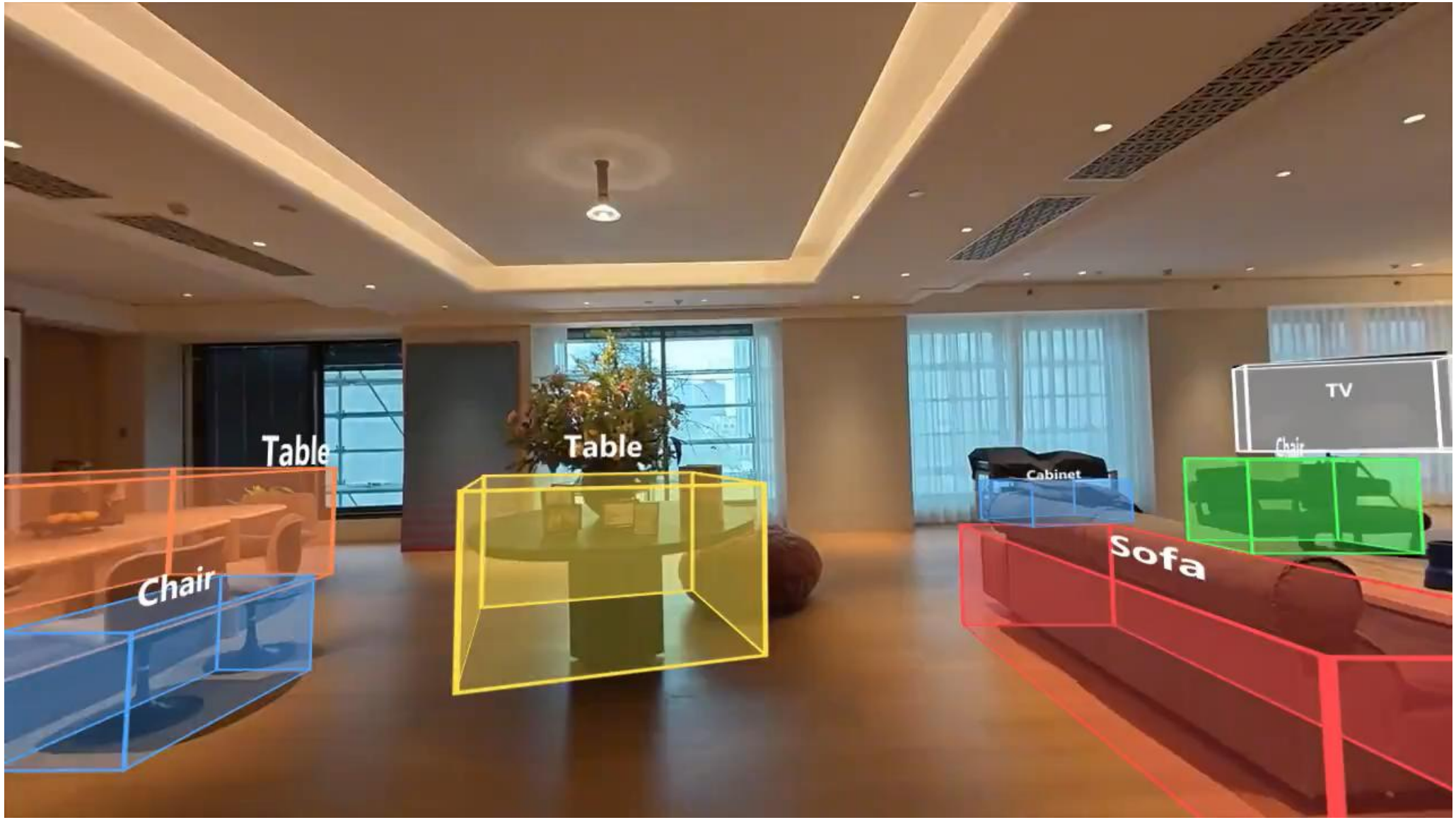


3D Reasoning

3D Planning

3D Planning

Zero-Shot Results



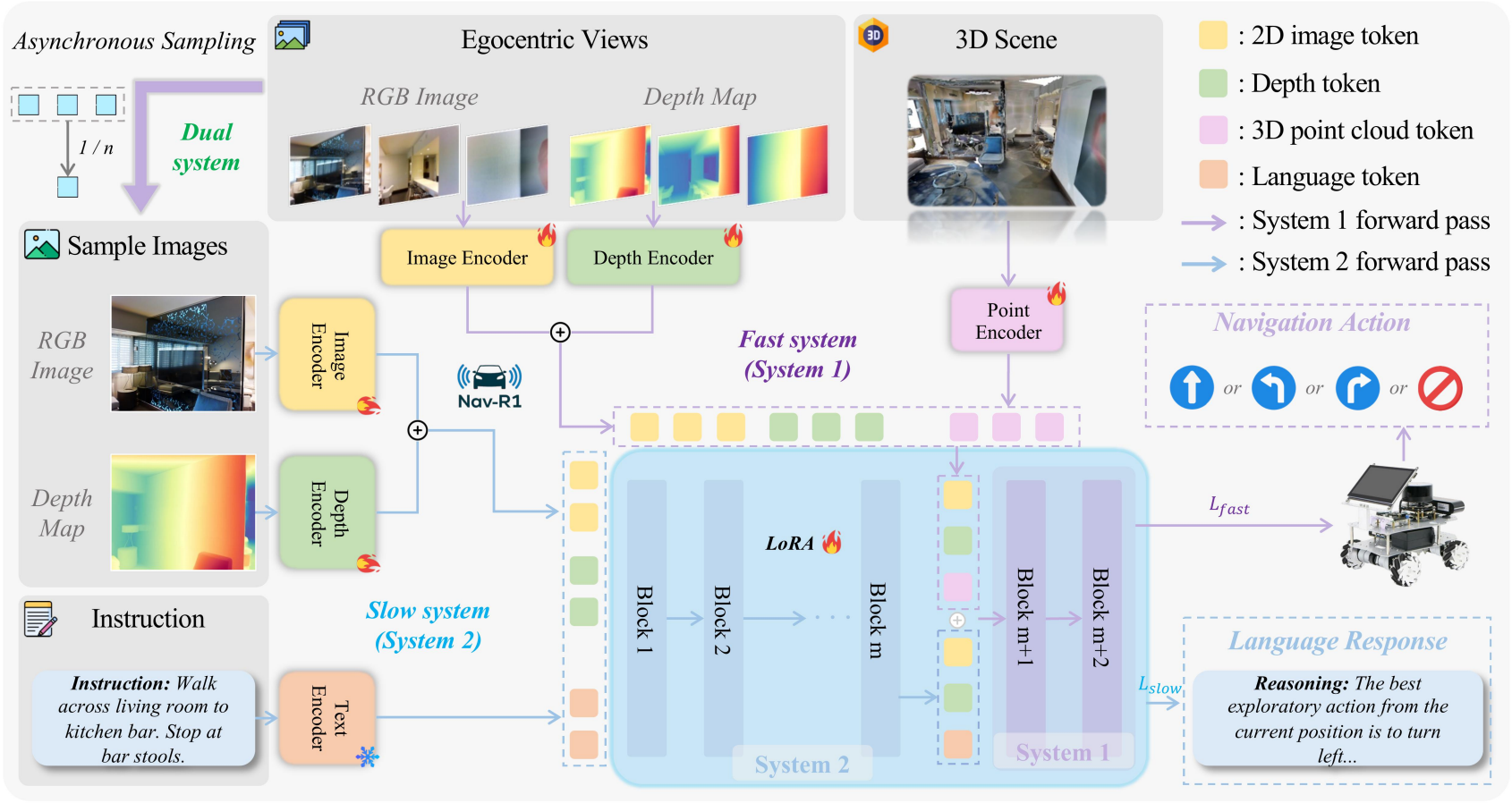
System and Memory: Nav-R1

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."

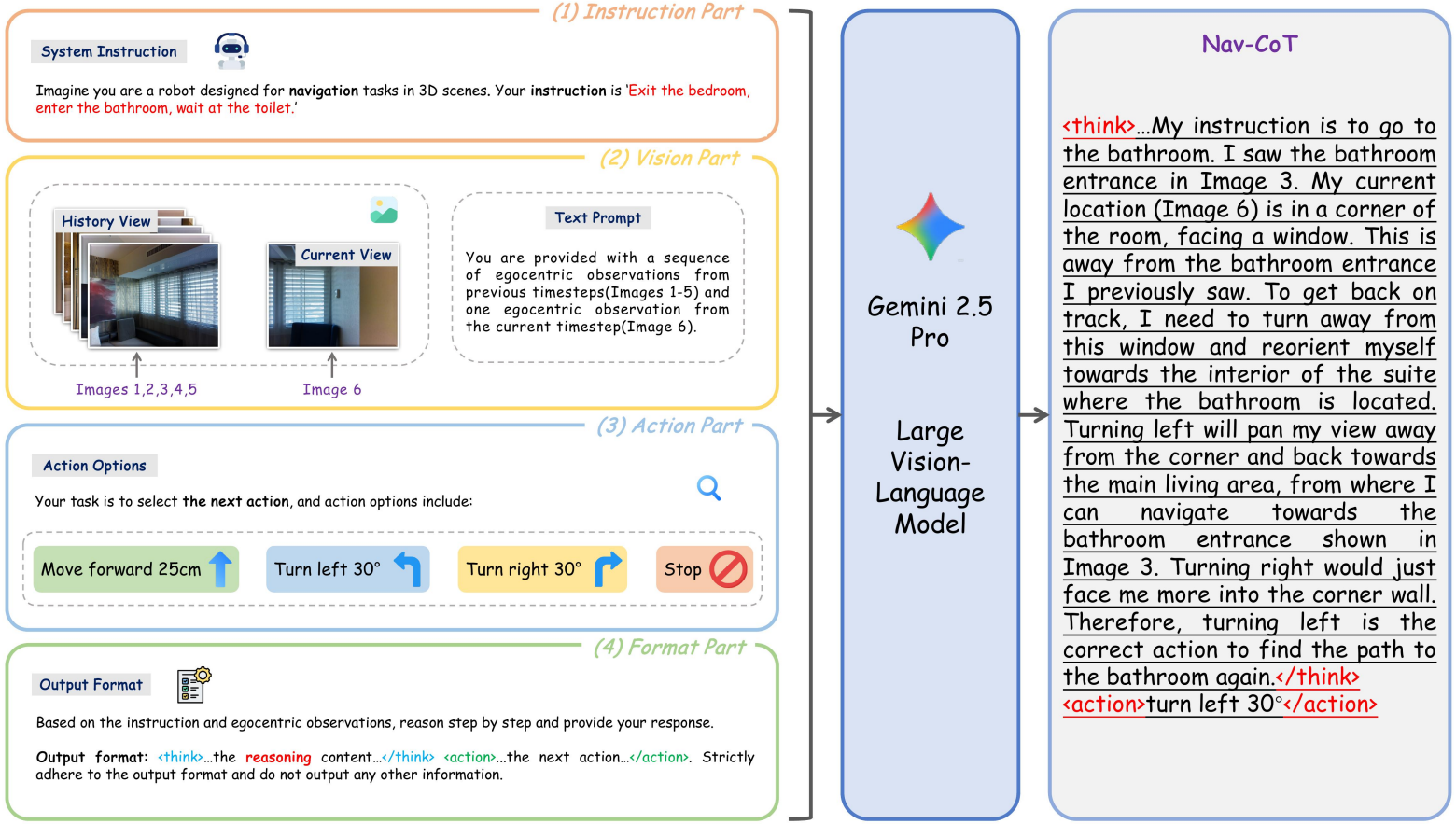
— Daniel Kahneman (Nobel Prize in Economics)

Fast-in-Slow: Nav-R1



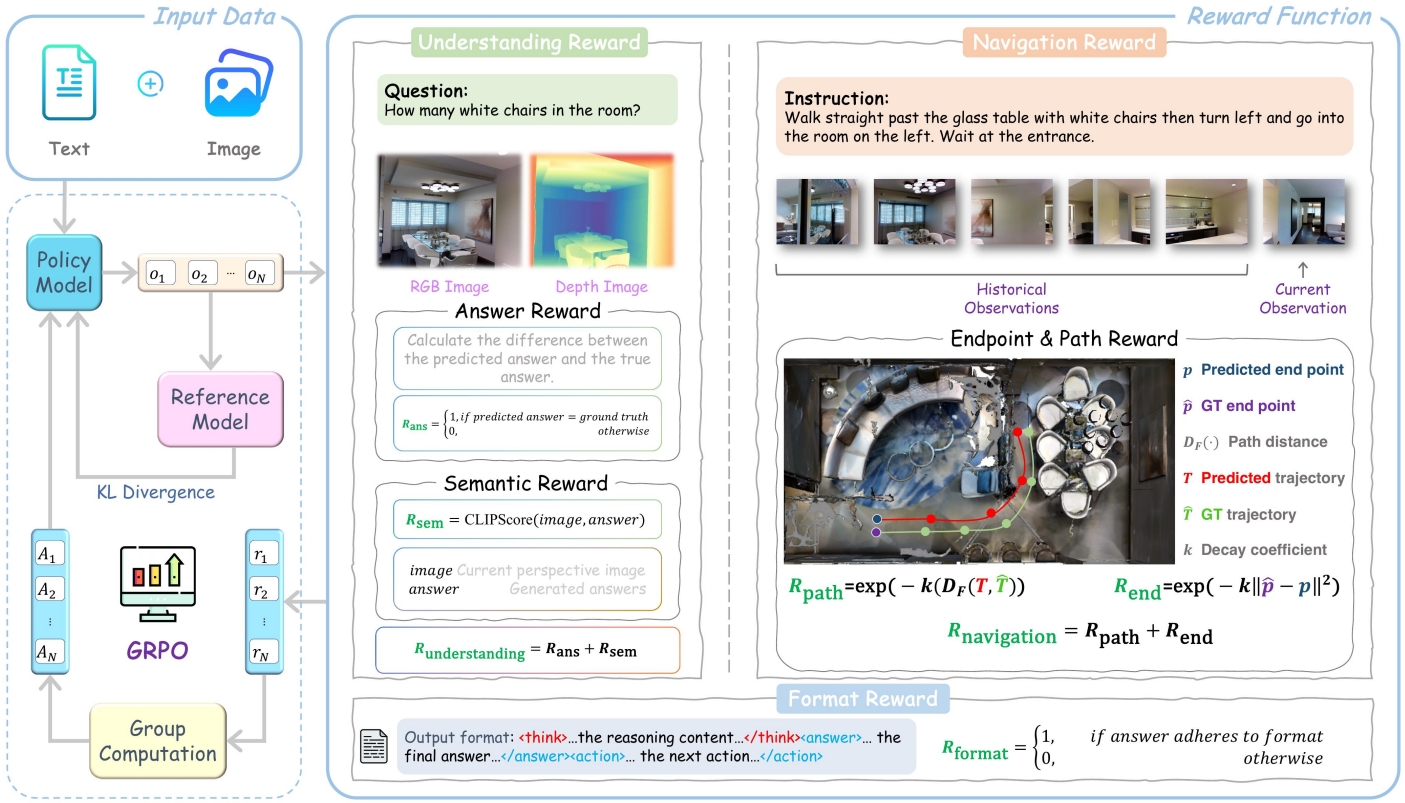
Nav-R1 features a Fast-in-Slow design that ensures rapid decision-making within long-horizon planning..

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	**39.45**	66.89	**55.34**	230.50	25.45	**48.34**	55.67	138.67	**23.56**	35.45	**60.02**
Nav-R1 (Ours)	**281.20**	39.34	**67.53**	55.12	**230.52**	**25.98**	47.11	**56.23**	**139.98**	23.20	**36.15**	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

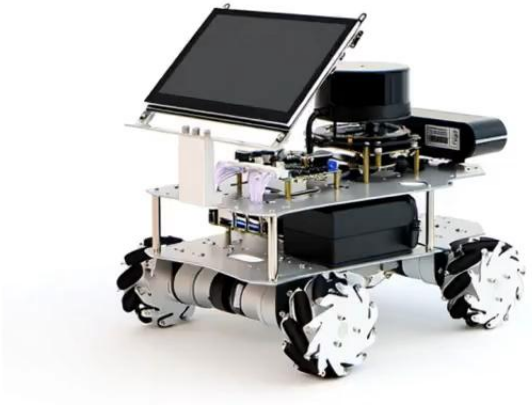
Nav-R1

北京大学 PEKING UNIVERSITY

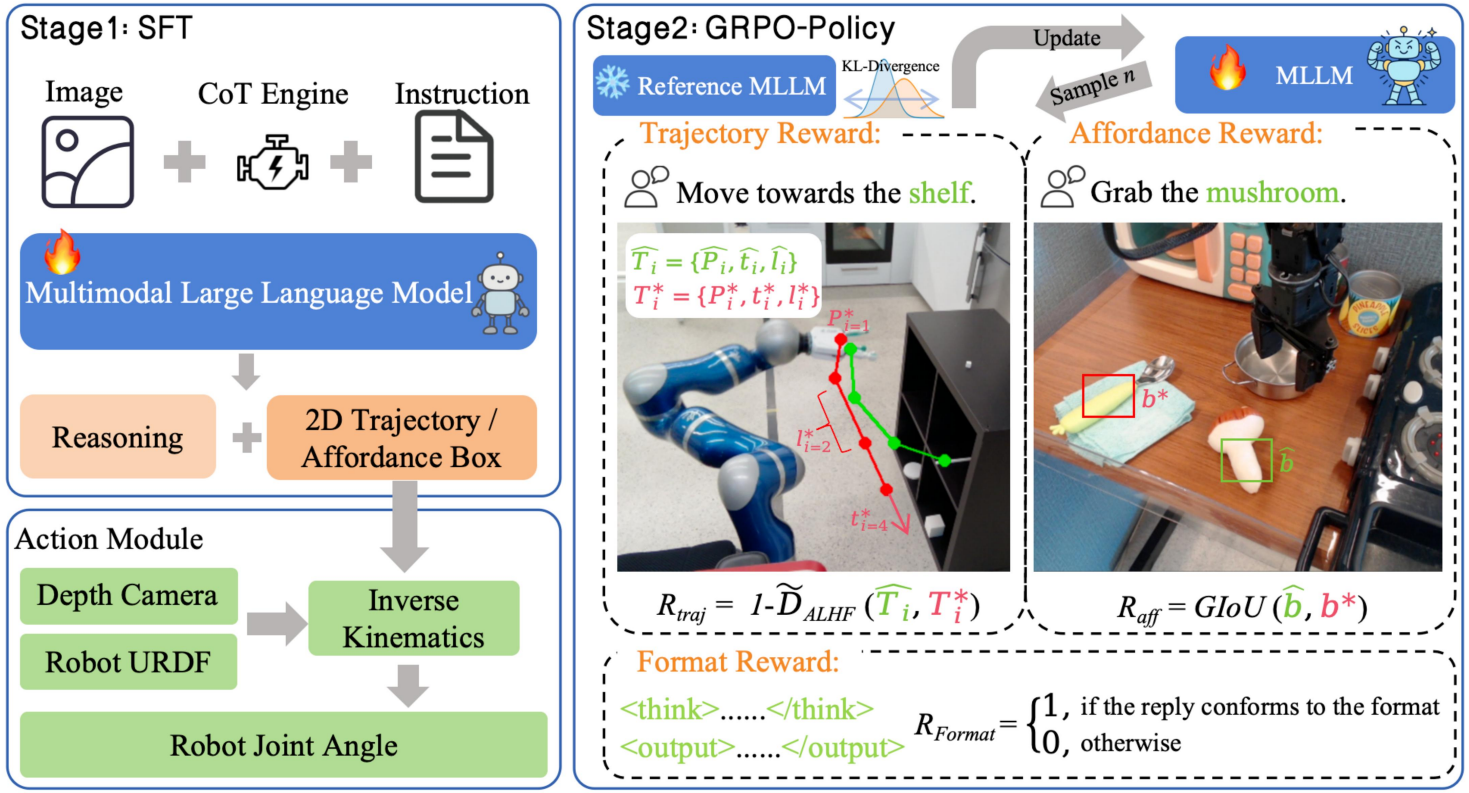
Nav-R1: Reasoning and Navigation in Embodied Scenes

Qingxiang Liu, Ting Huang, Zeyu Zhang, Hao Tang

Nav-R1

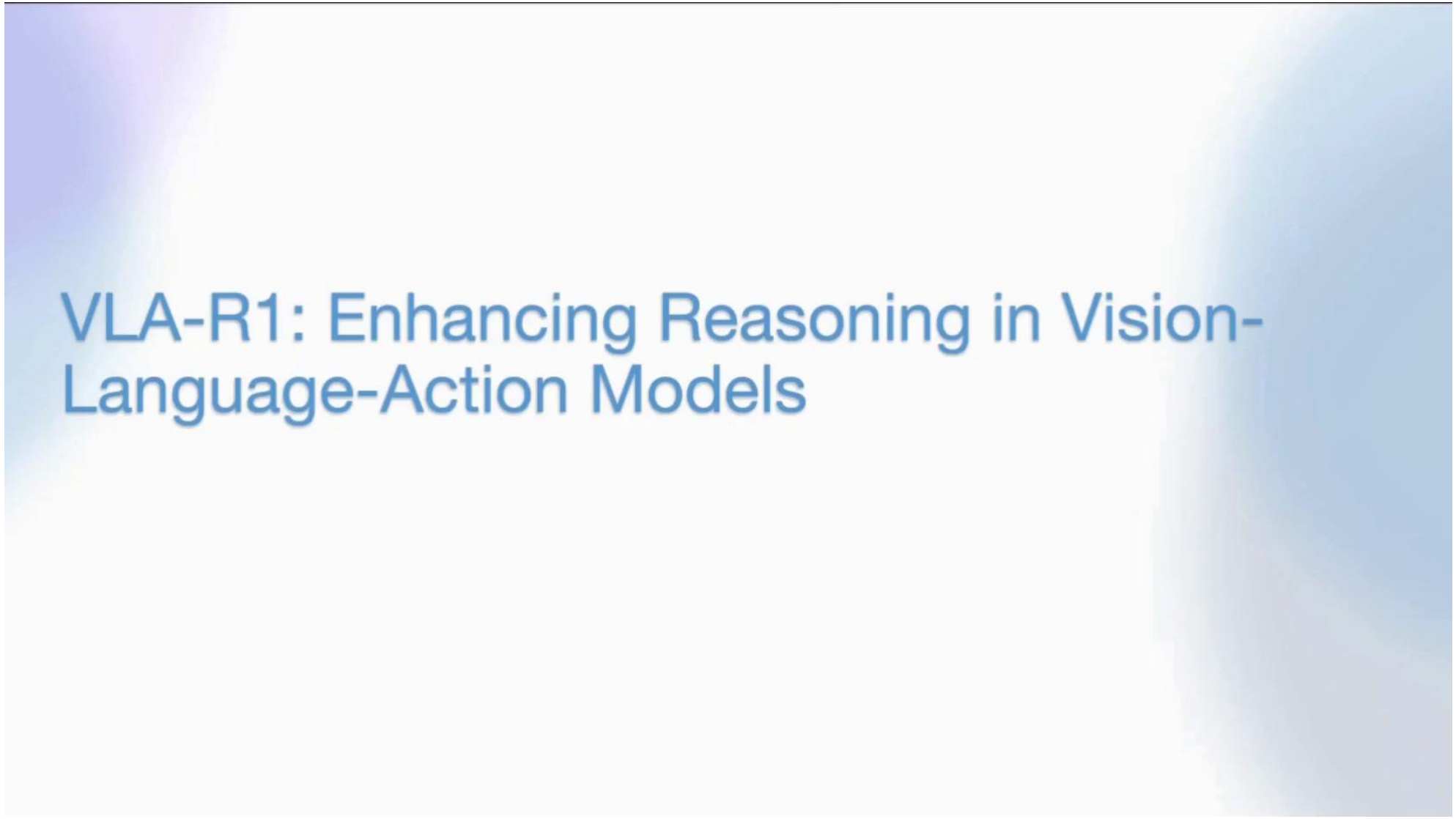


Similar Idea for Arm Robots: VLA-R1

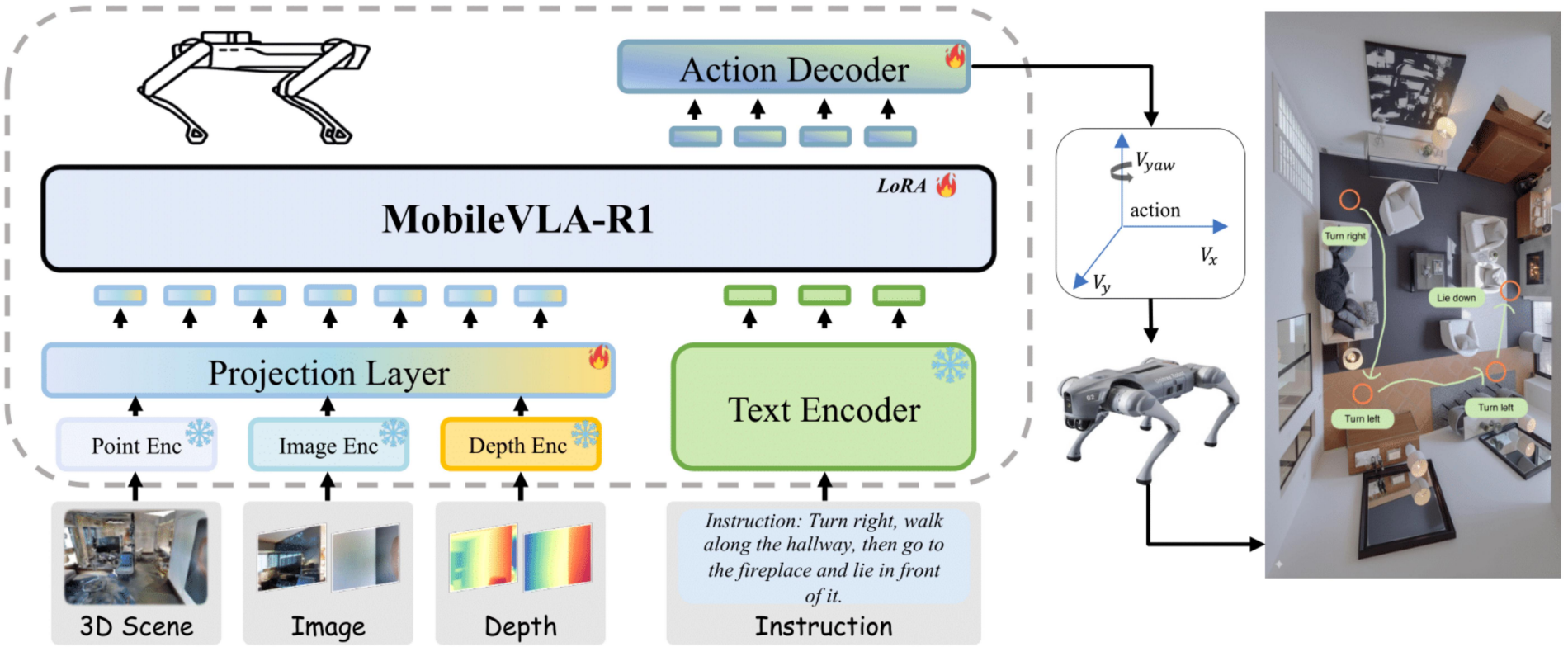


Training has two stages: **Stage 1** uses SFT with CoT supervision to learn reasoning over images and instructions; **Stage 2** refines reasoning and actions via RL with verifiable rewards (GRPO). **During inference**, a control stack converts outputs into joint-level robot commands.

Results: VLA-R1

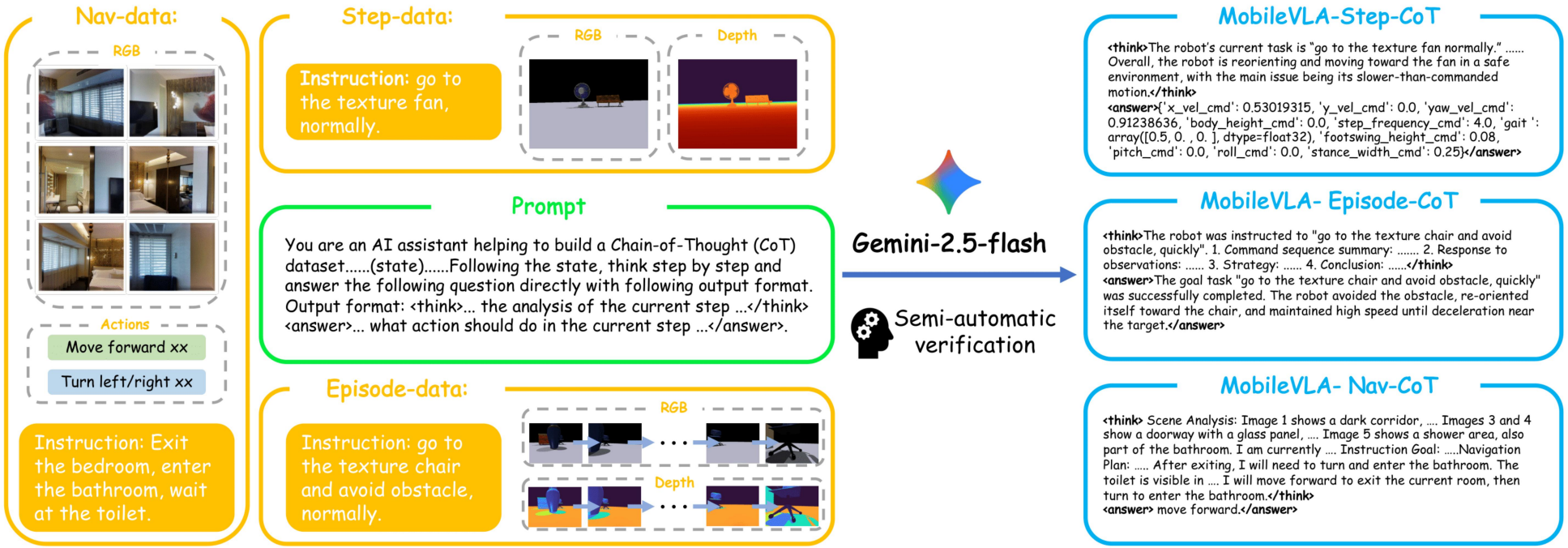


Similar Idea for Mobile Robots: MobileVLA-R1



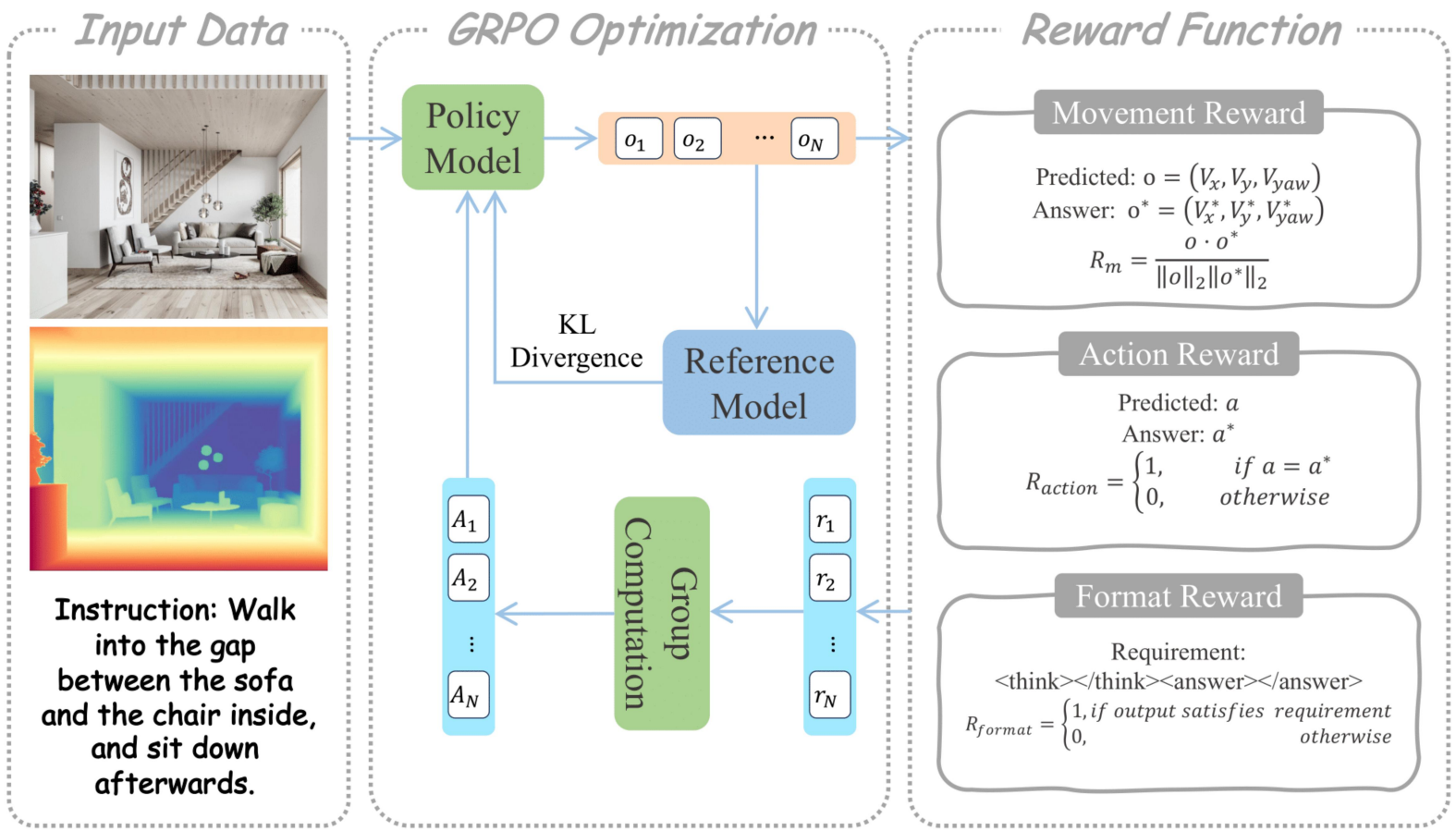
Architecture of MobileVLA-R1. MobileVLA-R1 is an end-to-end framework that integrates natural-language instructions with multimodal perception. It processes RGB, depth, and point cloud observations together with textual commands to generate continuous locomotion actions, enabling mobile robots to follow complex instructions and adapt to diverse environments in real time.

MobileVLA-R1 Data Engine



CoT Data Engine. We construct the MobileVLA-CoT-Episode-18K, Nav-38K, and Step-80K by defining navigation and step-level instructions, integrating RGB–Depth visual inputs, and specifying structured reasoning prompts. These inputs are fed into Gemini-2.5-Flash, which generates multi-granularity Chain-of-Thought (CoT) annotations with corresponding action outputs.

RLVR in MobileVLA-R1



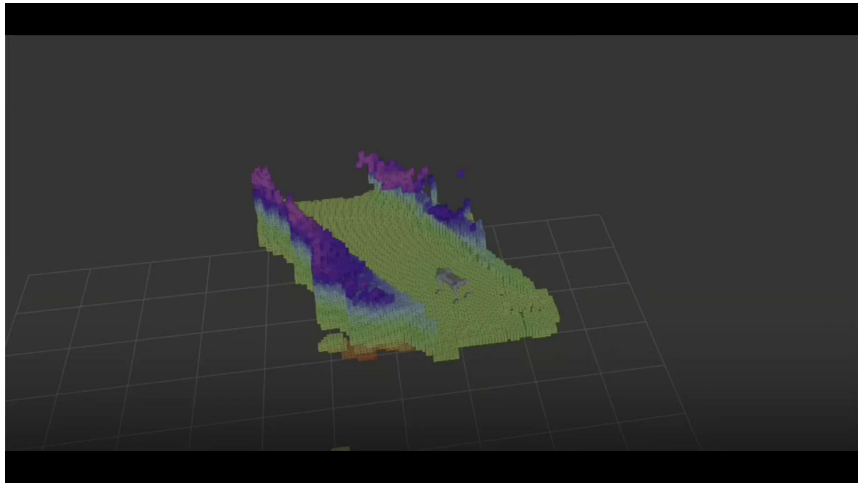
The pipeline of RL policy. The model generates N responses from a given input, rewards are then computed for each response. After normalizing and clipping, these rewards are conflated with a KL-divergence term, which prevents the model from over-updating, to update the policy.

Results: MobileVLA-R1

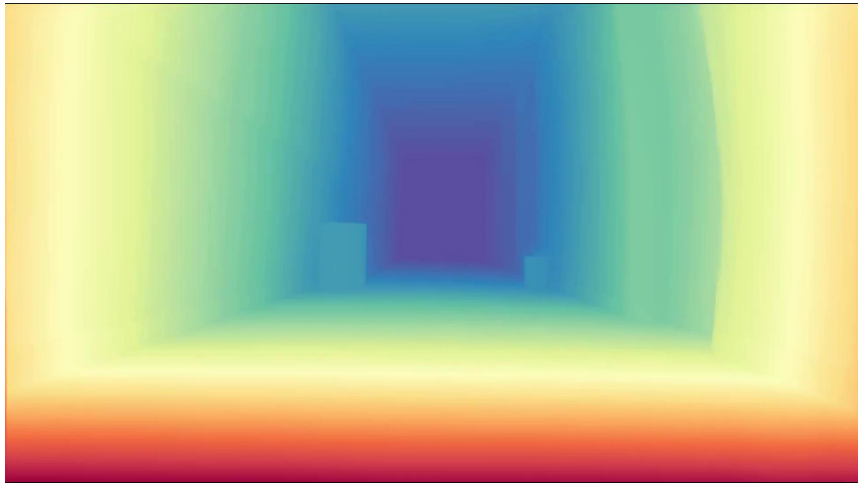
ExoView

EgoView

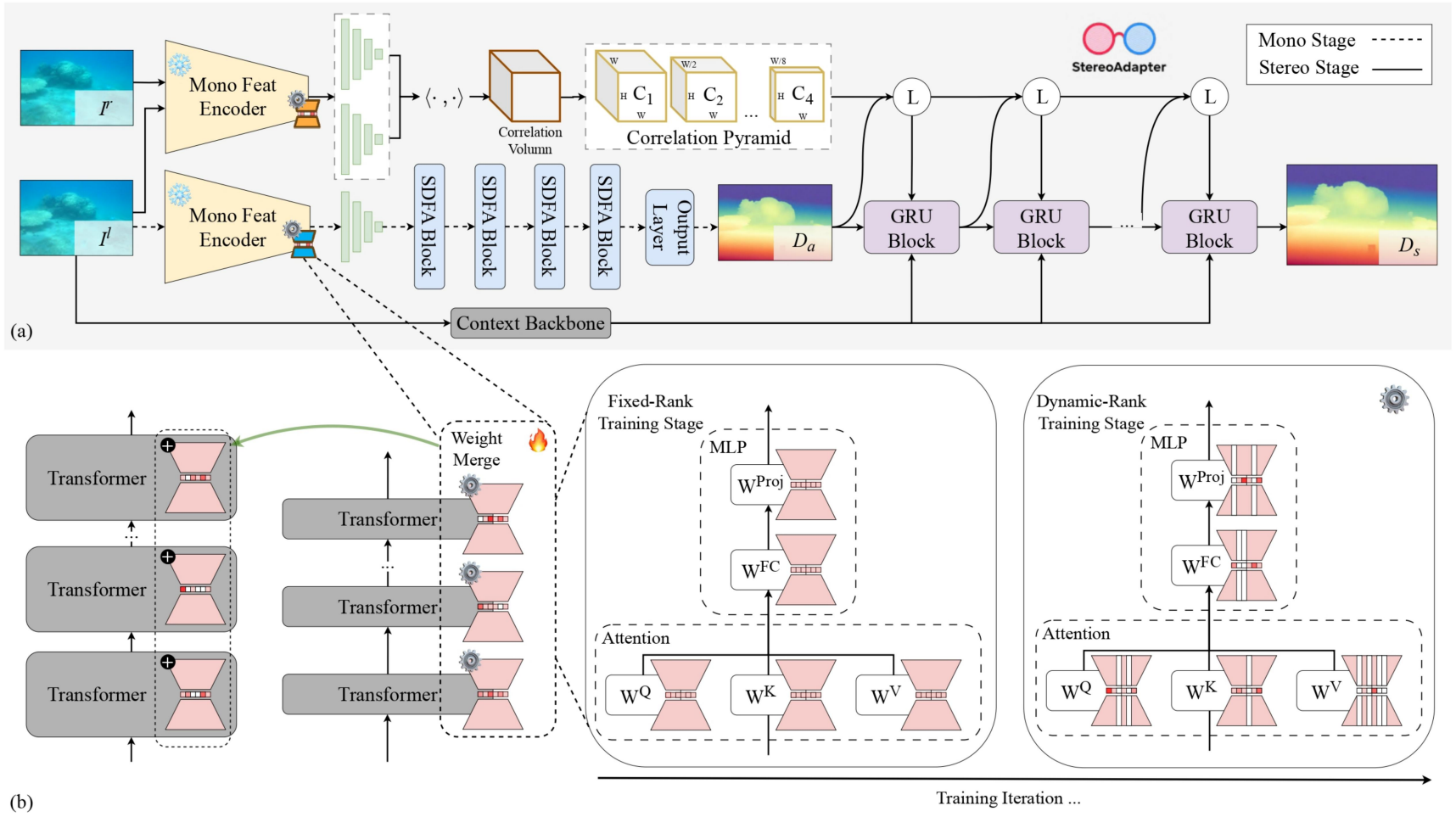
Point



Depth

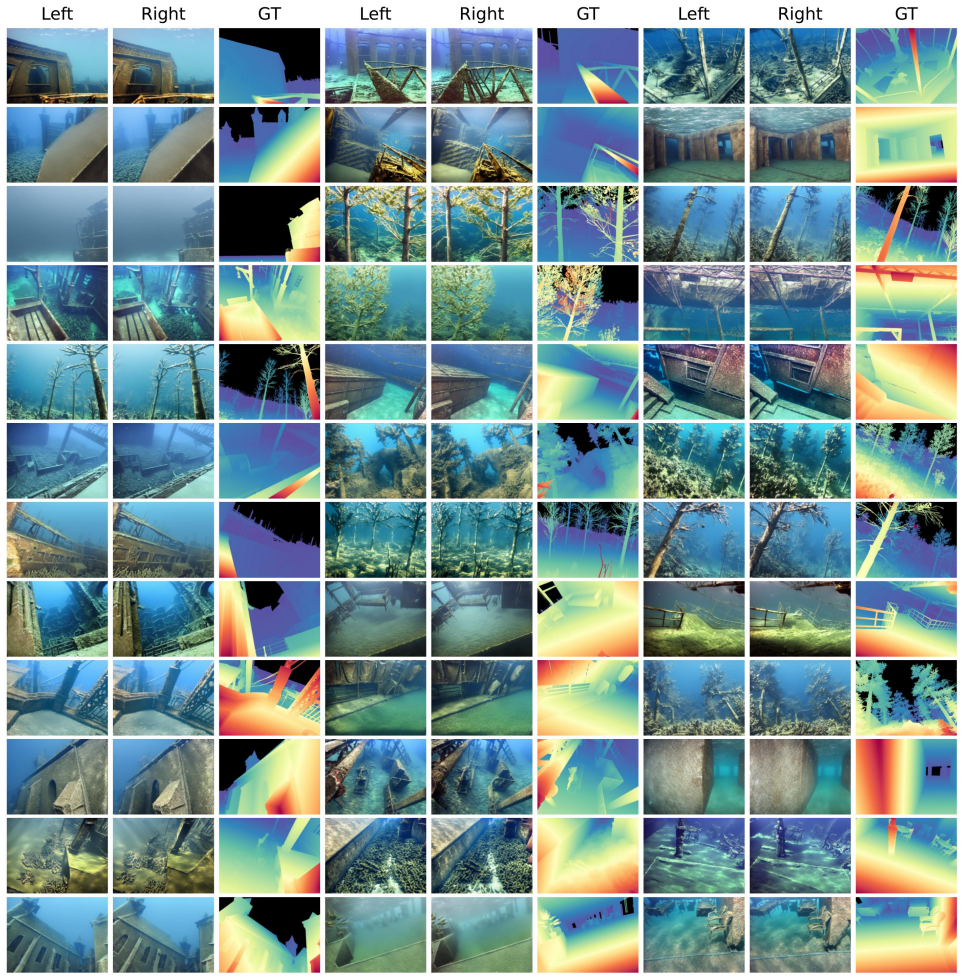


Bridging the Domain Gap in Post-Training: StereoAdapter (ICRA 2026)



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 (ICRA 2026)

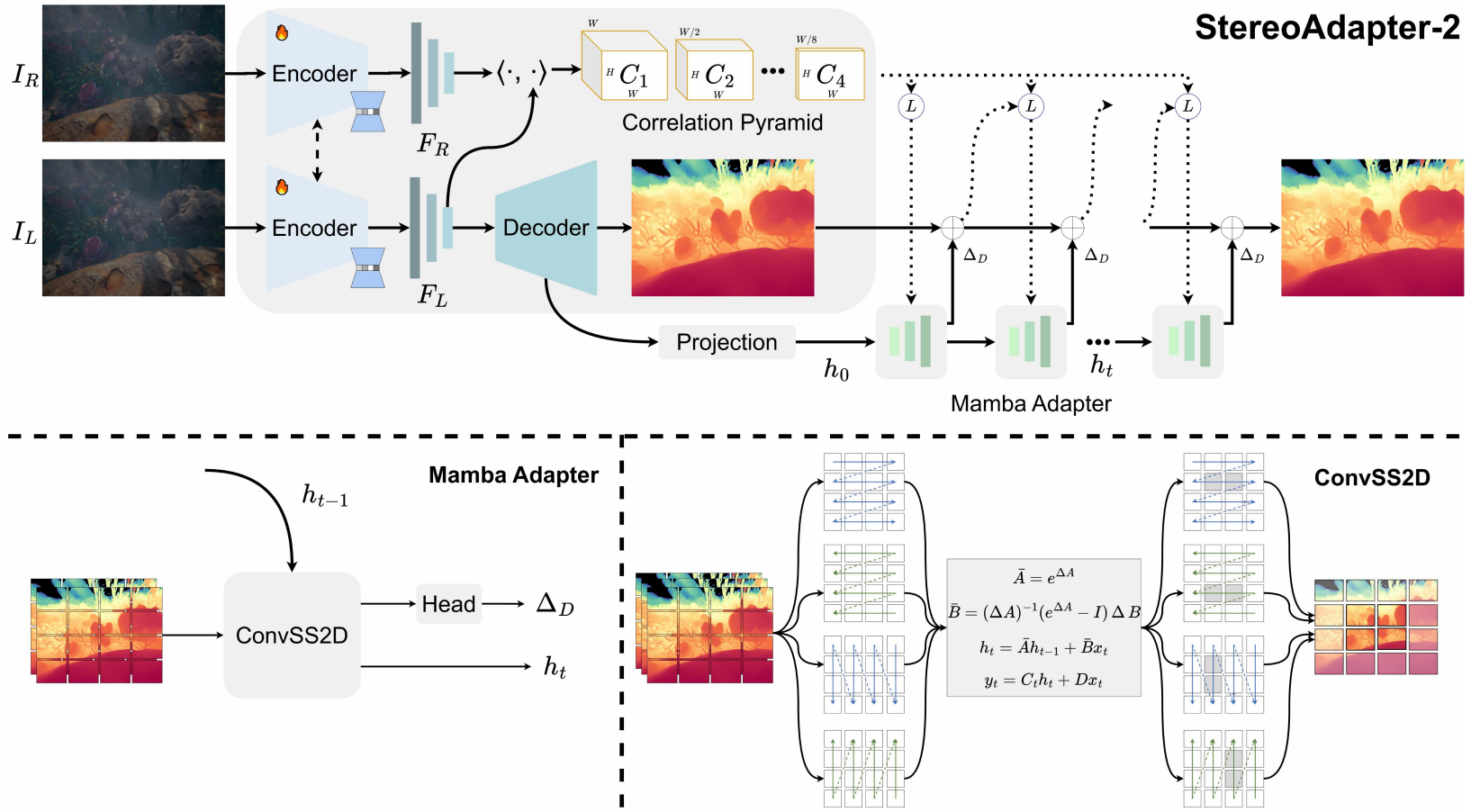
StereoAdapter

北京大学 PEKING UNIVERSITY

StereoAdapter: Adapting Stereo Depth Estimation to Underwater Scenes

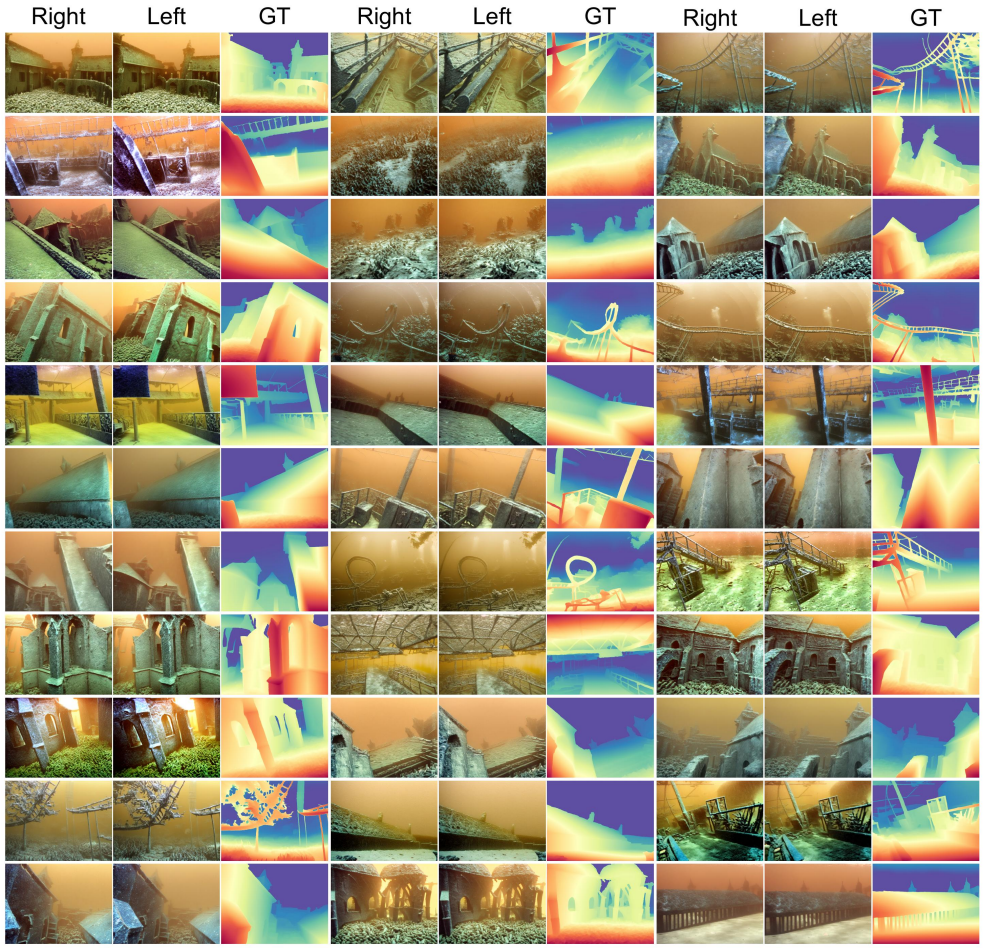
Zhengri Wu, Yiran Wang, Yu Wen, Zeyu Zhang, Biao Wu, Hao Tang

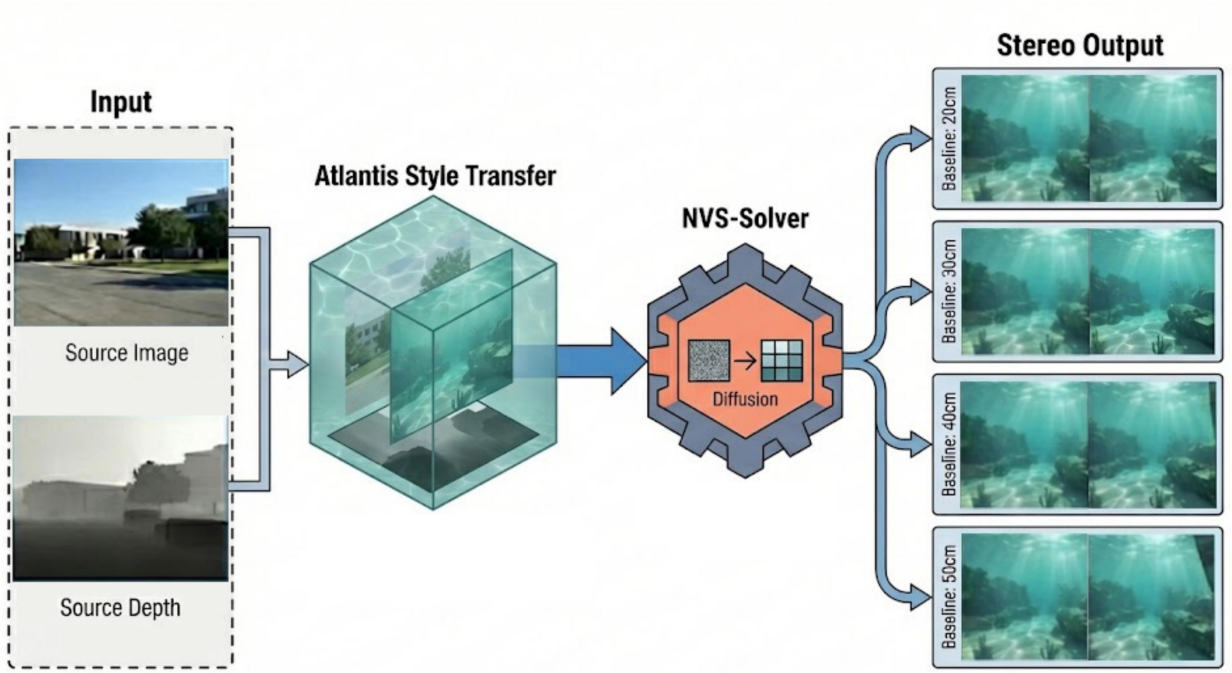
Re-design Decoders & Scaled-Up Synthetic Data: StereoAdapter-2 (2026)



Detailed architecture of the StereoAdapter-2. Our model iteratively refines disparity by integrating a Mamba Adapter. The refinement step is powered by the ConvSS2D operator, which enables adaptive and long-range spatial information propagation through multi-directional selective scanning.

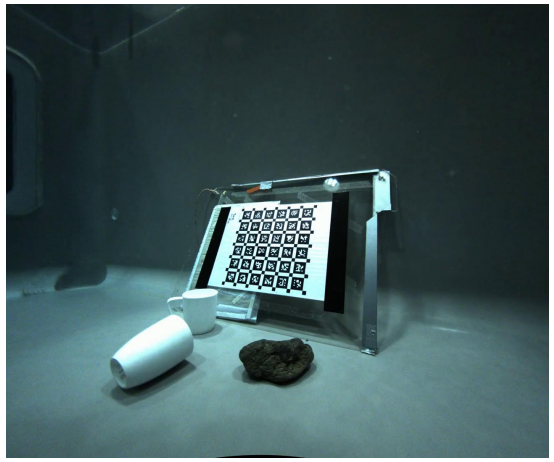
Scaled-Up Synthetic Data: UW-StereoDepth-80K



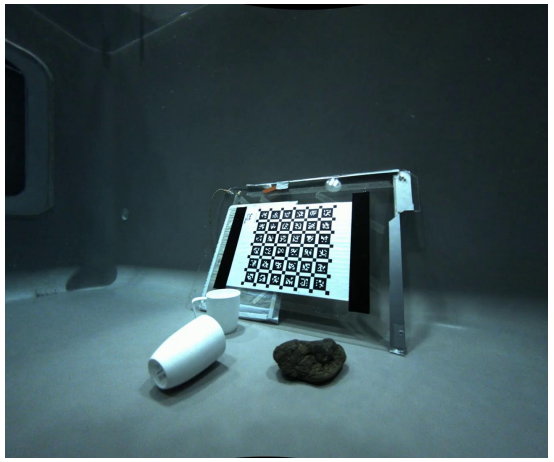


Data synthesis. Semantic-aware style transfer and geometry-consistent novel view synthesis rendering pipeline for the UW-StereoDepth-80K dataset.

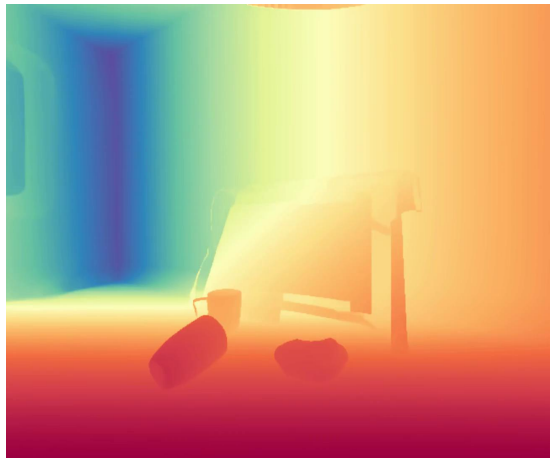
Results: StereoAdapter-2 (2026)



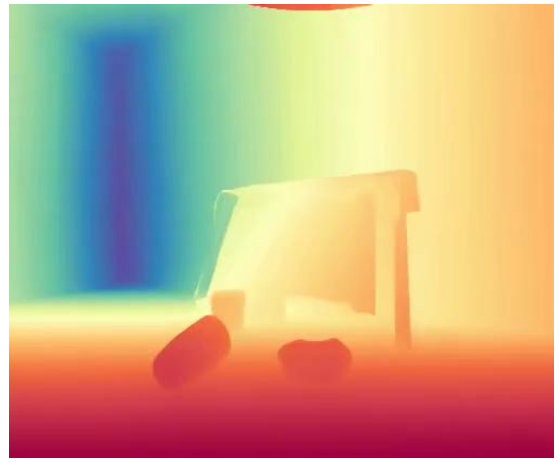
Right



Left



StereoAdapter-2



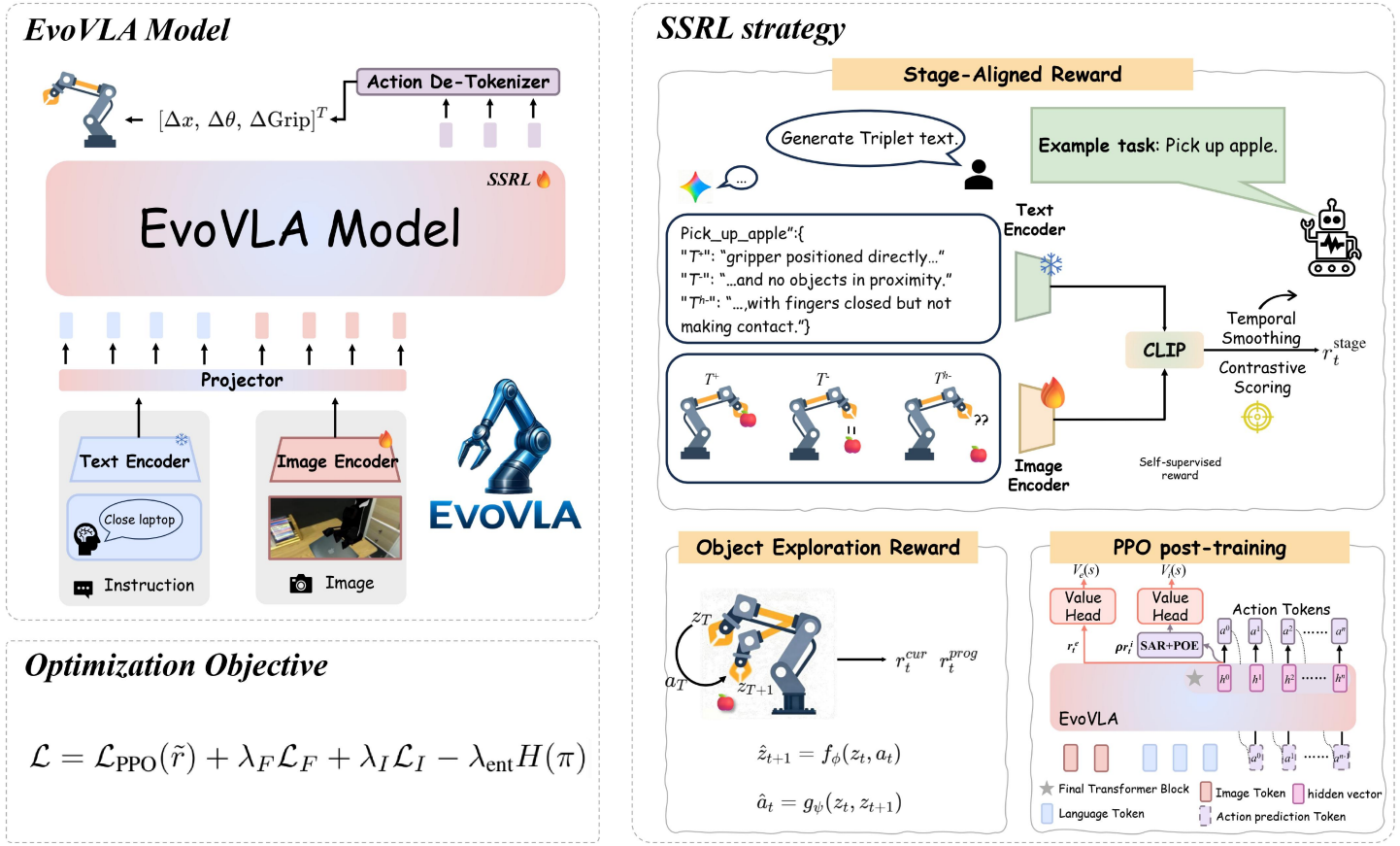
GT

Self-Evolving?

How can an embodied robot self-evolve when the reward function is misspecified?

- Online Reinforcement Learning, but still requires specified reward functions or goals...
- Self-Supervised Reinforcement Learning (SSRL)

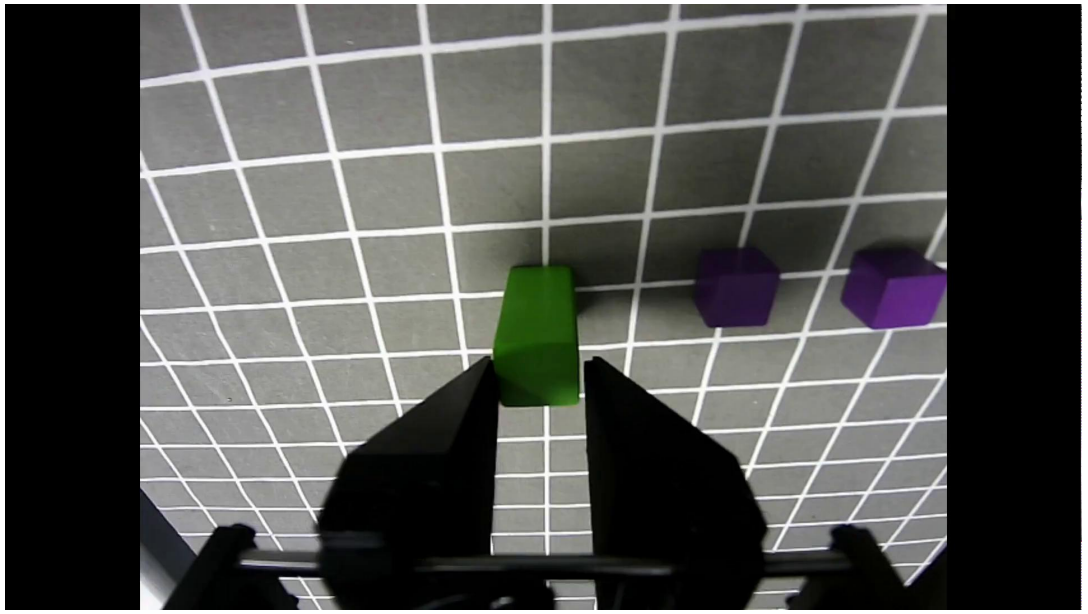
EvoVLA: Self-Supervised Reinforcement Learning (SSRL)



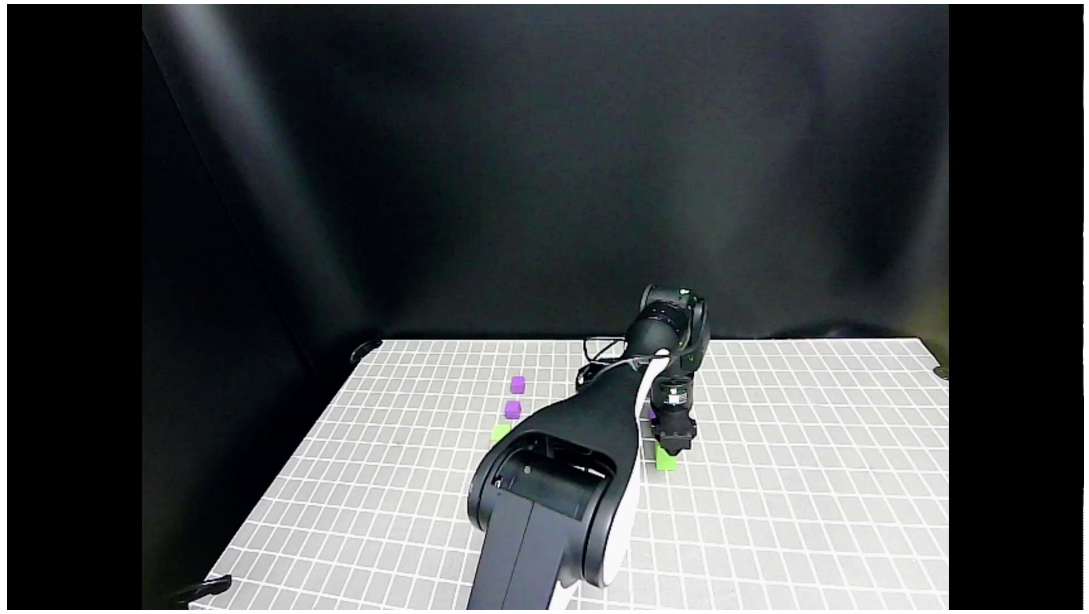
EvoVLA overview. Built on OpenVLA-OFT backbone, EvoVLA integrates three modules: Stage-Aligned Reward (SAR) with hard negatives and temporal smoothing, Pose-Based Object Exploration (POE) via world models, and Long-Horizon Memory with context selection and gated fusion. The framework couples with Discoverse-L for training and deploys to real robots.

Results: EvoVLA

Eye-in-hand



Eye-to-hand



build a bridge with green bars and fill with purple blocks

Takeaways

- Synthetic data and data-driven methods are the key to achieving scalability and generalizability.
- Do not abuse reinforcement learning for post-training; use RL only to adjust the foundation model's output.
- 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