SHRED: SELF-PLAY FOR HETEROGENEOUS ROAD-USERS WITHOUT EXPERT DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Realistic traffic simulation and autonomous planning both depend on modeling interactions among heterogeneous road users whose dynamics, constraints, and behaviors differ fundamentally. Existing approaches predominantly learn realistic interaction behavior from large-scale human driving demonstrations. In this work, we ask whether such behavior can instead emerge purely through self-play. We present SHRED, a framework that jointly trains vehicles, pedestrians, and cyclists through large-scale heterogeneous self-play from random initialization, without demonstrations or logged-trajectory replay. SHRED combines type-specific dynamics and objectives with procedural scenario generation and per-agent randomization, allowing diverse behaviors to emerge through closed-loop interaction. On the Waymo Open Sim Agents Challenge (WOSAC), SHRED jointly controls all three road-user types and outperforms both the closest demonstration-free self-play baseline and prior self-play methods that use demonstration-trained reference policies. Policies trained only on nuPlan map geometry retain comparable realism under zero-shot evaluation on Waymo. For planning, SHRED achieves state-of-the-art performance on the interaction-heavy InterPlan benchmark, to our knowledge. Controlled ablations show that heterogeneity in self-play improves both traffic realism and planning performance, with particularly large planning gains in interaction-heavy scenarios. Together, these results establish heterogeneous procedural self-play as a scalable, demonstration-free paradigm for learning realistic and interactive traffic behavior.

1 INTRODUCTION

Developing autonomous vehicles that operate safely alongside human road users requires scalable and realistic closed-loop traffic simulation. Real-world traffic, however, is inherently heterogeneous: beyond vehicles, it includes pedestrians, cyclists, and other agents with distinct physical dynamics, motion constraints, and behavioral patterns. These differences shape how agents respond to one another and how their interactions unfold over time. A pedestrian or scooter yielding at a crosswalk and a vehicle merging into traffic follow fundamentally different rules of engagement. Capturing such heterogeneous interactions is therefore central to realistic closed-loop simulation. It is equally central to autonomous planning, where the ego vehicle must anticipate how diverse agents will react to its own actions.

Recent advances in traffic simulation and autonomous planning have widely adopted data-driven learning, leveraging large-scale human driving datasets to model traffic behavior and train driving policies. Tokenized and generative simulation models reproduce realistic traffic behaviors and achieve strong performance on large-scale simulation benchmarks (Wu et al., 2024; Philion et al., 2024; Zhou et al., 2024). However, these approaches often rely on computationally expensive centralized models, limiting throughput in large-scale closed-loop simulation (Chang et al., 2026). In parallel, learning-based planning methods have increasingly leveraged large-scale driving datasets to learn scene representations, model interactions, and optimize driving decisions, achieving strong performance across challenging driving scenarios (Huang et al., 2023; 2024; Zheng et al., 2025; Tan et al., 2025). Despite their different objectives, both lines of work remain widely dependent on logged human driving data for learning realistic and effective interaction behaviors. Their behavior remains tied to the interaction distributions in logged datasets, which can make coverage of diverse heterogeneous interactions and cross-dataset transfer more challenging. Large-scale demonstration datasets are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

also expensive to collect and may not fully capture the diversity of reactive multi-agent behaviors observed in real-world traffic.

Self-play reinforcement learning offers a complementary direction: it can learn through online interaction without trajectory supervision, scales naturally to large numbers of agents, and enables diverse multi-agent interactions to emerge through closed-loop rollouts. Recent work has shown that self-play can produce robust driving policies at unprecedented scale (Cusumano-Towner et al., 2025) and achieve competitive realism when anchored to pretrained reference models (Chang et al., 2026). However, prior self-play systems focus on vehicle policies or homogeneous dynamics and rewards across road-user types despite fundamentally different motion characteristics. Scalable, fully reactive heterogeneous self-play across vehicles, pedestrians, and cyclists without relying on logged trajectory data or pretrained imitation-policy anchors therefore remains largely unexplored.

To address these challenges, we propose SHRED (**S**elf-Play for **H**eterogeneous **R**oad-Users without **E**xpert **D**emonstrations), a scalable framework for learning heterogeneous traffic behavior through large-scale self-play. SHRED jointly trains vehicles, pedestrians, and cyclists from random initialization, with type-specific dynamics, objectives, and action spaces that reflect their distinct modes of interaction. Procedural scenario generation and per-agent randomization of dynamics and reward parameters expose agents to diverse traffic configurations and behaviors throughout training. This enables a compact decentralized policy to learn realistic heterogeneous multi-agent behaviors entirely through self-play, without demonstration supervision.

We show that demonstration-free self-play can achieve both realistic heterogeneous traffic simulation and strong interactive planning. For simulation, we evaluate on the Waymo Open Sim Agents Challenge (WOSAC) (Montali et al., 2023) while jointly controlling vehicles, pedestrians, and cyclists. SHRED outperforms the closest demonstration-free self-play baseline, while controlled ablations show that heterogeneous self-play substantially improves traffic realism, including the behavior of vehicles themselves. For planning, we evaluate SHRED on InterPlan (Hallgarten et al., 2024), where it achieves state-of-the-art performance on interaction-heavy closed-loop scenarios. Complementary planning ablations show that heterogeneous self-play substantially outperforms vehicle-only self-play, with the largest gains on scenarios involving concentrated interactions with other road users. Beyond these benchmarks, policies trained on nuPlan maps transfer zero-shot to Waymo, while behavioral analyses show that self-play produces diverse, plausible alternatives to logged trajectories (Figure 1).

Our main contributions are as follows:

- We introduce SHRED, a demonstration-free heterogeneous self-play framework that jointly learns vehicle, pedestrian, and cyclist behaviors with distinct dynamics and objectives, entirely without demonstrations or logged-trajectory replay.
- We develop the key components for scalable heterogeneous self-play, including procedural scenario and goal generation, per-agent randomization of dynamics and objectives, and type-specific policy components that support joint learning across diverse road-user types.
- We show that demonstration-free self-play is effective across both simulation and planning: SHRED achieves strong sim-agent realism on WOSAC, state-of-the-art performance on the official 80-scenario InterPlan split to our knowledge, and zero-shot map transfer from nuPlan to Waymo.

2 Related Work

2.1 Data-Driven and Heterogeneous Traffic Simulation

Imitation from human driving logs has become the dominant paradigm for realistic traffic simulation. Tokenized autoregressive models (Wu et al., 2024; Philion et al., 2024; Zhou et al., 2024; Zhang et al., 2025) and diffusion-based approaches (Jiang et al., 2024) have driven strong performance on WOSAC (Montali et al., 2023), while closed-loop fine-tuning, promptable rollout, and RL/RLHF post-training further improve rollout realism and controllability (Zhang et al., 2025; Tan et al., 2024; Peng et al., 2024; Chen et al., 2025; Cao et al., 2024; Pei et al., 2026). RLFTSim (Ahmadi et al., 2026) fine-tunes a pretrained traffic simulator with a distribution-alignment reward. DecompGAIL (Guo et al., 2026) combines a SMART-based model with decomposed adversarial imitation learning. Both methods use a trajectory-model prior; SHRED instead studies heterogeneous recurrent self-play without one. Pedestrian and cyclist behavior has additionally been studied through rule-based crowd

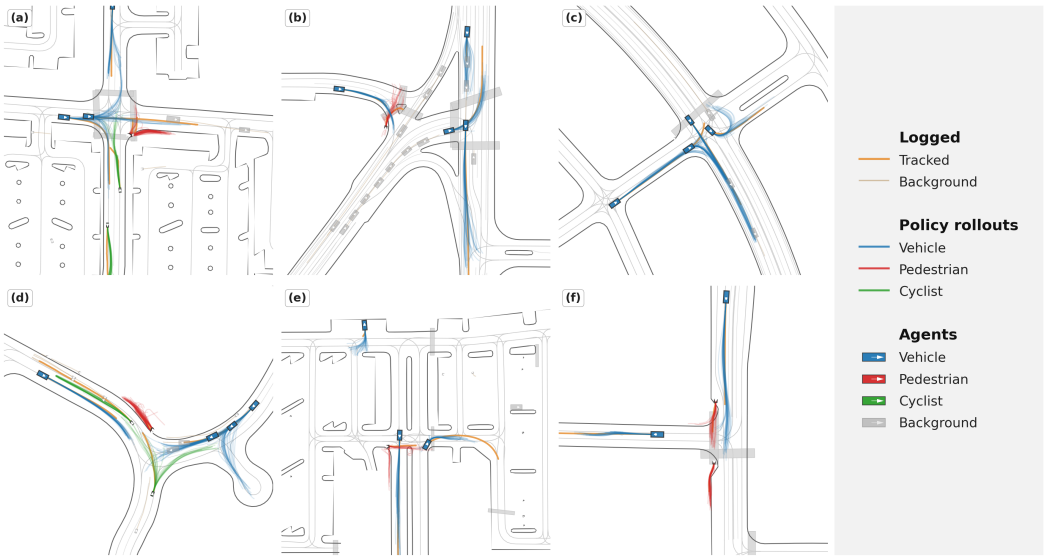


Figure 1: **Qualitative Rollout Distributions.** Qualitative trajectory distributions from SHRED across a diverse range of scenarios, from urban intersections to parking lots. Each panel shows 32 closed-loop policy rollouts (blue: vehicle, red: pedestrian, green: cyclist); orange traces show the ground-truth log trajectory. The panels show alternative maneuvers alongside the logged trajectory.

dynamics (Helbing & Molnár, 1995; Helbing et al., 2005; van den Berg et al., 2011), multimodal trajectory forecasting models (Alahi et al., 2016; Gupta et al., 2018; Salzmann et al., 2020; Xu et al., 2022; Rasouli & Kotseruba, 2023), and multi-agent driving simulators (Suo et al., 2021; Vinitsky et al., 2022; Gulino et al., 2023; Li et al., 2022; Kazemkhani et al., 2025; Rowe et al., 2024). However, realistic heterogeneous interaction behavior is typically learned implicitly from demonstrations, while learned closed-loop control remains largely vehicle-centric or relies on shared dynamics across all road users. Our work instead studies whether realistic heterogeneous traffic can emerge from demonstration-free self-play with type-specific dynamics, rewards, goals, and policy components for vehicles, pedestrians, and cyclists jointly.

2.2 Self-Play Reinforcement Learning for Driving

Self-play has driven landmark results in game playing (Silver et al., 2017; Vinyals et al., 2019; Berner et al., 2019; Jaderberg et al., 2019) and has recently been adopted for traffic simulation. Gigaflow (Cusumano-Towner et al., 2025) is the closest demonstration-free precedent, training a unified policy for vehicles, pedestrians, and cyclists through massively parallel self-play. However, all agents share the same jerk-actuated bicycle dynamics and goal-generation mechanism, with diversity induced primarily through conditioning variables inside a shared backbone. Other self-play systems explore complementary directions: Cornelisse et al. (2025) scale vehicle-only PPO from logged WOMD initial states, while HR-PPO (Cornelisse & Vinitsky, 2024) and SPACeR (Chang et al., 2026) regularize self-play with behavior-cloned or tokenized policy anchors. NOMAD (Wang et al., 2026) studies map-transfer through self-play, and GPUDrive (Kazemkhani et al., 2025) provides a GPU-resident substrate for large-scale simulation. In contrast, we jointly train vehicles, pedestrians, and cyclists from fully procedural initialization, without trajectory supervision, pretrained imitation-policy anchors, or logged-trajectory replay, using type-specific dynamics, rewards, and goal generation.

2.3 Heterogeneity in Multi-Agent Reinforcement Learning

Most heterogeneous multi-agent RL methods are developed for cooperative settings with shared rewards, including QMIX (Rashid et al., 2018), MAPPO (Yu et al., 2022), and HAPPO/HATRPO (Kuba et al., 2022; Zhong et al., 2024). Several of them rely on sequential per-agent updates (HAPPO/HATRPO) or separate per-agent policies, which become costly in traffic scenes containing dozens of agents with different dynamics and action spaces. Recent modular and role-conditioned approaches such as SePS (Christianos et al., 2021), ROMA (Wang et al., 2020), RODE (Wang et al., 2021),

and MoE-based RL architectures (Yang et al., 2020; Obando-Ceron et al., 2024) reduce interference through selective parameter sharing or latent routing, but remain primarily focused on cooperative domains.

Our setting is instead mixed-motive: each road user optimizes its own goal under type-specific rewards and dynamics without a shared team objective. We therefore organize policies by physical agent type rather than agent identity, sharing perception while maintaining separate recurrent, actor, and value heads for vehicles, pedestrians, and cyclists.

3 METHOD

3.1 HETEROGENEOUS AGENT MODELING

We consider heterogeneous traffic environments containing agent types $\mathcal{K} = \{\mathrm{VEH}, \mathrm{PED}, \mathrm{CYC}\}$, corresponding to vehicles, pedestrians, and cyclists. Each agent type follows its own dynamics model, reward structure, and discrete action space while interacting through a shared closed-loop environment. At each timestep, agent i selects an action according to a decentralized policy

$$a_t^i \sim \pi_\theta\big(o_t^i,\, g_t^i,\, k^i,\, C_{\mathrm{dyn}}^i,\, \alpha^i\big)\,,$$

where o_t^i denotes the local observation, g_t^i denotes goal and waypoint conditioning, k^i is the agent type, C_{dyn}^i denotes kinematic parameters, and α^i parameterizes the reward function.

Per-Type Dynamics and Actions. Each agent type follows its own continuous-control dynamics model. Vehicles use jerk-integrated Ackermann dynamics, pedestrians follow unicycle dynamics, and cyclists use kinematic bicycle dynamics. Actions are discretized separately for each type to reflect their distinct motion characteristics and control regimes. Vehicles operate over longitudinal and lateral jerk primitives, pedestrians use velocity and turning controls with an explicit stop action, and cyclists follow acceleration and steering controls. Appendix A gives the full dynamics and action grids.

Rewards. *Vehicle* rewards inherit the Gigaflow (Cusumano-Towner et al., 2025) formulation, while pedestrian and cyclist rewards introduce type-specific behavioral priors for heterogeneous traffic interaction. *Pedestrian* rewards remove lane-following objectives and instead encourage sidewalk, crosswalk, and waiting behavior through soft road-intrusion penalties and walking-speed regularization, allowing curb waiting, yielding, and occasional jaywalking while discouraging unrealistic motion. *Cyclist* rewards retain vehicle-like navigation objectives while introducing soft lane-edge preferences and relaxed lane-alignment constraints, encouraging curb-side riding without enforcing strict vehicle lane-centering behavior.

We also introduce asymmetric cross-type collision shaping: a vehicle's collision penalty doubles when the other party is a pedestrian. Detailed reward formulations, thresholds, and parameter ranges are provided in Appendix C.

3.2 PROCEDURAL HETEROGENEOUS SELF-PLAY

Recent self-play works often initialize agents directly from logged datasets or replay recorded trajectories during training and evaluation (Cornelisse et al., 2025; Chang et al., 2026). While effective for reproducing realistic traffic scenes, such approaches remain constrained by the finite scene layouts and trajectory horizons present in recorded data. To improve coverage and interaction diversity during heterogeneous self-play, we procedurally generate closed-loop training episodes through randomized initialization, goal generation, and agent parameterization.

Formally, each training episode is sampled from a factorized procedural distribution

$$\mathcal{D}_{\mathrm{ep}} = \mathcal{D}_{\mathrm{scene}} \prod_{k \in \mathcal{K}} \prod_{i \in \mathcal{N}_k} \mathcal{D}_{\mathrm{init}}^{(k)} \mathcal{D}_{\mathrm{goal}}^{(k)} \mathcal{D}_{b}^{(k)} \mathcal{D}_{C_{\mathrm{dyn}}}^{(k)} \mathcal{D}_{\alpha}^{(k)} \mathcal{D}_{\mathrm{mask}},$$

where each factor is parameterized independently per agent type. These factors jointly control the distribution of scene layouts, navigation objectives, agent behaviors, and goal observability encountered during self-play training.

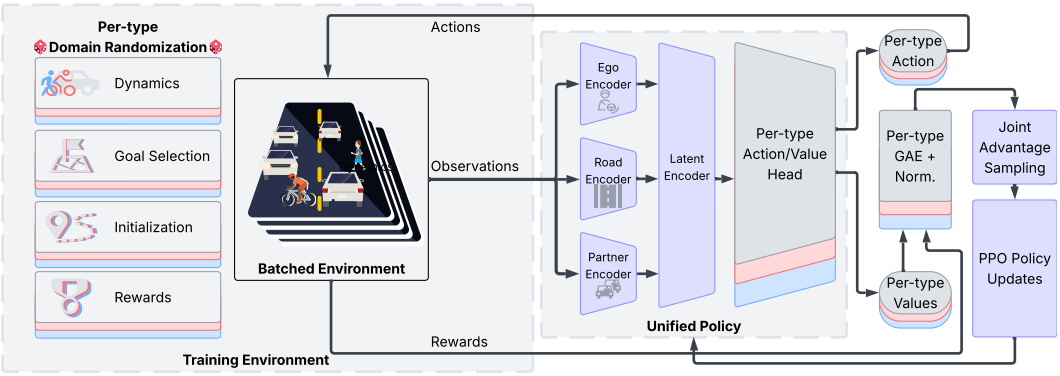


Figure 2: **SHRED Training Architecture.** Simulation produces mixed transitions across all agent types. The model consists of shared encoders followed by per-type LSTM cells, action heads, and value heads. Per-type rewards and value normalization accommodate distinct reward scales across road-user types. Each type uses its own dynamics model: jerk-integrated Ackermann dynamics for vehicles, unicycle dynamics for pedestrians, and kinematic bicycle dynamics for cyclists.

Per-Type Initialization ($\mathcal{D}_{\text{init}}^{(k)}$). Controlled agents are initialized procedurally from type-specific spawn surfaces. Vehicles and cyclists are sampled from lane centerlines, while pedestrians are initialized from crosswalks and road-edge regions. Agents are added sequentially through rejection sampling to ensure collision-free initial states, with randomized heading perturbations applied at spawn time to increase local interaction diversity under high environment densities. Because initialization is performed independently for each agent type, the resulting scenes span diverse heterogeneous compositions. These include dense crosswalk traffic, mixed cyclist–vehicle flows, and multi-type urban intersections, which are rare in recorded datasets.

Procedural Goal Generation. Rather than replaying logged destinations, goals are procedurally sampled from map topology conditioned on agent type. Vehicles and cyclists follow lane-graph connectivity, while pedestrian goals are generated along crosswalk and sidewalk topology. Unlike logged-goal initialization, where agents often terminate after reaching a single recorded destination, goals are continuously regenerated after completion during training. This decouples interaction horizon from the finite duration of recorded scenarios, allowing agents to experience substantially longer and more diverse closed-loop interactions within a single environment rollout.

Per-Agent Diversity. Behavioral diversity is induced through independent randomization of agent geometry, kinematic coefficients, and reward parameterizations. Each agent observes its own $(b^i, C^i_{\text{dyn}}, \alpha^i)$ but not those of surrounding agents, forcing the policy to model partner behavior as arising from a heterogeneous population rather than a fixed shared dynamics model.

Goal Dropout. To reduce over-reliance on explicit navigation conditioning, a subset of agents have their goal observations masked during training. These agents must instead act using local scene context and interaction history alone, while continuing to participate in fully reactive closed-loop self-play. This encourages the policy to produce robust and realistic behaviors even when long-horizon navigation intent is partially specified or absent.

3.3 Architecture and Algorithm Design Choices

Policy Architecture. The architecture of SHRED is illustrated in Figure 2. All agent types share encoders for ego state, road geometry, nearby agents, traffic signals, goals, and conditioning variables. The shared encoders produce a scene representation for each agent. Vehicles, pedestrians, and cyclists then use separate LSTM cells, actor heads, and value heads.

Value Normalization and Advantage Computation. Jointly optimizing heterogeneous agent populations introduces substantial optimization imbalance: different agent types exhibit distinct reward scales, motion statistics, and advantage distributions during self-play. As a result, naively optimizing all agents under a shared PPO objective can cause dominant agent classes or low-signal transitions to overwhelm the optimization process.

Table 1: **Comparison with demonstration-free self-play on WOSAC 2023.** SHRED outperforms Gigaflow in overall realism, with the largest gains in time-to-collision and linear speed. Without type-specific rewards and dynamics, SHRED falls below Gigaflow, showing that its advantage comes from heterogeneous modeling. Neither method observes logged destinations and goals are sampled from the map, following the official WOSAC setting. Gigaflow controls pedestrians with a scripted IDM-like policy, whereas SHRED learns all three road-user types.

Method	Realism ↑	Lin. Speed ↑	Lin. Accel. ↑	Ang. Speed ↑	Ang. Accel. ↑	Dist. Obj. ↑	TTC ↑	Dist. Rd. Edge ↑	Off-road ↑	Coll. ↑
Gigaflow (Cusumano-Towner et al., 2025)	0.619	0.261	**0.253**	**0.506**	**0.476**	**0.316**	0.808	0.545	**0.910**	**0.947**
SHRED	**0.631**	**0.353**	0.247	0.501	0.457	0.250	**0.894**	**0.607**	0.904	**0.947**
SHRED (homogeneous rewards and dynamics)	0.604	0.247	0.239	0.495	0.358	0.248	0.879	0.563	0.896	**0.947**

To stabilize heterogeneous self-play, we maintain independent PopArt normalization statistics (van Hasselt et al., 2016) together with per-type advantage standardization before PPO updates. Advantages are computed jointly over mixed heterogeneous rollouts using GAE with V-trace corrections (Espeholt et al., 2018). We further perform advantage-weighted prioritized replay, replacing hard advantage filtering used in prior large-scale self-play systems (Cusumano-Towner et al., 2025), where rollout segments are sampled proportionally to the magnitude of their estimated advantages. This increases the contribution of high-signal transitions during optimization while preserving balanced learning across heterogeneous agent populations.

4 Experiments

We evaluate demonstration-free heterogeneous self-play across both simulation and planning, focusing on its role in learning realistic traffic behavior and effective interactive planning.

4.1 Heterogeneous Traffic Simulation

Training and Evaluation Protocol. We train the recurrent heterogeneous policy with distributed mixed-precision self-play on 32 NVIDIA A100 GPUs. The policy shares scene encoders across road users and uses separate LSTM, actor, and value components for vehicles, pedestrians, and cyclists. Training uses map geometry, procedural initialization, and map-conditioned goals without trajectory supervision or a pretrained imitation-policy anchor. We evaluate WOSAC 2023 and WOSAC 2024 validation with 32 stochastic closed-loop rollouts per scenario. The WOSAC 2023 evaluation follows the leaderboard setting: agents receive the logged history and the map but no goal. Appendix E describes the evaluator verification and reports the WOSAC 2025 test evaluation separately.

Demonstration-Free Self-Play Comparison. Table 1 compares SHRED with Gigaflow, the closest demonstration-free self-play baseline, on WOSAC 2023. Gigaflow applies its learned policy to vehicles and cyclists during evaluation while controlling pedestrians with a scripted IDM-like policy (Cusumano-Towner et al., 2025). In contrast, SHRED learns policies for all three road-user types through self-play and achieves higher realism than Gigaflow (0.631 vs. 0.619) over 10,000 validation scenarios, with particularly strong improvement in interaction-based metrics: time-to-collision likelihood (0.894 vs. 0.808), while Gigaflow achieves higher distance-to-object likelihood.

To examine whether the gain arises from heterogeneous self-play rather than self-play alone, we additionally evaluate a homogeneous SHRED variant with shared dynamics, rewards, and policy components across road-user types, which is the joint homogeneous design of Table 2. The homogeneous variant scores 0.604, compared with 0.631 for the full heterogeneous policy. The largest degradations appear in motion realism, with linear-speed and angular-acceleration likelihood drop. Vehicles exhibit the largest per-type realism drop (0.641 to 0.614), suggesting that the benefits of heterogeneous modeling extend across road-user types.

Role of Heterogeneity. Table 2 examines how SHRED's different components of heterogeneous modeling contribute to simulation realism. Homogeneous dynamics give every road-user type the pedestrian unicycle model, and homogeneous rewards give every type the vehicle reward set. Homogenizing either dynamics or rewards causes modest degradation, while the fully homogeneous design produces the largest drop in realism. Vehicles benefit the most from heterogeneous modeling.

Table 2: **Ablation of heterogeneous dynamics and rewards on WOSAC 2024.** Overall and per-type realism for vehicles, pedestrians, and cyclists. Removing either component modestly reduces realism, while removing both causes the largest drop, most strongly for vehicles. Removing random initialization and goals also degrades realism.

Het. dyn.	Het. rew.	Overall ↑	Veh. ↑	Ped. ↑	Cyc. ↑
✓	✓	**0.721**	**0.742**	**0.724**	0.672
✗	✓	0.718	0.739	0.723	0.665
✓	✗	0.714	0.733	0.711	0.663
✗	✗	0.686	0.699	0.708	0.668
w/o rand. init. & goals		0.711	0.725	0.718	**0.676**

Table 3: **Realism and diversity on WOSAC 2024.** Unlike HR-PPO and SPACeR, which anchor self-play to demonstration-trained policies, SHRED uses no demonstrations, yet achieves the highest interactive score and higher realism than both, with greater rollout diversity. The nuPlan-trained policy retains comparable realism zero-shot on Waymo.

		WOSAC 2024				Diversity
Method	Demo Free	Realism ↑	Kinematic ↑	Interactive ↑	Map ↑	Mean dist. (m) ↑
SMART (CAT-K) (Wu et al., 2024; Zhang et al., 2025)	✗	0.752	0.453	0.715	0.815	6.97
HR-PPO (Cornelisse & Vinitsky, 2024)	✗	0.690	0.308	0.731	0.855	4.19
SPACeR (Chang et al., 2026)	✗	0.717	0.378	0.754	**0.863**	1.41
SHRED	✓	**0.721**	**0.402**	**0.778**	0.832	**5.38**
SHRED (zero-shot, nuPlan)	✓	0.715	0.395	0.776	0.821	4.93

Removing procedural initialization and goals lowers overall realism by 0.010, with the largest drop for vehicles.

Self-Play Realism, Diversity, and Map Transfer. Table 3 compares SHRED with prior self-play approaches on WOSAC 2024 across realism and trajectory diversity. We measure diversity as the mean pairwise Wasserstein distance between each agent's occupancy distributions across 32 rollouts, following the trajectory-space analysis of BITS (Xu et al., 2023). Notably, unlike HR-PPO and SPACeR, SHRED achieves competitive realism and greater trajectory diversity without demonstration data or a pretrained imitation-policy anchor. Moreover, the nuPlan-trained policy retains comparable realism under zero-shot evaluation on Waymo, suggesting that the learned traffic behaviors transfer across maps and datasets.

4.2 Interactive Ego Planning

In this section, we evaluate whether policies learned through demonstration-free heterogeneous self-play can support interactive ego planning. We use InterPlan (Hallgarten et al., 2024) because it emphasizes challenging interactions with other road users, particularly vulnerable road users (VRUs), through long-tail scenarios including jaywalking pedestrians, overtaking, dense lane changes, construction zones, and accident sites. The ego planner uses the SHRED policy trained on nuPlan maps. To isolate the effect of heterogeneous self-play, we compare against a separately trained vehicle-only self-play policy in the planning ablation.

Planning Performance. Table 4 presents representative baselines; Appendix E.3 reports the full comparison. Beyond learning realistic heterogeneous traffic behavior, SHRED also learns a robust ego planner, achieving the highest score among the compared methods on both InterPlan evaluation sets: 67.50 on the official 80-scenario split and 70.14 on the full 335-scenario evaluation. Notably, SHRED internalizes driving behavior through large-scale heterogeneous self-play in a compact 6.7M-parameter policy, orders of magnitude smaller than learned baselines such as the 7B-parameter LLaMA backbone used by HybridLLMPlanner, while achieving substantially higher performance (67.50 vs. 53.00). Even this 6.7M parameter count reflects a deliberate latency–memory tradeoff: fused heterogeneous LSTM weights increase parameter count and memory usage to enable lower-latency inference.

Table 4: **Interactive planning on InterPlan.** SHRED achieves the highest score on both the official 80-scenario split and the full 335-scenario evaluation. Comparison with a vehicle-only variant, trained with scripted pedestrians and no cyclists, showcases the benefit of heterogeneous self-play for interactive planning. Table 10 in Appendix E.3 gives the full comparison.

Planner	Type	InterPlan (80) ↑	Full Scale (335) ↑
IDM (Treiber et al., 2000)	Rule	31	—
SPDM (N_p=60) (Distelzweig et al., 2026b)	Rule	63.66	—
HybridLLMPlanner (Llama-7B) (Hallgarten et al., 2024)	Hybr.	53	—
PDM+PPO (Distelzweig et al., 2026a)	Hybr.	43.2	—
Diffusion Planner (Zheng et al., 2025)	IL	25.8	—
FlowDrive (Wang et al., 2025)	IL	36.96	—
Flow Planner (Tan et al., 2025)	IL	—	61.82
PPO (Distelzweig et al., 2026a)	RL	42.1	—
SHRED	RL	**67.50**	**70.14**
SHRED (vehicle-only)	RL	54.19	66.65

Planning Ablations. Table 4 asks whether heterogeneous self-play benefits the ego planner itself. We compare SHRED with a vehicle-only self-play policy trained on the same nuPlan maps with the same vehicle dynamics and rewards. The vehicle-only policy learns vehicle control through self-play while pedestrians follow scripted crossing and jaywalking behaviors and cyclists are absent, mirroring the general setup of prior vehicle-centric self-play approaches such as Gigaflow (Cusumano-Towner et al., 2025). SHRED improves the InterPlan score from 54.19 to 67.50 on the official 80-scenario split and from 66.65 to 70.14 on the full 335-scenario evaluation. The substantially larger gain on the official split, which concentrates challenging interactions with other road users, suggests that heterogeneous self-play is particularly beneficial for interactive planning.

4.3 Behavioral Analysis

Qualitative Evaluation of Data Diversity. A key question for any simulation policy is whether it reproduces only the specific behaviors present in the training data, or whether it generalizes to a broader distribution of plausible maneuvers. To assess this, we generate 32 closed-loop rollouts per scenario and overlay the resulting trajectory distributions against the ground-truth log (Figure 1). The policy distribution covers the log trajectory as a high-probability mode while placing meaningful probability mass on alternative maneuvers: unexecuted turns at intersections, alternative branches at forks, and a range of merge timings. Supplementary videos show planning around vulnerable road users and in dense interactive scenarios.

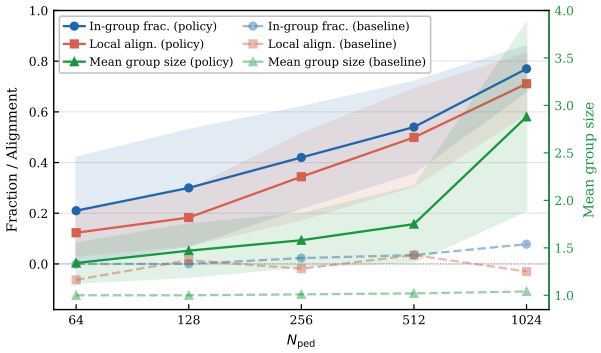


Figure 3: **Group Behavior Metrics.** Proximity clustering and directional alignment versus pedestrian population N_{ped}. Shaded bands: ± 1 standard deviation across scenarios. Left axis: in-group fraction f_{grp}; local alignment μ_{align}; right axis: mean group size $\bar{g}$. All three GBMs increase monotonically with crowd density for the policy while remaining near zero / unity for the stationary baseline.

Emergent Collective Behavior. We observe that structured collective pedestrian behavior emerges from purely decentralized, self-interested agents without any group reward or shared objective. To test this, we sweep the configured pedestrian population $N_{\text{ped}} \in \{64, 128, 256, 512, 1024\}$ while holding vehicles ($V = 16$) and cyclists ($C = 8$) fixed and evaluating over 32 validation scenarios of 60 s. We make no changes to the individual reward or goal structure; pedestrians independently optimize their own progress and safety objectives under random goal assignments.

We measure three *Group Behavior Metrics* (GBMs): i) *in-group fraction* f_{grp} (Küpper & Seyfried, 2023), the fraction of pedestrians that spend at least 50% of a rollout within 5 m of at least one co-moving companion (proximity-linked group); ii) *local alignment* μ_{align} (Luque Sánchez et al., 2020), per-frame mean cosine similarity between each pedestrian's velocity and those of its k-nearest neighbors, averaged over moving frames; and iii) *mean group size* $\bar{g}$ (Küpper & Seyfried, 2023), average headcount of identified proximity groups. A *stationary baseline*, pedestrians with random positions and orientations (used as proxy velocities) frozen for the entire rollout, provides a nonmoving reference.

Figure 3 shows increasing proximity clustering and directional alignment as pedestrian density rises, despite the absence of an explicit group objective. The stationary reference stays near its floor at every density, so static crowding alone does not produce the measured structure. The result illustrates aggregate structure under individual objectives, consistent with the emergence of crowd patterns in local-interaction models (Helbing & Molnár, 1995; Moussaïd et al., 2011).

5 Conclusion

We have presented SHRED, a framework for jointly training vehicles, pedestrians, and cyclists through self-play reinforcement learning with type-specific dynamics, rewards, and value normalization, from random initialization without trajectory supervision or logged-trajectory replay. On the Waymo Open Sim Agents Challenge, SHRED achieves strong realism without trajectory-supervised policy training, while policies trained on nuPlan maps retain comparable realism under zero-shot evaluation on Waymo. Beyond simulation, the learned policy also serves as an effective ego planner, achieving, to our knowledge, state-of-the-art performance on InterPlan. Controlled ablations further show that heterogeneous self-play improves both traffic realism and planning performance, with particularly large planning gains in interaction-heavy scenarios.

Our results position heterogeneous self-play as a complementary paradigm to data-driven approaches across both traffic simulation and autonomous planning. Imitation-based methods retain an advantage in overall distributional realism on WOSAC, while self-play removes dependence on demonstration data and enables behaviors to emerge through closed-loop interaction. Together, our results suggest that heterogeneous self-play provides a scalable, demonstration-free paradigm for learning interactive traffic behavior across both simulation and planning.

Limitations and Future Work. SHRED inherits the fidelity ceiling of its simulator and reward design: behaviors not captured by the dynamics or objectives, such as richer right-of-way conventions and traffic rules, may not emerge through self-play. Future work could address these limitations through richer traffic-rule objectives, automatically learned or calibrated rewards, and additional road-user types. A realism gap to imitation-based models also remains, particularly in map-based realism (0.832 vs. 0.863 for SPACeR), suggesting that lightweight data-driven priors could complement self-play where demonstrations are available. Finally, WOSAC measures agreement with logged behavior rather than behavioral validity, and we do not evaluate transfer to real-world driving. Evaluating learned behaviors against real-world interactions and using SHRED as interactive background traffic for downstream planner training and evaluation are important directions for future work.

6 AI Use Statement

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, interpreting results. We have not used generative AI tools for designing or providing feedback on research methodology or experiments. Helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, assisting with translation are not applicable to this work.

We also used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, creating artifacts, drafting parts of a research paper, summarizing or analyzing existing literature, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, and formatting references. We have reviewed all AI-assisted work. We

take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ehsan Ahmadi, Hunter Schofield, Behzad Khamidehi, Fazel Arasteh, Jinjun Shan, Lili Mou, Dongfeng Bai, and Kasra Rezaee. RLFTSim: Realistic and controllable multi-agent traffic simulation via reinforcement learning fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39734–39743, 2026.

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Yulong Cao, Boris Ivanovic, Chaowei Xiao, and Marco Pavone. Reinforcement learning with human feedback for realistic traffic simulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14428–14434, 2024.

Wei-Jer Chang, Akshay Rangesh, Kevin Joseph, Matthew Strong, Masayoshi Tomizuka, Yihan Hu, and Wei Zhan. SPACeR: Self-play anchoring with centralized reference models. In *International Conference on Learning Representations (ICLR)*, 2026.

Keyu Chen, Wenchao Sun, Hao Cheng, and Sifa Zheng. RIFT: Group-relative RL fine-tuning for realistic and controllable traffic simulation. *arXiv preprint arXiv:2505.03344*, 2025.

Filippos Christianos, Georgios Papoudakis, Arrasy Rahman, and Stefano V Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *Advances in Neural Information Processing Systems*, 2021.

Daphne Cornelisse and Eugene Vinitsky. Human-compatible driving partners through data-regularized self-play reinforcement learning. *Reinforcement Learning Journal*, 5:2320–2344, 2024.

Daphne Cornelisse, Aarav Pandya, Kevin Joseph, Joseph Suárez, and Eugene Vinitsky. Building reliable sim driving agents by scaling self-play. *arXiv preprint arXiv:2502.14706*, 2025.

Marco Cusumano-Towner, David Hafner, Alexander Hertzberg, Brody Huval, Aleksei Petrenko, Eugene Vinitsky, Erik Wijmans, Taylor W. Killian, Stuart Bowers, Ozan Sener, Philipp Kraehenbuehl, and Vladlen Koltun. Robust autonomy emerges from self-play. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11710–11737. PMLR, 2025.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning (CoRL)*, 2023.

Aron Distelzweig, Faris Janjoš, Andreas Look, Anna Rothenhäusler, Daniel Jost, Oliver Scheel, Raghu Rajan, Daphne Cornelisse, Eugene Vinitsky, and Joschka Boedecker. Beyond self-play and scale: A behavior benchmark for generalization in autonomous driving. *arXiv preprint arXiv:2605.10034*, 2026a.

Aron Distelzweig, Faris Janjoš, Oliver Scheel, Sirish Reddy Varra, Raghu Rajan, and Joschka Boedecker. Perfect prediction or plenty of proposals? what matters most in planning for autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026b. arXiv:2510.15505.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, pp. 1407–1416, 2018.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Ke Guo, Haochen Liu, Xiaojun Wu, and Chen Lv. DecompGAIL: Learning realistic traffic behaviors with decomposed multi-agent generative adversarial imitation learning. *arXiv preprint arXiv:2510.06913*, 2026. Accepted to ICLR.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2255–2264, 2018.

Marcel Hallgarten, Martin Stoll, and Andreas Zell. From prediction to planning with goal conditioned lane graph traversals. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

Marcel Hallgarten, Julian Zapata, Martin Stoll, Katrin Renz, and Andreas Zell. Can vehicle motion planning generalize to realistic long-tail scenarios? In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51 (5):4282–4286, 1995.

Dirk Helbing, Luboš Buzna, Anders Johansson, and Torsten Werner. Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transportation Science*, 39(1):1–24, 2005.

Zhiyu Huang, Haochen Liu, and Chen Lv. GameFormer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhiyu Huang, Peter Karkus, Boris Ivanovic, Yuxiao Chen, Marco Pavone, and Chen Lv. DTPP: Differentiable joint conditional prediction and cost evaluation for tree policy planning in autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castañeda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

Bernhard Jaeger, Daniel Dauner, Jens Beißwenger, Simon Gerstenecker, Kashyap Chitta, and Andreas Geiger. CaRL: Learning scalable planning policies with simple rewards. *arXiv preprint arXiv:2504.17838*, 2025.

Max Jiang, Yijing Bai, Andre Cornman, Christopher Davis, Xiukun Huang, Hong Jeon, Sakshum Kulshrestha, John Lambert, Shuangyu Li, Xuanyu Zhou, et al. SceneDiffuser: Efficient and controllable driving simulation initialization and rollout. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55729–55760, 2024.

Saman Kazemkhani, Aarav Pandya, Daphne Cornelisse, Brennan Shacklett, and Eugene Vinitsky. GPUDrive: Data-driven, multi-agent driving simulation at 1 million fps. In *International Conference on Learning Representations (ICLR)*, 2025.

Arne Kesting, Martin Treiber, and Dirk Helbing. General lane-changing model MOBIL for car-following models. *Transportation Research Record*, 1999(1):86–94, 2007.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Michael Küpper and Armin Seyfried. Identification of social groups and waiting pedestrians at railway platforms using trajectory data. *PLoS ONE*, 18(3):e0282526, 2023. doi: 10.1371/journal.pone.0282526.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Francisco Luque Sánchez, Isabelle Hupont, Siham Tabik, and Francisco Herrera. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion*, 64:318–335, 2020. doi: 10.1016/j.inffus.2020.07.008.

Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, Brandyn White, and Dragomir Anguelov. The Waymo open sim agents challenge. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 59151–59171, 2023.

Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz. How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences (PNAS)*, 108(17):6884–6888, 2011.

Johan Obando-Ceron, Ghada Sokar, Timon Willi, Clare Lyle, Jesse Farebrother, Jakob Foerster, Gintare Karolina Dziugaite, Doina Precup, and Pablo Samuel Castro. Mixtures of experts unlock parameter scaling for deep RL. In *International Conference on Machine Learning (ICML)*, 2024.

Muleilan Pei, Shaoshuai Shi, and Shaojie Shen. Advancing multi-agent traffic simulation via R1-style reinforcement fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhenghao Peng, Wenjie Luo, Yiren Lu, Tianyi Shen, Cole Gulino, Ari Seff, and Justin Fu. Improving agent behaviors with RL fine-tuning for autonomous driving. In *European Conference on Computer Vision*, pp. 165–181. Springer, 2024.

Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Amir Rasouli and Iuliia Kotseruba. PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Luke Rowe, Roger Girgis, Anthony Gosselin, Bruno Carrez, Florian Golemo, Felix Heide, Liam Paull, and Christopher Pal. CtRL-Sim: Reactive and controllable driving agents with offline reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2024.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, pp. 683–700, 2020.

Oliver Scheel, Luca Bergamini, Maciej Wolczyk, Błażej Osiński, and Peter Ondruska. Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *Conference on Robot Learning (CoRL)*, pp. 718–728, 2022.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. TrafficSim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10400–10409, 2021.

Shuhan Tan, Boris Ivanovic, Yuxiao Chen, Boyi Li, Xinshuo Weng, Yulong Cao, Philipp Krähenbühl, and Marco Pavone. Promptable closed-loop traffic simulation. In *Conference on Robot Learning (CoRL)*, 2024.

Tianyi Tan, Yinan Zheng, Ruiming Liang, Zexu Wang, Kexin Zheng, Jinliang Zheng, Jianxiong Li, Xianyuan Zhan, and Jingjing Liu. Flow matching-based autonomous driving planning with advanced interactive behavior modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.11083.

Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62:1805–1824, 2000.

Jur van den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research (Springer Tracts in Advanced Robotics)*, pp. 3–19. Springer, 2011.

Hado P van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Advances in Neural Information Processing Systems*, 2016.

Eugene Vinitsky, Nathan Lichtlé, Xiaomeng Yang, Brandon Amos, and Jakob Foerster. Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Lingguang Wang, Ömer Şahin Taş, Marlon Steiner, and Christoph Stiller. FlowDrive: Moderated flow matching with data balancing for trajectory planning. *arXiv preprint arXiv:2509.21961*, 2025.

Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9876–9886. PMLR, 2020.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations (ICLR)*, 2021.

Zilin Wang, Saeed Rahmani, Daphne Cornelisse, Bidipta Sarkar, Alexander David Goldie, Jakob Nicolaus Foerster, and Shimon Whiteson. Learning to drive in new cities without human demonstrations. *arXiv preprint arXiv:2602.15891*, 2026.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. SMART: Scalable multi-agent real-time motion generation via next-token prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 114048–114071, 2024.

Zhouchonghao Wu, Raymond Song, Vedant Mundheda, Luis E. Navarro-Serment, Christof Schoenborn, and Jeff Schneider. TADPO: Reinforcement learning goes off-road. *arXiv preprint arXiv:2603.05995*, 2026.

Danfei Xu, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. BITS: Bi-level imitation for traffic simulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. SocialVAE: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision (ECCV)*, pp. 511–528, 2022.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. In *Advances in Neural Information Processing Systems*, 2020.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative, multi-agent games. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 35, 2022.

Hang Yu, Ye Jin, Alessandro Canevaro, Julian Schmidt, Julian Jordan, Peizheng Li, Marc Kaufeld, Silvan Lindner, Johannes Betz, and Wilhelm Stork. G2DP: Diffusion planning with spatio-temporal grid guidance. *arXiv preprint arXiv:2606.26017*, 2026.

Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5422–5432, 2025.

Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Diffusion-based planning for autonomous driving with flexible guidance. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.15564.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25, 2024.

Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Ho. BehaviorGPT: Smart agent simulation for autonomous driving with next-patch prediction. In *Advances in Neural Information Processing Systems*, 2024.

A PER-TYPE DYNAMICS

This section gives the full continuous-time integrators for each of the three dynamics models summarized in §3.1. All models share a common kinematic state $s_t = (x, y, \theta, v_x, v_y, a_{\text{long}}, a_{\text{lat}}, \delta)$ and operate at $\Delta t = 0.1$ s. Per-type kinematic parameters and domain-randomization ranges appear in Table 5.

A.1 VEHICLE: JERK-INTEGRATED ACKERMANN ARC

The vehicle model operates at the *jerk* level: the policy selects longitudinal and lateral jerk values from a discrete grid $\dot{a}_{\text{long}} \in \{-15, -4, 0, 4\}$ m/s^3 and $\dot{a}_{\text{lat}} \in \{-4, 0, 4\}$ m/s^3, giving $|\mathcal{A}| = 12$ actions.

Acceleration Integration. Per-episode coefficients $C_{\text{throttle}}, C_{\text{steer}}$ scale the jerk inputs:

$$a_{\text{long}}^{(t)} = a_{\text{long}}^{(t-1)} + C_{\text{throttle}}\, \dot{a}_{\text{long}}\, \Delta t, \tag{1}$$

$$a_{\text{lat}}^{(t)} = a_{\text{lat}}^{(t-1)} + C_{\text{steer}}\, \dot{a}_{\text{lat}}\, \Delta t. \tag{2}$$

When longitudinal acceleration changes sign ($a_{\text{long}}^{(t-1)} \cdot a_{\text{long}}^{(t)} < 0$), it is snapped to zero, enabling smooth waiting and constant-velocity trajectories. A per-episode coefficient C_{acc} scales the upper acceleration bound before clamping:

$$a_{\text{long}}^{(t)} \leftarrow \text{clip}\Big(a_{\text{long}}^{(t)},\ -5.0,\ 2.5\, C_{\text{acc}}\Big), \qquad a_{\text{lat}}^{(t)} \leftarrow \text{clip}\Big(a_{\text{lat}}^{(t)},\ -4.0,\ 4.0\Big). \tag{3}$$

Velocity Update. Speed integrates by the trapezoidal rule and is clipped with a C_{vel}-scaled upper bound:

$$v^{(t)} = v^{(t-1)} + \tfrac{1}{2}\Big(a_{\text{long}}^{(t)} + a_{\text{long}}^{(t-1)}\Big)\, \Delta t, \qquad v^{(t)} \leftarrow \text{clip}\Big(v^{(t)},\ -2.0,\ 20.0\, C_{\text{vel}}\Big). \tag{4}$$

Velocity is snapped to zero on sign change, as with acceleration.

Steering Angle. The target steering angle is derived from the desired curvature $\kappa = a_{\text{lat}}^{(t)} \,/\, \max\big((v^{(t)})^2,\, \epsilon\big)$ with $\epsilon = 10^{-5}$:

$$\phi_{\text{target}} = \arctan(\kappa\, l_{\text{wb}}), \tag{5}$$

$$\Delta\phi = \text{clip}\Big(\phi_{\text{target}} - \phi^{(t-1)},\ -\dot{\delta}_{\max}\, \Delta t,\ \dot{\delta}_{\max}\, \Delta t\Big), \tag{6}$$

$$\phi^{(t)} = \text{clip}\Big(\phi^{(t-1)} + \Delta\phi,\ -\phi_{\max},\ \phi_{\max}\Big), \tag{7}$$

where $l_{\text{wb}} = 0.6 \times$ *length* is the wheelbase, $\dot{\delta}_{\max} = 0.6$ rad/s is the slew rate limit, and $\phi_{\max} = 0.55$ rad is the hard steering cap. After clamping, curvature and lateral acceleration are recomputed from the realized steering angle to maintain consistency: $\kappa \leftarrow \tan\phi^{(t)} / l_{\text{wb}}$, $a_{\text{lat}}^{(t)} \leftarrow (v^{(t)})^2\, \kappa$.

Position Update. The pose advances along a circular arc:

$$d = \tfrac{1}{2}(v^{(t)} + v^{(t-1)})\, \Delta t, \qquad \vartheta = d\, \kappa, \qquad \Delta x_{\text{local}} = \frac{\sin\vartheta}{\kappa}, \qquad \Delta y_{\text{local}} = \frac{1 - \cos\vartheta}{\kappa}, \tag{8}$$

rotated into the global frame by the current heading; heading updates as $\theta^{(t)} = \theta^{(t-1)} + \vartheta$. When $|\kappa| < 10^{-5}$ (nearly straight), the arc degenerates to a line segment $(\Delta x_{\text{local}},\, \Delta y_{\text{local}}) = (d,\, 0)$.

A.2 PEDESTRIAN: UNICYCLE

Pedestrian motion is modeled as a unicycle where heading is an *integrated state* and velocity is constrained to lie along the heading direction (zero slip). The action is a pair $(a_v,\, \omega)$, longitudinal acceleration and heading rate, selected from a nonuniform 7×7 grid plus a dedicated STOP action:

$$a_v \in \{-2.0,\, -0.8,\, -0.2,\, 0,\, 0.2,\, 0.8,\, 2.0\} \text{ m/s}^2,$$
$$\omega \in \{-2.0,\, -0.5,\, -0.1,\, 0,\, 0.1,\, 0.5,\, 2.0\} \text{ rad/s},$$

giving $|\mathcal{A}| = 50$ (49 grid + 1 stop). The grid concentrates resolution near zero where typical pedestrian accelerations and heading changes lie (Waymo WOMD p_{90}: $|a_v| \approx 1.5$ m/s^2, $|\omega| \approx 1.3$ rad/s). The STOP action sets both a_v and ω to zero and forces speed to zero, providing a clean discrete "stand still" primitive.

Integration. After clipping ($a_v \leftarrow \text{clip}(a_v, -2.0, 2.0)$, $\omega \leftarrow \text{clip}(\omega, -2.0, 2.0)$):

$$v^{(t)} = \text{clip}\Big(v^{(t-1)} + a_v\,\Delta t,\ 0,\ 3.0\Big)\,, \tag{9}$$

$$\theta^{(t)} = \theta^{(t-1)} + \omega\,\Delta t, \tag{10}$$

$$\mathbf{v}^{(t)} = v^{(t)}\Big(\cos\theta^{(t)},\ \sin\theta^{(t)}\Big)\,, \tag{11}$$

$$\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)} + \mathbf{v}^{(t)}\,\Delta t. \tag{12}$$

Speed is clipped to $[0, 3.0]$ m/s (walking/jogging cap). Body-frame accelerations for observation and reward parity with other models are derived numerically: $a_{\text{long}} = (v^{(t)} - v^{(t-1)})/\Delta t$, $a_{\text{lat}} = v^{(t)} \cdot (\theta^{(t)} - \theta^{(t-1)})/\Delta t$.

A.3 CYCLIST: KINEMATIC BICYCLE

Cyclists use a kinematic bicycle model where the action directly controls acceleration a and front-wheel steering angle δ. The discrete grid is 7×7 with **logarithmic spacing**:

$$a \in \{-5.0, -2.0, -0.5, 0, 0.5, 2.0, 5.0\}\ \text{m/s}^2,$$
$$\delta \in \{-0.20, -0.06, -0.02, 0, 0.02, 0.06, 0.20\}\ \text{rad},$$

giving $|\mathcal{A}| = 49$. The geometric ratio (≈ 3) concentrates roughly half of all actions within the low-magnitude regime where fine steering matters most for typical urban cycling, providing about $3\times$ finer resolution near zero for both acceleration and steering than a uniform 7-point grid with the same bounds.

Scaling and Clipping. Per-episode coefficients $C_{\text{steer}}, C_{\text{acc}}, C_{\text{vel}}$ scale the action before integration:

$$\delta \leftarrow C_{\text{steer}}\,\delta, \qquad a \leftarrow \text{clip}(\tilde{a},\ -5.0,\ 3.0)\,, \qquad \tilde{a} = \begin{cases} C_{\text{acc}}\,a & a > 0, \\ a & a \le 0, \end{cases} \tag{13}$$

where C_{acc} scales only forward acceleration (braking is not power-limited).

Velocity Update.

$$v^{(t)} = v^{(t-1)} + a\,\Delta t, \qquad v^{(t)} \leftarrow \text{clip}\Big(v^{(t)},\ -1.0,\ 10.0\,C_{\text{vel}}\Big)\,. \tag{14}$$

Velocity is snapped to zero on sign change.

Slip Angle and Yaw Rate. The slip angle β is computed via a $\tanh$-smoothed approximation that prevents numerical instability at large steering angles:

$$\beta = \tanh\big(\tfrac{1}{2}\tan\delta\big)\,, \qquad \dot{\theta} = \frac{v^{(t)}\cos\beta\,\tan\delta}{L}, \tag{15}$$

where $L = \max(\textit{length}, 0.1)$ m is the effective wheelbase, guarded against degenerate zero-length entities.

Position and Heading Update.

$$\theta^{(t)} = \theta^{(t-1)} + \dot{\theta}\,\Delta t, \tag{16}$$

$$\mathbf{v}^{(t)} = v^{(t)}\Big(\cos(\theta^{(t-1)} + \beta),\ \sin(\theta^{(t-1)} + \beta)\Big)\,, \tag{17}$$

$$\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)} + \mathbf{v}^{(t)}\,\Delta t. \tag{18}$$

Note that the velocity direction $\theta^{(t-1)} + \beta$ differs from the heading $\theta^{(t)}$ by the slip angle during turns. Lateral acceleration is derived as $a_{\text{lat}} = v^{(t)} \cdot \dot{\theta}$.

A.4 Domain Randomization Parameters

Each dynamics model is parameterized by per-episode coefficients $C_{\text{throttle}}, C_{\text{steer}}, C_{\text{acc}}, C_{\text{vel}}$ sampled uniformly from type-specific ranges at each episode reset (Table 5). Vehicle coefficients span a wide $[0.5,\ 1.5]$ band to encourage robustness across car-to-truck-scale dynamics; cyclist coefficients use a narrower $[0.7,\ 1.3]$ band reflecting the tighter physical envelope of bicycle kinematics; pedestrian coefficients are fixed at 1.0 since the unicycle model already covers the behaviorally relevant range. All four coefficients are included as observation features so that the policy can condition on its current kinematic regime.

Table 5: Per-type kinematic parameters and domain-randomization ranges. Coefficient ranges (C_*) are sampled uniformly per agent per episode; scalar values are fixed. Clip bounds are hard limits enforced by the dynamics integrator. Bounding-box ranges are sampled uniformly per agent at spawn, calibrated to the Waymo Open Motion Dataset distribution. Vehicles additionally draw a Bernoulli truck/car bodywork class that selects from an extended length range $[6.1,\ 12.2]$ m for trucks.

Parameter	Vehicle	Pedestrian	Cyclist
Dynamics model	jerk-Ackermann	unicycle	kin. bicycle
C_{throttle}	$[0.5,\ 1.5]$	1.0	1.0
C_{steer}	$[0.5,\ 1.5]$	1.0	$[0.7,\ 1.3]$
C_{acc}	$[0.5,\ 1.5]$	1.0	$[0.7,\ 1.3]$
C_{vel}	$[0.5,\ 1.5]$	1.0	$[0.7,\ 1.3]$
v clip (m/s)	$[-2.0,\ 20.0]$	$[0.0,\ 3.0]$	$[-1.0,\ 10.0]$
a_{long} clip (m/s^2)	$[-5.0,\ 2.5]$	$[-2.0,\ 2.0]$	$[-5.0,\ 3.0]$
a_{lat} clip (m/s^2)	$[-4.0,\ 4.0]$	$[-4.0,\ 4.0]$	$[-4.0,\ 4.0]$
$\lvert\delta\rvert$ cap (rad)	0.55	—	—
$\lvert\dot{\delta}\rvert$ slew (rad/s)	0.6	—	—
$\lvert\omega\rvert$ cap (rad/s)	—	2.0	—
Length (m)	$[3.5,\ 5.5]$	$[0.65,\ 1.0]$	$[1.4,\ 2.0]$
Width (m)	$[1.5,\ 2.5]$	$[0.65,\ 1.0]$	$[0.71,\ 1.0]$
Height (m)	$[1.2,\ 2.0]$	$[1.5,\ 1.9]$	$[1.5,\ 1.9]$

B Random Initialization Details

Map memory is partitioned into a **two-band entity layout**: index range $[0, N_{\text{log}})$ for log-replay entities and $[N_{\text{log}}, N_{\text{log}} + N_{\text{rand}})$ for randomly placed controlled agents. This layout is a simulator capability. The reported SHRED training run uses only the randomly placed controlled band and does not replay logged future trajectories.

For each type $k \in \mathcal{K}$ we draw up to N_k agents by **rejection sampling**: vehicles snap to lane segments with heading noise $\Delta\theta \sim \mathcal{U}(-\pi/7.2, \pi/7.2)$, initial speed $v_0 \sim \mathcal{U}(0, 2)$ m/s, and a Bernoulli truck/car bodywork draw; pedestrians sample a crosswalk with probability p_{cw} and otherwise a road edge; cyclists use the same lane-sampling pipeline as vehicles. Each candidate is rejected on collision or off-road, with at most 100 attempts per slot.

Per-Type Goal Generation. When an agent is randomly spawned (or reaches its current goal), a new goal is synthesized from map topology rather than copied from the log. The generation strategy is type-specific:

- **Vehicles** draw a target distance $d \sim \mathcal{U}(d_{\min}, d_{\max})$ and walk forward along the lane topology graph, randomly selecting successors at lane ends (up to 10 transitions), producing goals that follow the road structure. Candidate goals are filtered for heading alignment ($\geq \cos 45°$).
- **Pedestrians** spawn on either a crosswalk (goal at the opposite edge) or a road edge (goal along the edge polyline with a sidewalk-side lateral offset), with distance bounds $d \in [10, 30]$ m. Subsequent goals chain along the nearest road edge using the pedestrian's current heading for directionality.

- **Cyclists** follow the same lane-topology walk as vehicles, with distance bounds $d \in [20, 60]$ m.

To prevent the policy from learning to exploit goal-failure removal as a shortcut, agents that fail to obtain a valid goal are *not* removed from the scene. Instead, they receive an additional binary observation feature indicating the absence of an active goal, and the policy must learn to act sensibly under partial supervision.

C Per-Type Reward Details

Each agent's reward is a weighted sum of terms spanning safety, progress, comfort, and type-specific constraints. Let $r_t^k = \sum_i \alpha_i^k \, r_{i,t}$ denote the reward for an agent of type k at step t, where each weight α_i^k is sampled independently from a type-specific uniform range at episode reset (reward domain randomization). Table 6 lists the full per-type reward structure.

Comfort. The comfort term counts the number of kinematic channels that exceed a type-specific threshold:

$$N_{\text{comfort}} = \sum_{i \in \{a_l,\, a_t,\, j_l,\, j_t\}} \mathbb{1}\left[|x_i| > \tau_i^{(k)}\right],$$

with $r_{\text{comfort}} = -\alpha_{\text{comfort}} \, N_{\text{comfort}}$. Table 7 lists the thresholds, calibrated per type from an 80k-scenario analysis of the Waymo Open Motion Dataset at the p_{95}–$p_{97.5}$ percentile of real-data magnitudes.

Pedestrian Road Intrusion. Pedestrians receive a per-step jaywalking penalty when their center is on the road surface and *not* within a crosswalk polygon:

$$r_{\text{intrusion}} = -\alpha_{\text{intrusion}} \cdot \Delta t \cdot \mathbb{1}[\text{on_road}] \cdot \mathbb{1}[\neg\, \text{on_crosswalk}].$$

This term is zero for vehicles and cyclists, who are expected to occupy road lanes.

Type-Specific Lane and Speed Terms. Vehicles receive lane alignment ($\alpha_{l\text{-align}}$), lane centering ($\alpha_{l\text{-center}}$), velocity alignment ($\alpha_{v\text{-align}}$), and reverse penalties (α_{reverse}) that are absent for pedestrians and cyclists. Cyclists receive a softer version of lane alignment plus an edge-preference bonus (α_{edge}) that encourages riding near the road margin. Both pedestrians and cyclists receive a speed-limit penalty $r_{\text{speed}} = -\alpha_{\text{speed-lim}} \cdot \Delta t \cdot (v - v_{\text{lim}})/v_{\max}$ when $v > v_{\text{lim}}$, with v_{lim} sampled per agent from $[1.5,\ 2.5]$ m/s for pedestrians and $[5.0,\ 8.0]$ m/s for cyclists (Table 6). Similar reward design strategies have been explored in off-road (Wu et al., 2026) and urban autonomous driving settings (Jaeger et al., 2025).

D PopArt Weight Correction

Each per-type PopArt layer maintains running statistics (μ_k, σ_k, ν_k) updated by EMA over each epoch's batch returns:

$$\mu_k \leftarrow (1-\beta_k)\mu_k + \beta_k \, \bar{G}_k, \qquad \nu_k \leftarrow (1-\beta_k)\nu_k + \beta_k \, \overline{G^2}_k, \qquad \sigma_k \leftarrow \sqrt{\nu_k - \mu_k^2},$$

where $\beta_k = 1 - \rho^{T_{\text{bptt}}}$ with configurable decay $\rho = 0.9997$. To keep the real-scale value function unchanged when (μ_k, σ_k) shifts, the linear weights of Value^k are corrected:

$$W^k \leftarrow \frac{\sigma_k^{\text{old}}}{\sigma_k^{\text{new}}} \, W^k, \qquad b^k \leftarrow \frac{\sigma_k^{\text{old}} b^k + \mu_k^{\text{old}} - \mu_k^{\text{new}}}{\sigma_k^{\text{new}}},$$

ensuring no policy-gradient discontinuity at the commit step.

E WOSAC Evaluation Protocol and Additional Results

E.1 WOSAC 2024 Validation

Training remains procedural and does not use logged-trajectory replay.

Table 6: Per-type reward terms and domain-randomization ranges. Dashes mark terms that are disabled ($\alpha = 0$) for that type. *Const.* values are not randomized; ranges denote uniform sampling per agent per episode. All weights are multiplied by $\Delta t = 0.1$ s per step except collision and goal, which fire as one-shot events.

Reward term	Vehicle	Pedestrian	Cyclist
$\alpha_{\text{collision}}$	[2.0, 4.0]	[0.5, 2.0]	[1.0, 3.0]
$\alpha_{\text{coll-speed}}$	0.1	0.0	0.0
$m_{k'}$ (partner mult.)	[1, 2, 1]	—	—
α_{boundary}	[2.0, 4.0]	—	[0.5, 2.0]
α_{comfort}	[0.0, 0.1]	—	—
$\alpha_{l\text{-align}}$	[2.5e-4, 2.5e-2]	—	[5e-3, 2e-2]
$\alpha_{v\text{-align}}$	[0.0, 1.0]	—	[0.25, 0.75]
$\alpha_{l\text{-center}}$	[2.5e-4, 7.5e-3]	—	—
α_{reverse}	[2.5e-4, 7.5e-3]	—	—
$\alpha_{\text{intrusion}}$	—	[0.5, 2.0]	—
$\alpha_{\text{speed-lim}}$	—	[0.5, 2.0]	[0.25, 1.0]
v_{lim} (m/s)	—	[1.5, 2.5]	[5.0, 8.0]
α_{edge}	—	—	[0.01, 0.05]
δ_{goal} (m)	[2.0, 12.0]	[2.0, 12.0]	[2.0, 12.0]
α_{goal}	1.0	1.0	1.0

Table 7: Per-type comfort thresholds $\tau^{(k)}$, calibrated at the p_{95}–$p_{97.5}$ percentile of the Waymo Open Motion Dataset. Pedestrian and cyclist thresholds are calibrated but unused in the reported run, where $\alpha_{\text{comfort}} = 0$ for those types.

Type	τ_{a_l} (m/s^2)	τ_{a_t} (m/s^2)	τ_{j_l} (m/s^3)	τ_{j_t} (m/s^3)
Vehicle	3.0	3.0	5.0	5.0
Pedestrian	1.5	1.0	12.0	5.0
Cyclist	3.0	1.5	40.0	18.0

Table 8: Agreement between the GPU evaluator and the official WOSAC evaluator on 50 scenarios.

Metric	GPU evaluator	Official evaluator	Maximum difference
Distance to nearest object	0.258244	0.258244	1.2×10^{-7}
Time to collision	0.882676	0.882676	1.3×10^{-7}
Collision indication	0.850935	0.850935	8.0×10^{-8}
Distance to road edge	0.646314	0.646314	1.4×10^{-7}
Off-road indication	0.867736	0.867736	8.7×10^{-8}
Linear speed	0.236808	0.236271	3.7×10^{-3}
Linear acceleration	0.365495	0.367187	1.8×10^{-2}
Angular speed	0.366593	0.365808	4.0×10^{-3}
Angular acceleration	0.586135	0.585067	8.1×10^{-3}
Realism meta-score	0.686143	0.686108	3.5×10^{-5}

Table 9: WOSAC 2025 test evaluation for SHRED.

Metric	Value
Realism meta-score	0.722
Linear speed	0.162
Linear acceleration	0.337
Angular speed	0.343
Angular acceleration	0.574
Distance to nearest object	0.316
Collision	0.951
Time to collision	0.806
Distance to road edge	0.529
Off-road	0.907
minADE (m)	5.920

We compute the realism meta-score with the official weighted definition and prepend logged history when computing kinematic features. On 50 WOSAC scenarios, the GPU evaluator and the official evaluator agree within 3.5×10^{-5} on the realism meta-score.

E.2 WOSAC 2025 Test Evaluation

Table 9 reports the WOSAC 2025 test evaluation. The 2025 metric differs from the 2023 and 2024 metrics, so we report it separately.

E.3 Full InterPlan Comparison

Table 10 expands the representative comparison in Table 4 to include all baselines and their listed variants.

Table 10: Full comparison of interactive planners on InterPlan. InterPlan (80) is the official 80-scenario split, and Full Scale (335) evaluates all 335 scenarios. SHRED achieves the highest score on both evaluation sets. Vehicle-only self-play learns vehicle control with scripted pedestrians and no cyclists.

Planner	Type	InterPlan (80) ↑	Full Scale (335) ↑
IDM (Treiber et al., 2000)	Rule	31	—
IDM+MOBIL (Treiber et al., 2000; Kesting et al., 2007)	Rule	31	—
PDM-Closed (Dauner et al., 2023)	Rule	42	—
SPDM (N_p=15) (Distelzweig et al., 2026b)	Rule	42.00	—
SPDM (N_p=60) (Distelzweig et al., 2026b)	Rule	63.66	—
DTPP (Huang et al., 2024)	Hybr.	25	—
HybridLLMPlanner (GPT-3.5) (Hallgarten et al., 2024)	Hybr.	40	—
HybridLLMPlanner (Llama-13B) (Hallgarten et al., 2024)	Hybr.	48	—
HybridLLMPlanner (Llama-7B) (Hallgarten et al., 2024)	Hybr.	53	—
PDM+PPO (Distelzweig et al., 2026a)	Hybr.	43.2	—
FlowDrive w/ guidance + PDM scoring (Wang et al., 2025)	Hybr.	44.05	—
Urban Driver (Scheel et al., 2022)	IL	4	—
GC-PGP (Hallgarten et al., 2023)	IL	10	—
GameFormer (Huang et al., 2023)	IL	11	—
PDM-Open (Dauner et al., 2023)	IL	25	—
Diffusion Planner (Zheng et al., 2025)	IL	25.8	—
FlowDrive (Wang et al., 2025)	IL	36.96	—
Flow Planner (Tan et al., 2025)	IL	—	61.82
G2DP (Yu et al., 2026)	IL	—	61.74
PPO (Distelzweig et al., 2026a)	RL	42.1	—
SHRED	RL	**67.50**	**70.14**
SHRED (vehicle-only)	RL	54.19	66.65