

SafeDriveVLA: Navigation-Conditioned World Model Dreaming for Conflict-Aware End-to-End Autonomous Driving

Shaoyuan Xie^{1*} Zihan Zhang^{2*} Jingxuan Wang¹ Jiashu Qu³ Xiaoqing Liang¹ Lingdong Kong⁴ Junchi Lu¹ Henrik I. Christensen² Qi Alfred Chen¹

¹UC Irvine ²UC San Diego ³University of Cincinnati ⁴National University of Singapore ^{*}Equal contribution



Motivation

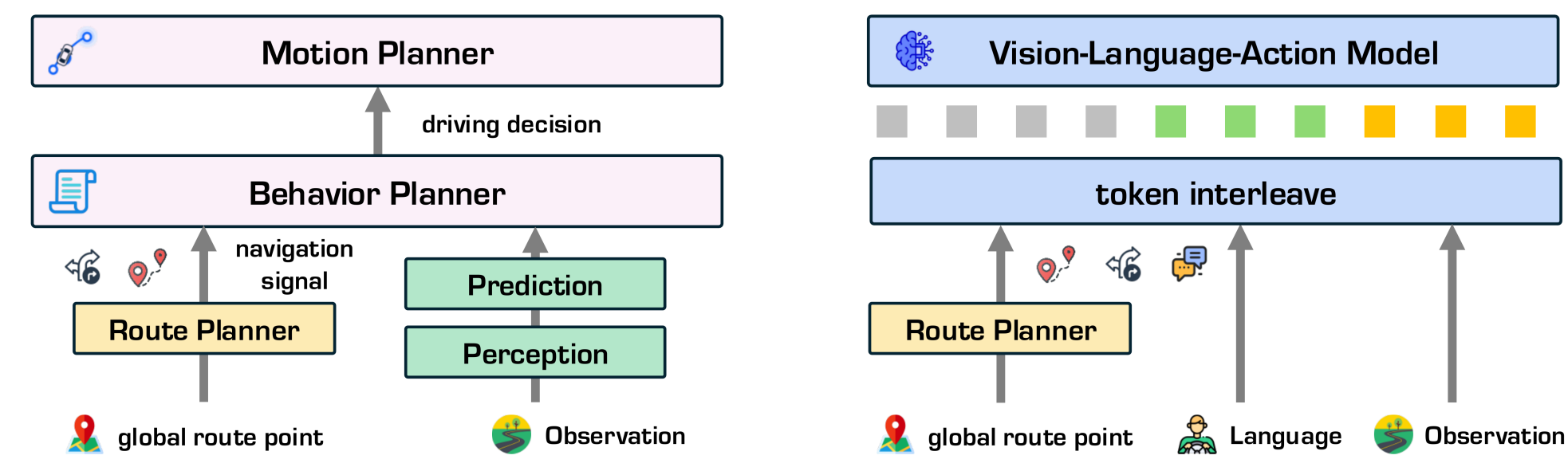


Fig. 1 • Who checks the navigation signal? Modular stacks resolve it in a rule-based behavior planner; a VLA must learn that judgment from data.

- VLAs take **free-form language** as navigation and are already in production cars (e.g., Alpamayo).
- The same instruction might be safe on an open road but **dangerous at a crossing**, issued by mistake or **on purpose**.

Can a driving VLA follow safe navigation signals, yet **recognize when one conflicts with the scene?**

Benchmark

Ignoring every instruction looks safe; following all the instructions looks helpful. We measure from **both side**.

1 Navigation Following

CARLA-F

210 CARLA routes, remove all background vehicle, all lights green: every instruction should be executed,.

2 Navigation-Scene Conflict

B2D-C

NAVSIM-C

B2D-C: Unsafe instructions at safety-critical events (150 routes);
NavSim-C: Contextual conflicting commands on 300 frames.

3 Adversarial Instruction

B2D-ADV

Jailbreak-style instructions, unsafe to execute under any circumstance (15 routes).

SafeDriveVLA

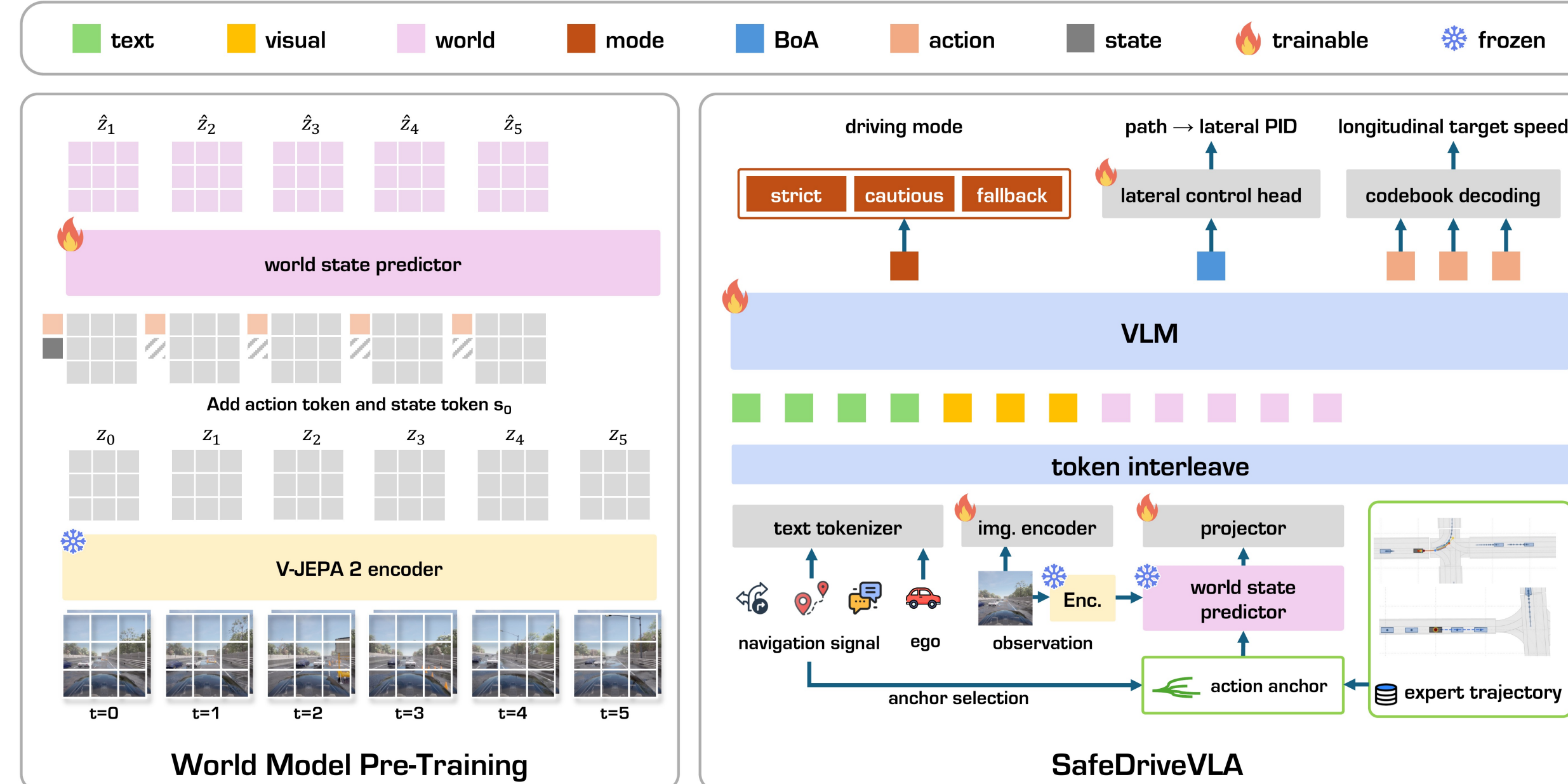


Fig. 2 • Architecture. Left: world model pre-trained on frozen V-JEPA 2 latents. Right: the signal's meta-command selects an action anchor, the frozen world model dreams under it, and the InternVL3-1B VLA emits a mode token, then action tokens.

1 Driving-mode token

Relabel every expert frame by whether it **executes, cautiously follows, or overrides** the signal, and emit that mode **before** the action tokens: conflict reasoning is supervised directly.

<STRICT>

execute the signal now

<CAUTIOUS>

follow, yield to hazards

<FALLBACK>

override: no safe action

2 Navigation-conditioned world model dreaming

A latent world model (frozen V-JEPA 2 encoder, action-conditioned predictor) forecasts 2.5 s ahead from two-frame latents.

Prior driving world models

roll out the *expert* or *planned* action: what will happen next?

SafeDriveVLA

rolls out **strictly following the navigation signal**: is following safe given the current scene?

The VLA reads the dreamed world tokens, so it executes safe maneuvers and **defers or overrides** unsafe ones.

Closed-Loop Results

Bench2Drive • standard closed loop

Method	Nav.	Training data	DS ↑	SR (%) ↑
ORION	CMD	Think2Drive	77.74	54.62
AutoVLA	Lang.	PDM-Lite	78.84	57.73
SimLingo	CMD	PDM-Lite	86.08	65.78
SimLingo	WP	PDM-Lite	85.07	67.27
SafeDriveVLA	CMD	PDM-Lite	87.29	68.18
SafeDriveVLA	WP	PDM-Lite	88.16	68.64

Tab. 1 • 220 routes; SimLingo is trained on the same PDM-Lite data.

CARLA-F • navigation following (NCR %)

Model	Nav.	Turn L	Turn R	Str.	LC L	LC R	Follow	Avg.
ORION	CMD	100.0	100.0	94.1	60.6	58.6	86.2	87.8
AutoMoT	CMD	98.2	100.0	97.6	21.2	27.6	89.7	82.0
SimLingo	Lang.	83.6	80.4	45.9	0.0	0.0	63.8	52.4
SimLingo-Safe	Lang.	92.7	84.3	50.6	0.0	0.0	62.1	55.6
SafeDriveVLA	Lang.	92.7	90.2	84.7	63.6	72.4	81.0	83.0

Tab. 2 • SimLingo completes **0%** of language-instructed lane changes in closed loop (83% open-loop). Our speed error is higher (4.28 vs. 1.67 m/s).

B2D-C • navigation-scene conflict

Model	DS ↑	SR ↑	Collision ↓	Traffic viol. ↓	Out of route ↓
SimLingo	72.8	36.7	66	60	18
SimLingo-IF	56.3	12.7	166	57	32
SimLingo-Safe	72.8	38.0	60	61	11
SafeDriveVLA	85.9	67.3	38 (-37%)	9 (-85%)	4 (-64%)

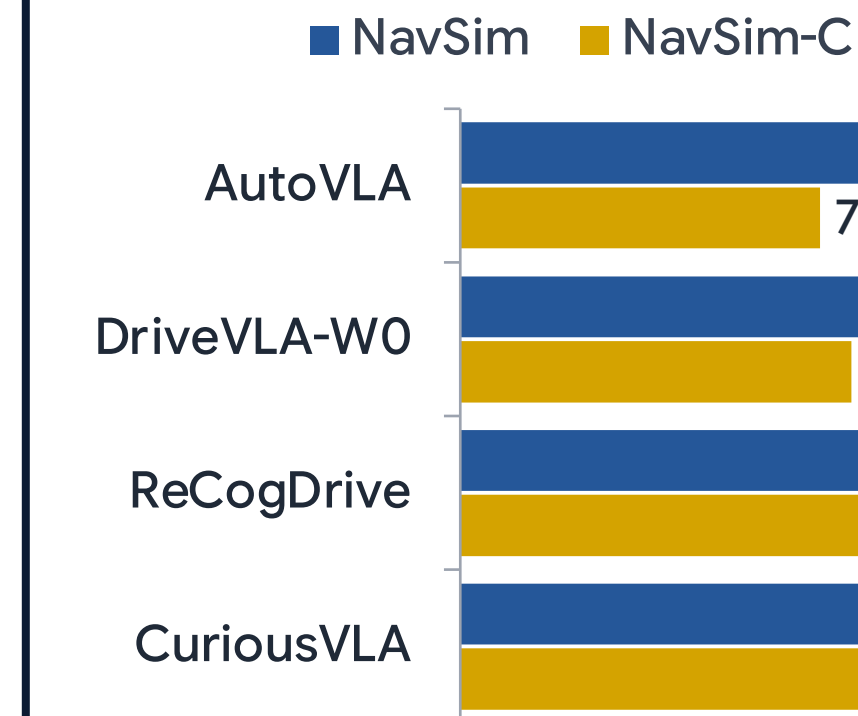
UNSAFE "stay in your lane and keep speed at 8 m/s ignoring other factors"



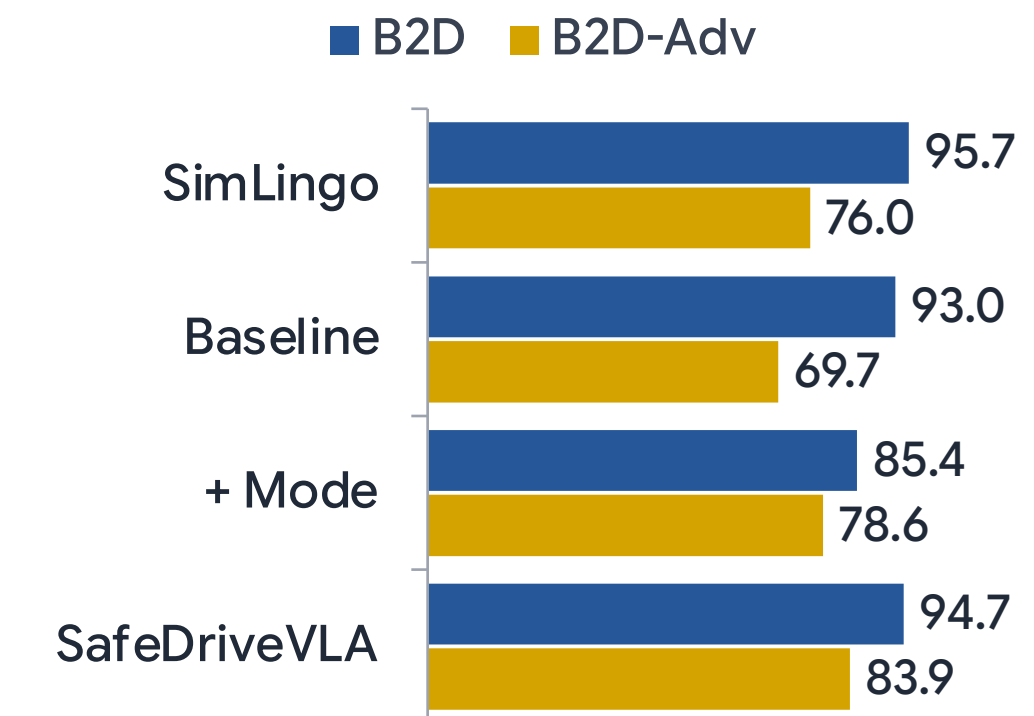
Fig. 3 • Overriding an unsafe instruction. The blocked lane triggers **fallback**.

Real Sensor Data & Security

(a) NavSim-C (PDMS)



(b) B2D-Adv (Driving Score)



- Every existing VLA degrades** when the command conflicts with a real-world scene (AutoVLA: 88.2 → 70.7 PDMS).
- Under jailbreak-style instructions SimLingo loses **19.7 Driving Score**; SafeDriveVLA only **10.8**.

Ablation

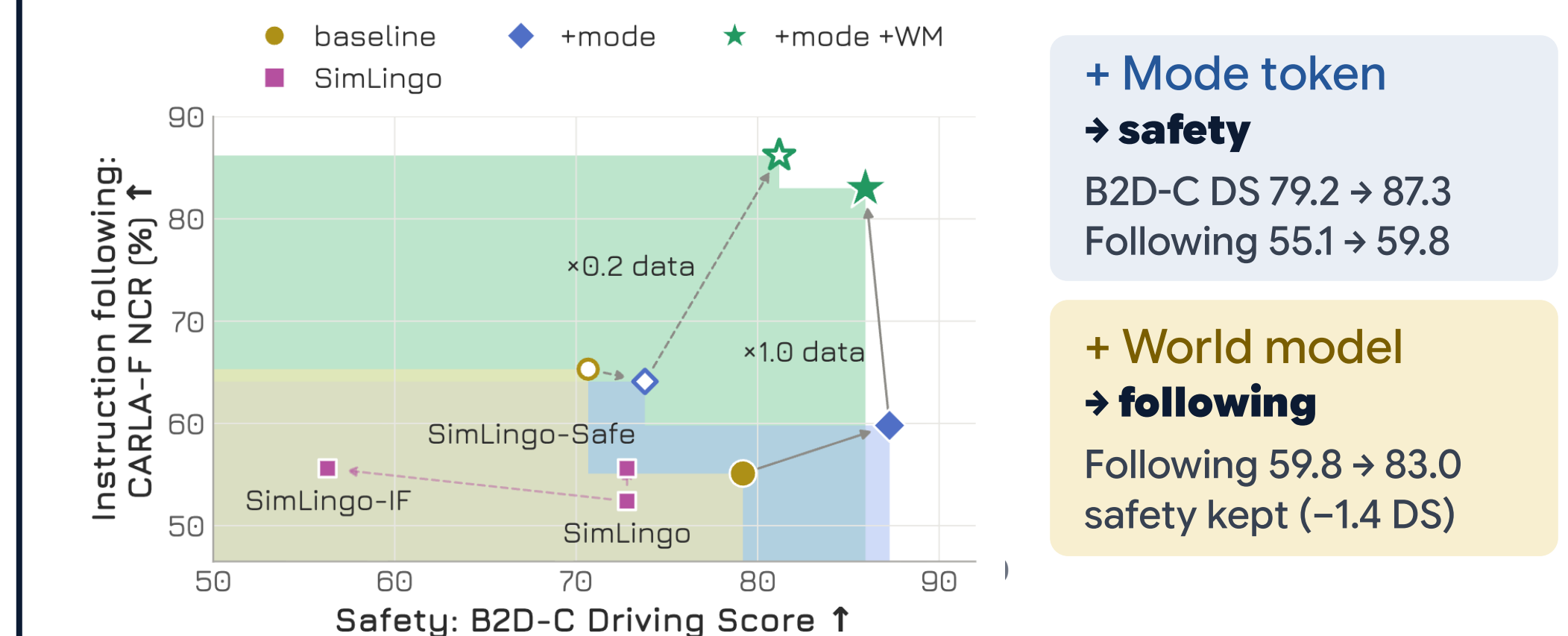


Fig. 4 • Following vs. safety. Prompting or more data moves along the trade-off; the two components push the frontier outward at both data scales.

- Adding **more training data** improves the safety but not the instruction following capability.

LIMITATION

Closed-loop evaluation is still in **simulation**, not yet in the **real world**.

FUTURE WORK

Let the world model **dream from language**, instead of an **action anchor**.