UTILITY-GUIDED RESIDUAL MULTI-AGENT REINFORCEMENT LEARNING FOR CLOSED-LOOP INTERACTIVE DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Closed-loop multi-agent driving without strong map structure (e.g., no lane markings or discrete lane IDs, and no fixed leader–follower roles) remains challenging for both classical motion planners and learning-based policies. Geometric interaction methods can under-specify objectives such as progress and comfort, while from-scratch Reinforcement Learning (RL) policies must discover useful behavior over large action spaces from limited safety-critical interaction data. We propose utility-guided residual multi-agent RL, where a discrete control scorer acts as an interpretable prior and a shared residual policy re-ranks that prior online. Every executed command remains an $\arg\max$ over the same discrete bicycle-kinematic candidate set, preserving the structured decision space and kinematic feasibility of the prior. The prior is a multi-term utility over one-step rollouts. We calibrate the prior independently on three naturalistic geometries (e.g., a weak-lane-discipline freeway, a no-lane-discipline urban roundabout, and an urban grid network) and evaluate closed-loop transfer across all three. Training follows a CaRL-style progress reward with multiplicative soft constraints and terminal contact; learned checkpoints use a common safety-first, then NAVSIM-style PDM selection rule, while deterministic inference applies a PDM-Closed accept-if-better gate. We benchmark against classical motion planners, geometric interaction methods, offline traffic-generation approaches, and matched online RL baselines under a common closed-loop evaluation protocol, and ablate residual parameterizations, the accept-if-better gate, forward collision lookahead, and traffic density. Under a matched low-budget training protocol on the freeway benchmark, the proposed residual achieves a higher PDM score and fewer collision events per episode than MAPPO and matched direct discrete RL. Ablations suggest that learned residual adaptation, not the accept-if-better gate, drives the gains; candidate-level residuals remain robust without collision lookahead, while the parameterized residual is stronger under dense interaction. These results support residual adaptation of a calibrated discrete utility prior as a sample-efficient approach to closed-loop interaction while retaining interpretable, kinematically feasible decisions.

1 INTRODUCTION

Interactive closed-loop driving without lane markings or discrete lane graphs challenges standard autonomy stacks. Map-centric planners and predictors lean on centerlines, lane IDs, and leader–follower roles that can be weakly defined when agents share continuous space inside a road boundary (Chandra & Kumar, 2004; Khan & Maini, 1999; Asaithambi & Shravani, 2017). Geometric multi-agent models (e.g., ORCA) handle pairwise feasibility but under-specify progress and human-like comfort. Unconstrained Multi-Agent RL (MARL) faces large action spaces, sparse safety rewards, and no structured prior. We therefore study closed-loop multi-agent motion generation where each agent selects bicycle controls every Δt from local observations to progress toward a goal while avoiding contact and leaving the road.

We evaluate residual learning and the paper-facing closed-loop benchmarks on a dense interactive freeway corridor reconstructed from high-resolution aerial trajectories near Jounieh, Lebanon Khoueiry et al. (2026). The same twelve-parameter prior family is also calibrated on a multi-arm

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.

roundabout and on the TGSIM Foggy Bottom urban grid (Appendix F); matched closed-loop runs on those two sites are reported in Appendix H.

We propose a utility-guided residual MARL architecture that keeps every executed move inside an interpretable discrete scorer under bicycle kinematics:

- **Discrete prior:** an additive control score U with directional, speed, proximity, collision, and path terms, an engineering trajectory/control cost.
- **Local observations:** each agent sees corridor/Frenet features and a fixed budget of nearest neighbors.
- **Residual policy:** a shared actor outputs a bounded additive residual on the discrete utility grid, so $\mathrm{score}(a) = U(a) + \alpha \tanh(\pi_\phi(o_i)[a])$.
- **Scored selection:** $a_i^\star = \arg\max_{a \in \mathcal{G}} \mathrm{score}(a)$ with shared road and car-to-car filters; zero residual recovers the frozen prior.

Related work. Our approach builds on structured motion planning van den Berg et al. (2011); Dauner et al. (2023), residual policy learning Johannink et al. (2019); Zhang et al. (2022); Sheng et al. (2024); Wang et al. (2025), and closed-loop multi-agent behavior generation Suo et al. (2021); Zhang et al. (2023a); Zhong et al. (2023); Rowe et al. (2024). Unlike residual controllers that directly modify control commands, we learn bounded corrections to a calibrated utility scorer over a fixed bicycle-kinematic action set. Further discussion is provided in Appendix A.

Contributions. (i) Fit a twelve-parameter discrete bicycle prior (with vehicle-scale collision kernels) on three naturalistic geometries. (ii) Learn a candidate-logit residual over the same bicycle grid rather than raw accelerations, and compare a nine-dimensional $\Delta\Theta$ residual with weights-only and σ-only masks. (iii) Align training with closed-loop practice: CaRL-style progress reward, NAVSIM-style PDM selection, and PDM-Closed accept-if-better at inference. (iv) Benchmark residual MARL against classical planners, matched online RL, offline generators, and the frozen prior; ablate the gate, lookahead, and density; report roundabout and Foggy Bottom transfer in the appendix.

2 METHODOLOGY

This work proposes a utility-guided residual MARL framework when discrete lane topology is unused. Unlike open-loop trajectory forecasting that imitates map-conditioned futures, and unlike planners that require a full lane graph, agents select controls online in continuous two-dimensional space through decentralized scoring of bicycle candidates plus residual re-ranking of that discrete prior. Rather than learning low-level accelerations from scratch, the RL layer learns residual corrections over the structured scorer. This preserves kinematic feasibility and interpretability while allowing context-dependent interaction strategies under crowding.

2.1 PROBLEM FORMULATION

Traffic is modeled as a decentralized multi-agent system operating in continuous two-dimensional space. Let $\mathcal{A} = \{1, 2, \ldots, N\}$ denote the set of traffic agents. Each agent $i \in \mathcal{A}$ is represented by a continuous kinematic state

$$s_i^t = \left[x_i^t, y_i^t, v_i^t, \theta_i^t\right], \tag{1}$$

where (x_i^t, y_i^t) denotes spatial position, $v_i^t \geq 0$ denotes speed magnitude, and θ_i^t denotes heading angle. Each agent observes a local neighborhood of at most K nearest neighbors within a perception radius R_p:

$$\mathcal{N}_i^t = \left\{ j \in \mathcal{A} \setminus \{i\} \;\middle|\; \|r_j^t - r_i^t\|_2 \leq R_p \right\}_{1:K}, \tag{2}$$

where $r_i^t = (x_i^t, y_i^t)$ and the subset retains the K closest neighbors by Euclidean distance. In all closed-loop experiments we set $K = 6$ and $R_p = 60\,\mathrm{m}$.

The objective of each agent is to generate feasible closed-loop motion that progresses along the corridor, avoids collisions, remains within drivable boundaries, and responds to neighboring agents.

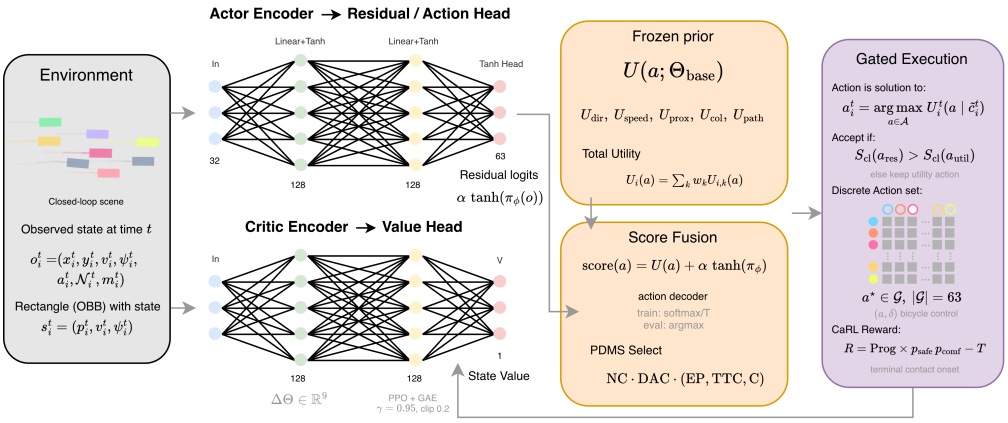

Figure 1: Overview of utility-guided residual MARL.

2.2 Discrete Control Prior

The nominal motion of each agent is a utility-maximization process over a discrete control-candidate set, i.e., a scored one-step planner. Throughout, $\mathcal{A}$ denotes the set of agents and $\mathcal{G}$ the discrete bicycle control grid (never overloaded for both roles). At each time step, agent i evaluates candidates generated from a finite grid of longitudinal acceleration and steering values, $\mathcal{G} = \{(a, \delta)\} = \mathcal{G}_{\text{acc}} \times \mathcal{G}_{\text{steer}}$, and selects the candidate that maximizes the discrete score after one-step kinematic bicycle rollout:

$$(a_i^*, \delta_i^*) = \arg \max_{(a,\delta) \in \mathcal{G}} U_i(s_i'(a, \delta); \Theta_i)\,, \tag{3}$$

where $s_i'(a, \delta)$ is the predicted next state obtained from the bicycle dynamics and Θ_i denotes the agent's discrete prior parameters. The total utility consists of five components:

$$U_i = U_{\text{dir}} + U_{\text{speed}} + U_{\text{dist}} - U_{\text{col}} - U_{\text{path}}. \tag{4}$$

The forms of U_{speed} and U_{col} are site-invariant. Direction, proximity, and path terms are evaluated in a site spatial frame: the main-text residual experiments use the freeway corridor/Frenet instantiation below, while the roundabout and Foggy Bottom calibrations replace that frame with a destination-plus-curb encoding (Appendix G).

2.2.1 Directional Utility

Directional utility encourages motion aligned with the local corridor tangent rather than an unconstrained destination vector. Let $\varepsilon_m = 10^{-6}$ m and define

$$\hat{m}_i = \begin{cases} \dfrac{r_i' - r_i}{\|r_i' - r_i\|_2}, & \|r_i' - r_i\|_2 \geq \varepsilon_m, \\ \hat{t}_i, & \|r_i' - r_i\|_2 < \varepsilon_m, \end{cases} \tag{5}$$

where $\hat{t}_i$ is the unit tangent of the corridor centerline at the ego agent's current projection. The second branch is the implemented stationary-candidate convention: a candidate with numerically zero displacement inherits the route direction and never divides by zero. In the destination-plus-curb instantiation, the analogous zero-displacement branch uses the current heading unit vector against the on-road direction to the destination. Then

$$U_{\text{dir}} = S_\theta\, \hat{m}_i^\top \hat{t}_i, \tag{6}$$

where $S_\theta > 0$ controls directional sensitivity.

2.2.2 Speed Incentive

Speed incentive rewards candidate rollout speeds that are close to a desired speed $v_{\text{des},i}$. For a candidate speed v_i', define the symmetric closeness ratio

$$\rho_i = \frac{\min(v_i', v_{\text{des},i})}{\max(v_i', v_{\text{des},i})} \in [0, 1]. \tag{7}$$

The speed utility is then

$$U_{\text{speed}} = S_v \, \frac{\rho_i}{1 + \rho_i^{(\xi_i - 1)/2}}, \tag{8}$$

where $S_v > 0$ scales the term and $\xi_i \in [1.1, 5]$ is a curvature parameter. Within this admissible interval the maximum occurs at $v_i' = v_{\text{des},i}$, and overspeeding and underspeeding are penalized symmetrically. Smaller ξ_i produces the stronger mismatch penalty; increasing ξ_i flattens the utility near $\rho_i = 1$.

2.2.3 Proximity Incentive

Proximity utility encourages progress toward a local reference point expressed in corridor Frenet coordinates. Let (s_i', n_i') be the longitudinal and lateral projections of the candidate next position onto the corridor, and let $(s_{\text{ref}}, n_{\text{ref}})$ be the corresponding projection of a local reference location (e.g., a short-horizon target state). The anisotropic effective distance is

$$d_{\text{eff}} = w_x |s_i' - s_{\text{ref}}| + w_y |n_i' - n_{\text{ref}}|. \tag{9}$$

A speed-dependent perception horizon, $H_p = \max(\kappa v_i,\, H_{\min})$, normalizes this distance, yielding

$$U_{\text{dist}} = \frac{S_d}{1 + (d_{\text{eff}}/H_p)^{\gamma}}, \tag{10}$$

where $S_d > 0$ caps the reward magnitude and $\gamma > 0$ controls how quickly the incentive decays for distant candidates.

2.2.4 Collision-Risk Utility

In weakly structured environments, collision risk cannot be reduced to hard geometric overlaps alone. Instead, collision risk is modeled by comparing the candidate next pose of the ego agent with short-horizon predictions of neighbors, expressed in the ego body frame and measured relative to vehicle footprints (length L, width W; defaults $4.5\,\text{m} \times 1.8\,\text{m}$).

For neighbor j, let $v_j \geq 0$ and θ_j denote its scalar speed and heading, and let $\mathbf{v}_j = v_j[\cos\theta_j, \sin\theta_j]^\top \in \mathbb{R}^2$ denote its velocity vector. The predicted future position after the candidate arrival time Δt is obtained by constant-velocity extrapolation,

$$\boldsymbol{\mu}_j = \mathbf{r}_j + \mathbf{v}_j \Delta t = \mathbf{r}_j + v_j[\cos\theta_j, \sin\theta_j]^\top \Delta t. \tag{11}$$

Let $\Delta \mathbf{r}_{i'j} = R(\theta_i')^\top (\mathbf{r}_i' - \boldsymbol{\mu}_j)$ be this relative vector in the candidate heading frame ($+x$ forward). Surface gaps beyond the same-size footprints are

$$g_\parallel = \max\big(0,\, |\Delta r_{i'j}^{(x)}| - L\big), \qquad g_\perp = \max\big(0,\, |\Delta r_{i'j}^{(y)}| - W\big), \tag{12}$$

so overlapping oriented boxes yield zero gap. Soft presence uses an anisotropic Gaussian on these gaps,

$$P_{j,i'} = \exp\left(-\frac{1}{2}\left[\left(\frac{g_\parallel}{\sigma_\parallel}\right)^2 + \left(\frac{g_\perp}{\sigma_\perp}\right)^2\right]\right), \tag{13}$$

with calibrated (and residual-adapted) scales $\sigma_\parallel, \sigma_\perp$. Aggregating over neighbors and capping cumulative risk yields

$$U_{\text{col}} = w_c \, \min\left(\sum_{j \in \mathcal{N}_i} P_{j,i'},\, 1\right), \tag{14}$$

where $w_c > 0$ controls collision aversion. Candidate selection additionally rejects bicycle rollouts whose predicted Oriented Bounding Box (OBB) overlaps a neighbor when a collision-free alternative exists, and speed alignment uses the nearest ahead neighbor as a car-following reference when one is present.

Calibration scores this kernel at the data/control interval Δt (one predicted neighbor pose). At closed-loop decision time, utility and residual proposals query the constant-velocity OBB predictor over $1.5\,\text{s}$ with four substeps.

2.2.5 PATH ADHERENCE AND BOUNDARY CONSTRAINTS

A path-adherence penalty discourages proximity to corridor edges and off-road intrusion:

$$U_{\text{path}} = w_\ell \big(1 - e^{-\beta\, \ell_i^2}\big), \tag{15}$$

Let c_i be the signed minimum center clearance to the two corridor boundaries (or to the nearest curb in a polygonal site), positive inside and negative outside, and let $b = 1.5\,\text{m}$ be the near-edge buffer. The implemented quantity is the boundary-buffer violation

$$\ell_i = \max(0, b - c_i) \tag{16}$$

Accordingly, ℓ_i is zero away from the edge, increases inside the buffer, and continues to increase with off-road penetration; it is not raw boundary clearance. The hard boundary filter separately checks the full vehicle footprint. The coefficient w_ℓ scales the penalty magnitude, and β controls its curvature.

2.3 KINEMATIC BICYCLE CANDIDATE GENERATION

Each control candidate (a, δ) is rolled out through a one-step kinematic bicycle model. Given the current ego state (x, y, v, θ) and wheelbase L,

$$v' = \text{clip}(v + a\Delta t, 0, v_{\max}),\ \theta' = \theta + \tfrac{v}{L}\tan\delta\,\Delta t,\ (x', y') = (x, y) + v'(\cos\theta', \sin\theta')\Delta t. \tag{17}$$

Thus the utility evaluation is performed over the resulting candidate next state (x', y', v', θ'), rather than over free speed/heading pairs.

2.4 RESIDUAL MULTI-AGENT REINFORCEMENT LEARNING

Fixed prior scores cannot fully capture adaptive yielding, temporary aggressiveness, or crowding-dependent interaction strategies. A residual multi-agent policy therefore re-ranks the same discrete bicycle candidates from the local observation.

Default residual (candidate logits). Let $U(a)$ be the frozen prior score of candidate $a \in \mathcal{G}$ ($|\mathcal{G}| = 63$ on the freeway 7×9 grid) after shared road and OBB filters. A shared actor π_ϕ emits a bounded residual on that grid:

$$\text{score}(a) = U(a) + \alpha \tanh\big(\pi_\phi(o_i)[a]\big), \qquad \alpha > 0. \tag{18}$$

Training samples a masked categorical distribution $\text{softmax}(\text{score}/T)$; deterministic evaluation selects $\arg\max_a \text{score}(a)$. A zero-initialized output layer recovers the prior $\arg\max$ at the start of training. Following PDM-Closed Dauner et al. (2023), deterministic rollouts execute a residual index only when its one-step closed-loop proposal score strictly improves on the utility proposal; otherwise the prior action is kept. Specifically,

$$S_c(a) = D(a)N(a)\,\frac{5P(a) + 5T(a) + 2C(a)}{12},\ a_{\text{exec}} = \begin{cases} a_{\text{res}}, & S_c(a_{\text{res}}) > S_c(a_{\text{prior}}), \\ a_{\text{prior}}, & \text{otherwise.} \end{cases} \tag{19}$$

Here D/N are one-step road/OBB feasibility, P progress, T clearance, and C in-band comfort; full definitions are in Appendix E. S_c is a local gate score, distinct from episodic PDMS.

Ablation residual ($\Delta\Theta$). As an ablation we also allow a nine-dimensional parameter residual

$$\Delta\Theta_i^t = \pi_\phi(o_i^t), \qquad \Theta_i^t = \text{clip}\big(\Theta_i^{\text{base}} + \Delta\Theta_i^t\big)\,, \tag{20}$$

so that motion is selected by $\arg\max_{(a,\delta)\in\mathcal{G}} U_i(s_i'(a,\delta); \Theta_i^t)$. Unless noted, reported residual MARL uses candidate logits, not $\Delta\Theta$.

2.5 Policy Observation Space

Each agent receives a decentralized observation in corridor Frenet coordinates for the ego and the ego body frame for neighbors:

$$o_i^t = \Big[s_i,\ n_i,\ v_i,\ \tilde{\theta}_i,\ \tilde{\theta}_{\mathrm{goal},i},\ d_{\mathrm{lower},i},\ d_{\mathrm{upper},i},\ s_{\mathrm{remain},i},\ \big\{\Delta x_{ij}^{\mathrm{body}},\ \Delta y_{ij}^{\mathrm{body}},\ \Delta v_{x,ij}^{\mathrm{body}},\ \Delta v_{y,ij}^{\mathrm{body}}\big\}_{j\in\mathcal{N}_i} \Big]. \tag{21}$$

Missing neighbors are zero-padded to budget K. Feature meanings and normalization scales are in Appendix E.

2.6 Residual Action Space

Under the default candidate-logit residual, the continuous actor output has dimension $|\mathcal{G}| = 63$ and is tanh-bounded by a per-coordinate scale α (default 0.5). Softmax temperature during training is small (T=0.005) so exploration stays near the prior mode while remaining categorical.

The $\Delta\Theta$ ablation uses a nine-dimensional continuous residual

$$\Delta\Theta_i^t = \big[\Delta S_v,\ \Delta S_\theta,\ \Delta S_d,\ \Delta w_c,\ \Delta\xi_i,\ \Delta\gamma,\ \Delta w_\ell,\ \Delta\sigma_\parallel,\ \Delta\sigma_\perp\big], \tag{22}$$

with the remaining utility parameters (w_x, w_y, β) held fixed. Residual actions are tanh-scaled and clipped to keep Θ_i^t within physically valid ranges.

2.7 Reward Function

Following CaRL Jaeger et al. (2025), progress is the only dense positive term and is multiplied by soft safety and comfort factors in $(0, 1]$ so discounted PPO cannot trade collisions for faster travel:

$$R_i = w_{\mathrm{prog}}\ \max(p_i, 0)\, p_{\mathrm{safe}}(o_i)\, p_{\mathrm{comf}}(a_i, \delta_i) - \lambda_{\mathrm{left}}\, \frac{\max(s_{\mathrm{goal}} - s_i, 0)}{L_{\mathrm{corr}}} + R_{\mathrm{boundary}}. \tag{23}$$

Here p_i is normalized station progress, p_{safe} soft-gates proximity/contact, p_{comf} soft-gates out-of-band accel/steer, and R_{boundary} penalizes leaving the corridor. Arrival receives a one-shot bonus; hard contact triggers a terminal onset penalty (simulation continues for metrics). Exact definitions and constants are in Appendix E.

2.8 Training Procedure

We train a shared-policy PPO residual with decentralized execution on the freeway corridor harness. Advantages are computed per agent trajectory with truncation-aware Generalized Advantage Estimation (GAE); collision-terminal agents end their GAE without bootstrap. Checkpoints must not regress versus the frozen prior on collision events, collision pair-steps, or off-road rate. Among admissible actors we maximize a NAVSIM-style Predictive Driver Model Score (PDMS) Dauner et al. (2024; 2023)

$$\mathrm{PDMS} = \mathrm{NC} \cdot \mathrm{DAC} \cdot \frac{5\,\mathrm{EP} + 5\,\mathrm{TTC} + 2\,\mathrm{C}}{12}, \tag{24}$$

where NC/DAC are multiplicative zeros on any collision event / off-road, EP is goal progress (or arrival), TTC penalizes unsafe Time-To-Collision (TTC) rate, and C is in-band comfort (accel / jerk / steer thresholds). Hyperparameters and simulator constants are in Appendix E.

3 Data and Domain

We use three naturalistic multi-agent recordings. The primary residual-learning domain is the aerial freeway corridor of Khoueiry et al. (2026), treated as a challenging interactive driving environment with weak discrete-lane cues. The other two sites are a multi-arm roundabout near the same city and the publicly released TGSIM Foggy Bottom urban street network Ammourah et al. (2025). Geometry, filters, sample counts, and reconstructed curb maps are given in Appendix F.

4 UTILITY CALIBRATION

The utility model in Eq. 4 contains twelve free parameters

$$\theta = \big(S_\theta,\, S_v,\, \xi_i,\, S_d,\, \gamma,\, w_x,\, w_y,\, w_c,\, w_\ell,\, \beta,\, \sigma_\parallel,\, \sigma_\perp\big).$$

Before any residual policy is trained, these parameters are calibrated from naturalistic trajectories so that discrete utility maximization under bicycle kinematics reproduces observed short-horizon motion. We fit the same prior family independently on all three sites. Residual MARL and the paper-facing closed-loop benchmarks freeze the freeway working prior; the roundabout and Foggy Bottom fits support the destination-plus-curb transfer experiments in Appendix H.

Calibration idea. Ego transitions are paired with neighbors inside a local radius and scored against the same finite acceleration–steering candidate grid $\mathcal{G}$. Calibration minimizes a weighted composite of (i) short-horizon closed-loop tracking loss on sampled trajectory windows, (ii) normalized mean rank of the observed-like candidate in $\mathcal{G}$, and (iii) a light one-step softmax negative log-likelihood. Random search with multiple restarts screens candidates by the cheap one-step terms, then scores a retained subset with closed-loop windows. Working parameters are chosen on held-out validation vehicles (the coordinate-wise median of a near-optimal trial cloud when that is more stable, otherwise the best-on-validation trial) and reported on a held-out test split.

Site spatial frames. On the freeway, directional, proximity, and path terms are evaluated in the local corridor/Frenet frame of the PCA tube: $\hat{t}_i$ is the tube tangent, U_{dist} is a Frenet distance to a short-horizon reference, and U_{path} saturates outside the lower/upper bounds. On the roundabout and Foggy Bottom sites, the same five-term family is retained, but those three terms are evaluated in a destination-plus-curb frame: each ego's last observed point is treated as a goal on the roadway polygon, heading and proximity follow the on-road path toward that goal, and path cost grows with off-road distance. U_{speed} and U_{col} keep the forms in Eq. 4 at every site.

Search bounds, held-out fits, and working parameters are reported in Tables 9–11.

5 EXPERIMENTAL SETUP FOR RESIDUAL LEARNING AND BENCHMARKING

All residual-learning and baseline comparisons use a shared closed-loop freeway harness with matched scenario seeds, spawn windows, and episode lengths.

Training & Baselines. The prior is frozen at the freeway Θ_{base} (Table 11). Default residual: candidate-logit PPO over 63 bicycle actions (Sec. 2.8); $\Delta\Theta$ uses the same recipe, with weights-only / σ-only masks at inference. Closed-loop scoring uses the $1.5\,\mathrm{s}$ OBB lookahead unless a no-lookahead ablation is stated. Residual, $\Delta\Theta$, matched discrete RL, and MAPPO each use 24 PPO updates, four 240-step episodes per update, 10 agents, and seeds $\{0, 1, 2\}$ (23,040 environment steps per seed). All online learned methods share the same checkpoint rule: safety non-regression versus the prior (collision events, pair-steps, off-road), then maximum PDMS. We compare ORCA (van den Berg et al., 2011), social force (Helbing & Molnár, 1995), DWA (Fox et al., 1997), MPPI (Williams et al., 2017), a Frenet lattice planner (Werling et al., 2010), MAPPO (Yu et al., 2022), matched direct discrete RL, local adaptations of CtRL-Sim (Rowe et al., 2024) and CTG++ (Zhong et al., 2023), the calibrated utility prior, and residual MARL. CtRL-Sim and CTG++ are trained offline on freeway caches and selected with the same safety/PDMS rule; training budgets are in Appendix E.

Metrics & Protocols. Primary metrics follow closed-loop motion-planning practice: mean collision *events*, collision pair-steps, off-road rate, arrival rate, goal progress, capped travel time, unsafe TTC rate, comfort bands (mean $|a|$, RMS jerk, mean $|\delta|$), and the scalar PDM score (PDMS) used for selection. The sparse benchmark and gate ablation (Tables 1 and 3) evaluate 20 matched scenes, 10 agents, and 240 steps at $\Delta t{=}0.5\,\mathrm{s}$, using three training seeds for learned models. Displayed row intervals are descriptive and not commensurate across frozen vs. learned rows; significance is instead determined via a paired two-level seed-and-scenario bootstrap (Appendix E). The *parameter residual, no-lookahead, and dense stress* suites (Table 2) use the same 20 scenes and seeds. The no-lookahead ablation replaces the $1.5\,\mathrm{s}/4$-substep OBB query with a one-step query, while the dense stress protocol spawns 16 agents in a tighter window ($s \in [20, 80]\,\mathrm{m}$). Matched transfer runs on urban sites and additional freeway metrics are in Appendices H–I.

Table 1: Gate vs. residual ablation (20 sparse scenes, 10 agents, 240 steps). Off-road = 0; CI protocol in Sec. 5.

MODEL	PDMS ↑	Coll. ↓	Arrival ↑	Jerk (m/s^3) ↓
Utility prior	0.944 [0.844, 0.995]	0.050 [0.000, 0.150]	1.000 [1.000, 1.000]	4.56 [4.13, 5.01]
Residual, no gate	0.976 [0.944, 0.993]	0.017 [0.000, 0.050]	1.000 [1.000, 1.000]	4.06 [3.62, 4.56]
Residual + gate	0.977 [0.944, 0.994]	0.017 [0.000, 0.050]	1.000 [1.000, 1.000]	4.11 [3.73, 4.56]
Random residual + gate	0.833 [0.664, 0.979]	0.200 [0.000, 0.500]	0.975 [0.935, 1.000]	4.54 [4.03, 5.31]
Direct discrete RL	0.059 [0.045, 0.086]	3.77 [3.20, 4.05]	0.952 [0.940, 0.970]	3.90 [3.30, 4.41]

Table 2: Parameter-residual, no-lookahead, and dense-stress ablations (20 scenes, 3 seeds). Sparse uses 10 agents; stress uses 16.

	Sparse, full LA		Sparse, no LA		Dense stress	
MODEL	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓
Utility prior	0.944	0.050	0.942	0.050	0.562	0.650
Residual MARL (logits)	0.977	0.017	0.975	0.017	0.574	0.583
Parameter residual $\Delta\Theta$	0.976	0.017	0.925	0.100	0.630	0.500
Weights only	0.978	0.017	0.955	0.033	0.571	0.633
σ only	0.976	0.017	0.926	0.067	0.597	0.600
Direct discrete RL	0.059	3.77	0.015	8.42	0.000	15.27
MAPPO	0.080	3.12	0.016	6.48	0.000	7.42

6 Results and Benchmarks

Table 1 isolates residual learning from the PDM-Closed accept-if-better gate under the same 20-scene, 10-agent, 3-seed protocol as the main benchmark. Under the matched 23,040-environment-step budget, direct discrete RL on the identical bicycle grid obtains PDMS 0.059 with 3.77 collisions. A randomly initialized residual with the *same* gate trails the trained method in mean PDMS (0.833 vs. 0.977) but the gap is not significant at 20 scenes (bootstrap p=0.076), so the gate alone is not a reliable substitute for learning. The gate is a myopic one-step S_c proxy, so random+gate can still underperform the prior on episodic PDMS. Removing the gate from the trained residual is essentially a no-op (PDMS 0.976 vs. 0.977). Versus the frozen prior, residual MARL improves mean PDMS (0.944 → 0.977) and collisions (0.050 → 0.017) and significantly reduces RMS jerk (p=0.049), but those safety/score gaps are not significant at 20 scenes; arrival is tied at 1.000.

Table 2 reports the $\Delta\Theta$ residual, its inference masks, the no-lookahead OBB setting, and 16-agent dense stress. On sparse full lookahead, $\Delta\Theta$ and its masks are essentially tied with candidate-logit residual (PDMS 0.976–0.978, collisions 0.017). Removing the 1.5 s lookahead leaves candidate-logit residual nearly unchanged (PDMS 0.975) while $\Delta\Theta$ drops to 0.925 with 0.100 collisions; the prior stays at 0.942 / 0.050. Under dense 16-agent stress, $\Delta\Theta$ is strongest in the utility family (PDMS 0.630, 0.500 collisions), with candidate-logit residual next (0.574 / 0.583) ahead of the prior (0.562 / 0.650). Those dense residual-versus-prior gaps are not significant at 20 scenes. Direct discrete RL and MAPPO remain near-zero PDMS and worsen with density.

A prior-vs-residual rollout on the freeway scene is in Appendix I (Figure 12). Table 3 reports the multi-seed sparse benchmark. Residual MARL has the highest PDMS (0.977) and the fewest collisions (0.017). Under the same low online-training budget, MAPPO and matched discrete RL reach the corridor quickly but collide (3.12 and 3.77 events); PDMS is near zero because any contact zeros the score. Among classical planners, ORCA leads on PDMS (0.777) with MPPI close behind (0.763), but both trail the utility family. The local CtRL-Sim and CTG++ adaptations record PDMS 0.009 and 0.119, with 2.68 and 1.88 collisions, respectively. Residual MARL is significantly better than MAPPO, direct discrete RL, and CtRL-Sim on PDMS and collisions (p<0.001), and significantly better than ORCA and MPPI on PDMS (p=0.006 / 0.014). Versus the frozen prior, PDMS and collisions move in the residual's favor but are not significant at 20 scenes (Table 1).

Table 3: Sparse freeway benchmark (20 scenes, 10 agents, 240 steps). CI protocol in Sec. 5

MODEL	PDMS ↑	Coll. ↓	TTC Score↑	Arrival ↑	Travel (s)	Jerk (m/s^3) ↓
ORCA	0.777 [0.608, 0.913]	0.150 [0.000, 0.300]	0.14 [0.00, 0.34]	0.975 [0.935, 1.000]	64.6 [63.3, 65.9]	0.87 [0.78, 0.96]
Social force	0.449 [0.249, 0.649]	1.45 [0.75, 2.15]	0.55 [0.24, 0.90]	0.795 [0.680, 0.900]	66.9 [65.0, 69.1]	0.87 [0.76, 0.98]
DWA	0.145 [0.031, 0.288]	1.25 [0.90, 1.60]	0.00 [0.00, 0.00]	0.545 [0.475, 0.615]	62.2 [60.8, 63.5]	2.71 [2.48, 2.93]
MPPI	0.763 [0.574, 0.910]	0.250 [0.050, 0.500]	0.05 [0.00, 0.12]	0.980 [0.950, 1.000]	62.6 [61.4, 63.7]	1.78 [1.68, 1.88]
Frenet planner	0.022 [0.000, 0.066]	2.80 [2.20, 3.45]	0.00 [0.00, 0.00]	0.325 [0.230, 0.425]	62.9 [60.3, 65.7]	0.91 [0.80, 1.06]
Utility prior (ours)	0.944 [0.844, 0.995]	0.050 [0.000, 0.150]	0.84 [0.68, 1.00]	1.000 [1.000, 1.000]	63.4 [60.5, 66.6]	4.56 [4.13, 5.01]
Direct discrete RL	0.059 [0.045, 0.086]	3.77 [3.20, 4.05]	0.00 [0.00, 0.00]	0.952 [0.940, 0.970]	35.3 [34.6, 36.2]	3.90 [3.30, 4.41]
MAPPO	0.080 [0.000, 0.192]	3.12 [2.25, 3.80]	0.00 [0.00, 0.00]	0.508 [0.170, 0.690]	37.5 [35.7, 39.9]	1.49 [0.99, 1.77]
CtRL-Sim	0.009 [0.000, 0.028]	2.68 [2.15, 3.30]	0.00 [0.00, 0.00]	0.248 [0.130, 0.370]	58.9 [54.1, 62.2]	1.95 [1.84, 2.13]
CTG++	0.119 [0.000, 0.269]	1.88 [0.80, 2.55]	0.04 [0.00, 0.13]	0.075 [0.000, 0.115]	71.0 [63.9, 78.0]	0.94 [0.63, 1.13]
Residual MARL (ours)	0.977 [0.944, 0.994]	0.017 [0.000, 0.050]	0.87 [0.84, 0.91]	1.000 [1.000, 1.000]	62.8 [62.1, 63.4]	4.11 [3.73, 4.56]

7 Conclusion

We investigated residual adaptation of a calibrated utility prior for closed-loop multi-agent driving under a limited online-training budget. On the sparse freeway benchmark, candidate-logit residual MARL achieves 0.977 PDMS and 0.017 collisions per episode, outperforming MAPPO and direct discrete RL at matched budgets. Relative to the frozen prior, mean PDMS increases and collision frequency decreases, although neither difference is statistically significant at 20 scenes. The ablations reveal distinct roles for learning, inference-time gating, and residual parameterization. Removing the accept-if-better gate leaves the trained policy's PDMS nearly unchanged (0.976 versus 0.977), indicating that the gate is not responsible for the observed performance of the trained residual. Candidate-logit and parameter residuals perform comparably under full collision lookahead but exhibit different sensitivities when the prediction horizon or interaction density changes. Without lookahead, candidate-logit adaptation retains higher PDMS (0.975 versus 0.925), whereas the parameter residual performs better under dense 16-agent interaction (0.630 versus 0.574). These results suggest that the choice of residual parameterization affects robustness to collision-prediction settings and traffic density.

8 Limitations

Prior calibration is weakly identified, yielding an acceptable parameter region rather than a unique $\Theta^\star$, and one-step ranking is moderately discriminative on urban sites (Appendix G). At 20 scenes, residual–prior performance gaps lack statistical significance, and dense stress still incurs collisions. Online RL comparisons use low budget to measure sample efficiency near initialization, not asymptotic MARL performance. CtRL-Sim and CTG++ are local harness adaptations, not full reproductions. Finally, evaluating broader generalization requires testing across more diverse datasets.

Reproducibility Statement

Code, configurations, and pre-trained checkpoints are included as supplementary material for peer review and will be open-sourced upon publication to maintain anonymity. Appendix E details all training budgets, seeds, and evaluation protocols.

Ethics Statement

This work collected aerial trajectory data and also used previously collected infrastructure-based trajectory data for scientific modeling. No personally identifying information is released.

AI USE STATEMENT

In this work, generative AI tools were used to assist with code refactoring and debugging, and to refine manuscript text for clarity. We have not used generative AI tools to generate synthetic datasets, design the research methodology, formulate mathematical claims, or produce final numerical tables. All scientific claims, experimental design choices, model implementations, and numerical results were produced and verified by the authors.

REFERENCES

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2669–2678, 2018. doi: 10.1609/aaai.v32i1.11797.

Rami Ammourah, Pedram Beigi, Bingyi Fan, Samer H Hamdar, John Hourdos, Chun-Chien Hsiao, Rachel James, Mohammdreza Khajeh-Hosseini, Hani S Mahmassani, Dana Monzer, et al. Introduction to the third generation simulation dataset: Data collection and trajectory extraction. *Transportation Research Record*, 2679(1):1768–1784, 2025.

Gowri Asaithambi and Gugulothu Shravani. Overtaking behaviour of vehicles on undivided roads in non-lane based mixed traffic conditions. *Journal of traffic and transportation engineering (English edition)*, 4(3):252–261, 2017.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Proceedings of the Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4709–4722, 2025. URL https://proceedings.mlr.press/v305/cao25a.html.

Satish Chandra and Upendra Kumar. Capacity estimation procedure for two-lane roads under mixed traffic conditions. *Journal of Transportation Engineering*, 130(5):737–743, 2004.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Proceedings of the Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1268–1281, 2023. URL https://proceedings.mlr.press/v229/dauner23a.html.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.15349.

Dieter Fox, Wolfram Burgard, and Sebastian Thrun. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33, 1997. doi: 10.1109/100.580977.

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*, 2023. URL https://arxiv.org/abs/2310.08710.

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

Jon Herman and Will Usher. SALib: An open-source Python library for sensitivity analysis. *Journal of Open Source Software*, 2(9):97, 2017. doi: 10.21105/joss.00097.

Bernhard Jaeger, Daniel Dauner, Jens Beißwenger, Simon Gerstenecker, Kashyap Chitta, and Andreas Geiger. Carl: Learning scalable planning policies with simple rewards. In *Proceedings of the Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*,

pp. 5301–5338, 2025. URL https://proceedings.mlr.press/v305/jaeger25a.html.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation*, pp. 6023–6029, 2019. doi: 10.1109/ICRA.2019.8794127.

Sarosh I Khan and Pawan Maini. Modeling heterogeneous traffic flow. *Transportation research record*, 1678(1):234–241, 1999.

Michel S. Khoueiry, Pedram Beigi, and Samer H. Hamdar. Towards universal autonomy: Modeling chaotic and non-lane based driving behavior using naturalistic trajectories. 2026. Under review.

Tibor Petrov, Wissam Sleiman, Pedram Beigi, Peter Počta, L'uboš Buzna, and Samer Hamdar. Simulation of vehicle-to-pedestrian communication for development of novel c-its services for vulnerable road users. *Transportation Research Procedia*, 93:674–679, 2026.

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

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2):259–270, 2010.

Zihao Sheng, Zilin Huang, and Sikai Chen. Traffic expertise meets residual rl: Knowledge-informed model-based residual reinforcement learning for cav trajectory control. *Communications in Transportation Research*, 4:100142, 2024. doi: 10.1016/j.commtr.2024.100142.

Jiamin Shi, Tangyike Zhang, Ziqi Zong, Shitao Chen, Jingmin Xin, and Nanning Zheng. Task-driven autonomous driving: Balanced strategies integrating curriculum reinforcement learning and residual policy. *IEEE Robotics and Automation Letters*, 9(11):9454–9461, 2024. doi: 10.1109/LRA.2024.3448237.

Tom Silver, Kelsey Allen, Joshua B. Tenenbaum, and Leslie Pack Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018. URL https://arxiv.org/abs/1812.06298.

Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

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

Alireza Talebpour, Hani S Mahmassani, and Samer H Hamdar. Multiregime sequential risk-taking model of car-following behavior: specification, calibration, and sensitivity analysis. *Transportation research record*, 2260(1):60–66, 2011.

Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research*, volume 70 of *Springer Tracts in Advanced Robotics*, pp. 3–19. Springer, 2011. doi: 10.1007/978-3-642-19457-3_1.

Pengcheng Wang, Chenran Li, Catherine Weaver, Kenta Kawamoto, Masayoshi Tomizuka, Chen Tang, and Wei Zhan. Residual-mppi: Online policy customization for continuous control. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.00898.

Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a Frenét frame. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 987–993. IEEE, 2010.

Grady Williams, Andrew Aldrich, and Evangelos A. Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 2017. doi: 10.2514/1.G001921.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15677.

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

Chris Zhang, James Tu, Lunjun Zhang, Kelvin Wong, Simon Suo, and Raquel Urtasun. Learning realistic traffic agents in closed-loop. In *Proceedings of the Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 800–821, 2023a. URL https://proceedings.mlr.press/v229/zhang23b.html.

Ruiqi Zhang, Jing Hou, Guang Chen, Zhijun Li, Jianxiao Chen, and Alois Knoll. Residual policy learning facilitates efficient model-free autonomous racing. *IEEE Robotics and Automation Letters*, 7(4):11625–11632, 2022. doi: 10.1109/LRA.2022.3192770.

Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. Trafficbots: Towards world models for autonomous driving simulation and motion prediction. In *IEEE International Conference on Robotics and Automation*, 2023b. URL https://arxiv.org/abs/2303.04116.

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *Conference on Robot Learning (CoRL)*, 2023. CTG++; PMLR 229.

APPENDICES

A RELATED WORK

A.1 STRUCTURED MOTION PLANNING AND INTERPRETABLE PRIORS

Classical motion planners provide explicit structure for balancing progress, collision avoidance, and dynamic feasibility. Reciprocal collision-avoidance methods such as ORCA formulate multi-agent navigation through local velocity constraints (van den Berg et al., 2011), while the Dynamic Window Approach (DWA) searches dynamically reachable velocities using goal-directed and obstacle-related objectives (Fox et al., 1997). Model Predictive Path Integral (MPPI) control extends sampling-based optimization to sequences of controls evaluated through forward dynamics (Williams et al., 2017). These methods provide structured alternatives to directly learned policies, but their objectives and interaction assumptions must be specified or tuned.

Structured planning has also been effective in autonomous driving. Prior work demonstrates the value of simple rule-based priors and trajectory scoring in closed-loop planning (Dauner et al., 2023). These findings motivate retaining explicit planning structure rather than learning the entire control policy from scratch. Our approach follows this principle but uses a utility scorer calibrated from naturalistic vehicle trajectories. The scorer evaluates bicycle-kinematic candidates using directional alignment, speed preference, proximity, collision risk, and corridor adherence. Geometry enters only as a continuous road boundary (curb / PCA-tube bounds): none of our sites supplies painted lane markings or a discrete lane graph, so agents negotiate longitudinally and laterally inside that boundary without lane IDs. Unlike a fixed local planner, the proposed framework allows these candidate preferences to be adapted through RL.

A.2 RESIDUAL POLICY LEARNING AND HYBRID CONTROL

Residual RL combines a conventional controller with a learned correction. Prior work augments a feedback controller with a learned residual for robotic manipulation (Johannink et al., 2019) and improves existing hand-designed and model-predictive policies through model-free residual learning (Silver et al., 2018). These studies establish residual learning as a mechanism for leveraging imperfect but useful control priors.

Related approaches have also been investigated in driving. Potential-field guidance has been combined with a learned residual policy for autonomous racing (Zhang et al., 2022); other methods integrate rule-based planning, residual RL, and curriculum learning for driving decisions (Shi et al., 2024), or incorporate traffic expertise into model-based residual RL for connected-vehicle trajectory control (Sheng et al., 2024; Petrov et al., 2026). More recently, Residual-MPPI customizes an existing continuous-control policy through online planning over residual actions (Wang et al., 2025).

Our formulation differs primarily in where the correction is applied. Rather than adding a residual directly to a continuous control command or using online trajectory optimization to customize an existing policy, we learn bounded additive corrections to the scores of a fixed set of bicycle-kinematic candidates. The calibrated utility prior remains frozen, and the learned policy changes the ranking of its feasible alternatives. A parameter-level residual is retained as an ablation to distinguish direct candidate-score adaptation from adaptation of the underlying utility coefficients.

A.3 INTERACTIVE MULTI-AGENT MOTION GENERATION

Learning realistic traffic behavior requires accounting for interactions that develop over multiple closed-loop decisions. TrafficSim (Suo et al., 2021) learns joint multi-agent behavior from demonstrations and uses differentiable simulation to address errors accumulated during rollout. Reinforcing Traffic Rules combines imitation and RL to balance realistic traffic behavior with rule compliance (Zhang et al., 2023a). TrafficBots introduces destination-conditioned multi-agent simulation with latent behavioral styles (Zhang et al., 2023b), while SMART represents agent motion through discrete tokens for scalable traffic generation (Wu et al., 2024). CTG++ generates multi-agent futures with a scene-level diffusion transformer and optional language-guided sampling losses (Zhong et al., 2023). CtRL-Sim instead trains a return-conditioned offline RL policy for reactive, control-

lable closed-loop agents (Rowe et al., 2024). We include local adaptations of both as paper-facing baselines in the shared bicycle harness (without CTG++'s original language stack).

RL provides another approach to interactive control. Multi-agent PPO is a competitive learning baseline in cooperative multi-agent environments (Yu et al., 2022). Waymax provides a data-driven simulation framework for training and evaluating interactive driving policies and highlights the potential for learned policies to overfit to simulated agents (Gulino et al., 2023).

Our objective is related to interactive traffic generation, but the learning problem is formulated differently. Instead of learning a joint trajectory distribution, a diffusion scene model, or a return-conditioned offline policy from scratch, agents independently evaluate kinematically feasible control candidates using local traffic observations and a shared residual policy. The calibrated prior provides a common action-scoring structure, while residual learning permits context-dependent changes in candidate preference. Evaluation therefore focuses on the resulting closed-loop interactions rather than open-loop trajectory prediction alone.

Rule-regularized and shielded MARL stacks that couple demonstrations or formal safety filters with learning (Zhang et al., 2023a; Shi et al., 2024; Sheng et al., 2024; Alshiekh et al., 2018) are complementary, but typically assume discrete lane graphs, rule sets, or shielding automata that are not available in our curb-only corridor setting. Adapting those baselines to continuous road-boundary geometry is outside the present scope; our online learning comparisons therefore use matched low-budget unconstrained MAPPO and direct discrete RL as sample-efficiency references rather than as stand-ins for modern rule-regularized MARL.

A.4 Safety-Aware Learning and Closed-Loop Evaluation

Safety constraints can be incorporated into learning through action filtering, reward design, or policy evaluation. Shielded RL monitors proposed actions and intervenes when they violate specified safety requirements (Alshiekh et al., 2018). In autonomous driving, CaRL demonstrates that emphasizing route completion while handling infractions through termination or multiplicative penalties can improve the scalability of PPO training (Jaeger et al., 2025).

Evaluation methodology is equally important. PDM-Closed demonstrates the effectiveness of structured proposal selection in closed-loop vehicle planning (Dauner et al., 2023). NAVSIM introduces a data-driven, non-reactive evaluation framework and the PDM score, which combines progress, safety, and comfort-related metrics (Dauner et al., 2024). Subsequent pseudo-simulation research further investigates how evaluation can capture recovery behavior without full sequential interactive simulation (Cao et al., 2025).

We adopt a CaRL-inspired reward structure and a NAVSIM-inspired scoring metric for checkpoint selection, but evaluate the learned controller in an interactive multi-agent closed-loop environment. At inference, an accept-if-better mechanism compares the residual proposal against the frozen prior and retains the prior action when the proposal does not improve the specified score. This mechanism is a proposal-selection rule, not a formal guarantee of collision-free execution. Its contribution must therefore be distinguished experimentally from the improvement attributable to residual learning itself.

A.5 Research Gap

What remains scarce is a single closed-loop multi-agent generator that is (a) calibrated to observed short-horizon choice, (b) kinematically feasible and interpretable, and (c) able to adapt that prior under crowding without collapsing into geometric interaction alone or an unconstrained speed-seeking RL policy. We address this with a calibrated discrete bicycle utility prior plus a candidate-logit residual under a shared closed-loop evaluation protocol; head-to-head comparison with rule-regularized or shielded MARL designed for lane-graph maps is left to future work.

B Sensitivity Analysis

Understanding parameter sensitivity is crucial to ensure that agent-based model produces reliable, realistic, and robust behavior under a wide range of conditions Talebpour et al. (2011). The sen-

sitivity analysis reveals how robust the model is to parameter variations and helps guide parameter tuning.

To systematically assess model robustness, we perform both local and global sensitivity analyses. Local analysis examines the effect of small variations in individual parameters on each component of the utility function, providing insight into immediate system responsiveness. Global analysis, using techniques such as Sobol indices, quantifies the contribution of each parameter including interactions across their entire plausible range. Together, these analyses reveal how sensitive the model is to parameter changes and guide informed tuning for reliable and realistic agent behavior.

C Local Sensitivity Analysis

Local sensitivity analysis quantifies how small parameter perturbations affect each component of the utility. We analyze the five utility terms from Equation 4 for directional utility, speed incentive, proximity incentive, collision-risk utility, and path adherence, separately.

Directional Alignment

Directional utility rewards candidate motion aligned with the local corridor tangent rather than a destination heading. Let $\hat{m}$ denote the piecewise unit move direction defined in Section 2.2, including the route-tangent fallback when candidate displacement is below 10^{-6} m, and let $\hat{t}$ be the local corridor unit tangent. Then

$$U_{\text{dir}} = S_\theta \, (\hat{m} \cdot \hat{t}). \tag{25}$$

$$\frac{\partial U_{\text{dir}}}{\partial S_\theta} = \hat{m} \cdot \hat{t}. \tag{26}$$

$$\frac{\partial U_{\text{dir}}}{\partial (\hat{m} \cdot \hat{t})} = S_\theta. \tag{27}$$

Thus S_θ linearly scales the directional reward. Utility is maximized when $\hat{m} \cdot \hat{t} = 1$ (perfect corridor alignment) and becomes negative when the candidate move opposes the corridor orientation. This remains valid for both positive- and negative-direction traffic because directionality is defined by corridor geometry.

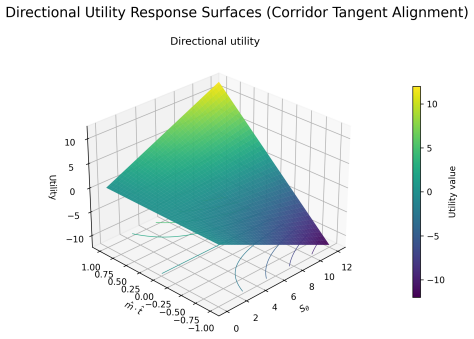

Figure 2: Directional utility U_{dir} versus corridor alignment $\hat{m} \cdot \hat{t}$ and directional sensitivity S_θ.

Speed Incentive

Speed utility rewards candidate bicycle-rollout speeds close to a desired speed. Define the symmetric closeness ratio

$$\rho = \frac{\min(v_{\text{cand}}, v_{\text{des}})}{\max(v_{\text{cand}}, v_{\text{des}})} \in [0, 1], \tag{28}$$

and

$$U_{\text{speed}} = S_v \frac{\rho}{1 + \rho^{(\xi_i - 1)/2}}. \tag{29}$$

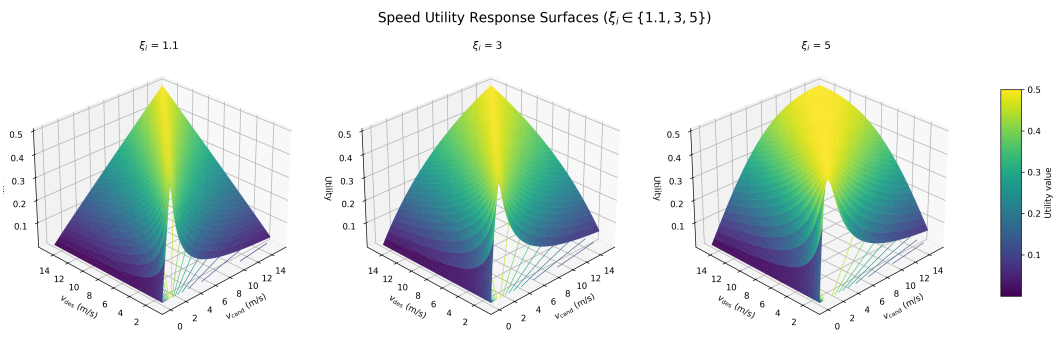

Figure 3: Speed utility U_{speed} versus candidate speed v_{cand} and desired speed v_{des} (with ξ_i slices).

The corresponding partial derivatives are

$$\frac{\partial U_{\text{speed}}}{\partial S_v} = \frac{\rho}{1+\rho^{(\xi_i-1)/2}}, \tag{30}$$

$$\frac{\partial U_{\text{speed}}}{\partial \xi_i} = -\frac{S_v\,\rho^{(\xi_i+1)/2}\ln\rho}{2\big(1+\rho^{(\xi_i-1)/2}\big)^2}. \tag{31}$$

For the calibrated range $\xi_i \in [1.1, 5]$, utility attains its maximum at $v_{\text{cand}} = v_{\text{des}}$ ($\rho = 1$). Overspeeding and underspeeding are penalized symmetrically. Here ξ_i is a curvature/shape parameter: smaller values impose the stronger mismatch penalty, whereas increasing ξ_i flattens the utility near $\rho = 1$.

Proximity Incentive

Proximity utility encourages progress toward a local corridor reference expressed in Frenet coordinates. The effective distance is

$$d_{\text{eff}} = w_x|\Delta s| + w_y|\Delta n|, \tag{32}$$

and

$$U_{\text{dist}} = \frac{S_d}{1+(d_{\text{eff}}/H_p)^{\gamma}}, \quad H_p = \max(\kappa v_i, H_{\min}). \tag{33}$$

The primary sensitivities are

$$\frac{\partial U_{\text{dist}}}{\partial S_d} = \frac{1}{1+(d_{\text{eff}}/H_p)^{\gamma}}, \tag{34}$$

$$\frac{\partial U_{\text{dist}}}{\partial \gamma} = -S_d\frac{(d_{\text{eff}}/H_p)^{\gamma}\ln(d_{\text{eff}}/H_p)}{\big(1+(d_{\text{eff}}/H_p)^{\gamma}\big)^2}, \tag{35}$$

$$\frac{\partial U_{\text{dist}}}{\partial d_{\text{eff}}} = -S_d\frac{\gamma(d_{\text{eff}}/H_p)^{\gamma}}{d_{\text{eff}}\big(1+(d_{\text{eff}}/H_p)^{\gamma}\big)^2}, \tag{36}$$

$$\frac{\partial U_{\text{dist}}}{\partial H_p} = S_d\frac{\gamma(d_{\text{eff}}/H_p)^{\gamma}}{H_p\big(1+(d_{\text{eff}}/H_p)^{\gamma}\big)^2}. \tag{37}$$

Here S_d scales importance, γ controls how quickly the reward decays outside the horizon, w_x and w_y weight longitudinal versus lateral Frenet error, and H_p softens normalized distance as speed increases.

Collision-Risk Penalty

The primary partial derivatives for the collision-risk penalty in Eq. 4 are

$$U_{\text{col}} = w_c\min\Big(\sum_j P_{j,i'},\,1\Big), \qquad \frac{\partial U_{\text{col}}}{\partial w_c} = \min\Big(\sum_j P_{j,i'},\,1\Big). \tag{38}$$

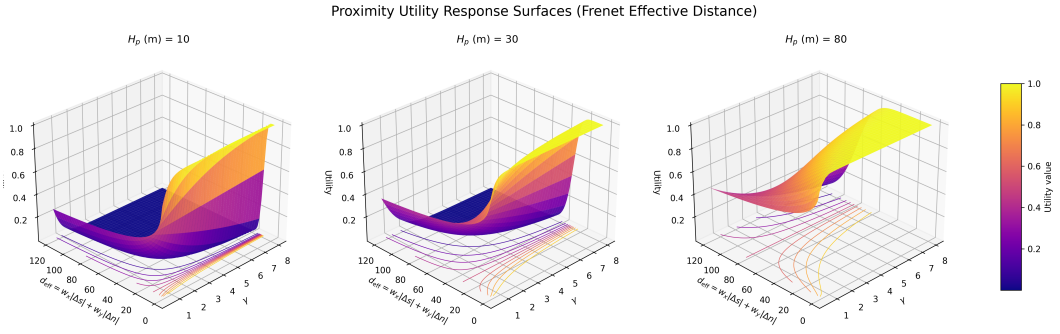

Figure 4: Proximity utility U_{dist} versus interaction exponent γ, effective Frenet distance d_{eff}, and perception horizon H_p.

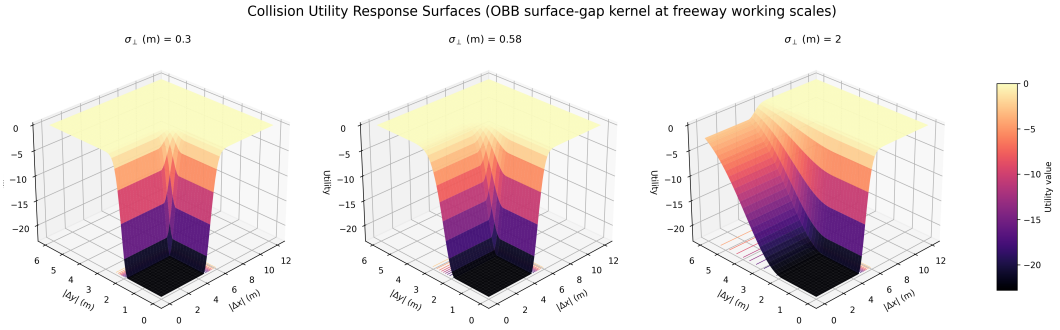

Figure 5: Collision utility $U_{\text{col}} = w_c P$ versus body-frame centre separations $|\Delta x|, |\Delta y|$ at $w_c = 22.8$, $\sigma_\parallel = 0.65\,\text{m}$, $L \times W = 4.5 \times 1.8\,\text{m}$ (this positive term is subtracted in Eq. 4). Panels vary $\sigma_\perp$; the flat well is the overlapping-footprint plateau.

with $P_{j,i'} = \exp(-\frac{1}{2}[(g_\parallel/\sigma_\parallel)^2 + (g_\perp/\sigma_\perp)^2])$ on OBB surface gaps. Increasing w_c makes the utility more sensitive to predicted overlap and therefore promotes more risk-averse behavior. Small $(\sigma_\parallel, \sigma_\perp)$ sharpen the risk field around the footprints; large scales broaden it.

Fig. 5 shows this kernel at the freeway working scales: unit presence on overlapping $L \times W$ boxes, then anisotropic decay on the surface gaps $(g_\parallel, g_\perp)$.

Path Adherence

The path-adherence term and its derivatives are

$$U_{\text{path}} = w_\ell \big(1 - e^{-\beta \ell_i^2}\big), \tag{39}$$

$$\frac{\partial U_{\text{path}}}{\partial w_\ell} = 1 - e^{-\beta \ell_i^2}, \tag{40}$$

$$\frac{\partial U_{\text{path}}}{\partial \beta} = w_\ell \ell_i^2 e^{-\beta \ell_i^2}, \tag{41}$$

$$\frac{\partial U_{\text{path}}}{\partial \ell_i} = 2\, w_\ell \beta \ell_i e^{-\beta \ell_i^2}. \tag{42}$$

Here ℓ_i is a non-negative corridor-boundary residual, w_ℓ scales the penalty magnitude, and β controls how quickly the penalty saturates near the boundary. Larger w_ℓ enforces stricter corridor adherence but may conflict with collision avoidance or progress when set excessively high.

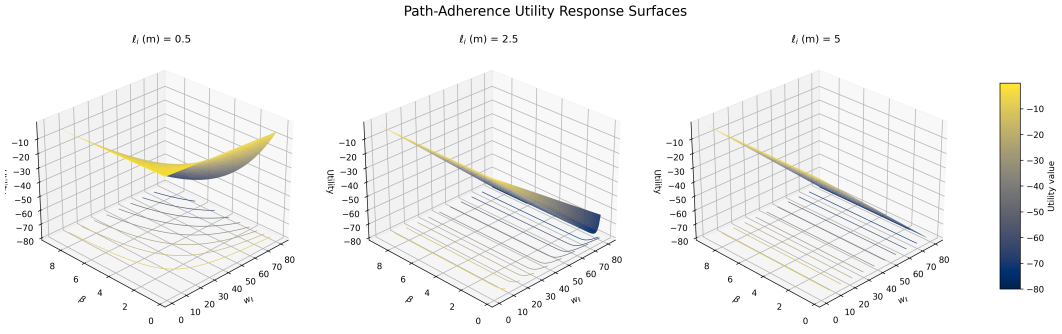

Figure 6: Path-adherence utility U_{path} versus path-deviation coefficient β and lateral deviation ℓ_i.

D Global Sensitivity Analysis

Global Sensitivity Analysis (GSA) attributes variance in a simulation-level output Y to uncertainty in the utility parameters, including interactions. This study uses the twelve-parameter top-cloud p_{05}–p_{95} box in Table 4, which contains θ_{rob}. Samples are drawn independently; they do not reproduce identifiability correlations such as S_v–w_c. Each evaluation runs a three-agent follow/merge scene for up to 80 steps ($\Delta t = 0.5\,\text{s}$) with the kinematic-bicycle candidate grid, the OBB surface-gap collision kernel, and OBB overlap counting.

The scalar output is

$$Y = \sum_{k \in \mathcal{A}} d_k^{\text{final}} + N_{\text{coll}}\, C_{\text{penalty}}, \tag{43}$$

where d_k^{final} is leftover destination distance if agent k has not arrived, N_{coll} counts pair-timesteps with overlapping oriented boxes, and $C_{\text{penalty}} = 10$. Lower Y is better. Across the 14336 evaluations, Y is heavy-tailed (mean 15.7, std. 18.9, max 407).

The sensitivity analysis is performed using the variance-based Sobol method Sobol' (1990); Saltelli et al. (2010). Let the model output be

$$Y = f(X_1, X_2, \ldots, X_D),$$

where X_j represents independent model parameters. The output variance is decomposed as

$$\text{Var}(Y) = \sum_{j=1}^{D} V_j + \sum_{1 \le j < k \le D} V_{jk} + \cdots + V_{12\ldots D}. \tag{44}$$

The first-order Sobol index is

$$S_j = \frac{V_j}{\text{Var}(Y)}, \tag{45}$$

while the total-order Sobol index is

$$S_{T_j} = 1 - \frac{\text{Var}_{\mathbf{X}_{\sim j}}\left(\mathbb{E}_{X_j}\left[Y \mid \mathbf{X}_{\sim j}\right]\right)}{\text{Var}(Y)}. \tag{46}$$

The first-order index quantifies the direct contribution of parameter X_j to the output variance, whereas the total-order index measures its overall contribution, including all interaction effects.

Sampling uses Sobol sequences in `SALib` (Herman & Usher, 2017) with $N = 1024$ and first- plus total-order estimators,

$$N_{\text{eval}} = N(D + 2) = 1024(12 + 2) = 14336.$$

Table 4: Sobol input box: freeway top-cloud p_{05}–p_{95} (contains θ_{rob}).

Parameter	Description	Range
S_θ	Directional weight	$[1.54, 9.13]$
S_v	Speed-incentive weight	$[0.10, 8.52]$
ξ_i	Speed-utility curvature	$[1.42, 4.78]$
S_d	Proximity-incentive weight	$[3.49, 11.5]$
γ	Proximity decay steepness	$[0.40, 7.22]$
w_x	Longitudinal Frenet weight	$[0.79, 9.15]$
w_y	Lateral Frenet weight	$[1.11, 9.46]$
w_c	Collision-risk penalty weight	$[2.59, 56.3]$
w_ℓ	Path-adherence penalty weight	$[2.74, 241]$
β	Path-adherence saturation rate	$[0.24, 9.01]$
$\sigma_\parallel$ (m)	Longitudinal kernel scale	$[0.62, 4.26]$
$\sigma_\perp$ (m)	Lateral kernel scale	$[0.45, 2.15]$

Table 5: Sobol indices on the calibrated twelve-parameter box ($N = 1024$).

Parameter	S_1	$S_{1,\mathrm{conf}}$	S_T	$S_{T,\mathrm{conf}}$
S_θ	-0.018	0.058	0.519	0.194
S_v	0.009	0.098	3.224	3.812
ξ_i	-0.022	0.057	0.568	0.209
S_d	0.005	0.031	0.248	0.164
γ	-0.023	0.059	0.407	0.167
w_x	-0.027	0.059	0.396	0.141
w_y	-0.029	0.076	0.462	0.273
w_c	0.077	0.057	0.649	0.212
w_ℓ	-0.006	0.080	0.675	0.320
β	0.009	0.061	0.541	0.265
$\sigma_\parallel$	0.037	0.069	0.684	0.241
$\sigma_\perp$	0.004	0.082	0.912	0.335

Table 4 lists the input box.

D.1 Sensitivity Analysis Results

Table 5 reports first-order and total-order indices. Small negative S_1 values are Monte Carlo noise around zero.

Most first-order indices are consistent with zero within their estimated confidence intervals. The collision-risk weight w_c has a small positive first-order estimate, while the remaining coordinates show limited individually identifiable effects. Total-order indices remain comparatively large for most coordinates ($S_T \in [0.25, 0.91]$, aside from an unstable S_v estimate), suggesting substantial interaction effects, although their confidence intervals remain broad. This pattern is consistent with weak parameter identifiability: independently combining calibrated parameters can produce off-manifold priors with substantially different overlap counts.

Collision-kernel coordinates participate strongly. w_c, $\sigma_\parallel$, and $\sigma_\perp$ have total-order mass 0.65, 0.68, and 0.91. Path terms w_ℓ and β also carry substantial S_T (0.68 and 0.54). Proximity weight S_d is the weakest total-order coordinate (0.25). The S_v total-order point estimate exceeds one with a very wide interval, and other S_T confidence bands remain broad: both reflect sampling error on a heavy-tailed Y (here $\mathrm{std}(Y) \approx 18.9$ with rare large overlap penalties) rather than a precise ranking.

We therefore do not replace Fig. 10 with a unique Sobol order. The GSA supports two statements about the calibrated box: main effects are small, and collision-kernel plus path-adherence coordinates participate strongly through interactions.

Table 6: Reward, boundary, and closed-loop simulator constants (freeway protocol, `run_id`= 2, `lane_kf`= 1).

Reward / boundary	**Value**	**Simulator timing / geometry**	**Value**
w_{prog} (progress)	1.0	Step Δt	0.5 s
Arrival bonus	8	Episode horizon $T_{\max}$	240 paper steps (120 dev)
Leftover coef. λ_{left}	0.08	Agents per episode N	10 (sparse) / 16 (dense stress)
Contact terminal / duration	1 / 0	Spawn s-range	$[20, 120]$ m sparse / $[20, 80]$ m dense
w_{safe} (soft safety)	1.0	Min. initial spacing	8.0 m sparse / 5.0 m dense
w_{smooth} (soft comfort)	1.0	Spawn lateral fraction	0.35 (sparse) / 0.55 (dense) of half-width
Soft floor η	0.1	Destination margin	uniform in $[5, 40]$ m from end
Comfort / gate $\lvert a\rvert$ band	2.4 m/s^2	Arrival tolerance	1.0 m along s
Gate $\lvert\delta\rvert$ band	0.3375 rad	Vehicle $L \times W$	4.5×1.8 m
C_{out}	20	Wheelbase	2.8 m
c_{depth}	3	$v_{\max}$	16 m/s
$d_{\max}$	5.0 m	$a_{\max}$ (actuator)	4.0 m/s^2
Accept-if-better	eval only	$\delta_{\max}$ (actuator)	0.45 rad
K (neighbor budget)	6		
R_p (perception radius)	60 m		
κ (prox. horizon)	2.0		
$H_{\min}$	5.0 m		
Candidate grid		**OBB lookahead**	
Accel. (m/s^2)	$\{-3, \ldots, 3\}$	Conflict horizon	1.5 s
Steering (rad)	9 values in $[-0.45, 0.45]$	Conflict substeps	4
Episodic PDMS bands			
Unsafe TTC threshold	1.5 s	TTC rate cap	0.05
PDMS mean $\lvert a\rvert$ band	2.4 m/s^2	PDMS RMS jerk band	8.37 m/s^3
PDMS mean $\lvert\delta\rvert$ band	0.10 rad	PDMS weights (P, T, C)	5:5:2

E Reproducibility and Training Configuration

Table 6–7 collect every constant needed to reproduce the closed-loop residual trainer (`RL/train_ppo.py`) and the shared benchmark harness (`Baselines/`). Values are identical for evaluation unless noted. Paper Tables 3, 1, and 2 use 24 PPO updates, four 240-step episodes per update, seeds $\{0, 1, 2\}$, and 20 evaluation scenes for residual MARL, $\Delta\Theta$, MAPPO, and matched discrete RL. This is 23,040 environment steps, or 230,400 agent transitions with 10 agents, per seed; sparse protocols use 10 agents and 240 steps, while dense stress uses 16 agents and a tighter spawn window. All online learned methods use the same checkpoint procedure: safety non-regression against the frozen prior on collision events, collision pair-steps, and off-road rate, followed by maximum PDMS among admissible checkpoints. CtRL-Sim and CTG++ are trained offline from recorded caches, then selected with the same closed-loop safety/PDMS rule before entering the shared harness. The CTG++ adaptation uses 10,000 optimizer steps per seed, retains candidates at steps 5,000 and 10,000 for closed-loop selection, and uses 20 reverse-diffusion steps at inference. The MAPPO baseline uses a separate centralized critic (joint corridor features concatenated with o_i); its PPO settings are listed in the last block of Table 7.

Reward term definitions. Let s_i^- and s_i be station before/after the transition. Normalized progress is

$$p_i = \text{clip}\left(\frac{\min(s_i, s_{\text{goal}}) - \min(s_i^-, s_{\text{goal}})}{v_{\max}\Delta t}, -1, 1\right).$$

For signed nearest-neighbor footprint gap g ($g \le 0$: OBB contact),

$$c(g) = \begin{cases} -16\big(1 - g/(1\,\mathrm{m})\big), & g \le 0, \\ -\exp\big(-g/(2\,\mathrm{m})\big), & g > 0, \end{cases} \qquad o_i = \min_{j \in \mathcal{N}_i} c(g_{ij})$$

(with $o_i = 0$ if $\mathcal{N}_i$ is empty), and

$$p_{\mathrm{safe}}(o_i) = \mathrm{clip}\Big(1 + \frac{w_{\mathrm{safe}} o_i}{16}, \eta, 1\Big), \quad \eta = 0.1.$$

Comfort is $p_{\mathrm{comf}} = 1$ if $|a_i| \le 2.4\,\mathrm{m/s^2}$ and $|\delta_i| \le 0.3375\,\mathrm{rad}$, else $\max(\eta, 1 - 0.5 w_{\mathrm{smooth}})$ (equals 0.5 at default $w_{\mathrm{smooth}} = 1$). With lower/upper clearances $c_i^{\mathrm{lo}}, c_i^{\mathrm{hi}}$ and $d_{\mathrm{out},i} = \max(0, -\min(c_i^{\mathrm{lo}}, c_i^{\mathrm{hi}}))$,

$$R_{\mathrm{boundary}} = \begin{cases} 0, & d_{\mathrm{out},i} = 0, \\ -(C_{\mathrm{out}} + c_{\mathrm{depth}} \min(d_{\mathrm{out},i}, d_{\max})), & d_{\mathrm{out},i} > 0, \end{cases}$$

using the constants in Table 6. Waiting has no additive bonus; contact duration cost is 0 and terminal onset is 1.

Accept-if-better gate terms. For candidate a, $D(a) \in \{0,1\}$ is one-step road feasibility (bicycle rollout and braking backup stay on-road); $N(a) \in \{0,1\}$ is OBB-safety under the shared 1.5 s/4-substep constant-velocity predictor; $P(a) = \mathrm{clip}((s'(a) - s)/(v_{\max}\Delta t), 0, 1)$; $T(a) = \min_j \mathrm{clip}(g_j(a)/(2\,\mathrm{m}), 0, 1)$ (or 1 if no neighbor), with g_j the footprint gap; $C(a) = 1$ iff $|a| \le 2.4\,\mathrm{m/s^2}$ and $|\delta| \le 0.3375\,\mathrm{rad}$. The accept-if-better rule is applied at deterministic evaluation only: rollouts execute a_{res} iff $S_c(a_{\mathrm{res}}) > S_c(a_{\mathrm{prior}})$ (Eq. 19); training trajectories are not rewritten, so PPO remains on-policy.

Episodic PDMS and checkpoint labels. Episodic PDMS uses $\mathrm{NC} \cdot \mathrm{DAC} \cdot (5\,\mathrm{EP} + 5\,\mathrm{TTC} + 2\,\mathrm{C})/12$. $\mathrm{TTC} = 1 - \mathrm{unsafe_ttc_rate}/0.05$ clipped to $[0,1]$, where a pair is unsafe if its disc-approximated TTC is below 1.5 s. $\mathrm{C} = 1$ iff mean $|a| \le 2.4\,\mathrm{m/s^2}$, RMS jerk $\le 8.37\,\mathrm{m/s^3}$, and mean $|\delta| \le 0.10\,\mathrm{rad}$ (Table 6); otherwise $\mathrm{C} = 0$. Failed exports that never improve on the frozen prior under the safety non-regression rule are labeled `utility_fallback` (the prior checkpoint is retained).

Observation features. Ego terms are station/lateral offset (s_i, n_i), speed, heading errors to the local tangent and goal, clearances to lower/upper corridor bounds, and remaining station. Neighbor terms are relative pose and velocity in the ego body frame for up to K nearest agents within R_p (Table 6).

Uncertainty and pairwise tests. Displayed row intervals are descriptive: frozen controllers resample scenarios; learned controllers use three seed-level means, so widths are not commensurate across row types. Model-vs-model p-values and difference intervals use a paired two-level bootstrap: resample the shared scene set jointly, and resample training seeds within each learned model (a frozen controller is the one-seed degenerate case).

F DATASETS

The main text evaluates residual MARL on a single dense interactive freeway corridor. The discrete prior itself is geometry-aware rather than lane-ID-based, so the same five-term family 4 can be instantiated on qualitatively different roadways if the directional, proximity, and path terms are evaluated in a site-appropriate spatial frame. This appendix documents three naturalistic environments used for that purpose (Fig. 7): the aerial freeway corridor that supports the main residual-learning experiments, a compact multi-arm roundabout near the same city, and the TGSIM Foggy Bottom urban street network in Washington, DC. Residual MARL, planner comparatives, and the dense-stress protocol in the main text use the freeway corridor; the two urban recordings are used for per-site prior calibration and for the matched closed-loop transfer runs in Appendix H.

Table 8 summarizes the calibration-ready samples. All three sources are subsampled to $\Delta t \approx 0.1$ s. For the two urban sites we drop any track that is shorter than 10 s, whose 85th-percentile speed is

Table 7: Residual PPO trainer and MAPPO baseline optimization settings.

Residual MARL (shared PPO)	**Value**	**MAPPO baseline**	**Value**
Actor / critic MLP	$d_{\text{obs}} \to 128 \to 128$	Actor MLP	$d_{\text{obs}} \to 128 \to 128$
Actor output	Candidate logits (63-D; $\Delta\Theta$ ablation)	Critic input	$6N$ joint + d_{obs}
Critic input	Local o_i only	Critic MLP	$d_{\text{state}} \to 128 \to 128 \to 1$
Hidden width	128	Hidden width	128
Learning rate	3×10^{-4}	Actor LR / critic LR	3×10^{-4} / 10^{-3}
PPO clip ϵ	0.2	PPO clip ϵ	0.2
Value coef. c_v	0.5	Value coef. c_v	0.5
Entropy coef. c_H	0.0	Entropy coef. c_H	0.01
GAE λ	0.95	GAE λ	0.95
Discount γ	0.95	Discount γ	0.95
PPO epochs / update	4	PPO / critic epochs	4 / 4
Minibatch size	512	Minibatch size	256
Episodes / update	4	Episodes / update	4
Gradient clip	0.5	Gradient clip	0.5
Policy updates	24 (paper tables)	Policy updates	24 (paper tables)
Residual scale α / temp. T	0.5 / 0.005	Init. $\log \sigma$ (actor)	-1.6
Contact terminal / duration	1 / 0	Contact terminal / duration	1 / 0
Selection	PDMS + safety non-regression	Selection	PDMS + safety non-regression
Observation normalization	**Scale**	**$\Delta\Theta$ ablation tanh range**	**Max $\lvert\Delta\rvert$**
s_i / remaining s	$/L_{\text{corridor}}$	$\Delta S_v, \Delta S_\theta$	1.5
n_i	$/8$	ΔS_d	2.0
v_i	$/16$	Δw_c	250
$\tilde{\theta}_i, \tilde{\theta}_{\text{goal},i}$	$/\pi$	$\Delta\xi_i, \Delta\gamma$	1.0
$d_{\text{lower}}, d_{\text{upper}}$	$/12$	Δw_ℓ	15
$\Delta x^{\text{body}}, \Delta y^{\text{body}}$	$/60$, $/12$	$\Delta\sigma_\parallel, \Delta\sigma_\perp$	1.5, 0.6
$\Delta v_x^{\text{body}}, \Delta v_y^{\text{body}}$	$/16$		

Θ clip ranges ($\Delta\Theta$ ablation only): $S_\theta, S_v, S_d \in [0.05, 12]$; $\xi_i \in [1.1, 5]$; $\gamma \in [0.05, 10]$; $w_c \in [0.01, 1200]$; $w_\ell \in [0.1, 300]$; $w_x, w_y \in [0.05, 15]$; $\beta \in [0.01, 12]$; $\sigma_\parallel \in [0.3, 6]$ m; $\sigma_\perp \in [0.2, 3]$ m. The parameters (w_x, w_y, β) remain fixed at their respective values in θ_{rob}.

Table 8: Calibration-ready trajectory samples. Counts are ego-track counts; durations and speeds are medians over ego tracks.

Site	**Tracks**	**Med. dur. (s)**	**Med. speed (m/s)**	v_{85}/v_{99} **(m/s)**	**Extent (m)**	**Geometry**
Lebanon Highway	368	57.8	8.15	11.6 / 16.3	505×433	PCA tubes (2 rec. $\times$ 2 streams)
Jounieh roundabout	479	24.4	1.88	2.70 / 3.89	56×31	Outer curb + 2 islands
TGSIM Foggy Bottom	2773	25.2	1.61	9.59 / 16.3	197×224	Union street curb

below 1 m/s, or whose travelled path is shorter than 8 m, so that parked and fragmentary objects never enter the ego pool or the neighbor set. The freeway calibration uses passenger vehicles (class 1) with valid kinematic transitions, as in Section 4. Urban calibrations use the dominant moving class at each site (class 2 at the roundabout; class 3 passenger cars at Foggy Bottom). Neighbor footprints default to $4.5\,\text{m} \times 1.8\,\text{m}$ unless a site file supplies length/width.

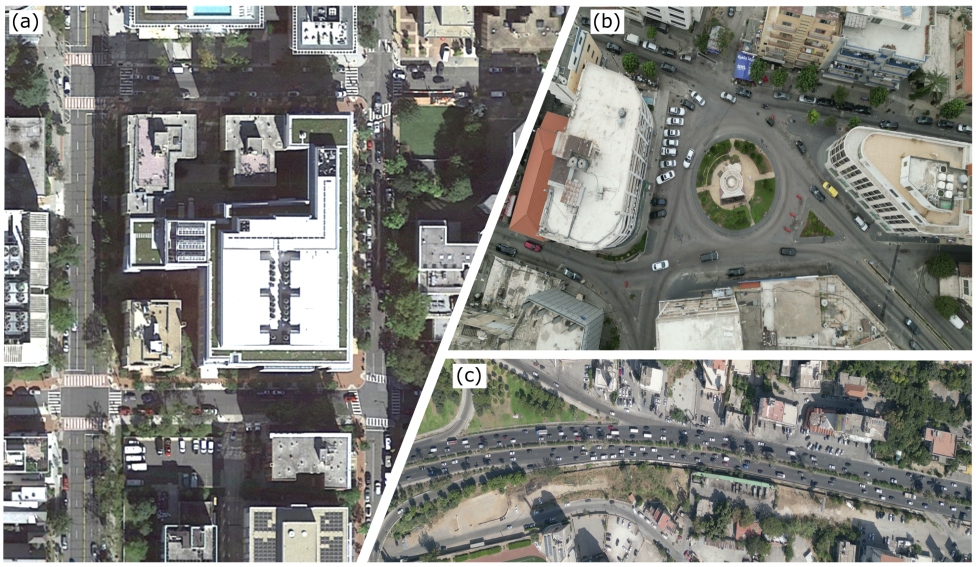

Figure 7: Aerial frame of the dataset environments; (a) TGSIM, (b) Roundabout, and (c) Freeway

F.1 Freeway corridor

The primary closed-loop domain is a multilane urban freeway near Jounieh, Lebanon, recorded by a near-stationary drone under weak discrete-lane cues (Fig. 7). Extraction details are given by Khoueiry et al. (2026). Two complementary recordings (`run_id`$\in \{1, 2\}$) cover the same curved corridor in opposite observational conditions; each recording is reconstructed into two Kalman-filter movement streams (`lane_kf`$\in \{1, 2\}$). Rather than relying on painted lane IDs, the spatial reference for scoring is a data-driven PCA tube per (`run_id`, `lane_kf`): a centerline for Frenet/direction terms and lower/upper bounds for U_{path}. Class-1 passenger trajectories (190,767 valid kinematic transitions from 368 tracks; median speed 8.15 m/s) form the ego pool for the main-text prior fit. This is the only environment on which the residual policy is trained and benchmarked in the main text; matched closed-loop transfer on the urban sites is deferred to Appendix H.

F.2 Jounieh Roundabout

A second Lebanon aerial collection covers a compact multi-arm roundabout (about $56\,\text{m} \times 31\,\text{m}$) rather than a linear corridor. Circulation wraps a large central island and a smaller splitter island; several entries and exits produce crossing and weaving at low speed (median 1.88 m/s; 85th percentile 2.70 m/s). A single roadway polygon, outer curb minus the two islands is used for every vehicle. After the duration/speed/path filter, 479 class-2 tracks remain for calibration. Because there is no unique downstream tangent, directional and proximity terms cannot be taken from a PCA corridor; they are instead destination-based (Section G).

F.3 TGSIM Foggy Bottom

The third environment is the publicly released TGSIM Foggy Bottom trajectory set Ammourah et al. (2025), collected with overlapping infrastructure cameras around the George Washington University campus in Washington, DC. The reconstructed street graph spans roughly $200\,\text{m} \times 230\,\text{m}$ of signalized and stop-controlled urban links with right-angle turns. Provided boundary polygons are unioned into one curb (exterior plus interior holes) that is shared by all IDs. The prepared file retains mixed road-user types after dropping parked/short tracks; we calibrate on class-3 passenger cars (2773 tracks). Speeds are strongly bimodal: the median (1.61 m/s) reflects queues and turning, while the 85th/99th percentiles (9.59 / 16.3 m/s) capture unconstrained urban running. As at the roundabout, each track's last observed position is treated as a destination on the roadway polygon.

Table 9: Search bounds used in utility calibration. Where a cell lists three values, they are Highway / Jounieh / TGSIM.

Parameter	Low	High
S_θ, S_v	0.05	10
ξ_i	1.1	5
S_d	0.05	12
γ, β	0.1 / 0.01	10
w_x, w_y	0.1	10
w_c	0.01	800 / 250 / 200
w_ℓ	0.1	300
$\sigma_\parallel$ (m)	0.5 / 0.3 / 0.5	5 / 5 / 6
$\sigma_\perp$ (m)	0.3 / 0.2 / 0.3	2.5 / 2.5 / 3

Table 10: Site-wise utility calibration summary. Test/validation L_{cl} are closed-loop tracking losses on held-out vehicles ($H = 8$). Rank is the mean one-step rank of the observed-like candidate for the working prior during search. Working source is the validation-selected rule.

Site	Frame	Test L_{cl}	Val L_{cl}	Rank	$\lvert\mathcal{A}\rvert$	Near-opt.	Working source
Lebanon Highway	corridor	0.787	0.681	11.0	63	10	best-on-val
Jounieh roundabout	destination + curb	1.375	1.553	29.7	63	31	best-on-val
TGSIM Foggy Bottom	destination + curb	1.241	1.033	28.0	63	3	robust-on-val

G Utility Calibration Results

This appendix expands Section 4: search settings, fit summaries, working parameters, parameter distributions, and sample multi-vehicle snapshots for the freeway corridor, the Jounieh roundabout, and the TGSIM Foggy Bottom urban grid.

Objective and search. Calibration minimizes the composite loss

$$J(\theta) = \lambda_{\mathrm{cl}}\, L_{\mathrm{cl}}(\theta) + \lambda_{\mathrm{trk}}\, \frac{\bar{r}(\theta)}{r_0} + \lambda_{\mathrm{nll}}\, L_{\mathrm{nll}}(\theta), \tag{47}$$

where L_{cl} is mean closed-loop tracking error on short windows ($H = 8$), $\bar{r}$ is the mean one-step rank of the observed-like candidate, $r_0 = |\mathcal{G}|$ normalizes rank, and L_{nll} is softmax NLL at temperature $T = 0.25$. We use $\lambda_{\mathrm{cl}} = 1$, $\lambda_{\mathrm{trk}} = 1$, and $\lambda_{\mathrm{nll}} = 0.1$. Parameters are drawn uniformly in the site boxes of Table 9. Every site uses 400 trials $\times$ 3 restarts with 240 closed-loop-scored candidates on the shared $7\times9 = 63$ action set. Urban sites additionally use per-track desired/max speeds (site p_{99} speed caps) and a 2 m destination hold.

Corridor versus destination-plus-curb. On the freeway, $\hat{t}_i$ is the local PCA-tube tangent and U_{dist} is a Frenet distance to a short-horizon reference; U_{path} saturates outside the lower/upper PCA bounds. On the roundabout and urban grid, each ego ID uses destination g_i equal to its last observed point. Heading and proximity follow the on-road path toward g_i on the site curb polygon, and path cost grows with off-road distance. Speed and collision kernels keep the same forms at every site.

Fit summary. Held-out tracking is strongest on the freeway (test $L_{\mathrm{cl}} = 0.787$), where a unique downstream tangent exists. Foggy Bottom is next (1.241), and the roundabout is the weakest of the three (1.375) under low-speed circulating geometry. One-step ranking of the observed-like candidate is informative on the freeway (mean rank 11.0 of 63) but near mid-pack on the urban sites (ranks 29.7 and 28.0; Table 10). Soft margins still concentrate near zero on urban sites, so the observed action is often a near-maximizer even when it is not uniquely ranked first. All three fits are weakly identified (near-optimal clouds of size 10, 31, and 3 out of 240 scored trials).

Table 11: Working utility parameters θ at each site, with mean/median of the closed-loop scored trial cloud in parentheses (mean / median). Highway and Jounieh working points are best-on-validation; TGSIM is robust-on-validation.

Parameter	Highway	Jounieh	TGSIM
S_θ	4.13 (5.02 / 5.02)	3.53 (4.75 / 4.63)	4.13 (4.97 / 4.90)
S_v	2.72 (4.73 / 4.63)	0.32 (4.72 / 4.70)	2.52 (4.91 / 4.93)
ξ_i	4.72 (2.97 / 2.91)	1.45 (3.01 / 2.96)	4.58 (3.08 / 3.02)
S_d	11.3 (5.96 / 6.15)	0.62 (5.88 / 5.97)	8.78 (5.93 / 6.07)
γ	0.56 (5.16 / 5.58)	3.31 (5.03 / 5.11)	0.80 (5.25 / 5.65)
w_x	4.85 (5.16 / 5.20)	9.85 (4.80 / 4.53)	0.56 (4.86 / 4.70)
w_y	2.45 (5.03 / 4.82)	1.50 (5.18 / 5.07)	2.45 (5.11 / 5.00)
w_c	22.8 (82.1 / 72.7)	51.8 (94.0 / 81.4)	3.26 (32.4 / 29.0)
w_ℓ	24.1 (138 / 122)	11.1 (43.7 / 37.0)	24.1 (92.0 / 79.6)
β	3.37 (4.92 / 4.87)	5.49 (4.29 / 3.60)	4.08 (4.69 / 4.68)
$\sigma_\parallel$ (m)	0.65 (2.82 / 2.76)	4.23 (2.71 / 2.66)	4.03 (3.27 / 3.11)
$\sigma_\perp$ (m)	0.58 (1.35 / 1.32)	0.49 (1.41 / 1.48)	1.25 (1.61 / 1.57)

G.1 Freeway corridor

Fig. 8 summarizes one-step and closed-loop diagnostics for the freeway prior. Fig. 9 overlays scored-trial KDEs across all three sites; Fig. 10 reports pairwise near-optimal tradeoffs for the freeway fit. Fig. 11 shows recorded-ego snapshot sequences on extracted curb outlines for all three sites.

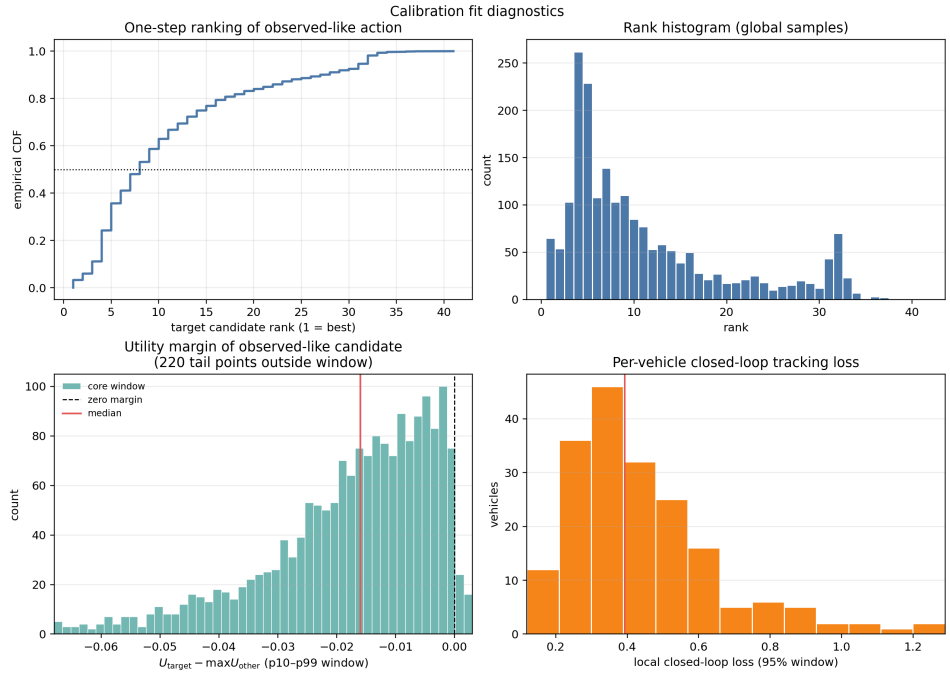

Figure 8: Freeway calibration fit diagnostics: empirical CDF and histogram of the observed-like candidate rank, utility margin relative to the best alternative, and per-vehicle closed-loop tracking loss.

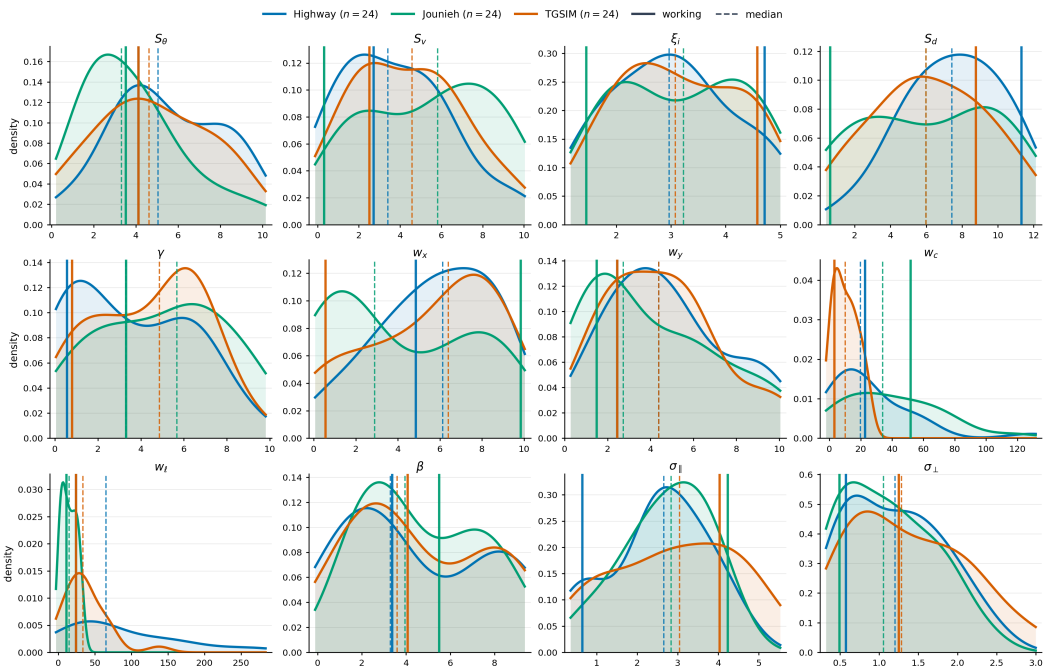

Figure 9: Utility-parameter distributions at all three sites from the best 10% of closed-loop scored trials by calibration objective (n=24 per site). Curves are Gaussian KDEs; solid vertical lines are each site's working prior and dashed lines are the cloud medians.

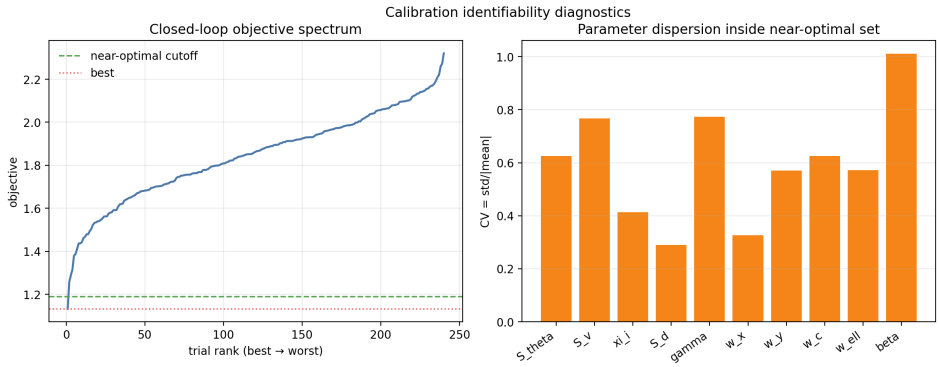

Figure 10: Identifiability diagnostics for the twelve-parameter freeway fit: pairwise tradeoffs in the near-optimal cloud and coefficient-of-variation across coordinates.

G.2 Jounieh roundabout

The destination-plus-curb encoding is shown in the middle row of Fig. 11: the ego enters the circulating roadway while neighboring vehicles occupy both the ring and the approach arms. Parameter distributions for this site appear with the other two fits in Fig. 9.

G.3 TGSIM Foggy Bottom urban grid

The bottom row of Fig. 11 shows a northbound ego approaching the northern T-intersection on the lane-marked curb map. Foggy Bottom parameter distributions appear in Fig. 9.

Scope. These results support transferring the five-term prior across freeway, roundabout, and urban-grid geometries *provided* the spatial frame is rewritten for the site. They do not show that

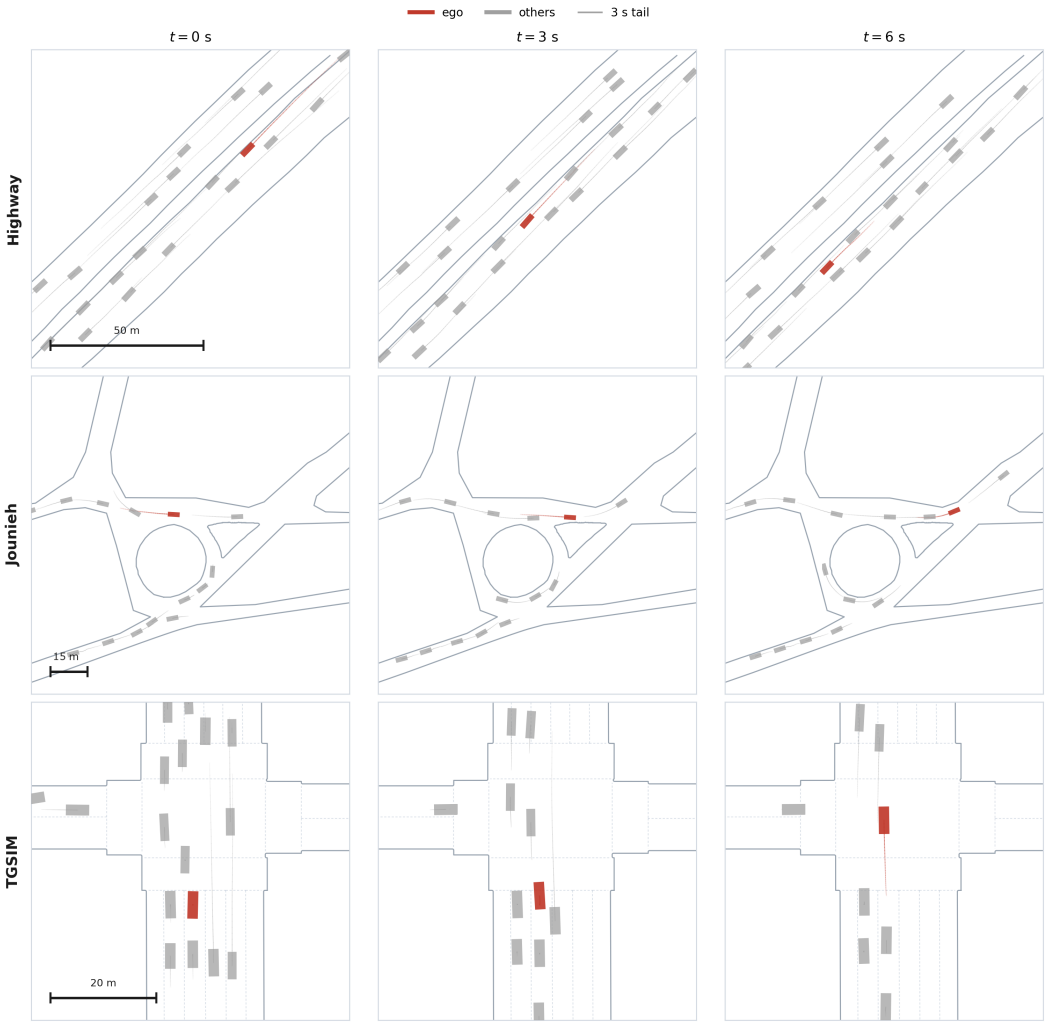

Figure 11: Recorded-ego snapshots on extracted curb outlines at $t = 0$, 3, and 6 s, with a 3 s tail.

one numerical θ is universal, and they are not a substitute for the residual-MARL evaluation in the main text.

H Additional Experiments

This appendix reports matched closed-loop sparse benchmarks, gate ablations, and parameter/stress suites on both additional sites (the Jounieh roundabout and TGSIM Foggy Bottom) including local CtRL-Sim and CTG++ adaptations on the sparse benchmarks. Each site uses its own frozen working prior from Table 11 under the destination-plus-curb spatial frame of Appendix G. Training budgets, seeds, metrics, and selection rules follow the freeway recipe in Section 5 unless a site-specific spawn or observation detail is noted. These transfer results extend empirical coverage to three distinct road geometries; the main-text optimization and ablation claims remain anchored to the freeway protocol.

H.1 Jounieh roundabout

Status. Training, selection, and the matched 20-scene suites are complete for residual MARL, $\Delta\Theta$, matched discrete RL, MAPPO, CtRL-Sim, and CTG++ (three seeds): sparse benchmark (Table 13), gate ablation (Table 14), and parameter/no-lookahead/dense-stress (Table 15). Table 12

Table 12: Jounieh roundabout *training validation* means (16 validation episodes, 6 agents, 80 steps). Learned rows average three seeds. These checkpoint-selection values are separate from the matched 20-scene test protocol in Table 13. Off-road rate is 0 for every row.

Model	PDMS ↑	Coll. ↓	Arrival ↑	Progress ↑
Utility prior	0.666	0.000	0.510	0.671
Residual MARL	0.718	0.000	0.507	0.675
Parameter residual $\Delta\Theta$	0.666	0.000	0.510	0.671
CtRL-Sim	0.491	0.094	0.120	0.363
Direct discrete RL	0.259	1.02	0.208	0.291
MAPPO	0.410	0.167	0.156	0.287

Table 13: Jounieh roundabout sparse closed-loop comparison (20 matched scenes; three seeds for learned models). Values are means; off-road rate is 0 for every row. Pairwise claims use the paired two-level seed-and-scenario bootstrap. Frenet travel time is undefined because arrival is 0.

Model	PDMS ↑	Coll. ↓	Progress ↑	Arrival ↑	Time ↓	Jerk (m/s^3) ↓
ORCA	0.740	0.100	0.771	0.542	8.47	3.92
Social force	0.224	0.850	0.541	0.317	6.83	3.54
DWA	0.406	0.200	0.554	0.467	11.99	2.42
MPPI	0.710	0.200	0.733	0.600	8.30	3.21
Frenet planner	0.492	0.000	0.328	0.000	—	1.35
Utility prior	0.602	0.100	0.713	0.450	11.24	3.74
Direct discrete RL	0.210	1.683	0.281	0.225	6.75	2.48
MAPPO	0.419	0.183	0.330	0.211	8.67	1.46
CtRL-Sim	0.472	0.100	0.370	0.136	9.31	2.16
CTG++	0.759	0.050	0.644	0.269	8.68	0.68
Residual MARL	0.671	0.067	0.714	0.500	11.75	3.82

retains the training-time validation readout used during checkpoint selection (16 episodes, 6 agents, 80 steps).

Validation readout. On the selection validation set, residual MARL improves mean PDMS over the frozen prior (0.718 vs. 0.666) with zero mean collision events and nearly matched arrival (0.507 vs. 0.510). The $\Delta\Theta$ checkpoint matches the prior on this readout. MAPPO, matched discrete RL, and the local CtRL-Sim adaptation trail on mean validation PDMS. These validation values document model selection only; matched test results are Tables 13–15.

Matched sparse benchmark. Table 13 reports the completed 20-scene sparse roundabout comparison. The local CTG++ adaptation attains the highest mean PDMS (0.759) with low contact (0.050), followed by ORCA (0.740), MPPI (0.710), and residual MARL (0.671). Relative to the frozen utility prior, residual MARL increases mean PDMS (0.602 $\rightarrow$ 0.671), decreases collisions (0.100 $\rightarrow$ 0.067), and increases arrival (0.450 $\rightarrow$ 0.500), but the paired two-level bootstrap does not resolve the PDMS or collision differences at 20 scenes (p=0.140 and 0.614). Residual MARL has significantly higher PDMS than direct discrete RL, MAPPO, CtRL-Sim, DWA, social force, and the Frenet planner; its PDMS difference from ORCA, MPPI, and CTG++ is not significant. These results show useful transfer, with a competitive offline local diffusion adaptation on this compact geometry.

Gate ablation. Table 14 isolates residual learning from the accept-if-better gate on the same matched 20-scene suite as Table 13. Residual+gate matches the sparse-benchmark residual (PDMS 0.671, 0.067 collisions) and improves mean PDMS over the prior (0.602 $\rightarrow$ 0.671); removing the gate leaves the trained residual unchanged on this suite. A random residual with the same gate trails the trained method (PDMS 0.614) and stays near the prior. Direct discrete RL collapses (PDMS 0.210, 1.683 collisions).

Parameter residual, no-lookahead, and dense stress. Table 15 mirrors the freeway/Foggy Bottom package. On sparse full lookahead (same matched scenes as Table 13), $\Delta\Theta$ and its masks collapse to the prior (PDMS 0.602), while candidate-logit residual matches the sparse-benchmark

Table 14: Jounieh roundabout gate vs. residual-learning ablation (same 20 matched scenes and three seeds as Table 13). Off-road rate is 0 for every row.

Model	PDMS ↑	Coll. ↓	Arrival ↑	Jerk (m/s^3) ↓
Utility prior	0.602	0.100	0.450	3.74
Residual, no gate	0.671	0.067	0.500	3.82
Residual + gate	0.671	0.067	0.500	3.82
Random residual + gate	0.614	0.067	0.428	3.30
Direct discrete RL	0.210	1.683	0.225	2.48

Table 15: Jounieh roundabout parameter-residual, no-lookahead, and dense-stress ablations (20 scenes, three seeds). Off-road rate is 0 in every cell.

	Sparse, full LA		Sparse, no LA		Dense stress	
Model	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓
Utility prior	0.602	0.100	0.594	0.050	0.660	0.000
Residual MARL (logits)	0.671	0.067	0.618	0.083	0.714	0.000
Parameter residual $\Delta\Theta$	0.602	0.100	0.594	0.050	0.660	0.000
Weights only	0.602	0.100	0.594	0.050	0.660	0.000
σ only	0.602	0.100	0.594	0.050	0.660	0.000
Direct discrete RL	0.210	1.683	0.145	2.00	0.142	1.85
MAPPO	0.419	0.183	0.195	1.27	0.385	0.400

residual (PDMS 0.671 / 0.067 collisions). Without OBB lookahead the same ordering holds at slightly lower scores. Under dense stress, residual MARL is strongest (PDMS 0.714, zero collisions) versus the prior and $\Delta\Theta$ family (0.660 / 0). Direct discrete RL and MAPPO remain far behind on every suite.

H.2 TGSIM Foggy Bottom

Protocol. Foggy Bottom uses the destination-plus-curb frame and the site working prior from Table 11. We report three matched suites parallel to the freeway: the sparse planner benchmark (Table 16), the gate vs. residual-learning ablation (Table 17), and the $\Delta\Theta$ / no-lookahead / dense-stress package (Table 18). Absolute travel times are much shorter than on the freeway corridor, and geometric avoiders are more competitive when destinations are close.

Sparse benchmark. Table 16 is the completed 20-scene sparse transfer (three seeds for learned models). ORCA attains the highest PDMS (0.924) with zero collisions. The frozen utility prior is next (0.872), and residual MARL follows (0.812) with near-zero contact (0.017). Under the same low online-training budget, residual MARL has higher PDMS and fewer collisions than MAPPO and matched discrete RL; it also exceeds the local CtRL-Sim (0.546) and CTG++ (0.571) adaptations on PDMS while keeping contact near zero.

Gate ablation. Table 17 isolates residual learning from the accept-if-better gate on the same matched 20-scene / 3-seed suite as Table 16. Under the matched low budget, direct discrete RL obtains PDMS 0.357 with 0.633 collisions. A randomly initialized residual with the same gate trails the trained method (PDMS 0.701 vs. 0.812). Removing the gate from the trained residual leaves PDMS unchanged at the sparse-benchmark residual value (0.812). Versus the frozen prior on this suite, residual MARL is lower in mean PDMS (0.872 $\rightarrow$ 0.812) with a small collision increase (0.000 $\rightarrow$ 0.017); arrival moves slightly against the residual (0.908 $\rightarrow$ 0.858).

Parameter residual, no-lookahead, and dense stress. Table 18 mirrors the freeway package. On sparse full lookahead (same matched scenes as Table 16), $\Delta\Theta$ / weights-only / σ-only remain at the prior (PDMS 0.872), while candidate-logit residual matches the sparse-benchmark residual (0.812 / 0.017 collisions). Removing the OBB lookahead leaves the utility-family rows near the prior; candidate-logit residual stays essentially tied with that suite (0.818 vs. 0.813). Under dense stress, candidate-logit residual leads (PDMS 0.850, 0.033 collisions) versus the prior (0.832 / 0.050). Direct discrete RL and MAPPO remain far behind.

Table 16: TGSIM Foggy Bottom sparse closed-loop comparison (20 scenes; three seeds for learned models). PDMS/collisions/arrival/time/jerk show means with descriptive row intervals; progress is the mean goal-progress fraction. Off-road rate is 0 for every model.

Model	PDMS ↑	Coll. ↓	Progress ↑	Arrival ↑	Time ↓	Jerk (m/s^3) ↓
ORCA	0.924 [0.885, 0.959]	0.000 [0.000, 0.000]	0.942	0.925 [0.892, 0.958]	11.5 [10.0, 13.0]	3.04 [2.77, 3.31]
Social force	0.790 [0.723, 0.856]	0.000 [0.000, 0.000]	0.896	0.858 [0.783, 0.925]	11.6 [9.9, 13.3]	2.93 [2.61, 3.29]
DWA	0.589 [0.478, 0.685]	0.100 [0.000, 0.250]	0.612	0.425 [0.350, 0.500]	12.1 [9.9, 14.4]	2.80 [2.61, 2.97]
MPPI	0.737 [0.607, 0.841]	0.100 [0.000, 0.250]	0.784	0.558 [0.458, 0.658]	11.6 [9.2, 14.1]	2.20 [2.05, 2.36]
Frenet planner	0.432 [0.334, 0.526]	0.100 [0.000, 0.300]	0.509	0.025 [0.000, 0.050]	9.0 [7.5, 11.5]	3.76 [3.27, 4.29]
Utility prior	0.872 [0.819, 0.920]	0.000 [0.000, 0.000]	0.909	0.908 [0.858, 0.958]	10.9 [9.5, 12.3]	4.04 [3.72, 4.35]
Direct discrete RL	0.357 [0.284, 0.494]	0.633 [0.200, 0.950]	0.313	0.225 [0.167, 0.275]	6.0 [4.9, 8.0]	2.47 [2.08, 2.78]
MAPPO	0.434 [0.345, 0.517]	0.467 [0.250, 0.800]	0.362	0.161 [0.125, 0.200]	7.8 [5.0, 10.8]	1.27 [1.14, 1.38]
CtRL-Sim	0.546 [0.477, 0.590]	0.183 [0.100, 0.350]	0.523	0.194 [0.183, 0.217]	11.9 [7.5, 15.7]	2.34 [2.20, 2.44]
CTG++	0.571 [0.536, 0.591]	0.267 [0.150, 0.450]	0.561	0.125 [0.067, 0.167]	8.8 [7.6, 10.3]	0.86 [0.74, 0.98]
Residual MARL	0.812 [0.799, 0.827]	0.017 [0.000, 0.050]	0.865	0.858 [0.842, 0.883]	11.1 [10.9, 11.2]	3.93 [3.79, 4.01]

Table 17: TGSIM Foggy Bottom gate vs. residual-learning ablation (same 20 matched scenes and three seeds as Table 16). Intervals are 95% bootstrap CIs where reported in the sparse benchmark. Off-road rate is 0 for every row.

Model	PDMS ↑	Coll. ↓	Arrival ↑	Jerk (m/s^3) ↓
Utility prior	0.872 [0.819, 0.920]	0.000 [0.000, 0.000]	0.908 [0.858, 0.958]	4.04 [3.72, 4.35]
Residual, no gate	0.812 [0.799, 0.827]	0.017 [0.000, 0.050]	0.858 [0.842, 0.883]	3.93 [3.79, 4.01]
Residual + gate	0.812 [0.799, 0.827]	0.017 [0.000, 0.050]	0.858 [0.842, 0.883]	3.93 [3.79, 4.01]
Random residual + gate	0.701 [0.659, 0.744]	0.017 [0.000, 0.050]	0.736 [0.667, 0.808]	3.84 [3.33, 4.20]
Direct discrete RL	0.357 [0.284, 0.494]	0.633 [0.200, 0.950]	0.225 [0.167, 0.275]	2.47 [2.08, 2.78]

Table 18: TGSIM Foggy Bottom parameter-residual, no-lookahead, and dense-stress ablations (20 scenes, three seeds). Off-road rate is 0 in every cell.

	Sparse, full LA		Sparse, no LA		Dense stress	
Model	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓	PDMS ↑	Coll. ↓
Utility prior	0.872	0.000	0.813	0.050	0.832	0.050
Residual MARL (logits)	0.812	0.017	0.818	0.050	0.850	0.033
Parameter residual $\Delta\Theta$	0.872	0.000	0.838	0.017	0.825	0.050
Weights only	0.872	0.000	0.827	0.033	0.832	0.050
σ only	0.872	0.000	0.853	0.000	0.821	0.050
Direct discrete RL	0.357	0.633	0.227	2.08	0.234	2.62
MAPPO	0.434	0.467	0.209	1.30	0.323	0.767

I ADDITIONAL FIGURES

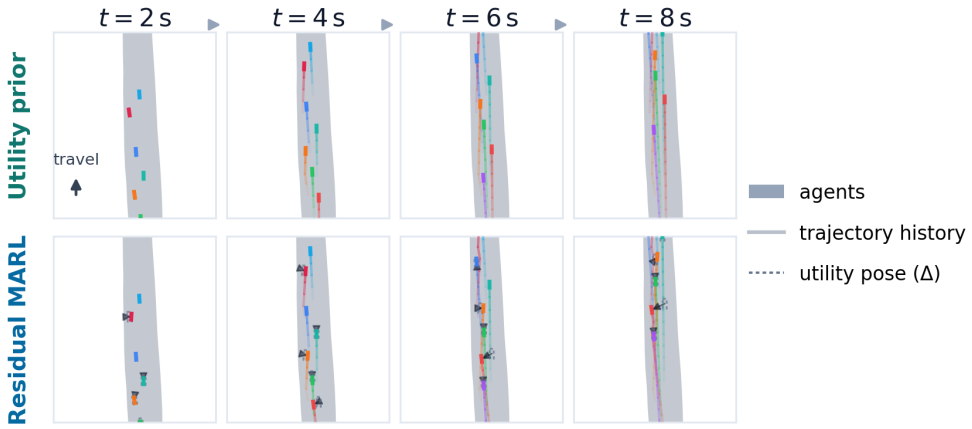

Figure 12: Closed-loop rollouts over time on the reconstructed freeway corridor (road-aligned view). Top: frozen utility prior. Bottom: residual MARL. Columns show the same scene at t=2, 4, 6, 8 s under a fixed camera spanning agent travel from the first to the last panel ($\pm$10 m). Agents share unique colors; translucent trails are trajectory history. On the residual row, dashed outlines and arrows mark poses that diverge from the utility proposal (accept-if-better Δ).

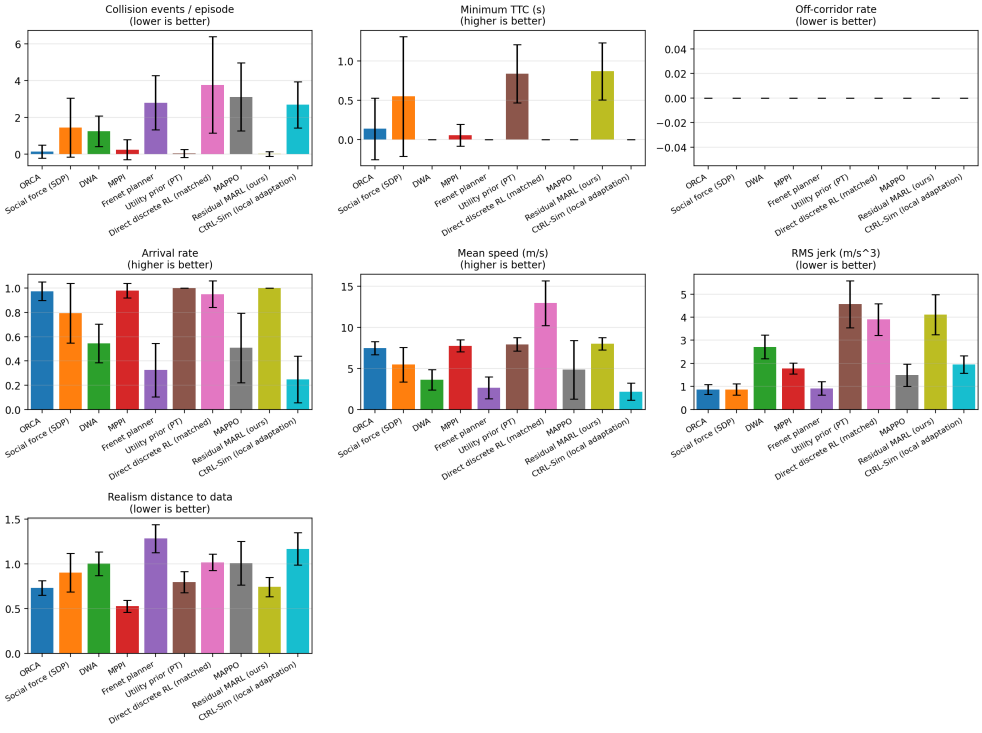

Figure 13: Sparse protocol metric panel including MAPPO and CtRL-Sim. Under the matched low online-training budget, residual MARL and the utility prior have higher PDMS and fewer collisions than MAPPO and matched discrete RL; the local CtRL-Sim adaptation also records lower PDMS and more collisions than residual MARL.

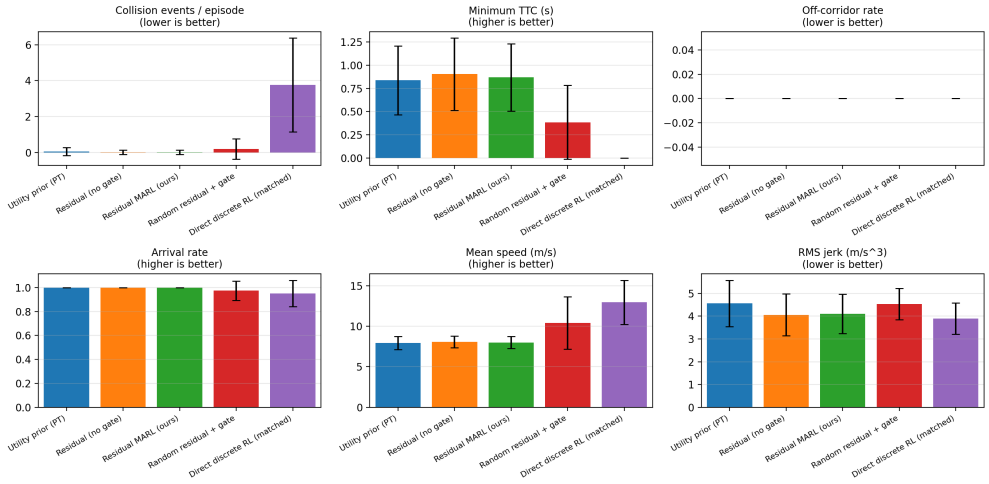

Figure 14: Gate vs. residual-learning ablation metrics. A random residual with the same accept-if-better gate does not recover the trained method, and can still underperform the frozen prior because one-step S_c is not episodic PDMS; removing the gate from the trained residual is essentially unchanged.

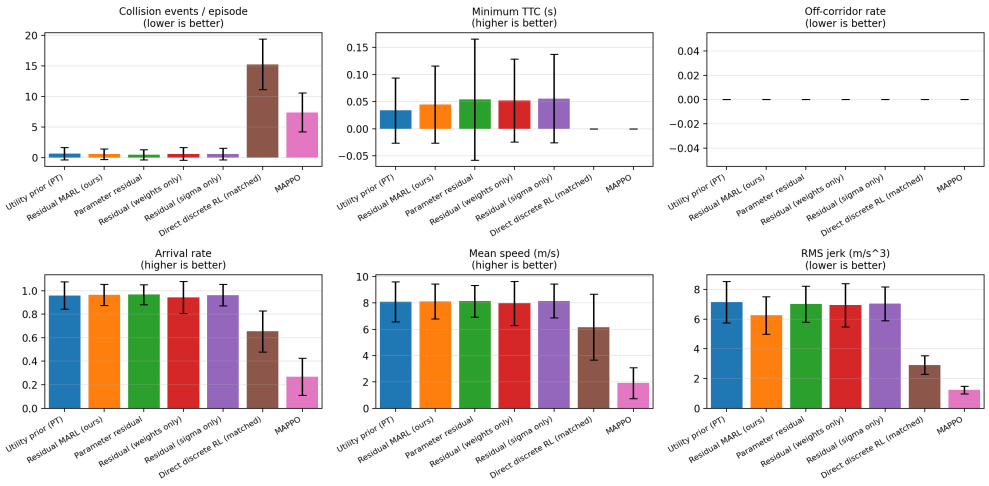

Figure 15: Dense 16-agent stress metric panel from the same protocol as Table 2. Candidate-logit and $\Delta\Theta$ residuals stay with the utility family; unconstrained MAPPO and matched discrete RL collide at high rates.

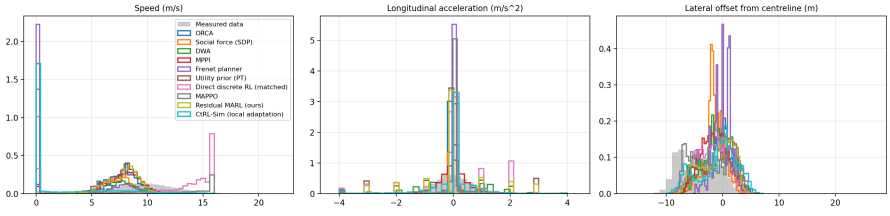

Figure 16: Realism panel on the sparse protocol: measured versus simulated speed, acceleration, and lateral-offset marginals. Residual MARL stays in the prior's distributional band; MAPPO and matched discrete RL put mass at the speed cap.