APEX: ROBUST DRIVING SELF-PLAY THROUGH ADAPTIVE POPULATION EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale driving self-play can learn robust policies through closed-loop interaction without human driving demonstrations. Yet existing methods use shared policies and fixed traffic-behavior distributions, causing driving policies to overfit to training-time behavioral regularities and generalize poorly to unseen traffic behaviors. We introduce Adaptive Population Exploration for Self-Play (APEX), which reframes driving self-play as a population-level multi-objective learning problem. Specifically, APEX separates the target driving policy, termed the focal policy, from a behavior-conditioned traffic policy, while making the distribution of traffic behaviors learnable. While the two policies interact and optimize their respective objectives, a population constructor uses focal-policy feedback to adapt the distribution of traffic conditions, thereby changing which self-interested behaviors the focal policy encounters without directly controlling vehicle actions. To focus this adaptation on policy weaknesses rather than intrinsic scene difficulty, we use the focal critic's initial value estimate as a baseline for realized focal performance. We show that this centering preserves the expected optimization direction of the population objective under our training protocol. By alternating policy learning and population adaptation, APEX tracks evolving weaknesses. Experiments show improved robustness and generalization to unseen traffic behaviors while maintaining competitive performance under standard self-play.

1 INTRODUCTION

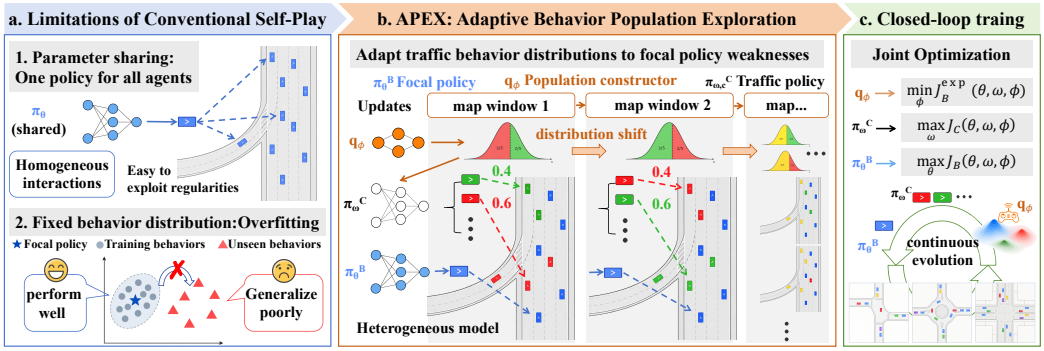


Figure 1: Limitations of conventional driving self-play, our solution, and its training process.

Large-scale simulation has made closed-loop self-play an increasingly practical paradigm for learning driving policies without relying on human driving demonstrations. Interaction-rich datasets such as the Waymo Open Motion Dataset provide realistic scenes and map context (Ettinger et al., 2021), while simulators including Nocturne, Waymax, and GPUDrive enable efficient multi-agent rollouts at increasing scale (Vinitsky et al., 2022; Gulino et al., 2023; Kazemkhani et al., 2025). Recent systems have pushed this paradigm further: GigaFlow demonstrates naturalistic driving through self-play over billions of simulated kilometers, while PufferDrive trains reliable simulation agents across thousands of scenarios (Cusumano-Towner et al., 2025; Cornelisse et al., 2025b). These advances greatly expand the amount and diversity of interactive experience available for policy learning. As

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.

self-play continues to scale, however, an increasingly important question is not only how much experience can be collected, but which traffic behaviors should generate that experience as the learner evolves.

More self-play experience does not by itself guarantee robustness to changes in traffic behavior. Scene generalization and behavioral generalization are distinct capabilities: a policy may perform well on unseen maps while still relying on behavioral regularities specific to its training partners. Recent evidence from BehaviorBench shows that policies learned through pure self-play can overfit to their training partners and generalize poorly when interacting with different traffic-agent behaviors (Distelzweig et al., 2026). This problem is especially relevant when large-scale training repeatedly samples from the same policy family or a fixed mixture of behavior styles, making frequently observed interaction patterns predictable and exploitable. More importantly, the behaviors that remain informative for training can change as the learner improves, whereas a fixed traffic distribution continues to allocate experience according to its original mixture. Scaling scenes and rollouts can therefore improve scene coverage without resolving the behavioral distribution shift between training and evaluation.

Existing methods address parts of this problem, but largely leave the composition of the training population outside the learning loop. One direction expands behavioral diversity through reward, context, or property conditioning: GigaFlow varies agent-level objectives, PHASE conditions a shared policy on agent context, and TerraZero randomizes agent properties and rewards (Cusumano-Towner et al., 2025; Qiu et al., 2026; Wu et al., 2026). Such approaches enlarge the family of behaviors that can be represented, but representational diversity is not the same as adaptive training exposure: the distribution from which behavior conditions are sampled is typically fixed or externally specified rather than adapted to the focal learner. A different direction introduces adversarial agents or traffic generators that adapt to the target policy and expose its weaknesses (Zhang et al., 2024; Gui et al., 2026; Nie et al., 2026). Yet directly adversarializing traffic creates a different tension: sparse adversaries have limited influence over the overall population, whereas making many surrounding vehicles explicitly optimize against the focal policy increasingly departs from ordinary traffic interactions. Real traffic participants primarily pursue their own driving objectives rather than collectively minimizing another driver's performance. This motivates adapting difficulty at the population level: instead of turning individual vehicles into adversaries, one can change how frequently different self-interested traffic behaviors appear during training.

To address these limitations, we introduce Adaptive Population Exploration for Self-Play (APEX), a population-level self-play framework that adapts the behavioral composition of training traffic to the evolving focal policy (Fig. 1(b)). APEX separates the target focal policy from a behavior-conditioned traffic policy, enabling heterogeneous self-interested traffic behaviors without relying on a single parameter-shared policy. A high-level population constructor further learns how frequently different traffic behaviors appear during training, replacing the fixed behavior distribution in conventional self-play with an adaptive one that continuously shifts training exposure toward the focal policy's evolving weaknesses. Importantly, the constructor changes the distribution of traffic behaviors rather than directly controlling vehicle actions, preserving self-interested traffic interactions while creating an evolving training curriculum (Fig. 1(c)). Since behavior conditioning is confined to the traffic side, the learned focal policy remains standalone and requires no behavior-condition assignment or population selection at test time. Together, APEX addresses both homogeneous interactions and fixed behavioral exposure in conventional self-play (Fig. 1(a)).

Our contributions are threefold:

- We formulate driving self-play as a population-level learning problem that separates focal-policy learning, self-interested traffic behavior learning, and adaptive traffic-population construction.
- We propose APEX, which adapts the training behavior distribution to the focal policy's evolving weaknesses through a learnable population constructor and alternating optimization, without requiring behavior-condition selection for the focal policy at test time.
- We introduce a critic-centered adaptation signal and provide theoretical analysis of its optimization properties. Experiments demonstrate improved robustness and generalization to unseen traffic behaviors while maintaining competitive standard self-play performance.

2 RELATED WORK

Scalable Driving Self-Play and Behavioral Diversity. Nocturne, Waymax, and GPUDrive have made closed-loop multi-agent driving simulation increasingly scalable (Vinitsky et al., 2022; Gulino et al., 2023; Kazemkhani et al., 2025). GigaFlow and PufferDrive further demonstrate large-scale self-play across massive simulated experience and diverse scenarios (Cusumano-Towner et al., 2025; Cornelisse et al., 2025b), with recent work extending this paradigm to procedural environments and end-to-end policies (Wu et al., 2026; Rowe et al., 2026). Behavioral heterogeneity can be introduced through agent-level objectives, context conditioning, or randomized properties, as in GigaFlow, PHASE, TerraZero, and TGPO (Cusumano-Towner et al., 2025; Qiu et al., 2026; Wu et al., 2026; Fan et al., 2025). Other studies incorporate learned references or limited human data to improve behavioral compatibility (Chang et al., 2026; Cornelisse et al., 2026). These approaches broaden the behaviors available to self-play, whereas APEX additionally separates behavior learning from population construction and adapts how frequently different conditioned behaviors are encountered during focal-policy training.

Adaptive Curriculum and Opponent Generation. Curriculum and environment-design methods adapt training challenges to the evolving learner. PAIRED and Prioritized Level Replay respectively generate or prioritize informative environments (Dennis et al., 2020; Jiang et al., 2021), while recent driving-specific methods extend adaptive curricula to multi-agent scenarios and traffic behaviors (Brunnbauer et al., 2025; Abouelazm et al., 2025; Sheng et al., 2026). Other approaches adapt explicit opponents: Asymmetric Self-Play learns a teacher that constructs challenging situations, Active Markov Games co-evolves ego and opponent strategies, and adversarial world modeling adapts traffic generators to expose planner vulnerabilities (Zhang et al., 2024; Gui et al., 2026; Nie et al., 2026). These methods primarily modify environments, scenarios, or opponent behaviors to increase training difficulty. In contrast, APEX adapts a distribution over conditions for a separately trained, self-interested traffic policy, shifting behavioral exposure without giving surrounding vehicles an explicit objective to cause focal failure.

3 APEX: ADAPTIVE POPULATION EXPLORATION FOR SELF-PLAY

We introduce APEX, an adaptive population-level framework for driving self-play. Sec. 3.1 formulates the population-level self-play problem. Sec. 3.2 presents adaptive population construction with critic-centered adaptation. Finally, Sec. 3.3 describes the alternating optimization procedure for policy learning and population adaptation.

3.1 PROBLEM FORMULATION

We model closed-loop driving as a finite-horizon partially observed Markov game $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i, \mathcal{O}_i, \Omega_i, R_i\}_{i\in\mathcal{N}}, P, \rho_0, H, \gamma)$. Here $\mathcal{N}$ is the set of controllable vehicle slots, $\mathcal{S}$ the joint scene-state space, $\mathcal{A}_i$ and $\mathcal{O}_i$ the action and observation spaces, Ω_i the local observation function, R_i the driving reward, P the closed-loop transition kernel, ρ_0 the initial-scene distribution, H the episode horizon, and γ the discount factor. A state $s_t \in \mathcal{S}$ contains the map and the vehicle states and goals needed to advance the simulator; $s_0 \sim \rho_0$ selects a logged map and its initialized traffic scene. At each step, agent i receives a local scene observation $\bar{o}_t^i = \Omega_i(s_t)$, comprising its own motion and goal features, nearby vehicles, and local road geometry, then takes $a_t^i \in \mathcal{A}_i$. The joint action $\mathbf{a}_t = (a_t^i)_{i\in\mathcal{N}}$ induces $s_{t+1} \sim P(\cdot \mid s_t, \mathbf{a}_t)$; individual episodes end at a terminal event or the horizon. A map window can contain multiple such episodes.

Within each window, we partition $\mathcal{N} = \mathcal{B} \dot\cup \mathcal{C}$ into focal and conditioned traffic agents. For $i \in \mathcal{B}$, set $o_t^i = \bar{o}_t^i$; the focal policy $\pi_\theta^B(a_t^i \mid o_t^i)$ optimizes the fixed driving reward $r_t^i = R_i^B(s_t, \mathbf{a}_t)$. Each $j \in \mathcal{C}$ receives a condition $\mathbf{c}_j \in \mathcal{C}_{\text{cond}}$ that shapes its own reward and partner visibility. A condition-dependent partner mask $M_{\mathbf{c}_j}$ produces $o_t^j = M_{\mathbf{c}_j}(\bar{o}_t^j)$, which, together with $\mathbf{c}_j$, feeds $\pi_\omega^C(a_t^j \mid o_t^j, \mathbf{c}_j)$, while its self-interested reward is $r_t^j = R_j^C(s_t, \mathbf{a}_t; \mathbf{c}_j)$. The two policies have separate parameters, objectives, rollout samples, and gradients; the focal policy does not require a behavior condition at test time. Varying $\mathbf{c}_j$ creates a family of self-interested traffic behaviors rather than a single opponent.

To determine how frequently different traffic behaviors are encountered, we introduce a parameterized population constructor q_ϕ, which defines a distribution over population descriptors. At the beginning of each map window, the constructor samples a shared descriptor ξ, which in turn parameterizes a distribution p_ξ over behavior conditions. Each conditioned traffic agent then independently draws its condition from this distribution:

$$\xi \sim q_\phi(\xi), \qquad \mathbf{c}_j \mid \xi \overset{\text{i.i.d.}}{\sim} p_\xi(\mathbf{c}), \quad j \in \mathcal{C}. \tag{1}$$

Let $\mathbf{c} = (\mathbf{c}_j)_{j\in\mathcal{C}}$ denote the joint condition assignment of the traffic population. The hierarchical sampling process in Eq. (1) induces the joint population distribution

$$p_\phi(\mathbf{c}) = \int q_\phi(\xi) \prod_{j\in\mathcal{C}} p_\xi(\mathbf{c}_j)\, d\xi. \tag{2}$$

Thus, q_ϕ controls which population compositions are encountered during training, while conditionally independent sampling from p_ξ preserves agent-level behavioral diversity within a window. The descriptor ξ remains fixed throughout the map window, whereas an individual traffic agent may resample its condition when starting a new episode. The marginalization and its population-level implication are detailed in Appendix A.1. For group $K \in \{B, C\}$, let G_K denote the mean per-agent discounted driving return, with each episode contributing $\sum_{t=0}^{T_i-1} \gamma^t r_t^i$ up to its terminal or time-limit step $T_i \le H$. Define $J_B = \mathbb{E}[G_B]$ and $J_C = \mathbb{E}[G_C]$, where the expectation covers map initialization, population and condition sampling, policy actions, and simulator dynamics.

The population constructor serves a different role: it should identify behavioral compositions that expose weaknesses of the current focal policy. Because a map window may contain repeated interaction under the same descriptor, later episodes can already reflect focal-policy adaptation to that population. We therefore evaluate the constructor using the focal policy's first exposure to each newly sampled descriptor. Let G_B^{exp} denote the mean focal return over the first focal episodes initiated after sampling ξ, and define

$$J_B^{\text{exp}}(\theta, \omega, \phi) = \mathbb{E}[G_B^{\text{exp}}]\,. \tag{3}$$

This objective evaluates only the population constructor; the focal and conditioned policies continue to optimize their driving objectives J_B and J_C. The exact return convention is given in Appendix B. In Sec. 3.2, we compare the realized first-exposure return with the focal critic's pre-exposure value estimate to construct the adaptation signal.

Definition 1 (population-level self-play). APEX trains the two driving policies and the population constructor with distinct objectives:

$$\max_\theta J_B(\theta, \omega, \phi), \qquad \max_\omega J_C(\theta, \omega, \phi), \qquad \min_\phi J_B^{\text{exp}}(\theta, \omega, \phi). \tag{4}$$

The focal policy learns its driving task and the conditioned policy learns to realize self-interested objectives under different conditions. Only the constructor is antagonistic to focal performance; it changes the frequency of traffic behaviors without rewarding individual traffic agents for causing focal failures. We next specify the population family and the signal used to adapt it.

3.2 Adaptive Population Construction with Critic-Centered Adaptation

Population parameterization. We instantiate the population constructor q_ϕ and the descriptor-conditioned behavior distribution p_ξ introduced in Sec. 3.1. A behavior condition

$$\mathbf{c} = (\alpha_{\text{col}}, \alpha_{\text{off}}, r_{\text{goal}}, r_{\text{resp}}, d_{\text{view}}, \beta_{\text{view}})$$

contains four coordinates controlling a traffic agent's own reward and two controlling its perception of surrounding agents. The constructor adapts the distribution over these conditions rather than directly controlling vehicle actions or deterministically assigning individual behaviors.

We write $\xi = (z_1, \ldots, z_6, h_1, \ldots, h_6)$, where each pair (z_ℓ, h_ℓ) parameterizes the mean and dispersion of one Beta distribution. Specifically,

$$q_\phi(\xi; \sigma) = \mathcal{N}(\mu_\phi, \sigma^2 I_{12}), \quad u_{j\ell} \mid \xi \sim \text{Beta}(a_\ell(z_\ell, h_\ell), b_\ell(z_\ell, h_\ell))\,, \quad \mathbf{c}_j \mid \mathbf{u}_j \sim \mathcal{T}(\cdot \mid \mathbf{u}_j), \tag{5}$$

where $j \in \mathcal{C}$, $\ell = 1, \ldots, 6$, and $\mathbf{u}_j = (u_{j1}, \ldots, u_{j6})$. The functions a_ℓ and b_ℓ produce the positive Beta shape parameters. Conditional on ξ, the normalized draws are independent across agents and condition coordinates. The mean μ_ϕ is learned, whereas σ is an externally scheduled exploration scale independent of ϕ; we suppress the scale argument when it is clear from context.

The sampling kernel $\mathcal{T}$ maps normalized draws to admissible behavior conditions, including auxiliary randomness for the view-angle distribution. The sigmoid parameterization of the Beta laws, condition bounds, and sampling convention are specified in Appendix B. The descriptor ξ remains fixed within a map window, while traffic agents may resample conditions after episode resets. By design, the constructor is scene-agnostic: q_ϕ is not conditioned on the reset scene and therefore learns a global training-population distribution rather than a scene-specific response.

Critic-centered adaptation. The realized focal return depends jointly on the reset scene and the sampled population descriptor. A low first-exposure return may therefore reflect an intrinsically difficult scene rather than an unfavorable traffic population. Optimizing against absolute return can give scene-dependent difficulty undue influence in a sampled update. We instead center the population signal using the focal critic available from policy learning.

Let $\mathbf{x} = s_0 \sim \rho_0$ denote the reset scene before the descriptor is sampled, and let $\mathcal{B}$ be the focal agents present at that reset. For the idealized constructor update, fix the driving policies and focal critic. Conditional on ξ and the reset scene, the rollout law has no additional dependence on ϕ, so the score-function identity gives

$$\nabla_\phi J_B^{\mathrm{exp}} = \mathbb{E}[G_B^{\mathrm{exp}} \nabla_\phi \log q_\phi(\xi)] \,. \tag{6}$$

Thus, sampled first-exposure returns provide a learning signal for the constructor, but do not separate population-induced degradation from scene-level difficulty.

The focal critic V_ψ^B, with parameters ψ, evaluates the initial focal observations before any conditioned traffic action is applied. We define the scene-dependent baseline and the *population advantage* by

$$b_\psi(\mathbf{x}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} V_\psi^B(o_0^i), \qquad A_{\mathrm{pop}} = b_\psi(\mathbf{x}) - G_B^{\mathrm{exp}}. \tag{7}$$

A positive advantage indicates that realized focal performance fell below the critic's pre-exposure estimate for the same reset scene. Larger values assign greater adaptation weight to descriptors associated with relative focal underperformance. This is a scene-centered learning signal rather than a causal estimate of a population's effect from a single rollout.

Proposition 1 (gradient invariance under critic centering). Fix the driving policies and critic. Assume that the reset-scene distribution is independent of ϕ and, conditional on $\mathbf{x}$, $b_\psi(\mathbf{x})$ is independent of the newly sampled descriptor ξ. Then

$$\mathbb{E}[b_\psi(\mathbf{x}) \nabla_\phi \log q_\phi(\xi)] = 0, \qquad \mathbb{E}[A_{\mathrm{pop}} \nabla_\phi \log q_\phi(\xi)] = -\nabla_\phi J_B^{\mathrm{exp}}. \tag{8}$$

Hence, critic centering preserves the expected adversarial optimization direction of the constructor.

Under the first-exposure protocol, ξ changes only the conditioned traffic population, whereas the initial focal observations are evaluated before any conditioned action can affect the scene. This establishes baseline independence for fixed policies and critic. Invariance does not require critic calibration; calibration affects the interpretation of the baseline as expected focal performance and its effectiveness as a control variate. The proof, collection conditions, and role of calibration are detailed in Appendices A.2–A.5.

3.3 Alternating self-play optimization

The objectives in Eq. (4) share closed-loop interactions but are updated on different timescales. We interleave online driving-policy adaptation with less frequent constructor updates from completed map-window records. Within each window, the descriptor remains fixed while the driving policies may continue to learn. Constructor updates affect subsequent windows, leaving active windows' descriptors unchanged.

Policy adaptation. At the beginning of window n, the simulator resets a scene $\mathbf{x}_n$ and samples $\xi_n \sim q_{\phi_n}(\cdot; \sigma_n)$, where ϕ_n and σ_n are the constructor parameters and exploration scale at collection time. Traffic conditions are drawn from p_{ξ_n} according to Eq. (1) and may be resampled from that same law upon traffic-agent resets. Both driving policies interact in the same scenes, but use separate trajectory batches and their own driving rewards for PPO updates (Schulman et al., 2017).

Before the first action, the focal critic supplies the baseline $b_n = b_{\psi_n}(\mathbf{x}_n)$ from Eq. (7), where ψ_n denotes its parameters at that evaluation. The first-exposure return yields $A_{\mathrm{pop},n} = b_n - G^{\mathrm{exp}}_{B,n}$. Later episodes may continue to train both driving policies, but are excluded from this target. Each eligible completed window therefore contributes one constructor record, while potentially supporting multiple policy-learning episodes.

Population adaptation. Record n retains $(\xi_n, \mu_n, \sigma_n, A_{\mathrm{pop},n})$, with $\mu_n = \mu_{\phi_n}$ and collection law $q_n(\xi) = \mathcal{N}(\xi; \mu_n, \sigma_n^2 I_{12})$. Once a record batch $\mathcal{D}$ reaches the collection threshold, we consume all available records and form the batch-standardized population advantages $\widetilde{A}_{\mathrm{pop},n}$. Each candidate likelihood uses its record's collection scale:

$$\rho_n(\phi) = \frac{q_\phi(\xi_n; \sigma_n)}{q_n(\xi_n)} = \frac{\mathcal{N}(\xi_n; \mu_\phi, \sigma_n^2 I_{12})}{\mathcal{N}(\xi_n; \mu_n, \sigma_n^2 I_{12})}. \tag{9}$$

The Gaussian constructor maximizes the clipped surrogate

$$\mathcal{L}^{\mathrm{clip}}_{\mathrm{pop}}(\phi) = \mathbb{E}_{n\sim\mathcal{D}}\left[\min\Big(\rho_n(\phi)\widetilde{A}_{\mathrm{pop},n}, \mathrm{clip}(\rho_n(\phi), 1-\epsilon_{\mathrm{pop}}, 1+\epsilon_{\mathrm{pop}})\widetilde{A}_{\mathrm{pop},n}\Big)\right]. \tag{10}$$

Here $n \sim \mathcal{D}$ denotes uniform averaging over the batch. Only ϕ is optimized; the recorded scales remain fixed. Larger advantages favor descriptors associated with relative focal underperformance, while clipping limits the incentive for further reweighting outside the prescribed ratio range.

For fixed driving policies and critic, raw advantages, and an unclipped on-policy update, Proposition 1 gives the adversarial direction $-\nabla_\phi J^{\mathrm{exp}}_B$. The practical surrogate additionally uses batch standardization, clipping, and records collected under evolving policies. Appendix B specifies the ratio and normalization convention. Appendix A.6 analyzes population tracking under stale driving-policy records, while Appendix A.7 relates the clipped surrogate locally to the ideal score-function update. Algorithm 1 summarizes the procedure; concurrently active windows are shown schematically as individual windows, with policy batches and constructor records pooled across windows.

Algorithm 1 APEX alternating policy and population updates

Input: $\pi^B_\theta, \pi^C_\omega, q_\phi$, threshold N
$\mathcal{D} \leftarrow \varnothing$
for each map window n **do**
  $\mathbf{x}_n \leftarrow \mathrm{ResetMap}()$ # Reset scene
  $\sigma_n \leftarrow \mathrm{ExplorationScale}();\ \mu_n \leftarrow \mu_\phi$ # Save collection parameters
  $\xi_n \sim \mathcal{N}(\mu_n, \sigma_n^2 I_{12})$ # Sample descriptor
  $\mathbf{c}_j \overset{\text{i.i.d.}}{\sim} p_{\xi_n}, \quad j \in \mathcal{C}$ # Sample traffic conditions
  $b_n \leftarrow |\mathcal{B}|^{-1} \sum_{i\in\mathcal{B}} V^B_\psi(o^i_0)$ # Before the first action
  while map window active **do**
    $(\mathcal{R}_B, \mathcal{R}_C) \leftarrow \mathrm{Rollout}(\pi^B_\theta, \pi^C_\omega; \xi_n)$ # Separate policy batches
    $\theta \leftarrow \mathrm{PPO}_B(\theta; \mathcal{R}_B)$ # Focal driving objective
    $\omega \leftarrow \mathrm{PPO}_C(\omega; \mathcal{R}_C)$ # Traffic driving objective
    $\mathbf{c}_j \sim p_{\xi_n}$ upon reset of traffic agent j # Keep ξ_n fixed
  end while
  $G^{\mathrm{exp}}_{B,n} \leftarrow \mathrm{FirstExposureReturn}(\mathrm{window})$ # Clip and bootstrap
  $A_{\mathrm{pop},n} \leftarrow b_n - G^{\mathrm{exp}}_{B,n}$ # Critic-centered signal
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\xi_n, \mu_n, \sigma_n, A_{\mathrm{pop},n})\}$ # One eligible window record
  if $|\mathcal{D}| \geq N$ **then**
    $\widetilde{A} \leftarrow \mathrm{Standardize}(\{A_{\mathrm{pop},n}\}_{n\in\mathcal{D}})$ # Center and scale
    $\phi \leftarrow \mathrm{PPO}_{\mathrm{pop}}(\phi; \mathcal{D}, \widetilde{A})$ # Maximize Eq. (10)
    $\mathcal{D} \leftarrow \varnothing$
  end if
end for

4 EXPERIMENTS

We structure our experiments around three questions. **(Q1)** Does adaptive population construction improve zero-shot generalization across traffic controllers that differ from the jointly trained traffic policy? **(Q2)** Are these gains attributable to adaptive population construction rather than behavioral randomization or the dual-policy training structure alone? **(Q3)** How does the learned traffic-behavior population evolve during training? Unless otherwise stated, a single focal checkpoint is evaluated against all traffic controllers, with no population constructor or behavior-condition selection used at test time.

4.1 EXPERIMENTAL SETUP

Datasets. We use maps and initial vehicle states derived from the Waymo Open Motion Dataset (Ettinger et al., 2021). Training uses the 10,000-map medium training collection provided by PufferDrive (Cornelisse et al., 2025a). Evaluation follows the two held-out 1,000-scene splits of Distelzweig et al. (2026): Interactive1k, which emphasizes interaction-heavy scenarios, and Random1k, which represents the broader validation distribution. All reported results are evaluated on the complete splits.

Benchmark. We evaluate closed-loop driving in PufferDrive (Cornelisse et al., 2025a) following BehaviorBench (Distelzweig et al., 2026). The benchmark contains eight traffic populations: IDM, PPO, SMART, expert trajectory replay, and four variants of a separately trained behavior-conditioned controller—Aggressive, Normal, Cautious, and Mix. The same traffic-controller checkpoints are used for all evaluated planners.

4.2 GENERALIZATION ACROSS TRAFFIC POPULATIONS

Table 1 evaluates one fixed focal checkpoint across eight traffic populations without test-time adaptation or population selection, covering rule-based, learned, and log-replay traffic.

	Traffic populations							
Planner	IDM	PPO	SMART	Expert	Aggr.	Normal	Caut.	Mix
				Interactive1k				
IDM	15.05	16.91	15.70	12.13	19.21	16.29	12.53	15.39
PDM	23.99	23.01	30.43	20.60	33.66	28.66	19.36	27.03
SMART	22.53	29.11	30.94	28.04	32.86	30.16	22.28	27.83
PPO	<u>38.73</u>	<u>51.54</u>	<u>41.69</u>	<u>29.40</u>	<u>49.17</u>	<u>43.32</u>	<u>39.15</u>	<u>43.44</u>
APEX (Ours)	**41.45**	**52.34**	**45.22**	**32.33**	**51.53**	**47.04**	**41.85**	**46.33**
				Random1k				
IDM	55.22	58.10	52.77	<u>50.65</u>	55.34	54.62	53.86	54.60
PDM	**63.67**	**65.18**	**60.86**	**56.67**	**64.62**	**62.61**	**60.52**	**61.89**
SMART	46.93	51.92	47.77	49.00	51.85	51.39	49.13	51.20
PPO	52.96	<u>61.21</u>	53.46	50.09	<u>59.03</u>	56.89	<u>55.48</u>	57.35
APEX (Ours)	<u>55.33</u>	60.91	<u>54.13</u>	49.85	58.97	<u>56.95</u>	55.46	<u>57.59</u>

Table 1: **Generalization across traffic populations on Interactive1k and Random1k.** Scores are scaled by 100; each entry uses the complete 1,000-scene split. Each planner uses one checkpoint across traffic populations, and PPO traffic uses the same official checkpoint for all planners. Bold and underline indicate the best and second-best results in each column and split.

Interaction-heavy generalization. On Interactive1k, APEX achieves the highest score across all eight evaluated traffic populations, increasing the mean score from 42.06 for PPO to 44.76. The gains extend beyond the conditioned traffic family, including $+2.72$ under IDM, $+3.53$ under SMART, and $+2.93$ under expert replay. APEX also improves consistently across Aggressive,

Normal, Cautious, and Mix traffic, whose mean score increases from 43.77 to 46.69. These results indicate that the gain is not restricted to a single traffic controller or behavior style.

Dependence on scene interactivity. On Random1k, APEX remains close to PPO (56.15 versus 55.81 on average), with the clearest gain under IDM (+2.37). PDM leads on this broader scene distribution, whereas APEX leads all eight Interactive1k columns. APEX's training reward does not directly optimize comfort, which contributes to the evaluation score; this objective mismatch may partly explain its lower aggregate score than PDM on Random1k. Together, the two splits suggest that APEX's main benefit lies in interaction-heavy scenarios, where variation in surrounding behavior is more consequential.

4.3 Ablation Study

We compare four training configurations to assess whether the gains of APEX persist beyond behavior conditioning and dual-policy training alone. PPO is the conventional self-play baseline; BC PPO introduces behavior conditioning; Fixed dual additionally separates focal and traffic policies but uses a fixed behavior distribution; APEX further adapts this distribution with the learned constructor. BC PPO uses the default driving condition and full configured perception at test time.

Table 2 reports collision and goal-completion rates, separating safety and progress from the benchmark's driving-quality terms.

Method	Training components			Collision ↓ / Goal completion ↑ (%)							
	π_θ^B	π_ω^C	q_ϕ	IDM	Self	SMART	Expert	Aggr.	Norm.	Caut.	Mix
Interactive1k											
PPO	✓	—	—	18.2/73.3	1.5/95.8	16.9/77.5	38.9/54.7	8.2/88.1	14.1/80.4	18.1/74.4	13.2/81.3
BC PPO	—	✓	—	11.8/71.2	1.0/94.9	12.0/78.2	33.9/54.4	5.5/89.1	13.1/78.8	14.6/72.4	12.3/79.1
Fixed dual	✓	✓	—	8.9/75.0	1.2/**96.2**	8.5/82.9	**28.5/61.4**	5.3/90.5	8.6/84.3	12.4/75.5	8.8/83.4
APEX (Ours)	✓	✓	✓	**5.8/78.9**	**0.8/96.2**	**8.1/83.5**	29.7/59.6	**3.5/92.1**	**6.5/86.8**	**9.0/78.9**	**6.4/85.0**
Random1k											
PPO	✓	—	—	9.5/85.0	1.1/96.3	8.9/86.2	16.8/79.3	3.6/92.9	6.0/90.5	7.4/88.6	5.7/91.1
BC PPO	—	✓	—	5.2/87.4	0.9/96.3	6.1/87.7	19.2/76.3	3.4/93.7	5.5/91.1	6.1/88.9	4.8/90.9
Fixed dual	✓	✓	—	3.7/88.9	0.6/97.0	4.3/89.4	15.7/80.2	2.0/**95.0**	3.8/**93.0**	4.7/90.9	3.6/93.1
APEX (Ours)	✓	✓	✓	**1.3/91.5**	**0.2/97.3**	**3.3/89.5**	**13.5/81.8**	**1.6/95.0**	**3.2**/92.9	**4.0/91.1**	**2.9/93.6**

Table 2: **Ablation of policy roles and adaptive population construction.** Each entry reports collision rate / goal-completion rate (%) on a complete 1,000-scene split. Collisions include both at-fault and other collisions. Checks indicate trained components; q_ϕ denotes a learned constructor, and BC PPO denotes behavior-conditioned PPO. Self uses each row's own evaluated policy as traffic; other traffic controllers are shared across rows. Bold and underline mark the best and second-best values separately for each metric, column, and split.

Compared with Fixed dual, APEX reduces the mean collision rate across the seven shared traffic populations excluding Self from 11.57% to 9.86% on Interactive1k and from 5.40% to 4.26% on Random1k. Goal completion also increases from 79.00% to 80.69% and from 90.07% to 90.77%, respectively. Because Fixed dual retains the focal/traffic policy separation while sampling the four view caps with fixed equal probabilities, these results provide evidence that adaptive population construction contributes beyond dual-policy training alone. APEX also preserves strong Self performance, suggesting that the cross-population safety gains do not require sacrificing standard self-play behavior.

4.4 What Does the Constructor Learn?

Figure 2 shows that APEX learns a modest but selective redistribution of training traffic. Over training, the probability of the 90° view cap increases from 26.69% to 29.50%, while that of 360°

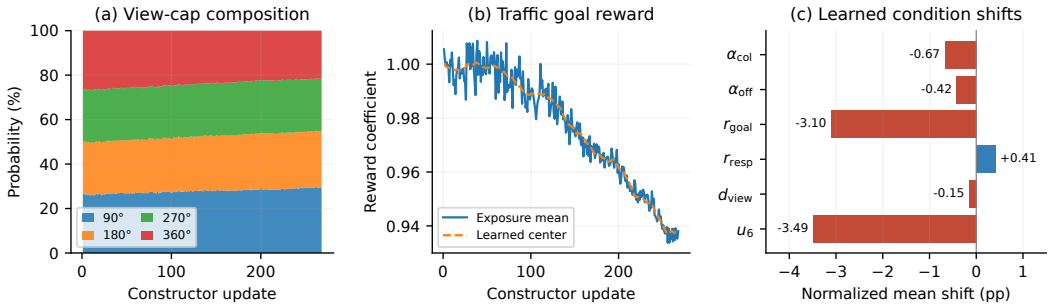


Figure 2: **Evolution of the training traffic population.** (a) View-cap probabilities obtained by averaging conditional Beta bin probabilities over recorded descriptors. Actual view angles are sampled uniformly below the selected cap. (b) Mean traffic goal-reward coefficient and the coefficient obtained by mapping the recorded Gaussian mean. Each point in (a,b) summarizes sampling-time records in one update batch from the evaluated seed-42 multi-exposure run (268 batches; 587,424 records). (c) First-to-last shifts in the six normalized condition means, with exploration fixed at $\sigma = 0.2$ and estimated using 100,000 shared Gaussian draws. The sixth coordinate u_6 selects the view cap; shifts are in percentage points of the normalized coordinate range.

decreases from 26.33% to 21.45%; the mean traffic goal-reward coefficient decreases from 1.006 to 0.938. Other condition dimensions change substantially less.

Because the exploration scale is externally annealed during training, changes in sampled exposure do not necessarily imply changes in the learned constructor. We therefore recompute the induced condition means at a common exploration scale, $\sigma = 0.2$. The largest learned shifts remain in the goal-reward and view-cap dimensions (-3.10 and -3.49 percentage points of their normalized ranges, respectively), indicating that the observed redistribution is not solely explained by exploration annealing. These results establish that the learned population evolves selectively during training, but do not by themselves establish causal tracking of focal-policy weaknesses.

5 Conclusion

We introduced APEX, a driving self-play framework that makes the surrounding traffic-behavior distribution an adaptive component of training. By separating the focal policy from behavior-conditioned traffic and learning a population constructor from critic-centered feedback, APEX adjusts behavioral exposure without directly adversarializing individual traffic agents.

Experiments on BehaviorBench show that APEX improves cross-population generalization most clearly in interaction-heavy scenarios and reduces collisions relative to fixed-sampling dual-policy training, while remaining competitive on the broader Random1k distribution and under self-play. Analysis of the learned constructor further reveals selective evolution of the training population, particularly along view-cap and goal-reward dimensions.

These results support adaptive behavioral exposure as a complementary direction to scaling self-play. Multi-seed evaluation and controlled frozen-policy comparisons of learned populations remain important for establishing the consistency of the gains and the extent to which adaptation tracks evolving focal-policy weaknesses.

AI use statement

Generative AI tools assisted with translation, language editing, manuscript organization and drafting, refinement of mathematical arguments and proofs, inspection of implementation details, and analysis and presentation of experimental results. The authors reviewed and revised the AI-assisted material and verified the mathematical claims, implementation descriptions, and reported results against the source code and experimental records. The authors retain responsibility for all methodological decisions, interpretations, and the final content of this work.

REFERENCES

Ahmed Abouelazm, Johannes Ratz, Philip Schoerner, and J. Marius Zöllner. Diverse and adaptive behavior curriculum for autonomous driving: A student-teacher framework with multi-agent rl. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12034–12041, 2025. doi: 10.1109/IROS60139.2025.11246550.

Axel Brunnbauer, Luigi Berducci, Peter Priller, Dejan Ničković, and Radu Grosu. Scenario-based curriculum generation for multi-agent driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1824–1830, 2025. doi: 10.1109/ICRA55743.2025.11128162.

Wei-Jer Chang, Akshay Rangesh, Kevin Joseph, Matthew Strong, Masayoshi Tomizuka, Yihan Hu, and Wei Zhan. SPACeR: Self-play anchoring with centralized reference models. In *International Conference on Learning Representations*, pp. 47010–47027, 2026.

Daphne Cornelisse, Spencer Cheng, Pragnay Mandavilli, Julian Hunt, Kevin Joseph, Waël Doulazmi, Aditya Gupta, and Eugene Vinitsky. PufferDrive: A fast and friendly driving simulator for training and evaluating RL agents. Software, version 2.0.0, 2025a. URL `https://github.com/Emerge-Lab/PufferDrive`.

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

Daphne Cornelisse, Julian Hunt, Zixu Zhang, Waël Doulazmi, Kevin Joseph, Jaime Fernández Fisac, and Eugene Vinitsky. Human-like autonomy emerges from self-play and a pinch of human data. *arXiv preprint arXiv:2606.19370*, 2026.

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

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13049–13061, 2020.

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

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

Jialin Fan, Ying Ni, Yuhao Yang, Wentao Zheng, Jie Sun, and Jian Sun. Learning to model diverse interactive traffic with driving tendency-guided policy optimization. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2047–2053, 2025. doi: 10.1109/IV64158.2025.11097406.

Yuan Gui, Hongchen Luo, Jiao Wang, and Liqi Qu. Beyond rule-based agents: Active Markov games for realistic multi-agent interaction in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10689–10698, 2026.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for

large-scale autonomous driving research. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7730–7742, 2023.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4940–4950, 2021.

Saman Kazemkhani, Aarav Pandya, Daphne Cornelisse, Brennan Shacklett, and Eugene Vinitsky. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *International Conference on Learning Representations*, pp. 19320–19336, 2025.

Tong Nie, Yuewen Mei, Junlin He, Yihong Tang, Jian Sun, and Wei Ma. World models as adversaries: Multi-agent self-play fine-tuning for robust motion planning. *arXiv preprint arXiv:2607.10630*, 2026.

Jinkai Qiu, Alessandro Saviolo, Chaojie Wang, Mingke Wang, and Xiaoyu Huang. Heterogeneous self-play for realistic highway traffic simulation. *arXiv preprint arXiv:2604.16406*, 2026.

Luke Rowe, Roger Girgis, Rodrigue de Schaetzen, Daphne Cornelisse, Alaap Grandhi, Felix Heide, Eugene Vinitsky, Christopher Pal, and Liam Paull. Scaling self-play for end-to-end driving. *arXiv preprint arXiv:2606.19641*, 2026.

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

Zihao Sheng, Zilin Huang, Yansong Qu, Yue Leng, Sruthi Bhavanam, and Sikai Chen. CurricuVLM: Towards safe autonomous driving via personalized safety-critical curriculum learning with vision-language models. *Transportation Research Part C: Emerging Technologies*, 185:105549, 2026. doi: 10.1016/j.trc.2026.105549.

Eugene Vinitsky, Nathan Lichtlé, Xiaomeng Yang, Brandon Amos, and Jakob Foerster. Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3962–3974, 2022.

Zhouchonghao Wu, Akshay Rangesh, Weixin Li, Wei-Jer Chang, Zachary Lee, Saeed Bonab, Tim Wang, and Wei Zhan. TerraZero: Procedural driving simulation for zero-demonstration self-play at scale. *arXiv preprint arXiv:2607.13028*, 2026.

Chris Zhang, Sourav Biswas, Kelvin Wong, Kion Fallah, Lunjun Zhang, Dian Chen, Sergio Casas, and Raquel Urtasun. Learning to drive via asymmetric self-play. In *European Conference on Computer Vision*, pp. 149–168, 2024. doi: 10.1007/978-3-031-73033-7_9.

A Theoretical details

Throughout this appendix, (θ, ω, ψ) are fixed during a population-constructor update.

A.1 Hierarchical population distribution

Proof of Eq. (2). Conditional on a shared descriptor $\xi \sim q_\phi$, the traffic conditions $\{\mathbf{c}_j\}_{j\in\mathcal{C}}$ are independent draws from p_ξ. Hence $p(\mathbf{c} \mid \xi) = \prod_{j\in\mathcal{C}} p_\xi(\mathbf{c}_j)$ and $p_\phi(\mathbf{c}, \xi) = q_\phi(\xi)p(\mathbf{c} \mid \xi)$. Marginalizing the descriptor gives

$$p_\phi(\mathbf{c}) = \int p_\phi(\mathbf{c}, \xi)\, d\xi = \int q_\phi(\xi) \prod_{j\in\mathcal{C}} p_\xi(\mathbf{c}_j)\, d\xi,$$

which proves Eq. (2).

Remark (population-level dependence). The conditions are independent only conditional on ξ. After marginalizing the shared descriptor, distinct agents' conditions are generally dependent: $p_\phi(\mathbf{c}_j, \mathbf{c}_k) = \int q_\phi(\xi)p_\xi(\mathbf{c}_j)p_\xi(\mathbf{c}_k)\, d\xi$ need not equal $p_\phi(\mathbf{c}_j)p_\phi(\mathbf{c}_k)$. The shared descriptor can therefore correlate behavior conditions across agents, while conditional sampling retains diversity within each sampled population.

A.2 Proof of Proposition 1

Throughout this subsection, the driving policies and critic are fixed. Write $J_B^{\exp}(\phi) = J_B^{\exp}(\theta, \omega, \phi)$, and let $\mathbf{x} \sim \rho_0$, $\xi \sim q_\phi(\cdot; \sigma)$, and τ denote the reset scene, descriptor, and subsequent first-exposure rollout. The external scale σ is fixed for this analysis. Conditional on $(\mathbf{x}, \xi)$, condition sampling and the rollout law have no further dependence on ϕ. Thus $J_B^{\exp}(\phi) = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\xi}\mathbb{E}_{\tau|\mathbf{x},\xi}[G_B^{\exp}(\tau)]$. Under the usual differentiability and integrability conditions for exchanging differentiation and expectation, applying the score-function identity to q_ϕ gives

$$\nabla_\phi J_B^{\exp}(\phi) = \mathbb{E}[G_B^{\exp} \nabla_\phi \log q_\phi(\xi)]\,. \tag{11}$$

The scale argument is suppressed below. Conditional baseline independence yields

$$\begin{aligned}
&\mathbb{E}_{\xi\sim q_\phi}[b_\psi(\mathbf{x})\nabla_\phi \log q_\phi(\xi) \mid \mathbf{x}] \\
&\quad = b_\psi(\mathbf{x}) \int q_\phi(\xi)\nabla_\phi \log q_\phi(\xi)\, d\xi \\
&\quad = b_\psi(\mathbf{x})\nabla_\phi \int q_\phi(\xi)\, d\xi = 0.
\end{aligned} \tag{12}$$

Averaging over reset scenes establishes the first identity in Eq. (8). Substituting $A_{\text{pop}} = b_\psi(\mathbf{x}) - G_B^{\exp}$ then gives

$$\begin{aligned}
\mathbb{E}[A_{\text{pop}}\nabla_\phi \log q_\phi(\xi)] &= \mathbb{E}[b_\psi(\mathbf{x})\nabla_\phi \log q_\phi(\xi)] \\
&\quad - \mathbb{E}[G_B^{\exp}\nabla_\phi \log q_\phi(\xi)] \\
&= -\nabla_\phi J_B^{\exp}(\phi).
\end{aligned} \tag{13}$$

Equivalently, the reset distribution is independent of ϕ, so

$$\mathbb{E}[A_{\text{pop}}] = \mathbb{E}_{\mathbf{x}}[b_\psi(\mathbf{x})] - J_B^{\exp}(\phi). \tag{14}$$

This supplies an objective-level interpretation of the same result. The identity concerns the idealized unclipped expectation with fixed policies and critic. Finite-batch standardization, PPO clipping, records from different policy snapshots, and policy drift during online collection are not part of this exact identity. Nor does it assert equality of $J_B^{\exp}$ and J_B or accuracy of the continuation estimate.

A.3 Variance-optimal descriptor-independent baseline

Proposition 1 establishes invariance for any admissible baseline independent of the descriptor conditional on the reset scene; it does not imply that the focal critic is variance-optimal. Let $s_\phi(\xi) = \nabla_\phi \log q_\phi(\xi)$ be the descriptor score. For a scalar baseline $b(\mathbf{x})$, consider the conditional second moment

$$\mathcal{V}(b \mid \mathbf{x}) = \mathbb{E}\left[\|s_\phi(\xi)\|_2^2 (G_B^{\exp} - b(\mathbf{x}))^2 \mid \mathbf{x}\right]. \tag{15}$$

Assume the required second moments are finite. Since the conditional estimator mean is invariant to b, minimizing this second moment also minimizes the trace of its conditional covariance. Differentiation with respect to $b(\mathbf{x})$ gives $\partial_b \mathcal{V} = -2\mathbb{E}[\|s_\phi(\xi)\|_2^2 (G_B^{\exp} - b(\mathbf{x})) \mid \mathbf{x}]$. When $\mathbb{E}[\|s_\phi(\xi)\|_2^2 \mid \mathbf{x}] > 0$, the unique minimizer is

$$b^\star(\mathbf{x}) = \frac{\mathbb{E}[\|s_\phi(\xi)\|_2^2 G_B^{\exp} \mid \mathbf{x}]}{\mathbb{E}[\|s_\phi(\xi)\|_2^2 \mid \mathbf{x}]}. \tag{16}$$

If the denominator is zero, the score vanishes almost surely conditional on $\mathbf{x}$, and the baseline choice is immaterial. Completing the square gives

$$\mathcal{V}(b \mid \mathbf{x}) = \mathcal{V}(b^\star \mid \mathbf{x}) + \mathbb{E}[\|s_\phi(\xi)\|_2^2 \mid \mathbf{x}](b(\mathbf{x}) - b^\star(\mathbf{x}))^2. \tag{17}$$

If the squared score magnitude and the return target are conditionally independent given $\mathbf{x}$, Eq. (16) reduces to $b^\star(\mathbf{x}) = \mathbb{E}[G_B^{\exp} \mid \mathbf{x}]$. The focal critic can approximate this conditional mean, but calibration alone does not establish optimality: in general, the optimal baseline weights returns by the squared descriptor-score magnitude.

A.4 INDEPENDENCE UNDER THE FIRST-EXPOSURE PROTOCOL

At the beginning of a map window, the simulator resets the map and vehicle states, producing $\mathbf{x}$. It then samples ξ and assigns traffic conditions from p_ξ. Focal agents receive the same fixed default condition for every descriptor. Before any vehicle action is applied, the focal critic evaluates the reset observations; these values are recorded when the first environment step begins. The focal-policy encoder does not use the appended condition entries.

Conditional on the reset scene and fixed policy and critic parameters, the initial focal observations and $b_\psi(\mathbf{x})$ are therefore unchanged by ξ. Evaluating the critic after descriptor sampling does not alter this independence, because no descriptor-dependent traffic action has yet affected the scene. The focal set in the baseline must also be fixed at reset: selecting its members using descriptor-dependent outcomes could violate independence. The first-exposure target uses these agents' completed first episodes or time-limit truncations.

The exact identity applies before finite-batch standardization and PPO clipping. Online collection may span policy updates between records or within a focal episode. Such records approximate the fixed-policy objective rather than forming an exact gradient estimator for a single policy snapshot. Restricting the target to first episodes does not itself freeze policy parameters.

For later episodes under the same descriptor, focal observations may already depend on preceding traffic actions and intervening policy updates. A baseline computed from those observations need not remain descriptor-independent. The argument therefore applies to first exposure, not automatically to the optional multi-episode window score.

A.5 INTERPRETATION OF CRITIC CENTERING

Proposition 1 establishes an optimization property. To clarify what centering removes from the sampled feedback, write $F(\mathbf{x}, \xi) = \mathbb{E}[G_B^{\mathrm{exp}} \mid \mathbf{x}, \xi]$ and let q_{ref} be a reference population distribution. Define the scene-dependent critic error $e_\psi(\mathbf{x})$ relative to its expected focal return by $b_\psi(\mathbf{x}) = \mathbb{E}_{\xi' \sim q_{\mathrm{ref}}} F(\mathbf{x}, \xi') + e_\psi(\mathbf{x})$. Then

$$\mathbb{E}[A_{\mathrm{pop}} \mid \mathbf{x}, \xi] = \mathbb{E}_{\xi' \sim q_{\mathrm{ref}}} F(\mathbf{x}, \xi') - F(\mathbf{x}, \xi) + e_\psi(\mathbf{x}). \tag{18}$$

If $F(\mathbf{x}, \xi) = d(\mathbf{x}) + \widetilde{F}(\mathbf{x}, \xi)$, where $d(\mathbf{x})$ is an additive scene-difficulty component shared by all descriptors, substitution gives

$$\mathbb{E}[A_{\mathrm{pop}} \mid \mathbf{x}, \xi] = \mathbb{E}_{\xi' \sim q_{\mathrm{ref}}} \widetilde{F}(\mathbf{x}, \xi') - \widetilde{F}(\mathbf{x}, \xi) + e_\psi(\mathbf{x}). \tag{19}$$

The additive component cancels exactly, but descriptor-dependent interactions and critic error remain. The focal critic is trained on the evolving policy's own data distribution, so its interpretation as the expected return under a fixed reference population is approximate. Neither this decomposition nor Proposition 1 identifies a population's causal effect from a single rollout. Centering supplies a scene-relative learning signal while preserving the expected constructor-gradient direction.

A.6 CONDITIONAL LOCAL POPULATION TRACKING UNDER POLICY DRIFT

We characterize policy drift across a constructor batch, separating stale policy effects from optimization and coverage errors. This is a conditional analysis, not a convergence guarantee for practical PPO. Let $u = (\theta, \omega)$ and $F(u, q) = \mathbb{E}[G_B^{\mathrm{exp}} \mid u, q]$ be expected first-exposure return under a descriptor distribution q. The critic and return convention are fixed in this analysis, and exploration-scale dependence is suppressed.

For a batch $\mathcal{D} = \{1, \ldots, M\}$, let u_n denote the driving-policy pair associated with record n and u_{cur} the pair at update time. For a metric d on policy pairs, define

$$\bar{F}_{\mathcal{D}}(q) = \frac{1}{M} \sum_{n=1}^{M} F(u_n, q), \qquad \Delta_{\mathcal{D}} = \max_{n \in \mathcal{D}} d(u_n, u_{\mathrm{cur}}). \tag{20}$$

Let $\mathcal{Q}_b \subseteq \mathcal{Q}^\star$ be the representable and larger admissible population classes. Define

$$g^\star(u) = \inf_{q \in \mathcal{Q}^\star} F(u, q), \qquad \delta_{\mathrm{cov}}(u) = \inf_{q \in \mathcal{Q}_b} F(u, q) - g^\star(u) \geq 0. \tag{21}$$

Assume these infima are finite.

Proposition 2 (local population tracking). Assume $F(\cdot, q)$ is L-Lipschitz under d, uniformly for $q \in \mathcal{Q}_b$. Suppose an update returns $q^+ \in \mathcal{Q}_b$ satisfying

$$\bar{F}_{\mathcal{D}}(q^+) \le \inf_{q \in \mathcal{Q}_b} \bar{F}_{\mathcal{D}}(q) + \epsilon_{\mathrm{opt}}. \tag{22}$$

Then

$$0 \le F(u_{\mathrm{cur}}, q^+) - g^\star(u_{\mathrm{cur}}) \le \epsilon_{\mathrm{opt}} + 2L\Delta_{\mathcal{D}} + \delta_{\mathrm{cov}}(u_{\mathrm{cur}}). \tag{23}$$

Proof. Write $g_b(u) = \inf_{q \in \mathcal{Q}_b} F(u, q)$. For every $q \in \mathcal{Q}_b$, uniform Lipschitz continuity implies

$$|F(u_{\mathrm{cur}}, q) - \bar{F}_{\mathcal{D}}(q)| \le \frac{L}{M} \sum_{n=1}^{M} d(u_n, u_{\mathrm{cur}}) \le L\Delta_{\mathcal{D}}. \tag{24}$$

Apply this inequality at q^+ and when comparing the two infima:

$$\begin{aligned} F(u_{\mathrm{cur}}, q^+) - g_b(u_{\mathrm{cur}}) &\le \bar{F}_{\mathcal{D}}(q^+) - \inf_{q \in \mathcal{Q}_b} \bar{F}_{\mathcal{D}}(q) + 2L\Delta_{\mathcal{D}} \\ &\le \epsilon_{\mathrm{opt}} + 2L\Delta_{\mathcal{D}}. \end{aligned}$$

Adding $g_b(u_{\mathrm{cur}}) - g^\star(u_{\mathrm{cur}}) = \delta_{\mathrm{cov}}(u_{\mathrm{cur}})$ proves Eq. (23).

The bound separates optimization error, driving-policy drift, and population coverage. Its approximate-minimization premise is an explicit assumption: optimizing Eq. (10) with sampled returns, centering, standardization, clipping, and mixed collection versions does not establish it. The snapshot model also abstracts from within-episode policy changes and critic drift; these effects are not bounded by $\Delta_{\mathcal{D}}$ alone.

A.7 Local direction of the clipped population surrogate

For record n, use $q_n(\xi) = \mathcal{N}(\xi; \mu_n, \sigma_n^2 I_{12})$ and ρ_n from Eq. (9). Suppose all collection means equal μ_{ϕ_k}, allowing different σ_n. Then $q_n(\xi) = q_{\phi_k}(\xi; \sigma_n)$ and $\rho_n(\phi_k) = 1$ for every record. With standardized advantages held fixed, clipping is locally inactive, giving

$$\nabla_\phi \mathcal{L}_{\mathrm{pop}}^{\mathrm{clip}}(\phi)\big|_{\phi=\phi_k} = \mathbb{E}_{n \sim \mathcal{D}}\left[\widetilde{A}_{\mathrm{pop},n}\ \nabla_\phi \log q_\phi(\xi_n; \sigma_n)\big|_{\phi=\phi_k}\right]. \tag{25}$$

A common exploration scale is unnecessary: each numerator uses its record's collection scale. With raw advantages and fixed driving policies and critic, Proposition 1 applies at each fixed scale. If the scales differ, averaging these expected estimators yields the negative gradient of the corresponding scale-averaged first-exposure objective, $M^{-1} \sum_n J_B^{\mathrm{exp}}(\phi; \sigma_n)$, where the scale dependence is made explicit. At a common scale this reduces to the original objective. This expectation statement treats the scales as fixed; it is not an assertion of conditional unbiasedness under arbitrary adaptive collection or stopping.

The implemented batch standardization changes raw advantages, and records may come from different driving-policy or constructor versions. When the collection means differ, there need not be one ϕ_k with $\rho_n(\phi_k) = 1$ for all records. Ratios remain well defined, but the common-parameter local identity need not hold. Equation (25) is therefore a local reference case, not an identity for every practical constructor step.

B Implementation details

Experimental training protocol. APEX trains a focal and a behavior-conditioned traffic policy with a 50:50 stratified allocation of agents. Both policies are recurrent and are trained from scratch on one GPU with seed 42. The total budget is four billion agent steps; the single-policy PPO and behavior-conditioned PPO comparisons use two billion steps. This matches the nominal focal allocation, while the dual-policy runs incur the additional cost of traffic-policy learning. We use 16 vectorized workers, PPO discount $\gamma = 0.98$, and GAE parameter $\lambda = 0.95$. The population constructor operates over six behavior conditions. The selected APEX configuration uses a continuous perception radius in $[20, 50]$ metres and view-angle caps of $90°$, $180°$, $270°$, and $360°$, with the actual angle sampled below the selected cap.

The reported APEX checkpoint uses the temporally discounted multi-exposure extension: later episode signals within a descriptor's map window are included with decay coefficient 0.8. This distinguishes the evaluated online protocol from the first-exposure special case analyzed in Proposition 1. The window lasts 910 simulator steps; its descriptor stays fixed while individual episodes and policy updates proceed. Return, normalization, and likelihood conventions are specified below.

Training and evaluation use collision shrink factors of 0.7 and 1.0, respectively. The fixed-cap dual-policy control independently samples the four view caps with equal probabilities. The single-policy conditioned baseline is evaluated with a fixed default driving condition and full configured perception. All trained baselines use our recorded training setups. The reported APEX checkpoint is the view-cap variant selected after comparison with the radius-and-view-cap variant.

Condition distribution. Write $\xi = (z_1, \dots, z_6, h_1, \dots, h_6)$ and let $s(a) = (1+e^{-a})^{-1}$. For each coordinate ℓ, the Gaussian descriptor determines Beta mean and dispersion parameters by

$$\begin{aligned} m_\ell &= \epsilon + (1-2\epsilon)s(z_\ell), \\ \varrho_\ell &= \varrho_{\min} + (\varrho_{\max} - \varrho_{\min})s(h_\ell), \\ \big(a_\ell(z_\ell, h_\ell), b_\ell(z_\ell, h_\ell)\big) &= \big(m_\ell, 1-m_\ell\big)(\varrho_\ell^{-1} - 1). \end{aligned} \tag{26}$$

The implementation uses $\epsilon = 0.02$ and $(\varrho_{\min}, \varrho_{\max}) = (0.02, 0.5)$. Initializing the descriptor mean at $m_\ell = 1/2$ and $\varrho_\ell = 1/3$ gives uniform Beta laws. The sampling kernel $\mathcal{T}$ in Eq. (5) combines deterministic coordinate mappings with independent view-angle sampling. In the selected view-cap configuration, the first five normalized draws map respectively to $[0, 3]$, $[0, 3]$, $[0, 2]$, $[0, 0.5]$, and $[20, 50]$ metres. The sixth draw selects one of four equal-width bins, giving a view cap of $90°$, $180°$, $270°$, or $360°$; an independent uniform draw sets the actual view angle between zero and that cap. Focal agents instead receive the fixed condition $(0.5, 0.5, 1, 0.25, 50, 360)$; their policy does not use the condition entries.

First-exposure constructor target. The independence argument in Proposition 1 applies to the `first_episode` reward mode. For each eligible focal agent i at a window reset, let T_i be its first episode length and let τ_i indicate a time-limit truncation rather than a true terminal. The focal set $\mathcal{B}$ is fixed at that reset. For the constructor objective and its gradient, the critic parameters ψ are held fixed; their dependence is suppressed in $J_B^{\exp}(\theta, \omega, \phi)$. The realized target is

$$G_B^{\exp} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\sum_{t=0}^{T_i - 1} \gamma^t \operatorname{clip}(r_t^i, -1, 1) + \mathbb{1}_{\{\tau_i\}} \gamma^{T_i} V_\psi^B(o_{T_i}^i) \right]. \tag{27}$$

This is the realized first-exposure return used in $A_{\text{pop}} = b_\psi(\mathbf{x}) - G_B^{\exp}$ and in Proposition 1. The critic parameters ψ are held fixed when analyzing the constructor objective; the bootstrap term specifies the return estimator and introduces no additional ϕ-dependence conditional on the reset scene and sampled descriptor. The environment accumulates the clipped discounted rewards and later subtracts the time-limit continuation value from $b_\psi - G_B^{\exp}$ once the next-state value is available. True terminals receive no continuation term. The simulator's default `legacy_window` mode also includes later episodes with a time-decay weight; its baseline need not be independent of the descriptor.

The PPO policy advantage is a different estimator. With matching discount, rewards, and truncation convention, $\widehat{A}_{0,1}^{\text{GAE}}$ telescopes to $G_B^{\exp} - V_\psi^B(o_0)$ for one episode. For $\lambda < 1$, intermediate value predictions remain in the GAE sum, so substituting $-\widehat{A}_{0,\lambda}^{\text{GAE}}$ changes the constructor target and does not inherit Proposition 1 automatically. In particular, $G_B^{\exp}$ is observable from the sampled rollout even though the expected return $J_B^{\exp}$ is not.

Constructor record and likelihood ratio. Each eligible completed window supplies $\mathcal{R}_n = (\xi_n, \mu_n, \sigma_n, A_{\text{pop},n})$, where $A_{\text{pop},n} = b_n - G_{B,n}^{\exp}$ and the remaining entries are the descriptor and its collection mean and scale. The collection law is $q_n(\xi) = \mathcal{N}(\xi; \mu_n, \sigma_n^2 I_{12})$. A candidate changes only the learned mean, retaining the recorded covariance as in Eq. (9). Thus its log-ratio is

$$\log \rho_n(\phi) = -\frac{\|\xi_n - \mu_\phi\|_2^2 - \|\xi_n - \mu_n\|_2^2}{2\sigma_n^2}. \tag{28}$$

The externally scheduled exploration scale receives no constructor gradient. Records can retain different means and scales because windows finish at different times. Only windows with both focal and conditioned agents and a valid first-exposure signal contribute a constructor record.

Population-advantage normalization. For a batch $\mathcal{D}$ of size M, define

$$\bar{A}_{\mathrm{pop}} = \frac{1}{M}\sum_{n\in\mathcal{D}} A_{\mathrm{pop},n}, \qquad s_A^2 = \frac{1}{M}\sum_{n\in\mathcal{D}} (A_{\mathrm{pop},n} - \bar{A}_{\mathrm{pop}})^2. \tag{29}$$

The Gaussian implementation uses the population standard deviation (denominator M), with a small-scale threshold $\eta_A = 10^{-8}$:

$$\widetilde{A}_{\mathrm{pop},n} = \begin{cases} (A_{\mathrm{pop},n} - \bar{A}_{\mathrm{pop}})/s_A, & M > 1 \text{ and } s_A > \eta_A, \\ A_{\mathrm{pop},n} - \bar{A}_{\mathrm{pop}}, & \text{otherwise.} \end{cases} \tag{30}$$

At zero variance the centered signal is zero. This finite-batch transformation is not part of the exact identity in Proposition 1. The update consumes the complete accumulated batch. Across each optimization epoch, chunk gradients are weighted by their fractions of that batch before one optimizer step; chunking therefore does not introduce intermediate mean updates within the epoch.

Alternative constructor parameterization. The anchored categorical-mixture variant samples a component m_n from weights w_n and then draws the descriptor around that component's fixed Gaussian anchor. Holding the anchors and the record's Gaussian scale fixed, the conditional Gaussian likelihood cancels in the joint-action ratio. With candidate weights w_ϕ, the PPO ratio becomes

$$\rho_n^{\mathrm{mix}}(\phi) = \frac{w_\phi(m_n)}{w_n(m_n)}. \tag{31}$$

This alternative changes component weights rather than the Gaussian mean of the main constructor.