ICML 2026 POSTER

DynVLA: Learning World Dynamics for Action Reasoning

Compact Dynamics Chain-of-Thought for safer and lower-latency autonomous driving VLA.

Shuyao Shang* · Bing Zhan* · Yunfei Yan · Yuqi Wang · Yingyan Li · Yasong An · Xiaoman Wang · Jierui Liu · Lu Hou · Lue Fan · Zhaoxiang Zhang · Tieniu Tan

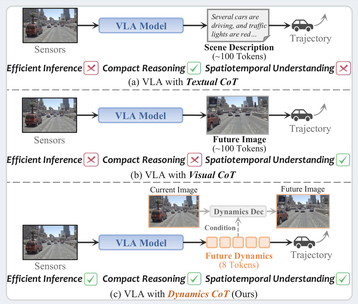
CASIA · UCAS · YINWANG Intelligent Technology · Nanjing University | * Equal contribution

Project Page

1 Why Reason About Dynamics?

Driving decisions depend on fine-grained spatiotemporal evolution. Textual CoT is too symbolic for physical interaction, while Visual CoT predicts redundant pixels and slows online inference in closed-loop tests.

Key idea: reason over compact world dynamics instead of long language traces or dense future images.



Dynamics CoT keeps planning-relevant future structure without dense future-frame generation.

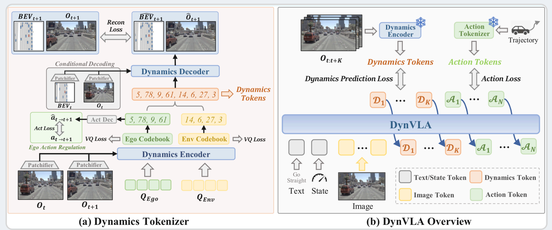
2 CoT Modes & Output

CoT	Grounding	Latency	Signal
Textual	coarse	3.04s	logic
Visual	dense	2.29s	pixels
Dynamics	**future**	**0.37s**	**tokens**

OBSERVE current + past front-view images → **THINK** dynamics tokens for future evolution → **ACT** FAST action token sequence

SFT teaches dynamics tokens before action tokens; RFT then optimizes the policy with trajectory and format rewards while preserving a valid output protocol.

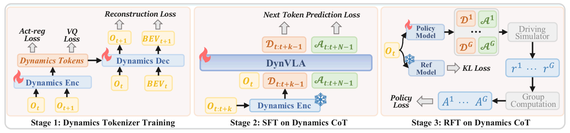
3 DynVLA Architecture ★ KEY



Dynamics Tokenizer. Adjacent observations become ego-centric and environment-centric dynamics tokens.

- **Ego branch:** regularized by ground-truth ego motion.
- **Environment branch:** captures surrounding agents and scene changes.
- **Cross-view consistency:** tokens jointly predict image futures and BEV trajectory states.

4 Training Pipeline



Four stages. Learn a tokenizer, train Dynamics CoT with SFT, then refine actions with GRPO-based RFT.

4+4	64	2s	16
ego/env tokens	codes per branch	future horizon	SFT dynamics tokens

The staged recipe keeps future-dynamics learning explicit before binding the trace to FAST action tokens.

5 Tokenization & Data

8	64	2s	700k
future dynamics tokens	codes per branch	future horizon	in-house frames

Compact dynamics align the planning trace with action learning while keeping the reasoning sequence short enough for closed-loop use.

6 Main Results ★ Headline

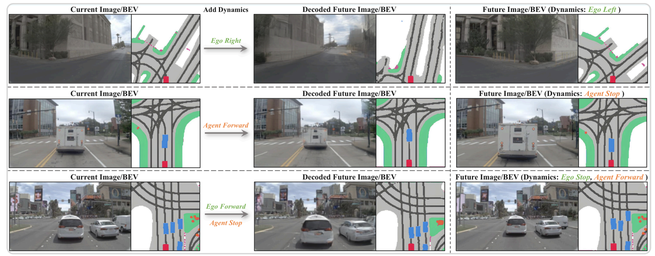
91.7	88.34	1.215	4.04
NAVSIM PDMS	Bench2Drive DS	In-house ADE	Collision per 10k

Benchmark	Metric	Best prior	DynVLA
NAVSIM	PDMS ↑	90.3	**91.7**
Bench2Drive	DS ↑	85.07	**88.34**
In-house	ADE ↓	1.344	**1.215**
In-house	Coll. ↓	5.13	**4.04**

DynVLA improves both public benchmark planning scores and large-scale in-house safety metrics.

Public and in-house gains point to the same effect: compact dynamics reasoning improves action quality before RFT.

7 Motion Transfer ★ Visual



Dynamics tokens transfer motion factors across image and BEV futures, showing controllable ego and agent evolution.

Ego	Agent	BEV
turn / stop / forward	stop / forward cues	spatially grounded future

Ego edit	Ego dynamics bend the decoded future road and BEV lane geometry coherently.
Agent edit	Agent dynamics move or stop nearby traffic while preserving the current scene context.
Composed	Mixed ego-agent tokens produce combined futures, not independent image artifacts.

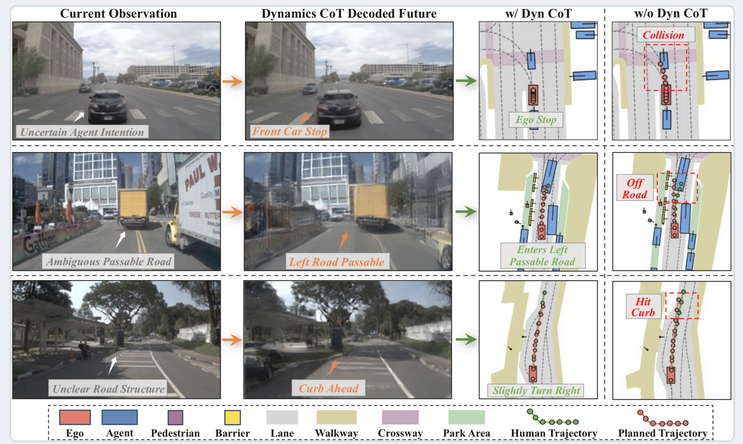
Adding dynamics changes decoded futures in the intended direction, making the reasoning trace visually inspectable.

8 Efficient CoT Design

CoT content	Lat. ↓	PDMS ↑
None	**0.20**	85.6
Scene desc.	3.04	85.3
Future image	2.29	86.3
Optical flow	2.29	86.4
Dynamics	**0.37**	**87.2**

Dynamics CoT keeps near-direct latency while outperforming text and image traces.

9 Qualitative Planning ★ Visual



Dynamics CoT anticipates stopping vehicles, free space, and geometry before decoding action tokens.

Stop	Pass	Curb
leading vehicle risk	newly free space	road-geometry constraint

Risk	Decoded futures expose hidden stops before the trajectory is selected.
Affordance	Open space and passable lanes become explicit planning evidence.
Constraint	Road geometry suppresses off-road or curb-collision plans.

The trace is compact enough to generate quickly and rich enough to guide downstream planning.

10 Takeaways

IDEA. Reason over dynamics, not text or pixels.

VISUAL. Motion transfer makes dynamics inspectable.

RESULT. 91.7 PDMS and 88.34 closed-loop DS.

PRACTICAL. 0.37s compact reasoning latency.

Contact: shuyao.shang@ia.ac.cn