D3D-VLP: Dynamic 3D Vision-Language-Planning Model for Embodied Grounding and Navigation

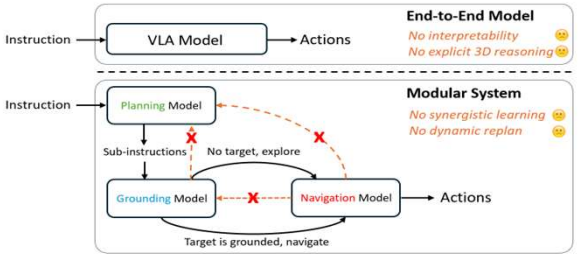
Zihan Wang[1], Seungjun Lee[1], Guangzhao Dai[2], Gim Hee Lee[1]
[1]National University of Singapore, [2]Nanjing University of Science and Technology

NUS National University of Singapore | School of Computing

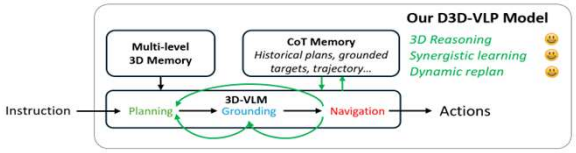
CVPR JUNE 3-7, 2026 DENVER COLORADO

1. The Embodied AI Dilemma

- **End-to-End Models:** Directly map instructions to actions, but lack interpretability and explicit 3D reasoning processes.
- **Modular Systems:** Assemble multiple specialized components, but ignore cross-component synergies and fail to dynamically replan based on real-time feedback.



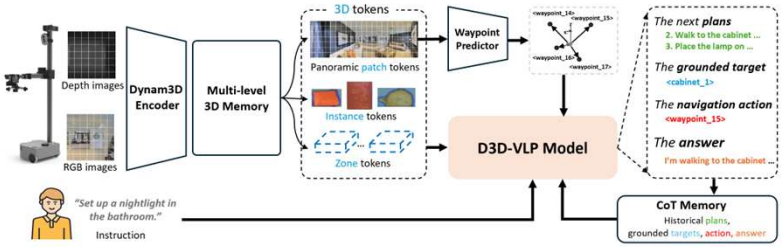
2. Our Solution: D3D-VLP



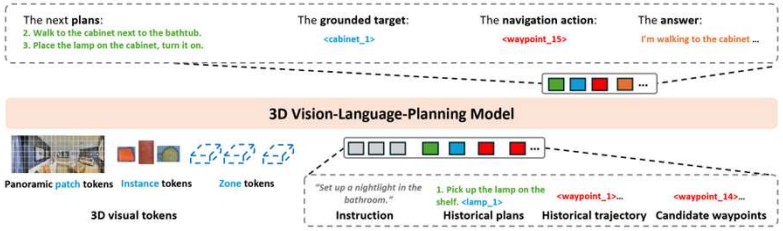
- Bridge the gap between interpretable modular systems and high-performance end-to-end models
- Unifies mluti-step planning, 3D grounding, and navigation within a single 3D-VLM + Chain-of-Thought pipeline.
- Supports explicit 3D reasoning, synergistic learning, and dynamic replanning

3D Reasoning | Synergistic Learning | Dynamic Replan

3. Dynamic 3D Chain-of-Thought (3D CoT)



- **Unified autoregressive model:** Planning, grounding, and navigation.
- **CoT Memory Loop:** Historical plans, grounded targets, and agent trajectories are fed back into the model's context.
- **Stateful Replanning:** Explicit memory makes the agent stateful, allowing perform dynamic replanning if a target is missing or a plan is blocked.



4. Synergistic Learning from Fragmented Supervision (SLFS)

Data Type	Planning	Grounding	Navigation	# Samples
1	✓	✓	✓	175K
2	✓	✓	×	161K
3	✓	×	✓	14K
4	×	✓	✓	5.8M
5	×	✓	×	2.3M
6	×	×	✓	1.6M

- **Hybrid Dataset:** A 10M-sample dataset containing only 175K fully annotated "gold" samples alongside 9.9M partially annotated samples.
- **Masked Autoregressive Loss:** Gradients from available annotations back-propagate through the shared 3D-VLM to implicitly supervise missing components.
- **Mutual Reinforcement:** Different CoT components mutually reinforce and implicitly supervise each other during training.

5. State-of-the-Art Performance

VLN Benchmarks (Higher is better)

	R2R-CE		REVERIE-CE		NavRAG-CE		HM3D-OVON	
	SR ↑	SPL ↑	SR ↑	SPL ↑	SR ↑	SPL ↑	SR ↑	SPL ↑
D3D-VLP (Ours)	**61.3**	**56.1**	**47.5**	**34.7**	**31.1**	**23.9**	**47.3**	**30.4**
Dynam3D	52.9	45.7	40.1	28.5	24.7	18.8	42.7	22.4
StreamVLN	56.9	51.9	–	–	–	–	–	–
InternVLA-N1	58.2	54.0	–	–	–	–	–	–

SG3D Long-Horizon Benchmark (Task-Level)

Method	Navigation			Grounding	
	t-SR ↑	s-SR ↑	SPL ↑	s-ACC ↑	t-ACC ↑
D3D-VLP (Ours)	**13.8**	**33.7**	**21.6**	**28.3**	**9.3**
Dynam3D-VisTA	9.3	26.4	15.4	21.4	4.2

121% relative improvement in SG3D task-level accuracy over Dynam3D-VisTA (4.2 → 9.3)!

Ablation: Why CoT Memory Matters

Setting	R2R-CE Nav.	SG3D Grounding (t-ACC↑)	
	SPL ↑	s-ACC ↑	t-ACC ↑
All pipeline (full)	**56.1**	**28.3**	**9.3**
w/o CoT Memory	**48.7**	**19.4**	**4.1**

w/o CoT Memory: SG3D t-ACC drops from **9.3** to **4.1**

Navigation SPL drops from **56.1** to **48.7**

This verifies the importance of the explicit memory feedback loop for long-horizon embodied tasks.

6. Real-world Manipulation

We validated D3D-VLP in zero-shot real-world environments using a Hello Robot Stretch 3.

The agent successfully managed long-horizon tasks requiring sub-tasks like navigation, grounding, grasping, and placing target objects.