π-Drive: Reinforcement Post-Training Turns a Manipulation VLA into a Real-Time Driving Policy

Mark Music Felipe Barbosa Alex Kim

{mmusic, fbarbosa, alexjkim}@stanford.edu

Project Overview

Motivation. Vision-language-action (VLA) models offer a promising path toward autonomous driving: a single policy maps visual observations, ego state, and language context directly into future trajectories. But driving policies must generate smooth long-horizon trajectories, reason about road geometry and safety, and replan in real time. Existing VLAs leave a gap between accuracy and deployability: driving-specific models such as Alpamayo-R1 achieve strong open-loop performance but depend on multi-camera inputs and reasoning-heavy inference [1], running at only about 0.5 Hz on the Jetson AGX Thor (NVIDIA's flagship edge GPU)—far below the 2–10 Hz needed for closed-loop control [2]. $\pi_{0.5}$ is the opposite trade-off: a compact flow-matching VLA built for robot manipulation [3], trained only on folding and placing tasks. It runs at 10.8 Hz on the same Jetson AGX Thor, fast enough for real-time control, but it was never trained to drive.

Objectives. Adapt a fast, manipulation-trained flow policy into a single-camera driving policy, and improve it through post-training (behavior cloning + group-relative RL).

Significance. A compact single-camera policy runs at control-rate on a single edge GPU, removing the multi-camera rigs and expensive multi-GPU inference that reasoning-heavy VLAs require—making on-vehicle deployment practical. We quantify this by rolling out the policy on an instrumented golf cart with direct steering.

Central Question

Can a fast flow-matching VLA become a reliable driving policy?

We introduce π-Drive, a driving-specialized version of $\pi_{0.5}$, trained through behavior cloning and improved with Flow-GRPO post-training.

Data & Metrics

Dataset:

NVIDIA PhysicalAI-AV [4]: Large-scale real-world driving; front-wide camera + ground-truth egomotion, filtered to daytime, right-hand traffic. 207K samples (176K train / 31K eval). *BC + RL.*

Cart teleoperation: Expert human-driven trajectories on the golf cart; egomotion (predicted via DPVO with IMU data) and front 120° camera. *BC only.*

Setup:

State: Front 120° camera, 2-D egomotion state, language navigation command.

Action: 6.4 s egomotion trajectory as (acceleration, curvature) at 10 Hz, generated by flow matching and integrated through unicycle kinematic model.

Reward (RL): $R = 0.5\,R_{\text{drive}} + 0.3\,R_{\text{cmd}} + 0.2\,R_{\text{ref}}$, with $R_{\text{drive}} = \text{DAC} \cdot (5\,\text{TTC} + 2\,\text{comfort} + 5\,\text{progress})/12$ (drivable-area gate × safety/comfort/progress), command match, and ADE guardrail.

Evaluation: Open-loop prediction on 200 held-out clips + closed-loop rollout on cart.

Metrics: *(Open-loop on 200 held-out PhysicalAI-AV clips)*

Mean / Median ADE: Average and typical displacement error across the trajectory.

Mean FDE: Displacement error at the end of the horizon.

Longitudinal jerk: RMS & peak $j = \dot{a}$, vs. human GT — ride smoothness.

Generalization: Qualitative closed-loop rollout on Jetson AGX Thor (campus driving).

Methods & Experiments

We aim to adapt a manipulation-trained flow-matching VLA into a fast autonomous-driving policy using NVIDIA PhysicalAI-AV driving trajectories [4]. Our framework combines behavior cloning, group-relative reinforcement learning, and a preference-optimization ablation to study whether $\pi_{0.5}$ can be converted into a deployable driving model.

Behavior Cloning: We fine-tune $\pi_{0.5}$ to generate future ego-trajectories, 64 timesteps of acceleration and curvature, a 6.4s horizon, through the flow-matching imitation objective, rather than robot manipulation actions.

Each example: A front-camera observation, ego state, and language command.

Navigation-intent labels generated by Gemini 2.5 Flash, with 30% prompt dropout (classifier-free guidance) to prevent overfitting to the labels.

Produces π-Drive-BC, our baseline driving policy and frozen reference model for post-training.

Flow-GRPO: We adapt GRPO to $\pi_{0.5}$'s flow-matching action head. Since the policy has no token log-probs, we reuse its flow-matching loss as a log-probability surrogate ($\log \pi \approx -\mathcal{L}_{\text{FM}}$): the step that makes RL on a flow policy possible [5].

For each scene, the policy samples 8 candidate trajectories.

Each trajectory is scored with a driving reward based on drivable-area compliance, time-to-collision, comfort, route progress, command following, and reference-path alignment.

Rewards are standardized within each group into advantages, so the policy improves without training a separate critic.

A KL penalty keeps the updated policy close to the behavior-cloned reference.

$$A_i = \frac{R_i - \mu_R}{\sigma_R}$$ group-relative advantage: group mean as baseline, no critic

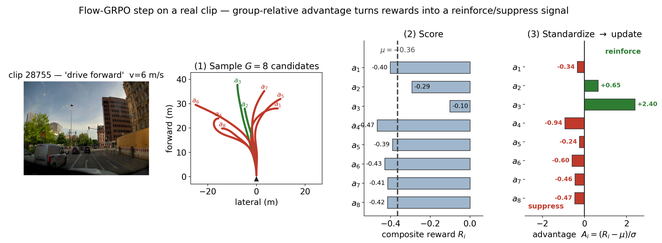

Figure 1: Example Flow-GRPO rollout. For a fixed driving scene, π-Drive samples multiple candidate trajectories, scores them with the composite driving reward, and reinforces the candidates with higher group-relative advantage while suppressing lower-scoring rollouts.

Preference Optimization Ablation: We also tested DPO, but it was not our main successful method [6, 7].

Preferences came from Cosmos-3-generated safety judgments [8].

Naive Flow-DPO was unstable: because this surrogate is unnormalized, DPO widens the margin by inflating the rejected action's flow loss, corrupting the shared velocity field.

An imitation anchor stabilizes it, but VLM-DPO still does not improve open-loop ADE; it optimizes scene-safety, not GT-proximity.

We therefore treat DPO as a diagnostic ablation and focus our main results on Behavior Cloning and Flow-GRPO.

Results

All models see the **same single front-camera feed** to isolate model quality from sensor advantage, giving a fair, apples-to-apples evaluation. We thus run Alpamayo-R1 front-only to match the input given to $\pi_{0.5}$ rather than its native four-camera setup.

Table 1: Open-loop accuracy on 200 fixed clips, identical single-camera input. ADE/FDE in m; lower better, best in bold.

Model	In	m.ADE	md.ADE	m.FDE	md.FDE
Alpamayo (front)	hist	4.23	**2.85**	12.40	8.75
$\pi_{0.5}$ GRPO	state	**3.58**	2.95	**10.19**	8.63
$\pi_{0.5}$ BC	state	4.13	3.48	11.04	**8.60**
$\pi_{0.5}$ Cosmos-DPO	state	4.69	4.02	12.31	10.63

Figure 2: Predicted paths vs GT for a night right turn (top) and daytime left turn (bottom), using a single front camera. Legend shows per-clip ADE. GRPO tracks GT closest; BC over-turns; DPO drifts the wrong way; front-only Alpamayo under-turns.

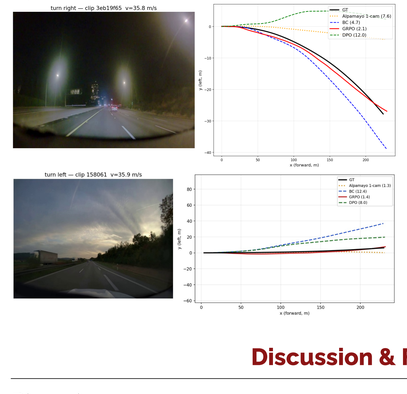

Table 2: Longitudinal jerk vs human GT (model/human ratio; 1.0 = human-like, <1.0 = smoother). GRPO closest on all four; best in bold.

Model	RMS m	RMS md	Pk m	Pk md
Alpamayo (front)	0.38	0.32	0.40	0.29
$\pi_{0.5}$ GRPO	**0.71**	**0.60**	**0.63**	**0.50**
$\pi_{0.5}$ BC	0.52	0.48	0.44	0.39
$\pi_{0.5}$ Cosmos-DPO	0.38	0.32	0.43	0.27

$\pi_{0.5}$ BC:

- Matches front-only Alpamayo; ∼3× smaller.
- Unstable on sharp turns (needs post-training).
- Over-damps jerk (ratios ∼0.5): smooth but timid.

$\pi_{0.5}$ GRPO:

- Top single-cam model on ADE and mean FDE.
- Fixes BC turn instability; tracks GT closest.
- Closest to human jerk; highest on all 4 ratios.

$\pi_{0.5}$ DPO:

- Underperforms BC/GRPO on ADE/FDE.
- VLM safety preferences weakly align with human-like GT tracking.
- Needs better preferences plus stronger BC anchoring.

Discussion & Future Work

Discussion:

- At equal input, a ∼3B flow policy rivals a 10B reasoning VLA at ∼20× the speed: the gap is sensors, not size, so improvements can come from data and post-training rather than scale.
- GRPO fixes the sharp turns where BC veers out of lane: reward-based post-training beats imitation alone.
- But not every signal helps: VLM-judge DPO optimizes scene-safety, not GT-proximity, the wrong lever for ADE.

Future Work:

Benchmark π-Drive on **Bench2Drive** with route completion, collision rate, and driving score.

Test whether Flow-GRPO's open-loop gains transfer to **closed-loop driving** on the Golf Cart.

Revisit DPO with stronger imitation anchoring and closed-loop safety metrics, since VLM safety preferences may not align with ADE-to-human-trajectory.

References

[1] Yan Wang et al. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.

[2] Daniel Dauner et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, 2024.

[3] Kevin Black et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[4] NVIDIA. NVIDIA PhysicalAI-Autonomous-Vehicles Dataset. Hugging Face dataset, 2025.

[5] Jie Liu et al. Flow-GRPO: Training flow matching models via online reinforcement learning. *arXiv preprint arXiv:2505.05470*, 2025.

[6] Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[7] Bram Wallace et al. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

[8] NVIDIA. Cosmos 3: Omnimodal world models for physical ai. Technical report, 2026.