Learning to Drive in New Cities Without Human Demonstrations

Zilin Wang [†1 2] **Saeed Rahmani** [†1 3] **Daphne Cornelisse** [4] **Bidipta Sarkar** [1 2] **Alexander David Goldie** [1 2]
Jakob Nicolaus Foerster [‡2] **Shimon Whiteson** [‡1]

Abstract

While autonomous vehicles have achieved reliable performance within specific operating regions, their deployment to new cities remains costly and slow. A key bottleneck is the need to collect many human demonstration trajectories when adapting driving policies to new cities that differ from those seen in training in terms of road geometry, traffic rules, and interaction patterns. In this paper, we show that self-play multi-agent reinforcement learning can adapt a driving policy to a substantially different target city using only the map and meta-information, *without requiring any human demonstrations from that city*. We introduce **NO** data **M**ap-based self-play for **A**utonomous **D**riving (NOMAD), which enables policy adaptation in a simulator constructed based on the target-city map. Using a simple reward function, NOMAD substantially improves both task success rate and trajectory realism in target cities, demonstrating an effective and scalable alternative to data-intensive city-transfer methods. Project Page: https://nomaddrive.github.io/

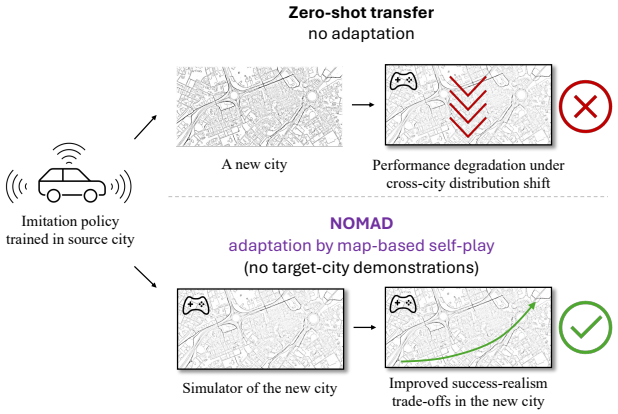

Figure 1. **City transfer in autonomous driving.** Top: Zero-shot deployment of an imitation policy trained in a source city into a new city leads to performance degradation due to cross-city distribution shift. Bottom: NOMAD adapts the same policy to the target city using only the target-city map and easily accessible meta-information, *without any human demonstrations*, by performing map-based self-play multi-agent reinforcement learning in a simulator of the new city. This adaptation substantially improves the policy's success rate and realism in the new city.

1. Introduction

Autonomous vehicles have improved dramatically over the past few years and now outperform humans in certain environments (Kusano et al., 2025; Di Lillo et al., 2024). However, they still operate reliably in only a small fraction of the global road network (Waymo, 2025b). For many current deployments, scaling and expansion remain a gradual, city-by-city effort (Waymo, 2025a). This is mainly because different cities, especially those in different countries or continents, vary in road geometry, traffic rules, and interaction patterns, i.e., the spatiotemporal behaviors induced by road topology, intersection design, and traffic density (Sun et al., 2023; Li et al., 2024a; Vasudevan et al., 2025). As a result, models trained with imitation learning in one city may exhibit degraded performance when deployed elsewhere (Feng et al., 2024), leading to passenger discomfort or even unsafe driving (Yasarla et al., 2025).

Current deployment pipelines address this problem by collecting new human demonstrations in target cities and fine-tuning policies on the new data (Leussink & Freed, 2025). However, this approach is slow and expensive, and hinders rapid expansion. If human demonstrations were not needed for geographic expansion, we could substantially reduce both the cost and the time needed to deploy autonomous vehicles at scale. This motivates a fundamental question:

Can we adapt driving policies to new cities without collecting additional human demonstrations from them?

Fortunately, although collecting driving trajectories in new cities is resource intensive, other forms of city information are more readily accessible. In particular, the lane-level

[†]Core Contributor,[‡]Equal Supervision [1]WhiRL, University of Oxford [2]FLAIR, University of Oxford [3]Delft University of Technology [4]NYU Tandon School of Engineering. Correspondence to: Zilin Wang <zilin.wang@lmh.ox.ac.uk>.

Workshop on Simulation for Autonomous Driving, CVPR 2026.

map and traffic meta-information, such as speed limits and traffic density, are prevalent and inexpensive.

Motivated by this observation, we introduce **NO** data **M**ap-based self-play for **A**utonomous **D**riving (NOMAD). As illustrated in Figure 1, NOMAD first constructs a simulator of the target city using its lane-level map and readily available traffic meta-information, such as speed limits and traffic density. This design is motivated by the hypothesis that much of the disparity in optimal driving behavior across cities is determined at the map level. A policy trained via imitation learning in the source city is then adapted through self-play multi-agent reinforcement learning (*self-play MARL*) within this simulator. During self-play, the policy interacts with other agents and explores feasible maneuvers and interaction dynamics induced by the target-city map, thereby acquiring experience that effectively substitutes for target-city demonstrations. Importantly, the adaptation relies on a *simple reward function*, avoiding city-specific reward engineering or manual tuning. We achieve it by anchoring the adaptation to the source-city policy via KL regularization, which preserves the behavioral prior learned from real human driving data in the source city while allowing deviations only where required by the target-city map and interaction structure. This selective adaptation substantially improves both success rate and behavioral realism in the target city.

Extensive closed-loop experiments demonstrate that NOMAD substantially improves zero-shot policy performance in unseen cities without relying on target-city human trajectories or complex reward engineering. In the Boston-to-Singapore transfer setting, NOMAD increases the success rate from 42% under zero-shot transfer to over 90%, while simultaneously improving trajectory realism from 0.679 to approximately 0.765, as measured by the WOSAC (Montali et al., 2023) realism metric in closed-loop evaluation. Furthermore, selected checkpoints produced by NOMAD form a Pareto frontier over success rate and realism, with each frontier checkpoint representing the best achievable policy for a particular success–realism trade-off. This enables practitioners to select deployment policies based on problem-specific requirements. We also conduct systematic analyzes of NOMAD, including the role of behavioral priors, the necessity of target-city map, comparison against policy with target-city demonstration access, generalization across diverse cities, and a sensitivity study on KL regularization strength. Overall, these results indicate that NOMAD substantially narrows cross-city generalization gaps, supporting scalable deployment of autonomous driving systems across diverse environments and highlighting the promise of self-play MARL for improving safety and robustness.

2. Related Work

2.1. Large-Scale Deployment of Autonomous Driving

Recent large-scale advances in autonomous driving have been driven by three complementary paradigms: the incorporation of Large Language/Vision-Language Models that leverage rich world knowledge to handle corner cases and improve generalization (Shao et al., 2024; Sima et al., 2024; Tian et al., 2024; Renz et al., 2025; Fu et al., 2025); World Models that act as data engines generating diverse traffic scenarios for training and evaluation (Hu et al., 2023; Russell et al., 2025; Ren et al., 2025; Wang et al., 2024); and simulation at scale providing infrastructure for closed-loop training and validation (Dosovitskiy et al., 2017; Montali et al., 2023; Caesar et al., 2021; Cao et al., 2025). However, a critical barrier remains: robustness under domain distribution shift. While early efforts addressed perception shifts across weather and sensor conditions (Sakaridis et al., 2018; Michaelis et al., 2019; Muhammad et al., 2022; Li et al., 2024b; Jeon et al., 2024), these scalable paradigms do not fully eliminate the need for explicit adaptation to new cities, where road geometry, traffic rules, and interaction patterns vary significantly—motivating the city-transfer setting discussed next.

2.2. City Transfer of Autonomous Driving

Transferring policies across diverse cities presents a more specific challenge. Vasudevan et al. (2025) demonstrate significant behavioral differences among different cities in the nuPlan dataset. UniTraj (Feng et al., 2024) shows that even state-of-the-art trajectory prediction models trained with large datasets struggle to generalize to new cities. Additionally, Yasarla et al. (2025) show that L2 error and collision rate both increase substantially when planners are transferred zero-shot to a new city.

These findings motivate the need for explicit cross-city adaptation mechanisms. TeraSim-World (Wang et al., 2025) proposes a pipeline to generate driving scenarios worldwide using the Cosmos-Drive world model (Ren et al., 2025). However, this approach still depends on world models that may not generate accurate scenarios for unseen cities. As an alternative, Vasudevan et al. (2025) propose AdaptiveDriver, a rule-based planner that adapts its behavior in new cities via interactions with a learned reactive world model. While their approach shows modest generalization to held-out cities within nuPlan, it still requires logged trajectories to calibrate the reactive world model in the new cities. Furthermore, the world model relies on IDM's (Kesting et al., 2010) single-agent car-following formulation, which limits the planner's ability to reason about complex multi-agent interactions. Moreover, RoCA (Yasarla et al., 2025) improves cross-domain robustness of end-to-end planners via a GP-based probabilistic to-

ken model and uncertainty-guided adaptation, achieved by prioritizing the most informative target data for finetuning. As another avenue to consider, LLaDA (Li et al., 2024a) uses Large Language Models to interpret text-based traffic codes and correct plan violations. While effective for explicit constraints like traffic rules, this approach struggles to capture geometry-dependent dynamics that are crucial for smooth trajectory planning but not formally codified in text. Although Wayve (2025) shows that large-scale foundation driving models can enable zero-shot transfer when the target city lies close enough to the training distribution, demonstration collection is necessary for cities with distinctly different traffic patterns. In contrast, NOMAD enables city transfer without collecting any logged trajectories from the target city. It adapts driving policies through experience gathered via self-play in simulator constructed from target-city map and meta-information, avoiding reliance on learned world models, target-city demonstrations, or industrial-scale training pipelines.

2.3. Self-Play for Autonomous Driving

Self-play is a powerful paradigm for training agents by allowing policies to improve through interactions with *copies of themselves*. CoPO (Peng et al., 2021) uses coordinated policy optimization to simulate self-driven particle systems in traffic by enabling agents to dynamically shift between cooperative and competitive behaviors. More recently, Cusumano-Towner et al. (2025) demonstrate that robust and naturalistic driving can emerge from self-play. A key challenge in self-play for driving, however, is maintaining human-like behavior while optimizing for task completion. Cornelisse & Vinitsky (2024) address this by regularizing self-play against a human reference policy. They also show that scaling self-play across thousands of Waymo scenarios yields reliable agents with over 99% goal completion and minimal collisions (Cornelisse et al., 2025). Similarly, SPACER (Chang et al., 2025) introduces a self-play framework that stabilizes RL training by anchoring decentralized agents to a centralized reference policy, mitigating non-stationarity while enabling human-like behavior learning. Beyond trajectory planning, self-play has been applied to asymmetric scenario generation (Zhang et al., 2024), where a teacher policy learns to create challenging yet solvable scenarios for a student, and to natural language communication for cooperative driving (Cui et al., 2025). While prior self-play methods target robustness and realism within a single domain, NOMAD repurposes self-play for cross-city transfer. By conducting self-play in map-based simulators of the target city, NOMAD enables adaptation without any target-city demonstrations.

3. Preliminaries and Problem Formulation

We model a target city C as a distribution over its map and traffic scenarios, $C = (\mathcal{M}_C, \Xi_C)$. Specifically, we sample a map segment $m \sim \mathcal{M}_C$, namely road layout and traffic direction, from a given region of the target city. Conditioned on m, we then sample a scenario $\xi = (s_0, \{g^i\}_{i=0}^{N-1}, m) \sim \Xi_C(\cdot|m)$, which specifies number of agents N, the initial state s_0, goal positions for all agents $\{g^i\}_{i=0}^{N-1}$, and the map segment m. Given a traffic scenario ξ, rolling out a policy π produces a trajectory $\tau \sim q_\pi(\cdot|\xi)$, where $q_\pi(\cdot|\xi)$ denotes the trajectory distribution induced by π in closed loop. We evaluate the performance of policy π with two metrics, success rate and realism, that are broadly accepted by the community (Li et al., 2022; Montali et al., 2023; Cornelisse* et al., 2025; Cusumano-Towner et al., 2025).

Success rate is the probability of a given vehicle reaching the goal-centered neighborhood within the horizon H:

$$\mathcal{S}(\pi, C) := \mathbb{E}_{\xi\sim\Xi_C}\left[\mathbb{E}_{\tau\sim q_\pi(\cdot|\xi)}\left[\mathbb{1}(\exists t \le H : d(p_t, g) \le \epsilon)\right]\right], \tag{1}$$

where p_t and g are the ego position at time t and the goal position, respectively. $\mathbb{1}(\cdot)$ is the indicator function. And ϵ is the error tolerance in reaching the goal.

Realism measures how well generated trajectories match the actual distribution of real-world driving behavior. Because the true distribution of real-world driving is unknown, we estimate realism using the likelihood of logged human trajectories under the trajectory distribution induced by π. Formally, we define:

$$\mathcal{R}(\pi, C) := \mathbb{E}_{\xi\sim\Xi_C}\left[\mathbb{E}_{\tau\sim p(\cdot|\xi)}\left[\log q_\pi(\tau|\xi)\right]\right], \tag{2}$$

where $p(\tau|\xi)$ denotes the distribution of logged human trajectories in the scenario ξ. Importantly, this realism metric requires access to target-city human trajectories *only for evaluation*. These trajectories are never used during training or adaptation, and serve solely as an offline benchmark for assessing behavior realism. Overall, we evaluate the policy π in city C by a multi-objective score

$$\mathbf{J}(\pi, C) = (\mathcal{S}(\pi, C), \mathcal{R}(\pi, C)). \tag{3}$$

City adaptation without demonstrations. Let π^0 be a planner trained using source-city data (e.g., via behavior cloning) and deployed zero-shot in target city C. Given target-city priors (the map $\mathcal{M}_C$ and meta-information $\mathcal{I}_C$ such as speed limits and traffic density), but no target-city demonstrations, our goal is to learn a *set* of adapted policies:

$$\Pi^+(C) = \text{Adapt}(\pi^0; \mathcal{M}_C, \mathcal{I}_C) = \{\pi^{+(1)}, \ldots, \pi^{+(\mathsf{N})}\} \tag{4}$$

that offer improved success–realism trade-offs in the target city. Namely, $\Pi^+(C)$ should contain policies that Pareto-dominate the zero-shot baseline:

$$\exists\, \pi^+ \in \Pi^+(C) \text{ s.t. } \mathbf{J}(\pi^+, C) \succ \mathbf{J}(\pi^0, C), \tag{5}$$

where $\succ$ indicates Pareto dominance. In practice, $\Pi^+(C)$ contains multiple candidate policies (e.g., checkpoints), from which we extract an empirical Pareto frontier. Ideally, we aim to expand the achievable Pareto frontier in the target city, yielding improved trade-offs across a wide range of success–realism preferences.

Deployment. At deployment time, a practitioner selects a single policy $\pi^{\text{deploy}} \in \Pi^+(C)$ according to application-specific preferences (e.g., prioritizing success rate or trajectory realism), corresponding to choosing an operating point along the induced Pareto frontier.

4. Multi-Agent Interaction Model

We formulate city adaptation problem as a partially observable stochastic game (POSG) (Hansen et al., 2004) defined by the tuple $(N, S, A, O, P, Z, r, H, \gamma, b_0)$, where each agent in the game is indexed by $i \in \{0, 1, \cdots, N-1\}$, and S is the set of possible states. P is the state transition function, deciding the probability of the next state s_{t+1} via joint action a_t at the state s_t. For agent i, A^i and O^i are the set of possible actions and observations, respectively. $Z^i(o_t^i|s_t, a_t, s_{t+1})$ is the observation function. $r^i(s_t, a_t, s_{t+1})$ is the scalar reward function. γ is the discount factor and b_0 is the probability of initial state s_0. We focus on homogeneous agents: all vehicles share the same observation space, action space, and reward function. This symmetry naturally supports a shared-parameter policy, which we exploit for scalable self-play training.

Observation space. At each timestep, each dynamic vehicle i receives a partial observation o_t^i consisting of its local ego state (e.g., velocity and goal) and a limited-range perception of nearby vehicles and road information (e.g., edges and lanes in a neighborhood). To facilitate learning and support parameter sharing, all observations are expressed in an ego-centric coordinate frame .

Action space. We adopt decentralized, memoryless policies $\pi(a_t^i|o_t^i)$, where temporal information such as velocity is encoded directly in the observation. At each step, the policy predicts discrete action increments in position and heading, $a_t^i = (\delta_x^i, \delta_y^i, \delta_h^i)$. These increments are defined in the ego vehicle's coordinate frame and are deterministically applied using kinematic pose updates. Each action dimension is bounded and uniformly discretized into multiple bins. A discretized action space improves training stability during reinforcement learning and supports multi-modal behavior naturally during imitation learning. In addition, it respects vehicle motion constraints and is commonly used in trajectory prediction and planning tasks (Philion et al., 2024; Cornelisse & Vinitsky, 2024; Wu et al., 2024; Zhang et al., 2025).

Reward function. Reward design in autonomous driving is non-trivial due to the difficulty of balancing safety, progress, and other factors (Knox et al., 2023). Moreover, precise reward engineering often requires substantial iterative tuning (Wurman et al., 2022) and can be brittle under transfer, since optimized policies may overfit to environment-specific reward proxies or simulator details, leading to degraded performance under distribution shift (Zhang et al., 2018; Pan et al., 2022). To demonstrate that NOMAD's effectiveness stems from the adaptation framework itself rather than reward tuning, we deliberately adopt a minimal reward formulation:

$$r^i(s_t, a_t, s_{t+1}) = w_g \mathbb{1}_{\mathcal{G}^i(s_{t+1})} + w_c \mathbb{1}_{\mathcal{C}^i(s_{t+1})} + w_o \mathbb{1}_{\mathcal{O}^i(s_{t+1})} \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. $\mathcal{G}^i(s_t)$, $\mathcal{C}^i(s_t)$, and $\mathcal{O}^i(s_t)$ represent if the goal is reached, a collision occurs, and the car drives offroad for agent i at timestep t, respectively. w_g, w_c, and w_o are corresponding weights, set as 1.0, -0.75, and -0.75, respectively, in all experiments.

5. NOMAD

We now introduce **NO** data **M**ap-based self-play for **A**utonomous **D**riving (NOMAD), which leverages map-based simulation and self-play MARL to adapt a driving policy to a new city without target-city demonstrations.

5.1. Overview

Figure 2 provides an overview of NOMAD. Let C be the city we want to transfer to, assuming the target-city map $\mathcal{M}_C$ and the city-specific meta-information $\mathcal{I}_C$ are accessible. $\mathcal{I}_C$ consists only of coarse information that is typically known a priori (e.g., speed limits and traffic density). Importantly, $\mathcal{I}_C$ is *global* to the city, independent of the specific map segment m, and low-dimensional, and thus does not require any trajectory data or city-specific learning.

Based on this, we construct a scenario generator $\Xi_\varphi(\xi|m, \mathcal{I}_C)$ with a learned or heuristic generic parameter φ to approximate the true distribution of scenarios $\Xi_C(\xi|m)$ in city C. The generator is used only to sample initial agent states and goal locations; it does *not* generate expert trajectories, driving behaviors, or policies. Any method that samples initial states and goals from the map can be used, and NOMAD, which focuses on map-based self-play rather than scenario generation, is agnostic to the specific scenario generation procedure. The generated scenarios $\xi = (s_0, \{g^i\}_{i=0}^{N-1}, m)$ are next loaded into a multi-agent data-driven autonomous driving simulator like GPUDrive (Kazemkhani et al., 2025), Waymax (Gulino et al.,

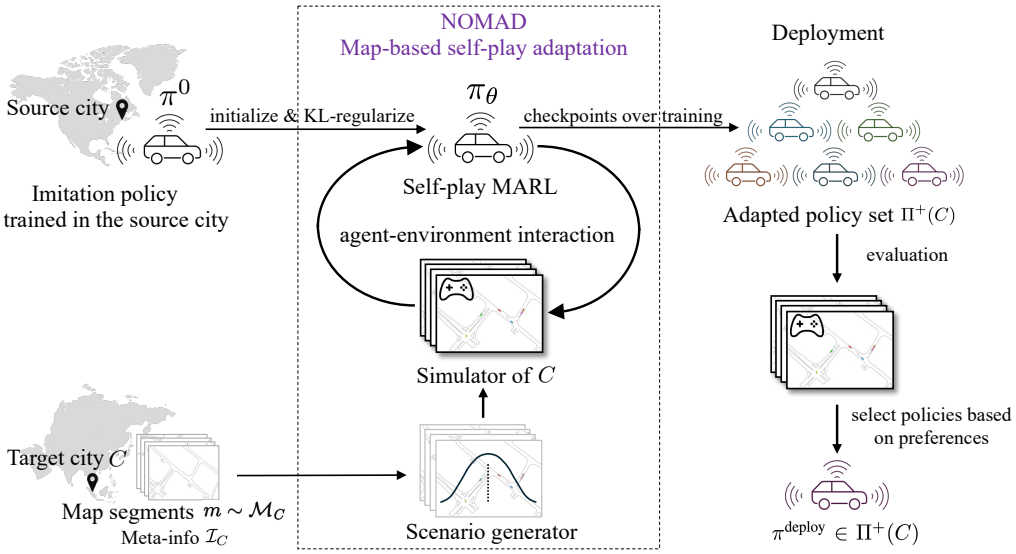

Figure 2. **NOMAD overview**. Starting from a source-city imitation policy π^0, NOMAD adapts it to a target city C using map segments $m \sim \mathcal{M}_C$ and meta-information $\mathcal{I}_C$. A scenario generator samples initial states and goals that are loaded in a data-driven multi-agent simulator, yielding a simulator of C. The policy π_θ is initialized from π^0 and optimized via KL-regularized self-play MARL. Training checkpoints produce an adapted policy set $\Pi^+(C)$, from which a deployment policy π^{deploy} is chosen based on practitioner preferences.

2023), or Nocturne (Vinitsky et al., 2022), yielding a simulator of the target city C.

Starting from an imitation learning policy π^0 from the source city, we initialize a policy π_θ and adapt it to the target city through KL-regularized self-play within the simulator. Training produces a set of adapted policies $\Pi^+(C)$ that achieve Pareto improvements over the zero-shot policy π^0. At deployment time, practitioners select a policy $\pi^{\text{deploy}} \in \Pi^+(C)$ along the induced Pareto frontier based on their preferences over success rate and realism.

5.2. Policy Adaptation via Regularized Self-Play

NOMAD solves the POSG described in Section 4 with regularized self-play. By requiring all agents to share a single policy $\pi_\theta(a_t|o_t)$ parameterized by θ, we can scale RL training to dense urban environments more effectively since the number of parameters is independent of the number of vehicles controlled. In addition, it also boosts training efficiency by allowing a single policy to learn from the diverse experiences collected by all agents simultaneously. We train π_θ with independent PPO (IPPO) due to its simplicity and efficacy (De Witt et al., 2020; Yu et al., 2022). IPPO treats each agent as an independent learner that optimizes a standard PPO objective,

$$\max_\theta \; J_{\text{PPO}}(\theta) = \mathbb{E}_{(o,a)\sim\pi_{\theta_{\text{old}}}}\Bigg[\min\Bigg(\frac{\pi_\theta(a|o)}{\pi_{\theta_{\text{old}}}(a|o)} A_{\pi_{\theta_{\text{old}}}}(o,a), \text{clip}\Bigg(\frac{\pi_\theta(a|o)}{\pi_{\theta_{\text{old}}}(a|o)}, 1-\epsilon, 1+\epsilon\Bigg) A_{\pi_{\theta_{\text{old}}}}(o,a)\Bigg)\Bigg], \tag{7}$$

using its own local observations and rewards, modeling the other agents as part of a non-stationary environment.

Given our minimal reward formulation, optimization without regularization could yield policies that complete goals but deviate from human-like driving patterns. To retain useful knowledge from the source city and mitigate reward hacking, we regularize the learned policy toward the prior π^0 using reverse Kullback–Leibler divergence. This choice penalizes assigning probability mass to actions deemed unlikely by the prior, thereby discouraging unnatural driving while allowing selective adaptation when required by the target city. The regularized objective is:

$$\max_\theta J(\theta) = J_{\text{PPO}}(\theta) - \lambda_{\text{KL}} \cdot \mathbb{E}_{\pi_\theta}\Bigg[\frac{1}{H}\sum_{t=0}^{H-1} D_{\text{KL}}\big(\pi_\theta(\cdot|o_t) \,\|\, \pi^0(\cdot|o_t)\big)\Bigg], \tag{8}$$

where D_{KL} is the KL divergence and λ_{KL} is a regularization weight, whose effects we analyze in Section 7.6.

6. Experimental Setup

Datasets and Simulator. We use nuPlan (Caesar et al., 2021), a large-scale dataset collected from the U.S. and Singapore. NuPlan stores the scenarios separately by cities and provides cross-continental coverage, which enables studying city transfer under large distribution shifts. For instance, Singapore and U.S. cities like Boston and Pittsburgh differ both in road topology and traffic handedness. In our experiments, we randomly select 4,000 traffic scenarios for each city, partitioned into 3,200 for training and 800 for test. For closed-loop traffic simulation, we use GPUDrive (Kazemkhani et al., 2025), a GPU-accelerated data-driven simulation tool suitable for fast and efficient RL training. We converted the nuPlan dataset into the format supported by GPUDrive, with 9-second scenarios dis-

cretized at 10 Hz.

Scenario Generation. We use a heuristic scenario generator $\Xi_\varphi(\cdot|m, \mathcal{I}_C)$ conditioned on the map, speed limits, and traffic density to produce feasible and realistic spawn and goal points for self-play simulation in the target city. For each map segment, we generate $K = 8$ distinct scenarios, yielding a total of $3{,}200 \times 8 = 25{,}600$ unique traffic scenarios for training.

Metrics. Evaluation is based on two core metrics, *success rate* and *realism*, defined in Section 3. For realism, we adopt the Waymo Open Sim Agents Challenge (WOSAC) evaluation metric (Montali et al., 2023). WOSAC aggregates weighted components over kinematics (e.g., speed and acceleration), interaction features (e.g., time-to-collision and collisions), and map compliance (e.g., road departures).

Training details and baselines. Our main experiments focus on policy transfer from Boston to Singapore, as this pair involves two different continents with differences in driving side road geometry, and traffic rules. We report the results for other city pairs in Section 7.5, however, unless stated otherwise, the results refer to a Boston-to-Singapore transfer by default. For the zero-shot baseline π^0, we train a trajectory planning model using behavior cloning (BC) using cross-entropy loss with trajectories from the source city. We also train its counterpart in the target city for comparison. Additional baselines include random and constant velocity policies, along with an oracle from logged demonstrations. We also report several diagnostic variants, including a variant with additional supervision beyond NOMAD. During adaptation, we initialize the network backbone and the actor head of π_θ with π^0 but learn the critic head from scratch. We set the interaction budget as 1 billion (approximately three days on a single NVIDIA A100) in all experiments to ensure a plateau of both the realism meta score and success rate.

7. Results

7.1. Main Results

Figure 3 (a) visualizes the empirical Pareto frontier of success rate versus realism meta score across different training checkpoints throughout self-play training in Boston-to-Singapore transfer . Each point represents an adapted policy (checkpoint) $\pi^+ \in \Pi^+(C)$ with the color shade indicating the number of interaction steps since the beginning of training. The yellow star denotes π_0, while the lime star shows behavior cloning using Singapore data. The modest performance of behavior cloning with 3,200 Singapore scenarios reflects the inherent data demands of imitation learning: empirical scaling analyzes in autonomous driving show that imitation-based policies are quite data hungry, with robustness and generalization improving only with substantially larger and more diverse datasets, while compounding errors dominate at smaller scales (Baniodeh et al., 2025) . By contrast, NOMAD achieves better success–realism trade-offs even without demonstrations in the target city, indicating that map-based self-play can be a more effective adaptation strategy than imitation on limited data even when some target-city trajectories are available.

Interestingly, the resulting Pareto frontier forms a clear upper envelope and dominates both the zero-shot transfer policy π^0 and behavior cloning trained on 3,200 Singapore scenarios. This dominance spans the entire spectrum of success–realism trade-offs, which means that regardless of the desired balance between these two objectives, NOMAD offers a policy that outperforms both baselines in both metrics simultaneously. This means that practitioners may select a deployment policy π^{deploy} anywhere along this frontier, from conservative settings that prioritize realism to aggressive settings that maximize task completion, while still benefiting from substantial improvements over zero-shot transfer. Surprisingly, the observed trade-off between success rate and realism is mild: substantial gains in success rate are achieved with only marginal reductions in realism (a 7-percentage-point gain in success rate corresponds to only a 0.012 decrease in realism). Therefore, in practice, one can simply select the checkpoint with the highest success rate, without requiring target-city trajectories to estimate realism meta scores.

To ground the frontier visualization with concrete values, Table 1 reports the success rate and realism meta score of representative baselines, the oracle policy, and the range spanned by the NOMAD Pareto frontier. NOMAD consistently achieves superior success–realism trade-offs compared to zero-shot transfer and other baselines, approaching the expert-derived upper bound without access to any target-city demonstrations.

Table 1. **Success rate and realism meta score for representative baselines and NOMAD in closed-loop evaluation for Singapore.** π_d^{expert} denotes an oracle policy obtained by inferring actions from logged expert trajectories and discretizing them to match the action space; its realism meta score serves as an approximate upper bound under the current POSG formulation.

Policy	Realism Meta Score	Success Rate
π_d^{expert}	**0.8056**	**88.26%**
Random	0.4074	4.30%
Constant Velocity	0.6147	19.45%
π^0	0.6795	42.25%
BC (Singapore)	0.7011	55.49%
NOMAD Frontier	**0.7570∼0.7697**	**87.28∼94.27%**

7.2. The Role of Behavioral Priors

NOMAD demonstrates that map-based self-play can effectively adapt a driving policy to a new city using only a simple reward function, which raises a natural question: *is an*

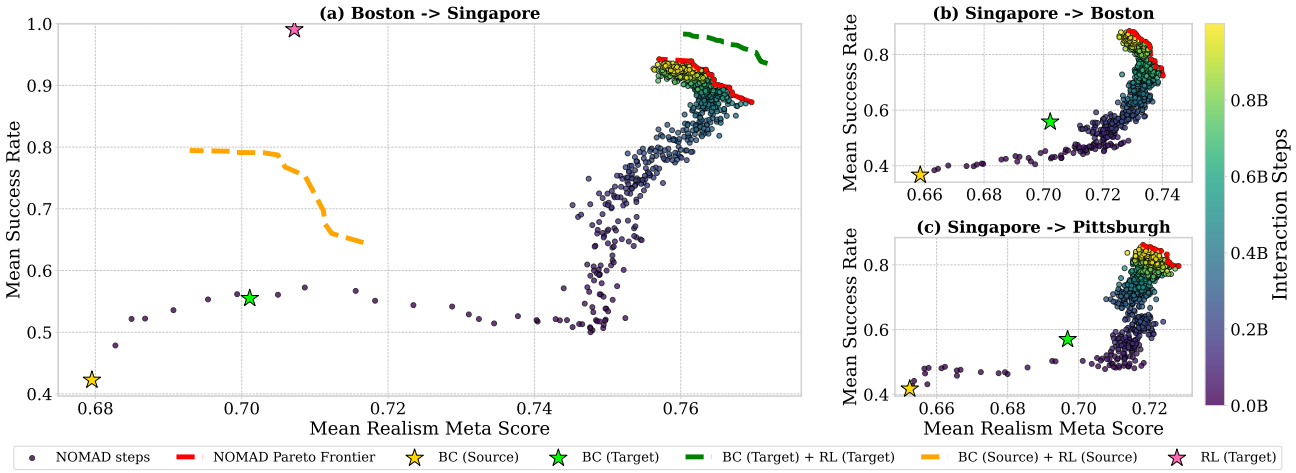

Figure 3. **Success–realism trade-offs under city transfer.** We plot mean success rate versus mean realism meta score over 5 runs for three transfer settings: (a) Boston-to-Singapore (primary), (b) Singapore-to-Boston, and (c)Singapore-to-Pittsburgh. Each dot is a NOMAD training checkpoint, colored by the cumulative number of interaction steps; the red dashed curve denotes the empirical Pareto frontier over NOMAD checkpoints. Stars and dashed curves denote reference policies and ablations: the zero-shot transfer behavior cloning policy from the source city π^0 (BC (Source)), behavior cloning with target-city demonstrations (BC (Target)), BC pretrained, self-play with logged target-city scenarios (BC (Target) + RL (Target)), BC pretrained, self-play with logged source-city scenarios (BC (Source) + RL (Source)), and RL from scratch in the target city with generated scenarios (RL (Target)).

imitation policy necessary to initialize and regularize policy learning? To answer this, we conduct an ablation study in which we train the policy purely via map-based self-play in the target city without regularization from scratch, using the same reward function and training budget. "RL (Target)" in Figure 3 shows the performance of the best checkpoint . Without an imitation policy, self-play can indeed achieve near-perfect success rates and improved realism meta score, compared to zero-shot transfer and behavior cloning policies in the target city. However, its realism meta score converges to a relative lower value. By contrast, NOMAD improves the success rate while maintaining a high realism meta score, ultimately achieving a substantially better success–realism trade-off.

To understand the source of this gap, we compare the kinematic metrics between these two methods in Figure 4. Self-play without pretraining and regularization fails to learn high kinematic scores. While NOMAD exhibits a mild decline in kinematic realism due to the absence of explicit realism rewards, its initialization from behavioral priors provides a superior starting point compared to training from scratch. Also, by anchoring the self-play process to this human-like prior, NOMAD maintains significantly higher kinematic scores throughout the adaptation horizon. These results highlight that self-play alone is insufficient for learning realistic driving behavior under minimal reward supervision. The pretrained BC policy both serves as an optimization warm-start and provides critical behavioral priors that constrains exploration to human driving patterns and mitigates reward hacking. This enables NOMAD to achieve superior realism without sacrificing task success.

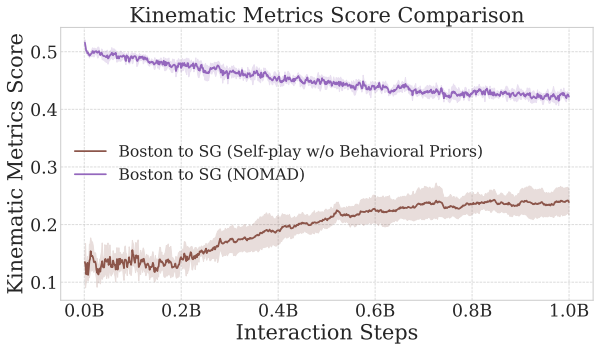

Figure 4. **Comparison of kinematic metrics between self-play with and without behavioral priors.** Self-play without behavioral priors struggles to learn kinematically realistic behavior, while NOMAD preserves substantially more realistic motion patterns despite lacking explicit kinematic rewards.

7.3. The Necessity of Target-City Map

To investigate the role of the target-city map in the adaptation process, we conduct behavior cloning followed by self-play with logged scenarios entirely in the *source* city (Boston) and evaluate the resulting policy zero-shot in the target city (Singapore). This baseline uses the same BC initialization and self-play training protocol as NOMAD, but never accesses the target-city map or scenarios during training.

The "BC (Source) + RL (Source)" in Figure 3 shows the Pareto frontier of source-city self-play . While this baseline improves upon the zero-shot transfer policy, its Pareto frontier remains significantly inferior to that of NOMAD. Specifically, self-play conducted in Boston saturates at substantially lower success rates and yields marginal gains in

realism when evaluated in Singapore.

These findings provide strong justification for map-based self-play in the *target* city. Self-play alone is not a silver bullet: improvements learned through interaction are largely city-specific and do not reliably transfer across distinct urban environments. Effective adaptation requires interaction dynamics to be grounded in the geometry, topology, and traffic structure of the deployment city, rather than relying solely on additional optimization in the source city.

7.4. NOMAD vs. Demonstration-Based Training

We further examine how closely NOMAD can approach the performance of a policy trained with access to target-city demonstrations. Specifically, we compare NOMAD, which uses neither target-city demonstrations nor logged scenarios, against a data-driven policy trained with access to 3,200 scenarios with human driving trajectories from Singapore. This policy is first pretrained via BC and subsequently trained using self-play on both generated and logged scenarios. The reward function, training budget, and overall training protocol are identical to those of NOMAD, with the only difference being that all training is conducted directly using target-city demonstrations.

This policy, denoted as "BC (Target) + RL (Target)" in Figure 3 , achieves only a modest improvement over NOMAD along the success–realism Pareto frontier. Notably, the performance gap between this policy and NOMAD is substantially smaller than the gap between NOMAD and zero-shot transfer. This indicates that map-based self-play provides the majority of the gains typically provided by target-city demonstrations.

7.5. Generalization to Other Cities

To verify that the effectiveness of NOMAD is not specific to a single target city, we evaluate cross-city transfer from Singapore to both Boston and Pittsburgh. Figure 3 (b) and (c) report the success–realism Pareto frontiers for these two transfer settings, respectively. These results demonstrate that NOMAD consistently delivers significant performance gains across different target cities. In all cases, the adapted policies substantially outperform the zero-shot baseline, despite the absence of any demonstrations from Boston or Pittsburgh. In particular, this generalization is achieved using a single reward function and a shared set of hyperparameters across all target cities, highlighting the robustness and scalability of NOMAD.

7.6. Sensitivity Study on KL Regularization Strength

We evaluate how KL divergence coefficient λ_{KL} in Equation 8 influences the balance between success rate and behavioral realism during city transfer. Figure 5 presents Pareto frontiers for different values of λ_{KL} in Boston-to-Singapore transfer. Higher KL coefficients constrain the policy to remain closer to the source-city prior, resulting in higher realism meta score but limiting the policy's ability to adapt to target-city geometries, manifested as lower success rates. The results reveal that the values of λ_{KL} at around 0.08 achieve the best balance. Importantly, even with relatively aggressive or conservative KL regularization choices, NOMAD can still substantially outperform the zero-shot baseline. In our experiments, we use the $\lambda_{\text{KL}} = 0.08$ for all city pairs.

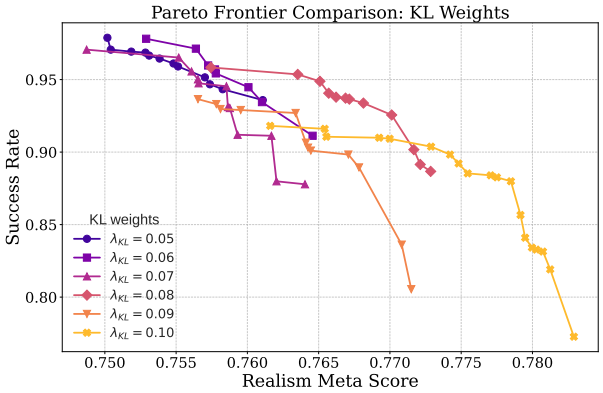

Figure 5. **Pareto frontiers of success rate versus realism meta score for different KL weights.** Smaller KL weights favor higher success at the cost of realism, while larger KL weights encourage more realistic behaviors but constrain success rate.

8. Conclusion and Future Work

This paper challenges a central assumption in city transfer in autonomous driving: that effective adaptation to a new city requires collecting human demonstrations in that city. We introduce NOMAD, a framework that adapts autonomous driving policies to new cities using only target-city map and its meta-information, without requiring any human demonstrations from them. Through extensive closed-loop evaluations, we show that NOMAD consistently expands the success–realism Pareto frontier, transforming brittle zero-shot transfer into a diverse set of policies that trade-off task success and behavioral realism. These results suggest that much of the disparity in optimal driving behavior across cities is determined at the map-level, which can be effectively addressed through map-based self-play.

Challenges remain. For example, differences in driving culture and social conventions across cities and countries continue to pose obstacles, highlighting the need for richer representations of interaction norms beyond geometric map structure. As traffic simulation continues to improve, we believe that scalable multi-agent self-play offers a viable and principled path toward robust, large-scale deployment of autonomous driving.

9. Acknowledgments

Compute for this project is graciously provided by the Isambard-AI National AI Research Resource, under the project "Robustness via Self-Play RL." Some experiments were also made possible by a generous equipment grant from NVIDIA. Zilin Wang is funded by a generous grant from Waymo. Saeed Rahmani is partially funded by the Transport & Mobility Institute at Delft University of Technology. Daphne Cornelisse is partially supported by the Cooperative AI Foundation and a Chishiki-AI SCIPE Fellowship. Bidipta Sarkar is supported by the Clarendon Fund Scholarship in partnership with a Department of Engineering Science Studentship for his Oxford DPhil. Alex David Goldie is funded by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems EP/S024050/1. Jakob Nicolaus Foerster is partially funded by the UKRI grant EP/Y028481/1 (originally selected for funding by the ERC). The authors thank Lukas Seier, Theo Wolf, Zengyuan Guo, Nathan Monette, Yulin Wang, Shashank Reddy, Juan Duque, Tingchen Fu, Ravi Hammond, Lu Li, and Darius Muglich for helpful discussions.

References

Baniodeh, M., Goel, K., Ettinger, S., Fuertes, C., Seff, A., Shen, T., Gulino, C., Yang, C., Jerfel, G., Choe, D., et al. Scaling laws of motion forecasting and planning–a technical report. *arXiv preprint arXiv:2506.08228*, 2025.

Caesar, H., Kabzan, J., Tan, K. S., Fong, W. K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., and Omari, S. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Cao, W., Hallgarten, M., Li, T., Dauner, D., Gu, X., Wang, C., Miron, Y., Aiello, M., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., and Chitta, K. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning (CoRL)*, 2025.

Chang, W.-J., Rangesh, A., Joseph, K., Strong, M., Tomizuka, M., Hu, Y., and Zhan, W. Spacer: Self-play anchoring with centralized reference models. *arXiv preprint arXiv:2510.18060*, 2025.

Cornelisse, D. and Vinitsky, E. Human-compatible driving partners through data-regularized self-play reinforcement learning. *arXiv preprint arXiv:2403.19648*, 2024.

Cornelisse*, D., Cheng*, S., Mandavilli, P., Hunt, J., Joseph, K., Doulazmi, W., Charraut, V., Gupta, A., Suarez, J., and Vinitsky, E. PufferDrive: A fast and friendly driving simulator for training and evaluating RL agents, 2025. URL https://github.com/Emerge-Lab/PufferDrive.

Cornelisse, D., Pandya, A., Joseph, K., Suárez, J., and Vinitsky, E. Building reliable sim driving agents by scaling self-play. *arXiv preprint arXiv:2502.14706*, 2025.

Cui, J., Tang, C., Holtz, J., Nguyen, J., Allievi, A. G., Qiu, H., and Stone, P. Talking vehicles: Cooperative driving via natural language. 2025.

Cusumano-Towner, M., Hafner, D., Hertzberg, A., Huval, B., Petrenko, A., Vinitsky, E., Wijmans, E., Killian, T., Bowers, S., Sener, O., et al. Robust autonomy emerges from self-play. *arXiv preprint arXiv:2502.03349*, 2025.

De Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Di Lillo, L., Gode, T., Zhou, X., Atzei, M., Chen, R., and Victor, T. Comparative safety performance of autonomous-and human drivers: A real-world case study of the waymo driver. *Heliyon*, 10(14), 2024.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Feng, L., Bahari, M., Amor, K. M. B., Zablocki, É., Cord, M., and Alahi, A. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *European Conference on Computer Vision*, pp. 106–123. Springer, 2024.

Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., and Bai, X. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025.

Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36:7730–7742, 2023.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pp. 709–715, 2004.

Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., and Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Jeon, M., Seo, J., and Min, J. Da-raw: Domain adaptive object detection for real-world adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2013–2020. IEEE, 2024.

Kazemkhani, S., Pandya, A., Cornelisse, D., Shacklett, B., and Vinitsky, E. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ERv8ptegFi.

Kesting, A., Treiber, M., and Helbing, D. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philosophical Transactions of the Royal Society A*, 368(1928):4585–4605, 2010. doi: 10.1098/rsta.2010.0084.

Knox, W. B., Allievi, A., Banzhaf, H., Schmitt, F., and Stone, P. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023.

Kusano, K. D., Scanlon, J. M., Chen, Y.-H., McMurry, T. L., Gode, T., and Victor, T. Comparison of waymo rider-only crash rates by crash type to human benchmarks at 56.7 million miles. *Traffic Injury Prevention*, pp. 1–13, 2025.

Leussink, D. and Freed, J. Waymo to begin data collection in tokyo with driver-operated test rides, April 2025. URL https://www.reuters.com/business/autos-transportation/waymo-begin-data-collection-tokyo-with-driver-operated-test-rides-2025-04-10/. Published April 10, 2025; updated April 10, 2025. Reporting by Daniel Leussink; editing by Jamie Freed.

Li, B., Wang, Y., Mao, J., Ivanovic, B., Veer, S., Leung, K., and Pavone, M. Driving everywhere with large language model policy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14948–14957, 2024a.

Li, L., Lyu, J., Ma, G., Wang, Z., Yang, Z., Li, X., and Li, Z. Normalization enhances generalization in visual reinforcement learning. In *AAMAS*, 2024b.

Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., and Zhou, B. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45 (3):3461–3475, 2022.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., and Brendel, W. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Montali, N., Lambert, J., Mougin, P., Kuefler, A., Rhinehart, N., Li, M., Gulino, C., Emrich, T., Yang, Z., Whiteson, S., et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36: 59151–59171, 2023.

Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., and De Albuquerque, V. H. C. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 22694–22715, 2022.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Peng, Z., Li, Q., Hui, K. M., Liu, C., and Zhou, B. Learning to simulate self-driven particles system with coordinated policy optimization. *Advances in neural information processing systems*, 34:10784–10797, 2021.

Philion, J., Peng, X. B., and Fidler, S. Trajeglish: Traffic modeling as next-token prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Z59Rb5bPPP.

Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J. Z., Chen, R., et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Renz, K., Chen, L., Arani, E., and Sinavski, O. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11993–12003, 2025.

Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., and Corrado, G. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Sakaridis, C., Dai, D., Hecker, S., and Van Gool, L. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 687–704, 2018.

Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S. L., Liu, Y., and Li, H. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15120–15130, 2024.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., and Li, H. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Sun, X., Jiang, Y., Burnett, G., Bai, J., and Bai, R. A cross-cultural analysis of driving styles for future autonomous vehicles. *Advanced Design Research*, 1(2):71–77, 2023.

Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., and Zhao, H. DriveVLM: The convergence of autonomous driving and large vision-language models. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=928V4Umlys.

Vasudevan, A. B., Peri, N., Schneider, J., and Ramanan, D. Planning with adaptive world models for autonomous driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14938–14945. IEEE, 2025.

Vinitsky, E., Lichtlé, N., Yang, X., Amos, B., and Foerster, J. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world. *Advances in Neural Information Processing Systems*, 35: 3962–3974, 2022.

Wang, J., Sun, H., Yan, X., Feng, S., Gao, J., and Liu, H. X. Terasim-world: Worldwide safety-critical data synthesis for end-to-end autonomous driving. *arXiv preprint arXiv:2509.13164*, 2025.

Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., and Lu, J. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Waymo. Bringing waymo to more people, sooner, August 2025a. URL https://waymo.com/blog/2025/08/bringing-waymo-to-more-people-sooner. Accessed: 2025-12-22.

Waymo. Where waymo is driving, December 2025b. URL https://waymo.com/rides/#rides-map. Accessed: 2025-12-22.

Wayve. The ai-500 roadshow: 500 cities and what we learned, 2025. URL https://wayve.ai/thinking/ai-500-roadshow-500-cities/. Accessed: 2025-01-07.

Wu, W., Feng, X., Gao, Z., and Kan, Y. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems*, 37:114048–114071, 2024.

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

Yasarla, R., Han, S., Cheng, H.-P., Liu, L., Mahajan, S., Bhattacharyya, A., Shi, Y., Garrepalli, R., Cai, H., and Porikli, F. Roca: Robust cross-domain end-to-end autonomous driving. *arXiv preprint arXiv:2506.10145*, 2025.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Zhang, C., Vinyals, O., Munos, R., and Bengio, S. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

Zhang, C., Biswas, S., Wong, K., Fallah, K., Zhang, L., Chen, D., Casas, S., and Urtasun, R. Learning to drive via asymmetric self-play. In *European Conference on Computer Vision*, pp. 149–168. Springer, 2024.

Zhang, Z., Karkus, P., Igl, M., Ding, W., Chen, Y., Ivanovic, B., and Pavone, M. Closed-loop supervised fine-tuning of tokenized traffic models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5422–5432, 2025.