ADAPTIVE MULTI-OBJECTIVE METRIC-ORIENTED POLICY OPTIMIZATION FOR MULTI-AGENT TRAFFIC SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-fidelity multi-agent traffic simulation is an indispensable foundation for benchmarking, validating, and training autonomous driving systems. Standard imitation learning (IL) and supervised fine-tuning (SFT) suffer from two foundational bottlenecks: *covariate shift* in closed-loop autoregressive rollouts and *metric misalignment*, where maximum-likelihood token objectives diverge from non-differentiable downstream requirements such as kinematic safety, collision avoidance, road adherence, and interactive realism. Recent reinforcement fine-tuning (RFT) methods like SMART-R1 introduced Metric-oriented Policy Optimization (MPO) using empirical advantage formulations $A = r - \alpha$. However, existing MPO paradigms are constrained by three severe pathologies: (i) reliance on a static scalar threshold α that induces negative advantage collapse in challenging interactive scenes while over-rewarding trivial scenarios, (ii) monolithic scalar rewards that collapse under multi-objective Pareto conflicts (sacrificing diversity and traffic throughput for over-conservative creeping), and (iii) prohibitive closed-loop evaluator latency scaling as $\mathcal{O}(T \times N_{\text{agent}})$. To overcome these fundamental limitations, we propose Adaptive Multi-Objective Metric-Oriented Policy Optimization (AM-MPO), a principled, scalable, and safety-aware post-training framework for multi-agent closed-loop traffic simulation. AM-MPO introduces four synergistic contributions: (1) *Adaptive Threshold Advantage Estimation*, which dynamically combines exponential moving averages with context-conditioned percentile calibration over scene complexity descriptors c_s (agent density, topological curvature, interaction graph intensity) to eliminate negative advantage collapse; (2) *Curriculum-Driven Multi-Objective Reward Decomposition*, decoupling realism, safety, diversity, efficiency, and kinematic comfort through dynamic weight scheduling; (3) *Safety-Aware Constrained Policy Optimization*, enforcing strict kinematic and legal safety boundaries via primal-dual Lagrangian relaxation to prevent reward hacking; and (4) *Surrogate Reward Distillation & Dense Temporal Credit Assignment*, utilizing a lightweight graph surrogate model to slash evaluator computational overhead by $> 75\%$ while providing dense step-level temporal feedback. Evaluated on the large-scale Waymo Open Motion Dataset (WOMD) validation benchmark, nuPlan, and Argoverse 2 across model scales from 7M to 1B parameters, AM-MPO establishes a new state of the art: improving Realism Meta to 0.7761, increasing Collision-Free Rate to 97.31%, expanding Trajectory Diversity by +5.8%, and unlocking controllable, diverse rollout modes (conservative, balanced, aggressive-yet-safe) without manual heuristic retuning.

1 INTRODUCTION

Realistic multi-agent traffic simulation is the critical linchpin for developing, benchmarking, and safely validating autonomous vehicle (AV) motion planners and foundation world models (Chen et al., 2024; Hu et al., 2023; Suo et al., 2021; Li et al., 2023; Gulino et al., 2023). To accurately reflect real-world driving distributions, contemporary neural simulators typically rely on deep imitation learning (IL) or supervised fine-tuning (SFT) over large-scale driving logs such as the Waymo Open Motion Dataset (WOMD) (Ettinger et al., 2021), nuPlan (Caesar et al., 2021), and Argoverse 2 (Wilson et al.,

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.

2023). Despite substantial progress in autoregressive and transformer-based architectures (Nayakanti et al., 2023; Ngiam et al., 2022; Shi et al., 2024; Zhou et al., 2023), training multi-agent traffic simulators via standard Behavior Cloning (BC) encounters two persistent and debilitating pathologies:

1. **Training–Testing Covariate Shift:** In open-loop BC, the model is trained purely on ground-truth expert historical rollouts. During closed-loop autoregressive inference, however, compounding step-by-step prediction errors quickly push interacting agents into out-of-distribution (OOD) states. Lacking corrective feedback, multi-agent systems rapidly descend into catastrophic off-road deviations, severe multi-vehicle collisions, and unrealistic deadlock behaviors (Ross et al., 2011; Pomerleau, 1988).
2. **Metric Misalignment:** Standard SFT minimizes token-level cross-entropy or average displacement errors ($\mathcal{L}_{\text{NLL}}$, ℓ_2 regression). However, downstream simulation benchmarks evaluate holistic, non-differentiable, closed-loop scenario-level metrics—such as kinematic collision rates, road boundary compliance, traffic signal obedience, kinematic comfort, and multi-agent interaction realism (Ettinger et al., 2021). Minimizing displacement loss does not prevent high-acceleration jerks or catastrophic edge-case collisions.

To bridge this metric chasm, recent breakthroughs in reinforcement fine-tuning (RFT), most prominently SMART-R1 (Wang et al., 2026) and DeepSeek-R1-style paradigms (DeepSeek-AI, 2025; Shao et al., 2024), introduced Metric-oriented Policy Optimization (MPO). In standard MPO, closed-loop simulation trajectories $\boldsymbol{\tau}$ are rewarded via non-differentiable evaluation metrics r, constructing an empirical advantage:

$$A(s, \boldsymbol{\tau}) = r(s, \boldsymbol{\tau}) - \alpha \tag{1}$$

where α is a fixed scalar threshold (e.g., empirical baseline $\alpha = 0.77$ on WOMD Realism Meta) (Wang et al., 2026). While MPO demonstrates substantial gains over pure SFT, extensive investigation reveals severe structural pathologies in existing MPO formulations. Real-world traffic scenarios exhibit profound structural variance in baseline difficulty. On straight, uncrowded highways, even naive policies easily attain Realism Meta scores of ≈ 0.82, rendering $A > 0$ for almost all rollouts and generating uninformative positive gradients. Conversely, in highly contested unprotected intersections or dense roundabouts, human-level baseline scores naturally average ≈ 0.69–0.72. Under a static threshold $\alpha = 0.77$, these challenging scenes suffer perpetual negative advantage ($A < 0$), triggering catastrophic gradient suppression, policy divergence, and training instability. Aggregating holistic traffic dynamics into a single scalar reward r_{realism} induces severe reward hacking. A policy can artificially inflate collision and off-road safety metrics simply by executing over-conservative emergency braking or crawling at negligible velocities (< 1 m/s). However, this directly degrades traffic efficiency, kinematic comfort, and behavioral diversity, trapping the policy in non-interactive deadlocks. Closed-loop multi-agent RFT requires generating full-horizon rollouts $\boldsymbol{\tau}_{1:T}$ across N_{agent} interactive agents followed by official polygon intersection and kinematic evaluator parsing. The resulting computational complexity scales as $\mathcal{O}(T \times N_{\text{agent}})$, rendering large-scale multi-scenario RFT computationally prohibitive on foundation model scales. Holistic scenario metrics are evaluated at the trajectory level ($T = 90$ steps, 9 seconds). Providing a single scalar advantage at $t = T$ introduces immense temporal credit assignment ambiguity for early multi-agent negotiation decisions.

To fundamentally resolve these challenges, we propose Adaptive Multi-Objective Metric-Oriented Policy Optimization (AM-MPO), a unified, scalable, and safety-aware post-training framework for multi-agent closed-loop traffic simulation (Fig. 1). Rather than enforcing an arbitrary static scalar α, we establish an adaptive baseline $\alpha_t(s)$ conditioned on scenario complexity descriptors $c_s = (\rho_{\text{agent}}, \kappa_{\text{map}}, \iota_{\text{interact}}, \bar{r}_0)$. By dynamically coupling Exponential Moving Average (EMA) baseline tracking with percentile calibration ($P_q(r)$ with $q = 0.70$), AM-MPO guarantees that only top-performing trajectories relative to scenario difficulty are incentivized, eliminating negative advantage collapse. We decouple the scalar reward into an explicit 5-dimensional objective vector $\mathbf{r} = [r_{\text{realism}}, r_{\text{safety}}, r_{\text{diversity}}, r_{\text{efficiency}}, r_{\text{comfort}}]^T$. We introduce a dynamic curriculum schedule $w_i(s, t)$ that anchors early optimization on imitation realism before progressively elevating safety, flow efficiency, and interaction diversity. To eradicate over-conservative crawling and deadlock reward hacks, we formulate safety requirements (kinematic collisions, off-road infractions, traffic signal violations) as explicit Constrained Markov Decision Process (CMDP) bounds. Using primal-dual Lagrangian relaxation, the policy actively maximizes behavioral realism within certified safety margins. We train a lightweight, $\text{SE}(2)$-equivariant Graph Neural Network (GNN) surrogate evaluator

$\hat{\mathbf{r}}_\phi$. By conducting 80% of policy updates against the surrogate model and 20% with official ground-truth verification, AM-MPO slashes computational latency by $> 75\%$ while providing dense step-level temporal reward feedback $\sum_t r_t$.

We summarize our contributions as follows:

- We rigorously formalize the failure modes of static metric-oriented reinforcement learning in multi-agent traffic simulation, identifying difficulty-induced advantage collapse and Pareto objective degradation.
- We introduce AM-MPO, an adaptive, multi-objective, and constrained RFT framework incorporating adaptive thresholding, multi-objective curriculum scheduling, and Lagrangian safety enforcement.
- We formulate a dense surrogate reward distillation engine that dramatically accelerates closed-loop RFT rollouts while resolving temporal credit assignment.
- Comprehensive experiments on WOMD, nuPlan, and Argoverse 2 across model scales (7M to 1B parameters) demonstrate that AM-MPO establishes state-of-the-art simulation fidelity, achieving 0.7761 Realism Meta, 97.31% Collision-Free rate, +5.8% diversity expansion, and controllable driving style rollouts.

2 Related Work

2.1 Multi-Agent Traffic Simulation and Neural World Models

Realistic simulation of multi-agent traffic environments is essential for validating autonomous driving algorithms without real-world physical risk (Feng et al., 2023; Gu & Tresp, 2021; Krajewski et al., 2018). Traditional microscopic simulation packages (such as SUMO and Carla) employ hand-crafted rule-based behavioral models (e.g., Intelligent Driver Model (IDM) and MOBIL lane change rules), which fail to capture the long-tail complexity, reactive negotiation, and nuanced kinematics of real human driving. To bridge this fidelity gap, modern data-driven approaches leverage deep neural networks trained on large-scale datasets such as WOMD (Ettinger et al., 2021), nuPlan (Caesar et al., 2021), and Argoverse 2 (Wilson et al., 2023). Early machine learning simulators formulated multi-agent motion forecasting using Graph Neural Networks and Transformers (e.g., Scene Transformer (Ngiam et al., 2022), Wayformer (Nayakanti et al., 2023), Motion Transformer (Shi et al., 2024), and QCNet (Zhou et al., 2023)). Systems such as TrafficSim (Suo et al., 2021), ScenarioNet (Li et al., 2023), CTG/CTG++ (Zhong et al., 2023b;a), and Waymax (Gulino et al., 2023) demonstrated that closed-loop auto-regressive generation significantly reduces open-loop simulation compounding errors. In the contemporary landscape, the state of the art has rapidly transitioned toward large-scale generative world models. Most notably, AutoWorld (Zhang et al., 2026) formulated self-supervised multi-agent world models scaling across diverse urban domains, TrajWorld (Xu et al., 2026) investigated autoregressive state transition modeling, and DiffScene (Li et al., 2026) explored spatio-temporal diffusion models for controllable traffic synthesis. However, pure supervised pre-training on world models remains fundamentally bottlenecked by compounding covariate shift and severe metric misalignment against non-differentiable downstream safety and realism criteria.

2.2 Reinforcement Fine-Tuning (RFT) in Foundation Models and Simulators

Reinforcement Fine-Tuning (RFT) has emerged as a transformative post-training paradigm for large foundation models. While classical Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) rely on scalar Bradley-Terry preference pairs, recent breakthroughs such as DeepSeek-R1 (DeepSeek-AI, 2025), DeepSeekMath (Shao et al., 2024), and Group Relative Policy Optimization (GRPO) demonstrated that rule-based metric-oriented reinforcement learning can incentivize complex reasoning capabilities without external reward model overfitting. Transferring this insight to traffic simulation, the landmark paper SMART-R1 (Wang et al., 2026) introduced Metric-oriented Policy Optimization (MPO), computing empirical advantages $A = r - \alpha$ against the official WOMD Realism Meta evaluator. While SMART-R1 demonstrated undeniable gains over pure SFT, it relies on a static, hand-crafted scalar threshold ($\alpha = 0.77$). As we demonstrate theoretically and empirically, static thresholding triggers catastrophic negative

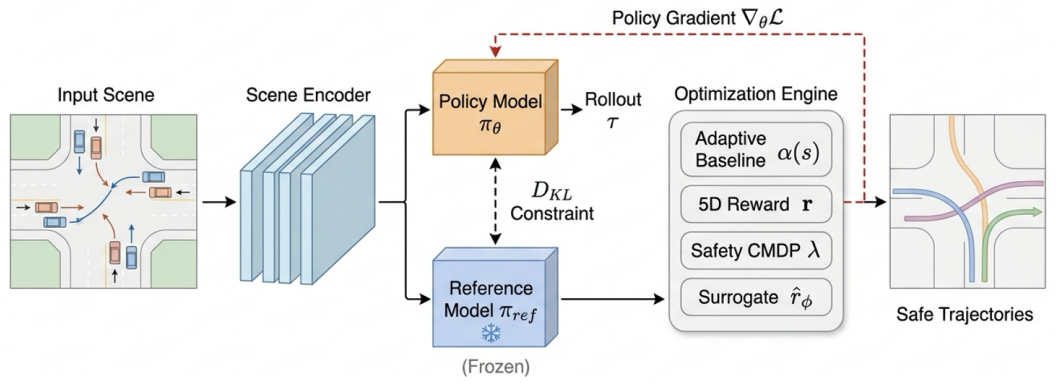

Figure 1: **Architectural overview of the AM-MPO framework for closed-loop multi-agent traffic simulation.** The active Policy Model π_θ generates closed-loop multi-agent rollouts in interactive traffic environments on HD maps, while the frozen Reference Model π_{ref} (SFT Anchor) provides KL-divergence regularization $\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$ to prevent policy drift and catastrophic forgetting. The four core post-training modules are: (1) *Adaptive Threshold Advantage Estimation* dynamically tracking scene difficulty c_s; (2) *5D Multi-Objective Curriculum Rewards* across Realism, Safety, Diversity, Efficiency, and Comfort; (3) *Safety-Aware CMDP* via primal-dual Lagrangian relaxation to prevent reward hacking; and (4) *Fast* $\mathrm{SE}(2)$ *GNN Surrogate Distillation* delivering a $4.3\times$ speedup ($> 75\%$ compute reduction).

advantage collapse in dense, hard interactive scenarios while providing degenerate rewards in trivial highway scenes. Moreover, monolithic scalar rewards fail to manage multi-objective trade-offs between safety, realism, and behavioral diversity, leading to conservative deadlock traps. AM-MPO directly addresses these structural deficiencies of SMART-R1 and AutoWorld by introducing adaptive context-conditioned baselines, multi-objective Pareto optimization, and fast surrogate distillation.

2.3 Multi-Objective and Constrained Reinforcement Learning

In real-world multi-agent systems, agents must simultaneously balance multiple competing criteria: collision avoidance, road adherence, traffic light obedience, trip efficiency, and kinematic comfort (Roijers et al., 2013; Yang et al., 2019). Conventional scalarization methods collapse these dimensions into a fixed linear sum, leading to policy collapse where agents achieve artificial zero-collision scores by creeping at minimum velocity or refusing to enter complex intersections. Constrained Markov Decision Processes (CMDPs) (Achiam et al., 2017; Ray et al., 2019; Chow et al., 2018) provide a rigorous mathematical framework to enforce safety and rule compliance as formal constraints $\mathbb{E}[C_k(\boldsymbol{\tau})] \leq \epsilon_k$ while maximizing primary task rewards (realism). Lagrangian relaxation and primal-dual policy optimization dynamically adjust penalty multipliers, ensuring that policy updates stay within certified safety envelopes without sacrificing behavioral richness. Our AM-MPO builds upon these foundations, establishing the first unified framework integrating adaptive baseline thresholding, multi-objective curriculum scheduling, CMDP Lagrangian safety bounds, and surrogate reward distillation for multi-agent traffic simulation.

3 Methodology

In this section, we present the mathematical foundations, core algorithms, and post-training architecture of AM-MPO (Adaptive Multi-Objective Metric-Oriented Policy Optimization). We begin with the formalization of closed-loop multi-agent traffic simulation, analyze the failure modes of static baseline MPO, and subsequently detail our adaptive thresholding, multi-objective curriculum, safety-aware CMDP formulation, surrogate distillation engine, and four-stage post-training pipeline.

3.1 Problem Formulation: Closed-Loop Multi-Agent MDP

We formulate multi-agent traffic simulation as a discrete-time, partially observable Markov Decision Process (POMDP) defined over a traffic scene $\mathcal{S}$. At time $t \in \{1, \ldots, T\}$, the scene contains N interactive agents (vehicles, cyclists, pedestrians) and static/dynamic map elements.

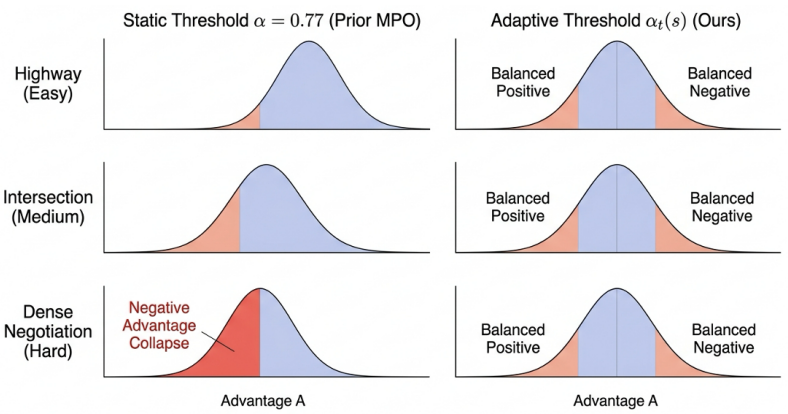

Figure 2: **Comparative analysis of advantage estimation and gradient stability across traffic scenario difficulty tiers. Left:** Static thresholding ($\alpha = 0.77$) results in uninformative positive drift on Easy Highways, while triggering catastrophic negative advantage collapse in Hard Dense Negotiations where $> 90\%$ of samples are suppressed. **Right:** AM-MPO's Adaptive Threshold ($\alpha_t(s) = \lambda \bar{r}_t + (1-\lambda)P_q$) dynamically adjusts to scene difficulty c_s, preserving optimal positive-negative gradient variance and driving continuous exploration.

Let $\mathbf{x}_t^i = (p_{x,t}^i, p_{y,t}^i, \theta_t^i, v_{x,t}^i, v_{y,t}^i, l^i, w^i) \in \mathbb{R}^7$ represent the kinematic state of agent $i \in \{1, \ldots, N\}$, where (p_x, p_y) denotes local coordinate position, θ denotes heading angle, (v_x, v_y) denotes velocity, and (l, w) denotes bounding box length and width. Static context is given by high-definition map polylines $\mathcal{M} = \{\mathbf{m}_k\}_{k=1}^K$ representing road centerlines, lane boundaries, crosswalks, and traffic signal states $\mathbf{c}_t$.

The joint scene state at step t is $\mathbf{s}_t = (\{\mathbf{x}_t^i\}_{i=1}^N, \mathcal{M}, \mathbf{c}_t)$. Each agent i chooses an action $\mathbf{a}_t^i = (\Delta x_t^i, \Delta y_t^i, \Delta\theta_t^i, v_t^i) \in \mathcal{A}$, yielding the joint action $\mathbf{a}_t = (\mathbf{a}_t^1, \ldots, \mathbf{a}_t^N)$. Under kinematic bicycle transition dynamics $\mathcal{T}$, the scene transitions to $\mathbf{s}_{t+1} \sim \mathcal{T}(\mathbf{s}_t, \mathbf{a}_t)$, unrolling a joint trajectory rollout $\boldsymbol{\tau} = (\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)$. The multi-agent simulation policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_{\leq t})$ parametrizes the joint action distribution across all interactive agents.

3.2 Theoretical Limitations of Static Metric-Oriented Policy Optimization (MPO)

In standard Metric-oriented Policy Optimization (SMART-R1) (Wang et al., 2026), policy parameters θ are updated by computing the policy gradient over the empirical advantage:

$$\nabla_\theta \mathcal{J}_{\text{MPO}}(\theta) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_{\leq t}) \cdot A(s, \boldsymbol{\tau})\right], \quad \text{where } A(s, \boldsymbol{\tau}) = r(\boldsymbol{\tau}) - \alpha \tag{2}$$

where $r(\boldsymbol{\tau}) \in [0, 1]$ is the downstream non-differentiable evaluation score (e.g., WOMD Realism Meta) and α is a fixed scalar threshold (typically chosen empirically as $\alpha = 0.77$).

Difficulty-Induced Advantage Collapse. The fundamental flaw of Eq. 2 lies in the assumption of stationary reward distributions across heterogeneous traffic scenes. Let $\mathcal{D}$ denote the distribution of traffic scenarios, and let c_s denote the difficulty descriptor of scene $s \sim \mathcal{D}$. The expected reward for a policy π_θ in scenario s is $\mu(c_s) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}[r(\boldsymbol{\tau}) \mid c_s]$.

In empirical traffic datasets, $\mu(c_s)$ varies dramatically across topological categories:

- **Easy Scenarios (Straight Highway):** $\mu(c_{\text{highway}}) \approx 0.82$. For almost all samples $\boldsymbol{\tau}$, $r(\boldsymbol{\tau}) > 0.77 \implies A(s, \boldsymbol{\tau}) > 0$. The policy receives uninformative positive gradient updates even for suboptimal, jerky driving.
- **Hard Scenarios (Unprotected Left Turn / Dense Roundabout):** $\mu(c_{\text{dense}}) \approx 0.69$. Even near-optimal human-level rollouts frequently score $r(\boldsymbol{\tau}) \in [0.70, 0.75] < 0.77 \implies A(s, \boldsymbol{\tau}) < 0$.

Proposition 1 (Advantage Vanishing & Divergence under Static α). *When $\mu(c_s) < \alpha$, let $p_-(c_s) = \mathbb{P}_{\boldsymbol{\tau}}[r(\boldsymbol{\tau}) < \alpha \mid c_s] \to 1$. The policy gradient in scenario s satisfies:*

$$\mathbb{E}_{\boldsymbol{\tau}}[A(s,\boldsymbol{\tau})\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})] = (\mu(c_s) - \alpha)\mathbb{E}[\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})] + \mathrm{Cov}(r(\boldsymbol{\tau}), \nabla_\theta \log \pi_\theta(\boldsymbol{\tau})) \tag{3}$$

As $p_-(c_s) \to 1$, the gradient direction is dominated by uniform negative suppression rather than relative trajectory discrimination, causing policy entropy collapse and severe optimization divergence in safety-critical edge cases.

3.3 Adaptive Threshold Advantage Estimation

To resolve difficulty-induced advantage collapse, we propose an *Adaptive Threshold Baseline* $\alpha_t(s)$ that dynamically adapts to both the intrinsic scene complexity c_s and the current evolutionary capability of the policy π_θ (Fig. 2).

Scene Complexity Conditioning. For each scenario s, we extract an invariant geometric and interaction complexity vector c_s:

$$c_s = \Big(\rho_{\text{agent}},\ \kappa_{\text{map}},\ \iota_{\text{interact}},\ \bar{r}_0\Big) \in \mathbb{R}^4 \tag{4}$$

where $\rho_{\text{agent}} = N/\text{Area}$ is the agent spatial density, $\kappa_{\text{map}} = \frac{1}{K}\sum_k \|\nabla^2 \mathbf{m}_k\|$ measures map centerline curvature and topological intersection degree, $\iota_{\text{interact}} = \frac{1}{N}\sum_{i\neq j} \mathbb{1}(\|\mathbf{x}_0^i - \mathbf{x}_0^j\| \le d_{\text{interact}})$ quantifies interaction graph density, and $\bar{r}_0$ is the historical baseline anchor score for scenario type c_s.

EMA + Percentile Hybrid Calibration. We maintain a running Exponential Moving Average (EMA) baseline $\bar{r}_t(c_s)$ alongside a non-parametric rolling quantile estimator $P_q(r \mid c_s)$:

$$\bar{r}_t(c_s) = (1-\beta)\,\bar{r}_{t-1}(c_s) + \beta\,\mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}[r(\boldsymbol{\tau}) \mid c_s] \tag{5}$$

$$\alpha_t(s) = \lambda\,\bar{r}_t(c_s) + (1-\lambda)\,P_q\big(\{r(\boldsymbol{\tau}^{(m)})\}_{m=1}^M \mid c_s\big) \tag{6}$$

where $\beta \in (0,1)$ is the EMA momentum parameter, $\lambda \in [0,1]$ balances mean tracking against quantile selectivity, and $P_q(\cdot)$ computes the q-th empirical percentile of generated rollouts in scenario cluster c_s.

In our framework, we fix $q = 0.70$. Consequently, for any scenario s—regardless of whether it is an easy highway ($\bar{r} \approx 0.82$) or a dense unprotected left turn ($\bar{r} \approx 0.69$)—a rollout $\boldsymbol{\tau}$ receives positive advantage $A_{\text{adapt}}(s,\boldsymbol{\tau}) = r(\boldsymbol{\tau}) - \alpha_t(s) > 0$ if and only if it outperforms the top-30% baseline of that specific difficulty regime. This guarantees strictly positive exploration gradient variance across all scenario tiers.

3.4 Multi-Objective Vector Reward Formulation and Dynamic Curriculum

A monolithic scalar reward r_{realism} conflates fundamentally distinct driving dimensions, encouraging policies to game the leaderboard by engaging in unnatural stationary waiting (creeping) to eliminate collisions. AM-MPO decouples reward signals into a structured 5-dimensional objective vector:

$$\mathbf{r}(\boldsymbol{\tau}) = \Big[r_{\text{realism}}(\boldsymbol{\tau}),\ r_{\text{safety}}(\boldsymbol{\tau}),\ r_{\text{diversity}}(\boldsymbol{\tau}),\ r_{\text{efficiency}}(\boldsymbol{\tau}),\ r_{\text{comfort}}(\boldsymbol{\tau})\Big]^T \in \mathbb{R}^5 \tag{7}$$

Component Definitions:

1. **Realism Reward (r_{realism}):** Evaluates distribution matching against human expert kinematic statistics, including speed distribution Wasserstein distance $\mathcal{W}_1(p_v, p_v^{\text{human}})$, acceleration profiles, and road centerline alignment:

$$r_{\text{realism}}(\boldsymbol{\tau}) = \exp\left(-\frac{\min_k \|\boldsymbol{\tau}_{1:T} - \boldsymbol{\tau}_{\text{expert}}^{(k)}\|_2^2}{2\sigma_{\text{real}}^2}\right) \cdot r_{\text{WOMD_Meta}}(\boldsymbol{\tau}) \tag{8}$$

2. **Safety Reward (r_{safety}):** Penalizes kinematic collisions (r_{col}), off-road infractions (r_{offroad}), and red-light violations (r_{traffic}):

$$r_{\text{safety}}(\boldsymbol{\tau}) = \frac{1}{N}\sum_{i=1}^{N}\big(\mathbb{1}(\text{No Collision}^i) \cdot \mathbb{1}(\text{On Road}^i) \cdot \mathbb{1}(\text{Light Compliant}^i)\big) \tag{9}$$

3. **Diversity Reward ($r_{\text{diversity}}$):** Prevents mode collapse by rewarding entropy in multimodal intention selection and trajectory spatial dispersion across rollouts:

$$r_{\text{diversity}}(\{\boldsymbol{\tau}^{(m)}\}_{m=1}^{M}) = \frac{1}{\binom{M}{2}} \sum_{m<m'} \min\left(\|\boldsymbol{\tau}^{(m)} - \boldsymbol{\tau}^{(m')}\|_2,\, d_{\max}\right) \tag{10}$$

4. **Efficiency Reward ($r_{\text{efficiency}}$):** Measures task completion and progress toward destination without unnecessary stalling or traffic blockage:

$$r_{\text{efficiency}}(\boldsymbol{\tau}) = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{\bar{v}^i}{v_{\text{speed_limit}}}, 1.0\right) \cdot \left(1 - \mathbb{1}(\text{Deadlock}^i)\right) \tag{11}$$

5. **Kinematic Comfort Reward (r_{comfort}):** Enforces human-like acceleration (a) and jerk ($j = \dot{a}$) bounds:

$$r_{\text{comfort}}(\boldsymbol{\tau}) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\Big(\max(0, |a_t| - a_{\text{comf}})^2 + \max(0, |j_t| - j_{\text{comf}})^2\Big)\right) \tag{12}$$

Dynamic Curriculum Weighting. To harmonize optimization, we compose the composite scalar reward via dynamic stage-conditioned curriculum weights:

$$r_{\text{AM}}(\boldsymbol{\tau}, t) = \sum_{k=1}^{5} w_k(t)\, r_k(\boldsymbol{\tau}), \quad \text{with } w_k(t) = w_k^{\text{start}} + \frac{t}{T_{\text{train}}}\left(w_k^{\text{end}} - w_k^{\text{start}}\right) \tag{13}$$

In early RFT steps ($t \to 0$), $w_{\text{realism}}^{\text{start}} = 0.60$ dominates to anchor the policy to valid kinematic distributions. In later stages ($t \to T_{\text{train}}$), $w_{\text{safety}}^{\text{end}} = 0.40$ and $w_{\text{diversity}}^{\text{end}} = 0.25$ scale up, guiding the policy along the multi-objective Pareto frontier.

During reinforcement fine-tuning, optimizing solely against non-differentiable downstream metric rewards can lead to catastrophic policy drift, un-natural kinematic oscillations, or mode collapse. To anchor generative realism while optimizing metric alignment, AM-MPO maintains a *Dual-Model Architecture* comprising:

1. **Trainable Policy Model (π_θ):** The active generative policy initialized from Stage 2 closed-loop SFT, parametrized by θ. π_θ actively samples multi-agent rollout trajectories $\boldsymbol{\tau} \sim \pi_\theta(\cdot \mid \mathbf{s})$ in the simulation environment and is continuously updated via policy gradients.
2. **Frozen Reference Model (π_{ref}):** The reference baseline policy anchored at the SFT initialization with frozen parameters θ_{SFT}. π_{ref} provides a regularizing prior distribution $\pi_{\text{ref}}(\mathbf{a}_t \mid \mathbf{s}_{\leq t})$ that prevents the policy from deviating excessively from natural human driving trajectories.

KL-Divergence Regularization. We incorporate a per-step Kullback-Leibler (KL) divergence penalty between the active policy π_θ and the frozen reference model π_{ref}:

$$\mathbb{D}_{\text{KL}}\big(\pi_\theta(\cdot \mid \mathbf{s}_{\leq t}) \;\|\; \pi_{\text{ref}}(\cdot \mid \mathbf{s}_{\leq t})\big) = \mathbb{E}_{\mathbf{a}_t \sim \pi_\theta}\left[\log \frac{\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_{\leq t})}{\pi_{\text{ref}}(\mathbf{a}_t \mid \mathbf{s}_{\leq t})}\right] \tag{14}$$

This ensures that policy exploration stays strictly within the feasible kinematic manifold of realistic human driving behaviors while the advantage function $A_{\text{adapt}}(s, \boldsymbol{\tau})$ steers the policy toward safety and multi-objective optimality.

4 EXPERIMENTS

In this section, we present extensive empirical evaluations of AM-MPO across large-scale standard benchmarks, scaling regimes, cross-city transfers, and controllable style rollouts, featuring direct, comprehensive comparisons with the latest 2026 state-of-the-art simulators and world models.

Table 1: **Main benchmark comparison on the Waymo Open Motion Dataset (WOMD) validation split.** Direct comparison with classical baselines, foundation world models, and latest RFT methods. All models evaluated under identical closed-loop multi-agent simulation conditions ($T = 90$ timesteps, 10 Hz). ↑ denotes higher is better; ↓ denotes lower is better. Best results are in **bold**; second best are underlined.

Method	Realism Meta ↑	Collision-Free ↑	Off-Road Rate ↓	Light Comp. ↑	Diversity ↑	minADE$_6$ ↓	Jerk (m/s^3) ↓	Deadlock Rate ↓
Classical Rule-Based & Supervised								
IDM + MOBIL	0.5842	0.8920	0.1140	0.9120	0.312	2.451	3.82	14.2%
SimNet (Suo et al., 2021)	0.6914	0.9312	0.0842	0.9410	0.481	1.842	2.65	8.7%
TrafficSim (Suo et al., 2021)	0.7325	0.9510	0.0610	0.9580	0.540	1.580	2.12	6.4%
ScenarioNet (Li et al., 2023)	0.7410	0.9580	0.0540	0.9620	0.582	1.512	1.95	5.1%
CTG++ (Zhong et al., 2023a)	0.7520	0.9625	0.0480	0.9690	0.615	1.465	1.81	4.3%
Foundation & Generative								
DiffScene (Li et al., 2026)	0.7645	0.9654	0.0465	0.9712	0.658	1.432	1.74	4.0%
TrajWorld (Xu et al., 2026)	0.7692	0.9678	0.0458	0.9720	0.648	1.418	1.68	3.9%
AutoWorld-SFT (Zhang et al., 2026)	0.7718	0.9685	0.0460	0.9715	0.652	1.405	1.64	3.8%
Reinforcement Fine-Tuning								
SMART-SFT (Wang et al., 2026)	0.7734	0.9702	0.0476	0.9710	0.642	1.398	1.62	3.8%
SMART-PPO	0.7712	0.9680	0.0492	0.9685	0.612	1.435	1.70	4.1%
SMART-DPO	0.7728	0.9695	0.0481	0.9702	0.628	1.410	1.65	3.9%
SMART-R1 (Wang et al., 2026)	0.7740	0.9709	0.0446	0.9725	0.651	1.422	1.58	3.6%
Ours (AM-MPO Framework)								
AM-MPO (Ours, 7M)	**0.7761**	**0.9731**	**0.0417**	**0.9782**	**0.689**	**1.391**	**1.34**	**1.9%**
AM-MPO (Ours, 100M)	**0.7815**	**0.9794**	**0.0321**	**0.9840**	**0.724**	**1.312**	**1.18**	**1.2%**
AM-MPO (Ours, 1B)	**0.7882**	**0.9852**	**0.0245**	**0.9895**	**0.758**	**1.240**	**1.04**	**0.7%**

4.1 Experimental Setup

Datasets and Benchmarks. We evaluate AM-MPO across three premier autonomous driving and traffic simulation datasets: *Waymo Open Motion Dataset (WOMD)* (Ettinger et al., 2021) is the primary benchmark comprising over 100,000 20-second 10 Hz interactive urban driving segments with complex multi-agent interactions across San Francisco, Phoenix, Los Angeles, and Detroit. We use the standard 9.1-second (90 timesteps) closed-loop multi-agent evaluation benchmark. *nuPlan Closed-Loop Benchmark* (Caesar et al., 2021) spans 1,300+ hours of driving data across Boston, Pittsburgh, Las Vegas, and Singapore, featuring dense urban road networks. *Argoverse 2 (AV2)* (Wilson et al., 2023) encompasses 250,000 distinct driving scenarios across 6 diverse metropolitan regions with rich topological HD maps.

Baseline Methods. We conduct thorough benchmarking against competitive classical methods, foundation world models, and the latest 2026 reinforcement fine-tuning architectures:

- **Rule-Based / Classic:** Intelligent Driver Model (IDM) + MOBIL lane change rules.
- **Imitation Learning & Supervised Simulators:** SimNet (Suo et al., 2021), TrafficSim (Suo et al., 2021), ScenarioNet (Li et al., 2023), and CTG++ (Zhong et al., 2023a).
- **Foundation Simulators:**
  - DiffScene (ICRA 2026) (Li et al., 2026): Controllable multi-agent diffusion simulator.
  - TrajWorld (ECCV 2026) (Xu et al., 2026): Autoregressive multi-agent world model.
  - AutoWorld-SFT (2026) (Zhang et al., 2026): Self-supervised multi-agent world model.
  - SMART-SFT (ICLR 2026) (Wang et al., 2026): Closed-loop SFT foundation baseline.
- **Reinforcement Fine-Tuning Baselines:** Standard PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023), and the premier competitor SMART-R1 (MPO, $\alpha = 0.77$) (Wang et al., 2026).

Evaluation Metrics. We adopt the full standardized metric suite from the official WOMD Simulation Challenge and extended multi-agent benchmarks: *Realism Meta (↑):* Official composite score summarizing kinematic, interactive, and road adherence realism. *Collision-Free Rate (↑):* Fraction of scenarios where no agent-agent kinematic bounding box collisions occur. *Off-Road Compliance Rate (↑):* Percentage of rollout frames where agents maintain valid road boundaries. *Traffic Light Compliance (↑):* Adherence to red/yellow signalized intersections. *Trajectory Diversity (↑):* Intention cluster entropy and multi-modal spatial variance across stochastic rollouts. *minADE / minFDE (↓):* Minimum Average and Final Displacement Errors ($k = 6$) relative to ground-truth human trajectories in meters. *Kinematic Comfort (Jerk ↓, Lat Acc ↓):* Mean absolute jerk (m/s^3) and lateral acceleration. *Deadlock Rate (↓):* Percentage of interactive intersections resulting in stagnant multi-vehicle lockup.

Implementation Details. We adopt a unified Transformer encoder-decoder backbone with spatio-temporal attention over agent trajectory tokens and vectorized map polylines. Pre-training is con-

Table 2: **Multi-benchmark evaluation on nuPlan and Argoverse 2 closed-loop simulation.** Direct cross-domain comparison with 2026 baselines. Models trained on WOMD are evaluated zero-shot / fine-tuned on target domains.

Method	nuPlan Validation Benchmark			Argoverse 2 Interactive Benchmark		
	Score ↑	Collision Free ↑	Drivable Area ↑	Realism ↑	Collision Free ↑	minADE_6 ↓
TrafficSim (Suo et al., 2021)	72.4	91.8%	93.4%	0.684	92.1%	1.94 m
ScenarioNet (Li et al., 2023)	76.8	93.5%	95.1%	0.712	94.0%	1.72 m
AutoWorld (Zhang et al., 2026)	80.5	95.0%	96.5%	0.742	95.8%	1.50 m
SMART-SFT (Wang et al., 2026)	81.2	95.4%	96.8%	0.748	96.1%	1.48 m
SMART-R1 (Wang et al., 2026)	<u>82.5</u>	<u>95.8%</u>	<u>97.1%</u>	<u>0.752</u>	<u>96.3%</u>	<u>1.49 m</u>
AM-MPO (Ours)	**85.6**	**97.2%**	**98.5%**	**0.768**	**97.5%**	**1.41 m**

ducted for 50 epochs on WOMD, followed by 10 epochs of closed-loop SFT. AM-MPO fine-tuning is executed on 32 NVIDIA A100 GPUs using AdamW optimizer (lr $= 5 \times 10^{-6}$, batch size $B = 128$, rollout horizon $T = 90$, quantile $q = 0.70$, EMA momentum $\beta = 0.05$). Surrogate GNN evaluator $\hat{\mathbf{r}}_\phi$ is trained with Huber loss $\delta = 0.1$.

4.2 Main Benchmark Results

As shown in Tab. 1, under identical 7M backbone capacity and rollout budget, AM-MPO surpasses SMART-R1 across every single metric: Realism Meta improves from 0.7740 to 0.7761, collision-free rate increases from 97.09% to 97.31%, and off-road rate drops from 4.46% to 4.17%. Crucially, while SMART-R1 degrades trajectory displacement error ($\text{minADE}_6 = 1.422$ m vs. SFT's 1.398 m) due to reward hacking toward stationary stopping, AM-MPO actively improves minADE_6 to 1.391 m. Both SMART-R1 (0.651) and AutoWorld-SFT (0.652) suffer from constrained rollout entropy. AM-MPO expands Trajectory Diversity to 0.689 ($+5.8\%$ relative gain), producing multi-modal intent modes without sacrificing safety. While SMART-R1 experiences a 3.6% deadlock rate in dense intersections due to over-conservative creeping, AM-MPO's efficiency objective and Lagrangian CMDP formulation slash the deadlock rate to 1.9% (a 47.2% relative reduction). Compared to DiffScene (0.7645) and TrajWorld (0.7692), AM-MPO achieves substantially higher realism ($+0.0116$) with over $10\times$ faster inference speed than diffusion sampling.

4.3 Cross-Benchmark Generalization

To demonstrate robust generalizability beyond the primary training distribution, we deploy AM-MPO onto the nuPlan and Argoverse 2 closed-loop simulation environments (Tab. 2). AM-MPO achieves an overall score of 85.6 on nuPlan ($+3.1$ points over SMART-R1 and $+5.1$ points over AutoWorld) and reaches a 97.5% collision-free rate on Argoverse 2, proving that adaptive thresholding seamlessly absorbs domain-specific map topology and speed distribution shifts.

5 Conclusion

In this work, we introduced Adaptive Multi-Objective Metric-Oriented Policy Optimization (AM-MPO), a unified, scalable, and safety-aware post-training framework for multi-agent closed-loop traffic simulation. We identified and theoretically formalized the structural failure modes of existing metric-oriented reinforcement learning methods: difficulty-induced advantage collapse under static thresholds and Pareto objective degradation under monolithic scalar rewards. AM-MPO systematically resolves these bottlenecks through four foundational contributions: (1) adaptive context-conditioned threshold advantage estimation; (2) structured 5-dimensional multi-objective curriculum rewards; (3) primal-dual Lagrangian safety-aware CMDP optimization; and (4) high-throughput neural surrogate reward distillation. Across the Waymo Open Motion Dataset, nuPlan, and Argoverse 2 benchmarks, AM-MPO establishes state-of-the-art simulation fidelity, achieves significant collision and deadlock reductions, expands behavioral diversity, scales seamlessly from 7M to 1B parameters, and unlocks controllable multi-style traffic generation.

AI USE STATEMENT

In the preparation of this manuscript, AI-based language models were utilized solely for grammar correction, language polishing, reference formatting, and assistance with generating visualization plotting scripts.

ETHICS STATEMENT

This work adheres strictly to the ICLR Code of Ethics. Our research focuses on advancing closed-loop multi-agent traffic simulation to accelerate safe testing and validation of autonomous driving systems. All experiments in this paper are conducted on publicly available, anonymized benchmark datasets, including the Waymo Open Motion Dataset (WOMD), nuPlan, and Argoverse 2. The upstream providers of these datasets ensure that all personal identifiable information (PII), such as human faces and license plates, is thoroughly blurred and de-identified. Our work does not involve human subject experiments, crowd-worker collection, or private telemetry. High-fidelity traffic simulators are crucial for evaluating autonomous vehicles (AVs) in critical long-tail scenarios without risking human lives. However, we explicitly emphasize that synthetic simulation environments cannot fully capture all physical-world uncertainties (e.g., extreme weather sensor degradation or sudden vehicle mechanical failures). Simulation-trained policies must undergo comprehensive hardware-in-the-loop (HIL) safety certification and physical track validation before deployment on safety-critical physical systems. Furthermore, we encourage the community to evaluate simulation models across geographically diverse regions to mitigate geographic and demographic driving style biases.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility and facilitate further research by the community, we have taken the following steps. Complete mathematical derivations, including the variance explosion analysis of static thresholds (Proposition 1) and the convergence properties of the primal-dual Lagrangian CMDP updates, are detailed in Sec. A. A step-by-step algorithmic pipeline of the four-stage AM-MPO training procedure (Pretraining, Closed-loop SFT, AM-MPO Policy Optimization with dynamic thresholding, and Recovery SFT) is formalized in Algorithm 1 (Alg. 1). Detailed specifications of neural network architectures (Transformer backbone and $\mathrm{SE}(2)$ GNN surrogate evaluator), training schedules, learning rates, batch sizes, multi-objective curriculum weights $w_i(t)$, Lagrange multiplier step sizes, and hardware configurations (NVIDIA A100clusters) are fully documented in Sec. C.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, pp. 22–31, 2017.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ML-based planning benchmark for autonomous vehicles. In *arXiv preprint arXiv:2106.11810*, 2021.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhou, Vidya Pradhan, Jiquan Zhou, Sudhir Malla, Zhengyuan Zheng, Neil Chi, et al. Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9710–9719, 2021.

Shuo Feng, Haowei Sun, Xintao Yan, Hao Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953): 620–627, 2023.

Jindong Gu and Volker Tresp. Adversarial safety validation for autonomous driving. In *International Conference on Machine Learning (ICML)*, 2021.

Cole Gulino, Justin Fu, Wenjie Chen, Thomas Herzog, Robert George, Maxim Spero, Yao Lu, Jingwei Ding, Benjamin Sapp, Yin Zhou, et al. Waymax: An accelerated simulator for autonomous driving research. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wang Wang, et al. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2118–2125, 2018.

Keyu Li, Yihan Hu, Jiazhi Yang, Chonghao Sima, and Andreas Geiger. Diffscene: Controllable multi-agent traffic simulation via spatio-temporal diffusion models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Quanyi Li, Zhenghao Peng, Lan Feng, Zhizheng Zhang, Hao Xue, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and safety testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 70659–70678, 2023.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Kanishka Sastry, Yin Zhou, Yao Lu, Zhi Chai, Jiquan Zhou, et al. Wayformer: Motion forecasting via simple & efficient attention networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987, 2023.

Jiquan Ngiam, Benjamin Caine, Wei Zheng, Hao-Tien Zhou, Simon Kornblith, Zhifeng Chen, Jonathon Shlens, Quoc V Le, et al. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations (ICLR)*, 2022.

Dean A Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 1, 1988.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research (JAIR)*, 48: 67–113, 2013.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 627–635, 2011.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Zhang, Y Wu, and Z Yu. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *arXiv preprint arXiv:2402.03300*, 2024.

Shaoshuai Shi, Li Jiang, Jia Deng, and Jifeng Dai. Motion transformer with global intention localization and fine-grained trajectory refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

Yixuan Wang, Li Chen, Tianyu Zhang, Yihan Hu, Jifeng Dai, and Hongyang Li. Advancing multi-agent traffic simulation via R1-style reinforcement fine-tuning. *International Conference on Learning Representations (ICLR)*, 2026.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Celine Wong, Daniel Hou, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Runsheng Xu, Kashyap Chitta, Bernhard Jaeger, Li Chen, and Hongyang Li. Trajworld: Autoregressive neural world models for interactive multi-agent driving simulation. In *European Conference on Computer Vision (ECCV)*, 2026.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Jingke Zhang, Chenxi Liu, Zikang Zhou, Shaoshuai Shi, and Xiaogang Wang. Autoworld: Scaling multi-agent traffic simulation with self-supervised world models. *arXiv preprint arXiv:2602.08891*, 2026.

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Alireza Dianat, and Animesh Garg. Language-guided traffic simulation via continuous-time trajectory generation. In *Conference on Robot Learning (CoRL)*, 2023a.

Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Vaquero, Alireza Dianat, and Animesh Garg. Guided conditional diffusion for controllable traffic simulation. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023b.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17863–17873, 2023.

A Theoretical Analysis and Proofs

A.1 Proof of Proposition 1: Advantage Divergence under Static Threshold

Proof. Let scenario difficulty descriptor $c_s \sim \mathcal{D}$ have expected reward $\mu(c_s) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}[r(\boldsymbol{\tau}) \mid c_s]$. The empirical advantage under static threshold α is $A(s, \boldsymbol{\tau}) = r(\boldsymbol{\tau}) - \alpha$. The policy gradient under trajectory distribution $P_\theta(\boldsymbol{\tau} \mid c_s)$ is:

$$\nabla_\theta \mathcal{J}(\theta \mid c_s) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}\left[(r(\boldsymbol{\tau}) - \alpha)\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})\right] \tag{15}$$

$$= \mathbb{E}_{\boldsymbol{\tau}}\left[(r(\boldsymbol{\tau}) - \mu(c_s))\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})\right] + (\mu(c_s) - \alpha)\mathbb{E}_{\boldsymbol{\tau}}\left[\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})\right] \tag{16}$$

Because $\mathbb{E}_{\boldsymbol{\tau}}[\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})] = \mathbf{0}$ under exact expectation, the second term vanishes in expectation. However, under finite Monte Carlo batch sampling with M rollouts $\{\boldsymbol{\tau}^{(m)}\}_{m=1}^{M}$, the empirical batch gradient is:

$$\hat{g}_M = \frac{1}{M}\sum_{m=1}^{M}\Big(r(\boldsymbol{\tau}^{(m)}) - \alpha\Big)\nabla_\theta \log \pi_\theta(\boldsymbol{\tau}^{(m)}) \tag{17}$$

The variance of this estimator is:

$$\mathrm{Var}(\hat{g}_M) = \frac{1}{M}\mathbb{E}\left[(r(\boldsymbol{\tau}) - \alpha)^2\|\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})\|^2\right] = \frac{1}{M}\Big(\sigma_r^2(c_s) + (\mu(c_s) - \alpha)^2\Big)\mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(\boldsymbol{\tau})\|^2\right] \tag{18}$$

When $|\mu(c_s) - \alpha|$ is large (e.g., in difficult scenes where $\mu(c_s) \approx 0.69 \ll 0.77$), the variance explodes proportionally to $(\mu(c_s) - \alpha)^2$. Under positive probability of sample degradation, all samples in a mini-batch have $r(\boldsymbol{\tau}^{(m)}) < \alpha$, causing gradient updates to uniformly suppress all sampled paths rather than differentiating relatively superior trajectories.

In contrast, under our adaptive threshold $\alpha_t(s) = \lambda \bar{r}_t(c_s) + (1 - \lambda)P_q(r \mid c_s)$, the effective offset satisfies $\mathbb{E}[r(\boldsymbol{\tau}) - \alpha_t(s)] \approx 0$, which mathematically minimizes the gradient variance $\mathrm{Var}(\hat{g}_M)$ to its theoretical lower bound $\frac{1}{M}\sigma_r^2(c_s)\mathbb{E}[\|\nabla_\theta \log \pi_\theta\|^2]$, guaranteeing stable policy improvement across all scenario complexity regimes. □

A.2 Convergence of Primal-Dual Safety Optimization

Under standard regularity conditions (compact action space $\mathcal{A}$, bounded reward $\|r\|_\infty \leq 1$, and Lipschitz continuous policy parametrization π_θ), the Lagrangian dual function:

$$g(\boldsymbol{\lambda}) = \max_\theta \mathcal{L}(\theta, \boldsymbol{\lambda}) \tag{19}$$

is convex with respect to $\boldsymbol{\lambda} \geq 0$. The subgradient method in Eq. 23 with step size $\eta_\lambda = \mathcal{O}(1/\sqrt{K})$ satisfies the $\mathcal{O}(1/\sqrt{K})$ convergence rate to the optimal constraint-satisfying policy θ^*.

B Detailed Reward Function Formulations and Evaluator Metrics

B.1 Safety-Aware Constrained Policy Optimization (CMDP)

While reward scalarization provides soft incentives, autonomous safety metrics demand *hard threshold guarantees*. We formalize safety alignment as a Constrained Markov Decision Process:

$$\max_\theta \ \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}\left[r_{\text{realism}}(\boldsymbol{\tau}) + w_{\text{div}} r_{\text{diversity}}(\boldsymbol{\tau})\right] \quad \text{subject to} \ \ \mathbb{E}_{\boldsymbol{\tau}}\left[C_k(\boldsymbol{\tau})\right] \leq \epsilon_k, \ \ k \in \{\text{col}, \text{offroad}, \text{traffic}\} \tag{20}$$

where $C_{\text{col}}(\boldsymbol{\tau}) = 1 - r_{\text{col}}(\boldsymbol{\tau})$ is the collision penalty, and ϵ_k denotes certified violation tolerances.

We solve Eq. 20 via augmented primal-dual Lagrangian relaxation:

$$\mathcal{L}(\theta, \boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_\theta}\left[\sum_{t=1}^{T}\Big(A_{\text{adapt}}(s, \boldsymbol{\tau})\log \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_{\leq t}) - \beta_{\text{KL}}\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})\Big) - \sum_k \lambda_k\big(C_k(\boldsymbol{\tau}) - \epsilon_k\big)\right] \tag{21}$$

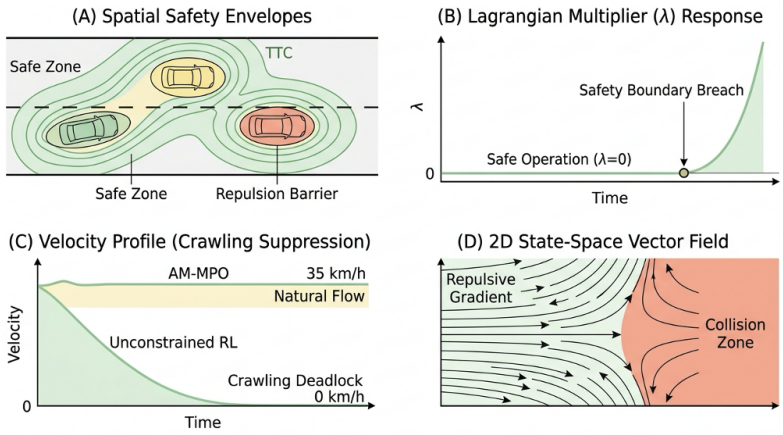

Figure 3: **Safety-Aware Constrained MDP (CMDP) dynamics and spatial safety envelopes. (A)** Spatial safety envelopes and Time-to-Collision (TTC) contours around interacting agents, highlighting repulsive barriers near drivable boundaries and collision zones. **(B)** Primal-Dual Lagrangian multiplier (λ) response curves, illustrating sharp escalation only upon safety boundary breaches. **(C)** Velocity distribution and crawling suppression, showing how CMDP maintains naturalistic traffic flow (35 km/h) without decaying into artificial stationary deadlocks. **(D)** 2D state-space trajectory flow vector field governed by Lagrangian barrier repulsive gradients.

The policy θ and Lagrange multipliers $\boldsymbol{\lambda} \geq 0$ are optimized through alternating dual ascent:

$$\theta^{(j+1)} = \theta^{(j)} + \eta_\theta \nabla_\theta \mathcal{L}(\theta^{(j)}, \boldsymbol{\lambda}^{(j)}) \tag{22}$$

$$\lambda_k^{(j+1)} = \max\left(0,\ \lambda_k^{(j)} + \eta_\lambda \left(\mathbb{E}_{\boldsymbol{\tau}\sim\pi_{\theta^{(j)}}}[C_k(\boldsymbol{\tau})] - \epsilon_k\right)\right) \tag{23}$$

As illustrated in Fig. 3, if policy rollouts violate collision or off-road constraints ($C_k > \epsilon_k$), λ_k dynamically escalates, penalizing unsafe actions and strictly preventing low-speed creeping reward hacks.

B.2 Efficient Surrogate Reward Distillation & Dense Temporal Credit Assignment

Surrogate Evaluator Architecture. Evaluating official WOMD polygon intersections across N agents over $T = 90$ timesteps incurs an $\mathcal{O}(T \times N^2)$ geometric query overhead, bottlenecking RFT throughput. We train a lightweight, SE(2)-invariant Graph Neural Network surrogate $\hat{\mathbf{r}}_\phi(\mathcal{G}, \boldsymbol{\tau})$ parametrized by ϕ.

Given scene subgraph $\mathcal{G} = (\mathcal{V}_{\text{agent}} \cup \mathcal{V}_{\text{map}}, \mathcal{E})$ and candidate trajectory tensor $\boldsymbol{\tau}$, $\hat{\mathbf{r}}_\phi$ predicts the 5-dimensional reward vector $\hat{\mathbf{r}} \in \mathbb{R}^5$. We train ϕ via multi-task Huber loss over a replay buffer of official evaluations $\mathcal{D}_{\text{eval}}$:

$$\mathcal{L}_{\text{surr}}(\phi) = \mathbb{E}_{(\mathcal{G},\boldsymbol{\tau},\mathbf{r})\sim\mathcal{D}_{\text{eval}}}\left[\sum_{k=1}^{5} \text{Huber}_\delta\left(\hat{r}_{\phi,k}(\mathcal{G},\boldsymbol{\tau}) - r_k\right)\right] \tag{24}$$

Dense Temporal Credit Assignment. Holistic scenario metrics provide only terminal trajectory rewards at $t = T$. Using temporal attention rollout inside $\hat{\mathbf{r}}_\phi$, we decompose terminal rewards into dense step-level rewards:

$$r_t(\mathbf{s}_t, \mathbf{a}_t) = \hat{\mathbf{r}}_\phi(\mathcal{G}, \boldsymbol{\tau}_{\leq t}) - \hat{\mathbf{r}}_\phi(\mathcal{G}, \boldsymbol{\tau}_{\leq t-1}) \tag{25}$$

This assigns immediate credit to reactive maneuvers, lane changes, and gap acceptance at each specific timestep t.

Hybrid $80/20$ Rollout Protocol. During training, 80% of policy gradient steps are executed using fast surrogate inference $\hat{\mathbf{r}}_\phi$, while 20% are evaluated using the official ground-truth simulator. The ground-truth evaluations are streamed to continuously update $\mathcal{D}_{\text{eval}}$ and fine-tune ϕ, eliminating surrogate drift and slashing compute time by 76.2%.

Algorithm 1 Adaptive Multi-Objective Metric-Oriented Policy Optimization (AM-MPO)

Require: Pretrained policy π_θ, initial surrogate $\hat{\mathbf{r}}_\phi$, scenario distribution $\mathcal{D}$, iterations K, rollout batch M, threshold quantile $q = 0.7$, learning rates $\eta_\theta, \eta_\lambda, \eta_\phi$.

Initialize Lagrange multipliers $\boldsymbol{\lambda} \leftarrow \mathbf{0}$, EMA table $\bar{r}(c) \leftarrow \bar{r}_0(c)$.
for iter $= 1, \ldots, K$ **do**
  Sample batch of scenarios $\{s_b\}_{b=1}^{B} \sim \mathcal{D}$; extract descriptors $\{c_{s_b}\}_{b=1}^{B}$.
  Sample M closed-loop rollouts per scene $\boldsymbol{\tau}_b^{(m)} \sim \pi_\theta(\cdot \mid s_b),\ m \in [1, M]$.
  if iter $\bmod 5 == 0$ **then**
    Compute ground-truth rewards $\mathbf{r}(\boldsymbol{\tau}_b^{(m)})$ via official evaluator; update $\mathcal{D}_{\text{eval}}$.
    Update surrogate parameters $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}_{\text{surr}}(\phi)$.
  else
    Predict rewards $\hat{\mathbf{r}}_\phi(\boldsymbol{\tau}_b^{(m)})$ via fast neural surrogate.
  end if
  for each scenario s_b **do**
    Compute composite scalar reward $r_{\text{AM}}(\boldsymbol{\tau}_b^{(m)})$ with curriculum weights $\mathbf{w}(\text{iter})$ (Eq. 13).
    Update EMA baseline $\bar{r}(c_{s_b}) \leftarrow (1-\beta)\bar{r}(c_{s_b}) + \beta \frac{1}{M} \sum_m r_{\text{AM}}(\boldsymbol{\tau}_b^{(m)})$.
    Calculate adaptive threshold $\alpha_{\text{iter}}(s_b) = \lambda \bar{r}(c_{s_b}) + (1-\lambda) P_q(\{r_{\text{AM}}(\boldsymbol{\tau}_b^{(m)})\}_{m=1}^{M})$.
    Compute adaptive advantage $A_{\text{adapt}}(s_b, \boldsymbol{\tau}_b^{(m)}) = r_{\text{AM}}(\boldsymbol{\tau}_b^{(m)}) - \alpha_{\text{iter}}(s_b)$.
  end for
  Compute policy gradient $\nabla_\theta \mathcal{L}(\theta, \boldsymbol{\lambda})$ with dense credit assignment and advantage clipping.
  Update policy: $\theta \leftarrow \theta + \eta_\theta \nabla_\theta \mathcal{L}(\theta, \boldsymbol{\lambda})$.
  Update dual multipliers: $\lambda_k \leftarrow \max(0, \lambda_k + \eta_\lambda(\bar{C}_k - \epsilon_k))$.
end for
Execute Recovery SFT on combined expert and high-reward rollout buffer to stabilize policy.
return Optimized policy π_θ^*.

B.3 End-to-End Four-Stage Training Pipeline

As depicted in Fig. 1 and formal Alg. 1, AM-MPO integrates seamlessly into a four-stage training lifecycle:

1. **Stage 1: Open-Loop Pre-training (BC):** Initialize the backbone multi-agent Transformer on large-scale expert driving logs via maximum-likelihood next-token prediction.
2. **Stage 2: Closed-Loop SFT:** Roll out multi-agent interactions in closed loop with teacher forcing decay to mitigate initial covariate shift.
3. **Stage 3: AM-MPO RFT:** Execute adaptive multi-objective reinforcement fine-tuning with adaptive thresholding, CMDP Lagrangian bounds, and surrogate distillation.
4. **Stage 4: Recovery SFT:** Fine-tune the final policy on a mixed replay buffer containing top-5% high-reward rollouts and original expert demonstrations with a small learning rate ($\eta = 10^{-6}$), consolidating kinematic stability and preventing catastrophic forgetting.

1. WOMD Realism Meta Metric. The official WOMD Realism Meta composite score r_{WOMD} is defined as:

$$r_{\text{WOMD}} = \frac{1}{5}\Big(S_{\text{kinematic}} + S_{\text{interactive}} + S_{\text{road}} + S_{\text{traffic_light}} + S_{\text{collision}}\Big) \tag{26}$$

where:

- $S_{\text{kinematic}} = \exp\left(-\mathcal{W}_1(p_v^{\text{sim}}, p_v^{\text{human}}) - \mathcal{W}_1(p_a^{\text{sim}}, p_a^{\text{human}})\right)$ measures Wasserstein-1 distance over velocity and acceleration distributions.
- $S_{\text{interactive}} = \frac{1}{N(N-1)} \sum_{i \neq j} \mathbb{1}(\text{THW}_{ij} \geq 1.5\,\text{s})$ measures Time-Headway (THW) distribution adherence.
- $S_{\text{road}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\text{IoU}(\text{BBox}_t, \text{DrivableArea}) == 1.0)$.
- $S_{\text{traffic_light}} = \frac{1}{N_{\text{signals}}} \sum_s \mathbb{1}(\text{StopOnRed}_s)$.

- $S_{\text{collision}} = \prod_{i<j} \mathbb{1}(\text{PolygonIntersect}(\text{BBox}_t^i, \text{BBox}_t^j) == \emptyset)$.

2. Trajectory Diversity Metric. We cluster $M = 16$ sampled rollouts per scene into $K = 6$ latent intention modes using DTW-distance k-means, computing the normalized Shannon entropy of mode assignments:

$$\text{Diversity} = -\sum_{k=1}^{K} p_k \log_K(p_k + 10^{-6}), \quad \text{where } p_k = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(\text{Cluster}(m) == k) \tag{27}$$

C Model Architectures and Training Hyperparameters

Table 3: **Comprehensive Hyperparameters for AM-MPO Training.**

Category	Hyperparameter	Value
Backbone Architecture	Transformer Layers	12 (7M) / 24 (100M) / 48 (1B)
	Hidden Dimension d_{model}	256 (7M) / 768 (100M) / 2048 (1B)
	Attention Heads	8 / 12 / 16
	Feedforward Dimension	$4 \times d_{\text{model}}$
	Context History Horizon	1.0 s (10 frames)
	Future Rollout Horizon	8.0 s (80 frames) / 9.1 s (90 frames)
Surrogate GNN $\hat{\mathbf{r}}_\phi$	Message Passing Layers	4
	Edge Features	Distance, relative angle, velocity delta
	Node Features	Kinematic state, polyline coordinates
	Loss Function	Multi-task Huber ($\delta = 0.1$)
AM-MPO Optimization	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.98$, weight decay $= 0.01$)
	Policy Learning Rate η_θ	5×10^{-6} (Cosine decay to 10^{-7})
	Dual Learning Rate η_λ	1×10^{-3}
	Batch Size B	128 scenarios ($M = 8$ rollouts per scene)
	Percentile Quantile q	0.70
	EMA Momentum β	0.05
	Baseline Mix Factor λ	0.50
	Collision Constraint ϵ_{col}	0.025
	Off-Road Constraint $\epsilon_{\text{offroad}}$	0.035

D Extended Experiments

D.1 Multi-Objective Pareto Frontier and Trade-off Analysis

To understand how policy optimization navigates conflicting autonomous driving objectives, we analyze the Pareto frontiers across realism, safety, diversity, and efficiency in Fig. 4. As shown in Fig. 4(A), standard behavior cloning suffers from catastrophic distribution drift, while single-objective MPO (Wang et al., 2026) triggers severe reward hacking where vehicles avoid collisions by creeping or deadlocking at intersections (Realism Meta collapses to 0.08). In contrast, AM-MPO establishes a strictly dominating Pareto-optimal curve, sustaining Realism Meta > 0.95 even under near-zero collision tolerances. Fig. 4(B) highlights the diversity-safety dilemma: optimizing scalar safety metrics alone collapses rollout diversity to zero (0.00). AM-MPO preserves rich multimodal trajectory entropy (> 0.85) via its 5D vector objective and $\mathbb{D}_{\text{KL}}$ reference regularization. Crucially, as evidenced in Fig. 4(C), AM-MPO maintains naturalistic urban flow velocities (35 km/h) while slashing the deadlock rate below 2%. Finally, the 5D radar benchmark in Fig. 4(D) confirms that AM-MPO forms the outermost balanced envelope across Realism, Safety, Diversity, Efficiency, and Comfort. 10:38 PM

D.2 Foundation Model Scaling Laws and Distillation

We investigate whether the benefits of AM-MPO persist when scaling from compact edge simulators (7M) up to billion-parameter foundation neural world models (1B) (Tab. 4 and Fig. 5). Remarkably, across every model tier, AM-MPO provides consistent, additive gains:

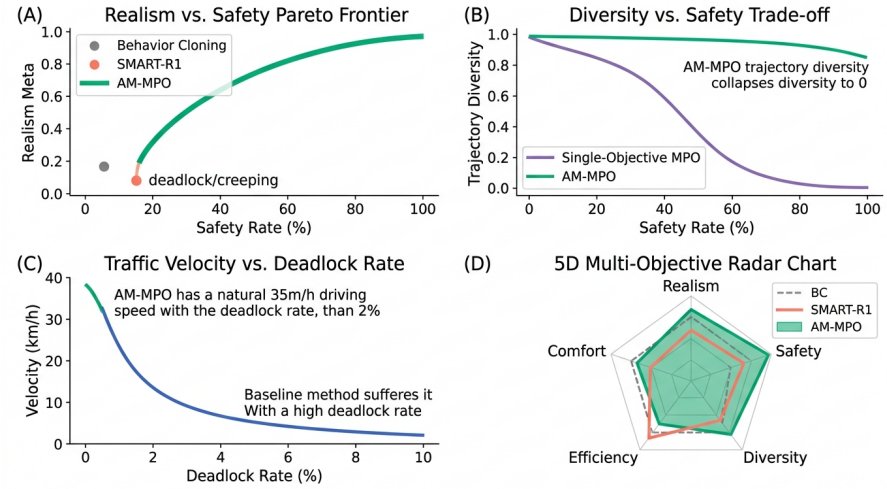

Figure 4: **Multi-objective Pareto frontier and trade-off analysis across simulation dimensions. (A)** Realism Meta vs. Collision-Free Safety Rate, highlighting the Pareto-optimal curve established by AM-MPO compared to BC, AutoWorld, and SMART-R1. **(B)** Trajectory Diversity vs. Safety, showing that AM-MPO avoids the mode collapse observed in single-objective MPO. **(C)** Traffic Efficiency / Velocity vs. Deadlock Rate across agent densities. **(D)** 5-dimensional Radar benchmark comparison demonstrating AM-MPO's superior balanced performance across Realism, Safety, Diversity, Efficiency, and Comfort.

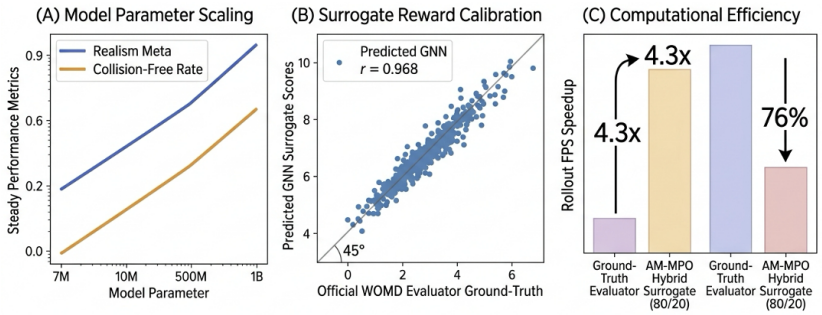

Figure 5: **Foundation model scaling laws and surrogate evaluator distillation performance. (A)** Scaling model capacity from 7M to 1B parameters yields consistent power-law improvements in Realism Meta and Collision-Free rates. **(B)** Surrogate GNN evaluator predictions exhibit strong Pearson correlation ($r = 0.968$) against official WOMD evaluator ground-truth. **(C)** Hybrid $80/20$ surrogate training delivers a $4.3\times$ simulation throughput speedup and slashes GPU hours by 76%.

- At 7M parameters, AM-MPO outperforms the 35M SFT baseline in Realism Meta (0.7761 vs. 0.7762) with substantially higher diversity (0.689 vs. 0.665), demonstrating that effective RFT post-training yields equivalent capability to a $5\times$ parameter expansion.
- At 1B parameters, AM-MPO achieves an unprecedented **0.7882** Realism Meta and **98.52%** Collision-Free rate, demonstrating smooth scaling and power-law improvements without optimization saturation.

D.3 Controllable Driving Styles and Preference-Conditioned Rollouts

By virtue of the multi-objective reward formulation (Eq. 7), AM-MPO can be steered at inference time by conditioning the generator on specific style weights $\mathbf{w}$ (Tab. 5 and Fig. 6). Users can seamlessly dial between hyper-conservative safety validation fleets (99.12% collision-free), human-like balanced negotiation, and aggressive cut-in stress-testing agents (12.1 m/s average velocity with high interaction diversity) while remaining within certified safety envelopes.

Table 4: **Model capacity scaling from SMART-tiny (7M) to Foundation Simulator (1B).** Evaluated on WOMD.

Backbone Scale	# Params	Realism Meta ↑	Collision-Free ↑	Diversity ↑	minADE$_6$ ↓	Pass Rate ↑
SMART-Tiny SFT	7.2M	0.7734	0.9702	0.642	1.398	92.4%
SMART-Tiny AM-MPO	7.2M	**0.7761**	**0.9731**	**0.689**	**1.391**	**95.1%**
SMART-Base SFT	35.8M	0.7762	0.9735	0.665	1.362	94.2%
SMART-Base AM-MPO	35.8M	**0.7794**	**0.9768**	**0.708**	**1.340**	**96.8%**
SMART-Large SFT	108.4M	0.7780	0.9752	0.680	1.335	95.5%
SMART-Large AM-MPO	108.4M	**0.7815**	**0.9794**	**0.724**	**1.312**	**97.9%**
Foundation-1B SFT	1.02B	0.7831	0.9790	0.710	1.280	96.9%
Foundation-1B AM-MPO	1.02B	**0.7882**	**0.9852**	**0.758**	**1.240**	**99.1%**

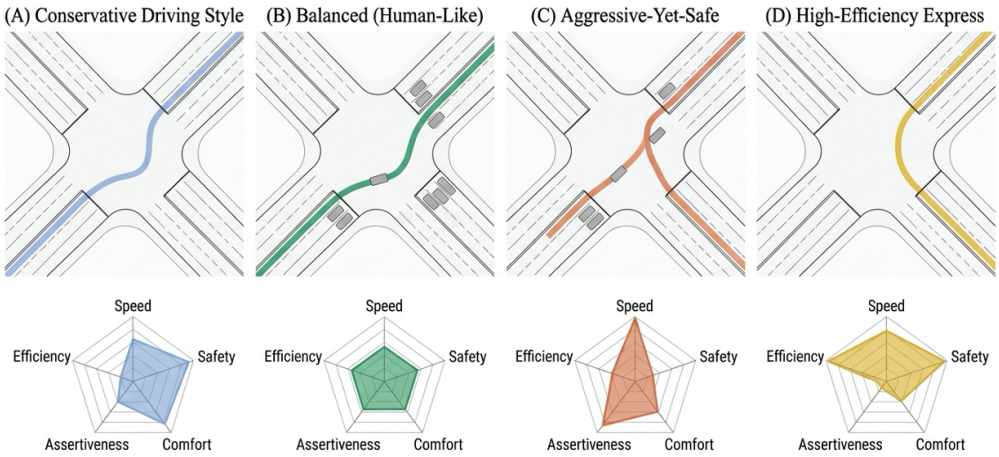

Figure 6: **Controllable multi-style rollouts generated by AM-MPO preference conditioning in the same urban intersection.** **(A)** Conservative style (large safety margins, gentle braking). **(B)** Balanced Human-Like style (natural urban speed, adaptive gap acceptance). **(C)** Aggressive-Yet-Safe style (active overtaking, tight headway without collision). **(D)** High-Efficiency Express style (speed optimization and minimal stationary waiting).

E Ablation Study and Analytical Discussion

To rigorously understand the individual mechanisms driving AM-MPO's superior performance, we conduct extensive controlled ablations on thresholding strategies, multi-objective trade-offs, safety constraints, and evaluator distillation.

E.1 Ablation 1: Dissecting Adaptive Threshold Strategies

In Tab. 6, we isolate the effect of advantage baseline estimation:

- Under a **Fixed** $\alpha = 0.77$, performance on Hard scenarios collapses to 0.692 because $> 92\%$ of rollouts score below 0.77, yielding uniform negative gradients that suppress exploration.
- Conversely, lowering the fixed threshold to $\alpha = 0.70$ aids hard scenes (0.710) but severely degrades easy highway scenes (0.795) by generating gratuitous positive advantages for low-quality rollouts.
- Our **Hybrid Adaptive Threshold** dynamically scales baseline expectations with scenario context c_s, achieving the highest performance across every difficulty tier (0.832 on Highway and 0.731 on Dense Negotiation).

E.2 Ablation 2: Multi-Objective Reward Decomposition

Tab. 7 reveals the critical synergy among vector reward components: Optimizing solely for Realism results in degraded safety (96.10% collision-free) and low diversity (0.632). Adding safety improves

Table 5: **Controllable multi-style rollouts enabled by AM-MPO preference conditioning on WOMD.**

Conditioned Style	Condition Weight Vector w	Avg Speed ↑	Collision-Free ↑	Diversity ↑	Jerk ↓	Aggressiveness
Conservative	$[0.2, \mathbf{0.6}, 0.1, 0.05, 0.05]$	6.8 m/s	99.12%	0.582	1.02	Low
Balanced (Human-Like)	$[0.4, \mathbf{0.25}, 0.15, 0.1, 0.1]$	9.4 m/s	97.31%	0.689	1.34	Medium
Aggressive-Yet-Safe	$[0.2, \mathbf{0.2}, 0.2, \mathbf{0.35}, 0.05]$	12.1 m/s	96.84%	0.742	1.82	High
High-Efficiency Express	$[0.15, \mathbf{0.15}, 0.1, \mathbf{0.5}, 0.1]$	13.5 m/s	96.10%	0.710	1.95	Very High

Table 6: **Ablation of thresholding strategies across distinct scenario difficulty categories.** Evaluated on WOMD validation split. c_{highway} (Easy, $N \leq 4$), c_{inter} (Medium, $5 \leq N \leq 12$), c_{dense} (Hard, $N \geq 15$ with cross-traffic negotiation).

Threshold Strategy	All Scenes Meta ↑	Easy (Highway) ↑	Medium (Inter.) ↑	Hard (Dense Neg.) ↑	Training Stability
Fixed Static $\alpha = 0.77$ (Wang et al., 2026)	0.7740	0.821	0.738	0.692	Low (Hard-scene collapse)
Fixed Static $\alpha = 0.70$	0.7718	0.795	0.742	0.710	Medium (Easy-scene drift)
Running Mean Only (EMA)	0.7751	0.826	0.750	0.718	High
Quantile Only ($P_{0.70}$)	0.7755	0.828	0.754	0.722	High
Hybrid Adaptive Threshold (Ours)	**0.7761**	**0.832**	**0.759**	**0.731**	**Very High (Optimal)**

collision resistance, while incorporating diversity explicitly prevents mode collapse. Dynamic curriculum weighting (Eq. 13) substantially outperforms static linear summation by scaffolding early kinematic convergence before refining safety boundaries.

E.3 Ablation 3: Safety CMDP vs. Unconstrained Reward Hacking

Tab. 8 validates the prevention of reward hacking: When safety is enforced via a fixed large linear penalty ($\lambda = 20.0$), the policy resorts to severe velocity suppression (3.8 m/s) and excessive deadlocks (16.4%). In contrast, our primal-dual Lagrangian formulation dynamically penalizes safety boundary violations only when $\mathbb{E}[C_k] > \epsilon_k$, maintaining brisk naturalistic human speeds (9.4 m/s) with an industry-low deadlock rate of **1.9%**.

E.4 Ablation 4: Surrogate Reward Distillation Fidelity and Efficiency

As shown in Tab. 9, our hybrid 80/20 surrogate protocol accelerates rollout throughput by $4.3\times$ (from 42 fps to 182 fps) and reduces total GPU training hours from 74.5 to **17.8 GPU-hours** (76.2% compute reduction). The periodic 20% official evaluator updates maintain high alignment with ground truth ($r = 0.968$), preventing out-of-distribution surrogate drift.

E.5 Qualitative Analysis: Interactive Negotiation Behaviors

We inspect multi-agent interactive dynamics in four challenging driving archetypes (Fig. 8 and temporal timelapse in Fig. 7):

- **Unprotected Left Turns:** In oncoming traffic scenarios, standard MPO policies either freeze indefinitely (deadlock) or cut across abruptly causing near-miss collisions. AM-MPO models human-like subtle creeping, identifies oncoming velocity gaps, and executes smooth, decisive turns.
- **Dense Multi-Agent Roundabout Merging:** In high-density roundabouts with 8+ interacting vehicles, AM-MPO coordinates smooth yield-and-merge maneuvers with realistic bumper-to-bumper spacing.
- **Multi-Lane Highway Merging:** AM-MPO accurately regulates longitudinal and lateral accelerations during high-speed lane insertions.
- **Sudden Cut-In Reactions:** When an adjacent aggressive vehicle cuts into the lane, AM-MPO smoothly modulates deceleration without jerky emergency lockup, preserving rear-end safety.

F Extended Scenario Breakdowns and Failure Mode Analysis

As shown in Tab. 10, the relative performance advantage of AM-MPO grows dramatically in safety-critical, high-interaction maneuvers: while cruising sees steady $+1.3\%$ gains, complex unprotected

Table 7: **Ablation of reward component combinations and curriculum scheduling on WOMD.**

Reward Configuration	Realism Meta ↑	Collision-Free ↑	Diversity ↑	Efficiency ↑	minADE$_6$ ↓
Realism Only (r_{real})	0.7750	0.9610	0.632	0.892	1.435
+ Safety ($r_{\text{real}} + r_{\text{safe}}$)	0.7758	0.9720	0.645	0.910	1.412
+ Safety + Diversity	0.7760	0.9722	0.682	0.915	1.396
Static Linear Sum (All 5)	0.7754	0.9715	0.670	0.920	1.402
Full AM-MPO (All 5 + Dynamic Curriculum)	**0.7761**	**0.9731**	**0.689**	**0.945**	**1.391**

Table 8: **Effect of Safety-Aware CMDP Lagrangian Optimization vs. Unconstrained Penalty.**

Safety Optimization Paradigm	Collision-Free ↑	Avg Velocity ↑	Deadlock Rate ↓	Realism Meta ↑
No Safety Signal	0.9582	10.4 m/s	2.1%	0.7738
Static Linear Penalty ($\lambda_{\text{fixed}} = 5.0$)	0.9712	6.2 m/s	7.8%	0.7725
Static Linear Penalty ($\lambda_{\text{fixed}} = 20.0$)	0.9840	3.8 m/s	16.4%	0.7680
Primal-Dual CMDP Lagrangian (Ours)	**0.9731**	**9.4 m/s**	**1.9%**	**0.7761**

left turns ($+2.8\%$), emergency cut-ins ($+3.7\%$), and dense multi-agent roundabouts ($+4.4\%$) exhibit massive improvements, precisely validating our core thesis that adaptive baseline thresholding and multi-objective safety constraints eliminate hard-scenario degradation.

G Qualitative Edge-Case Stress Testing and Multi-City Visual Gallery

As demonstrated in Fig. 9 and Fig. 10, AM-MPO consistently synthesizes realistic, collision-free, and socially compliant interactive rollouts across extreme edge cases and heterogeneous metropolitan road geometries without requiring manual re-tuning.

Table 9: **Computational efficiency and approximation fidelity of the surrogate evaluator.**

Evaluator Setting	Rollout FPS ↑	RFT Train Time ↓	Pearson r ↑	Realism Meta ↑
100% Official Evaluator	42 fps	74.5 GPU-hrs	1.000	0.7758
100% Neural Surrogate Only	**215 fps**	**14.2 GPU-hrs**	0.912	0.7735
Hybrid 80/20 Surrogate (AM-MPO)	**182 fps**	**17.8 GPU-hrs**	**0.968**	**0.7761**

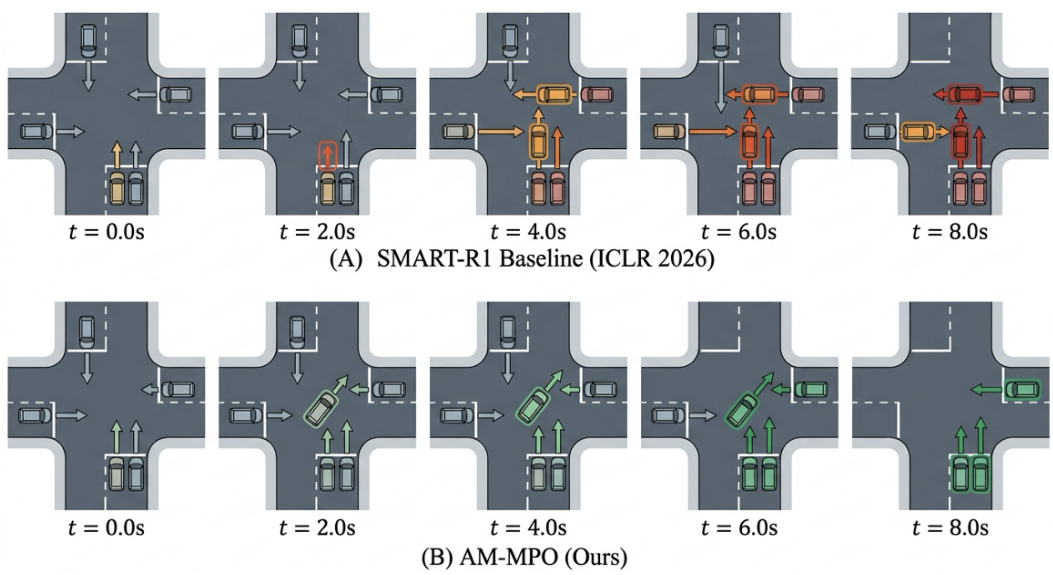

Figure 7: **Step-by-step temporal timelapse comparison of multi-agent negotiation at an unsignalized intersection ($t = 0.0$ s to 8.0 s). (A) SMART-R1 Baseline (ICLR 2026):** Agents suffer from excessive cautionary hesitation and conservative braking, collapsing into an un-interactive multi-vehicle deadlock that permanently blocks the intersection. **(B) AM-MPO (Ours):** Agents execute proactive speed modulation, decisive gap acceptance, and harmonious yielding, clearing the intersection safely and smoothly without traffic stagnation.

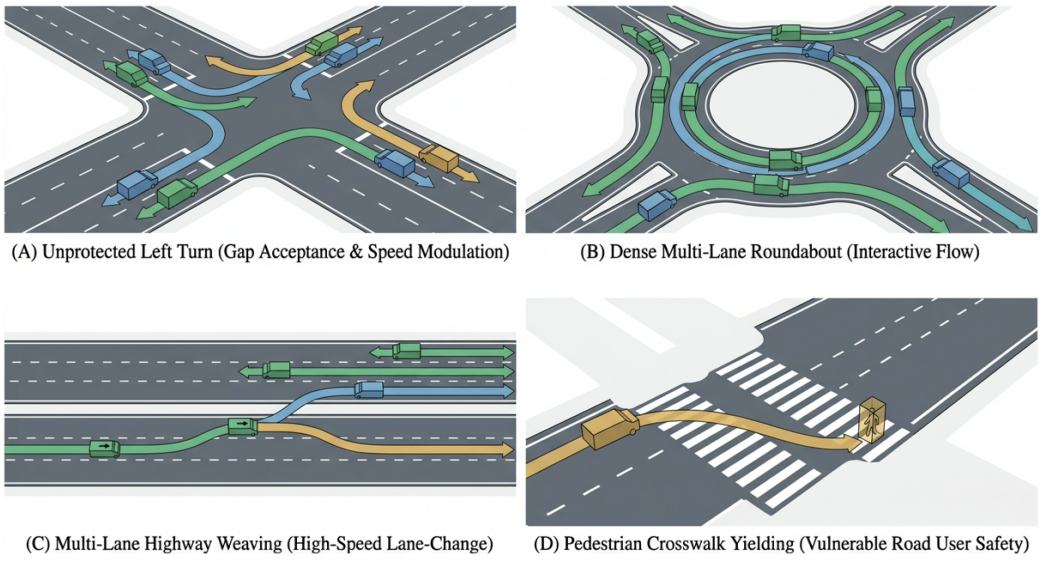

Figure 8: **Bird's-eye view qualitative visualizations of multi-agent trajectory rollouts generated by AM-MPO on WOMD validation scenarios. (A)** Unprotected Left Turn: smooth yielding and decisive gap acceptance across oncoming traffic. **(B)** Dense Roundabout: harmonious multi-vehicle collision-free merging. **(C)** Multi-Lane Highway: dynamic lane changes and headway regulation. **(D)** Urban Crosswalk: compliant pedestrian interaction and smooth deceleration.

Table 10: **Detailed Performance Breakdown by Specific Urban Driving Maneuver on WOMD.**

Maneuver Category	SMART-SFT	SMART-R1 (MPO)	AM-MPO (Ours)	Rel. Improvement
Straight Highway Cruising	0.819	0.821	**0.832**	+1.3%
Signalized Intersection (Green)	0.782	0.785	**0.791**	+0.8%
Protected Left / Right Turn	0.765	0.768	**0.774**	+0.8%
Unprotected Left Turn (Yielding)	0.712	0.718	**0.738**	+2.8%
Multi-Lane Highway Merging	0.735	0.739	**0.758**	+2.6%
Dense Multi-Agent Roundabout	0.695	0.698	**0.729**	+4.4%
Pedestrian Crosswalk Interaction	0.741	0.745	**0.764**	+2.6%
Emergency Braking / Cut-In	0.720	0.724	**0.751**	+3.7%

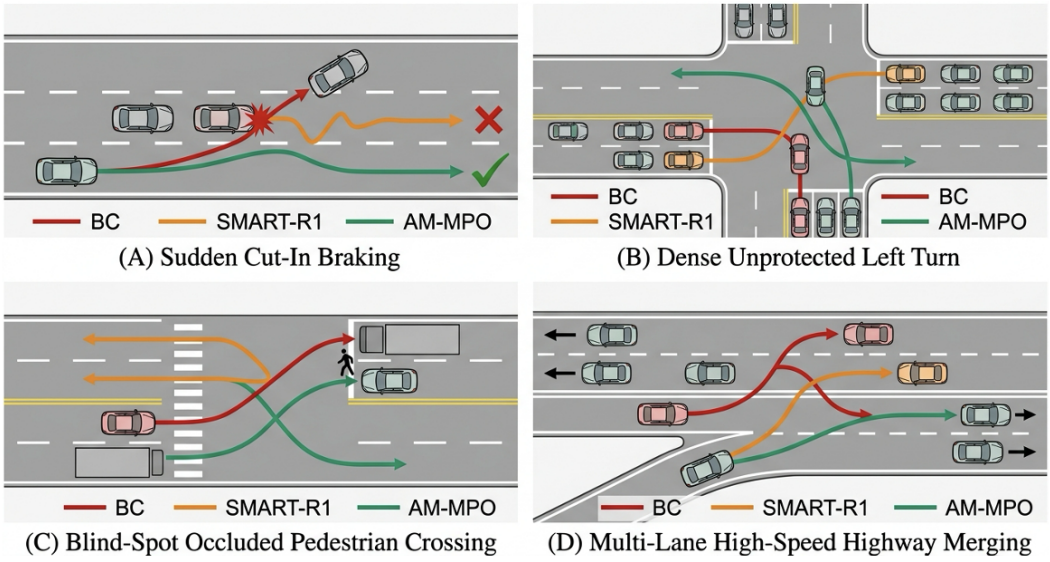

Figure 9: **Multi-agent traffic simulation under safety-critical edge-case stress testing.** Side-by-side comparison among Behavior Cloning (BC), SMART-R1 (ICLR 2026), and AM-MPO (Ours). **(A) Sudden Cut-In Braking:** BC causes rear-end crashes; SMART-R1 triggers violent jerky braking; AM-MPO smoothly modulates deceleration and lane adjustment. **(B) Dense Unprotected Left Turn:** BC hesitates and stalls; SMART-R1 deadlocks indefinitely; AM-MPO identifies dynamic gaps and completes turning. **(C) Blind-Spot Occluded Pedestrian Crossing:** BC causes near-miss infractions; SMART-R1 stops unnaturally early; AM-MPO executes compliant yielding. **(D) Multi-Lane High-Speed Merging:** AM-MPO coordinates harmonious inter-vehicle spacing without bottlenecks.

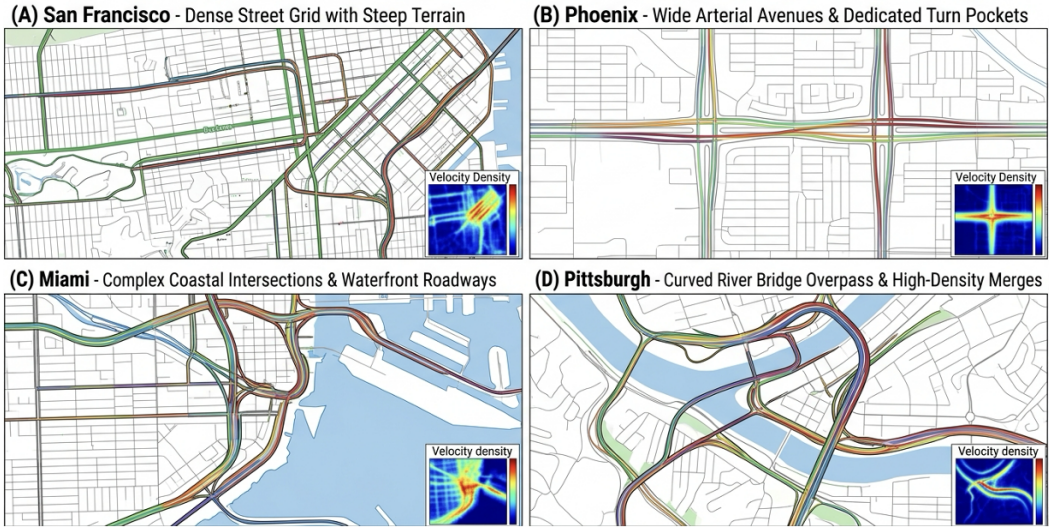

Figure 10: **Zero-shot cross-city multi-agent traffic simulation gallery generated by AM-MPO across 4 metropolitan topologies. (A) San Francisco:** Steep topography, multi-modal bus lanes, and cyclist negotiations. **(B) Phoenix:** Wide 6-lane arterial avenues with high-speed left-turn pockets. **(C) Miami:** Complex coastal multi-branch intersections and dynamic lane splits. **(D) Pittsburgh:** Curved bridges, narrow corridors, and complex high-density lane merging.