HIERARCHICAL SAFETY-ORIENTED DENSE RL WITH REFINED CREDIT ASSIGNMENT FOR TOKENIZED MULTI-AGENT TRAFFIC SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-fidelity simulation of safety-critical traffic interactions is essential for developing and evaluating autonomous driving systems. Tokenized autoregressive models achieve strong average realism in multi-agent traffic simulation through large-scale supervised pretraining, but safety-critical events such as collisions are rare in natural driving data, providing limited supervision for learning these interactions. Uniformly sampled reinforcement learning (RL) post-training also allocates many updates to ordinary scenarios, agents, and timesteps that provide little safety information, limiting the efficiency of safety-oriented learning. We propose a hierarchical safety-oriented dense learning framework for RL post-training of tokenized multi-agent traffic simulation models. Episode-level densification prioritizes scenarios whose current-policy probe produces a collision, while state-level masks localize updates to vehicle agent-time positions with a collision or physical near miss in at least one same-scenario rollout, retaining aligned safe outcomes for comparison. Our refined credit assignment retains physical rewards at agent-step granularity and compares aligned agent-time positions across same-scenario rollouts, yielding critic-free group-relative advantages with explicit temporal-agent resolution. Experiments with SMART on the Waymo Open Motion Dataset show that, across three training seeds, our method reduces mean collision rate from 3.394% to 2.914% relative to the uniformly sampled RL baseline using the same refined credit assignment, while maintaining a comparable average displacement error, achieving a favorable balance between safety-oriented alignment and per-rollout trajectory fidelity.

1 INTRODUCTION

Realistic and reactive traffic simulation is essential for developing and evaluating autonomous driving systems. Compared with real-world testing, simulation enables controlled and repeatable evaluation over substantially more traffic interactions without incurring the same safety risks and data-collection costs. Data-driven simulators improve upon non-reactive log replay and rigid rule-based agents by learning interactive behavior from large-scale driving data. The Waymo Open Sim Agents Challenge (WOSAC, Montali et al. (2023)) has further accelerated this development by establishing a standardized closed-loop simulation task.

Tokenized autoregressive models such as MotionLM (Seff et al. (2023)), Trajeglish (Philion et al. (2023)), and SMART (Wu et al. (2024)) discretize continuous motion and formulate multi-agent behavior generation as next-token prediction. Large-scale supervised pretraining allows these models to capture diverse motion patterns, while closed-loop supervised fine-tuning (Zhang et al. (2025)) and reinforcement learning (RL) post-training (Pei et al. (2025); Wang et al. (2026); Ahmadi et al. (2026)) further optimize behavior under model-induced rollout distributions. Nevertheless, both paradigms remain constrained by the scarcity of safety-critical interactions. Supervised objectives are dominated by frequent nominal behavior, whereas uniformly sampled RL spends most updates on ordinary scenarios, agents, and timesteps. Consequently, rare collisions and near misses provide insufficient training signals despite their importance to reliable closed-loop simulation.

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.

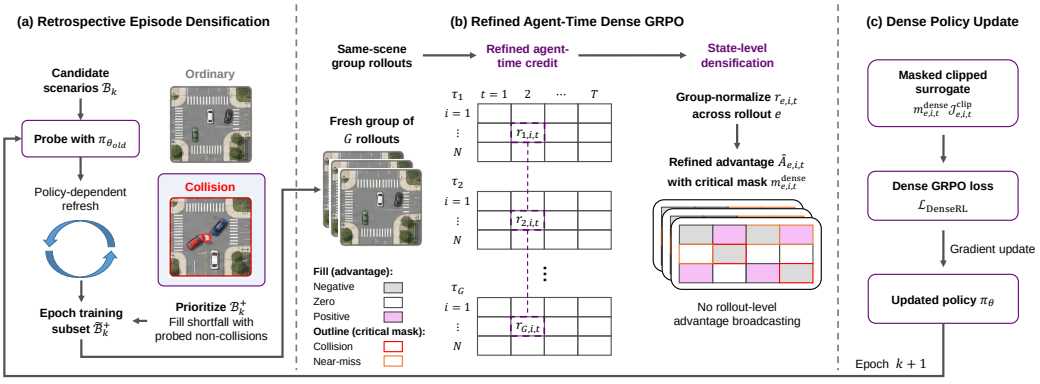

Figure 1: Hierarchical Dense RL: scenarios whose probes collide form $\mathcal{B}_k^+$. Probed noncollision scenarios fill the epoch training subset $\widetilde{\mathcal{B}}_k$ when needed. Fresh group of rollouts retain joint rollouts for aligned agent-time credit and episode-consistent state masking.

The central challenge is therefore to concentrate closed-loop optimization on interactions that provide meaningful safety information. This problem is related to the Curse of Rarity (CoR) (Liu & Feng (2024)) studied in Dense Learning (Feng et al. (2026)), where rare informative samples are overwhelmed by frequent samples with negligible contributions to the target policy gradient. In multi-agent traffic simulation, this sparsity is hierarchical: only a small fraction of scenarios contain safety-critical interactions, and only a subset of agents and generation steps participate in each interaction. Multi-agent densification is particularly challenging because informativeness must be identified jointly across the scenario, agent, and temporal dimensions while preserving the coupled multi-agent dynamics.

As summarized in Figure 1, *Hierarchical Dense RL* (*Dense RL* for short) adapts Dense Learning to safety-oriented post-training of tokenized multi-agent traffic simulators. Scenario selection prioritizes current-policy probe collisions. State masks retain vehicle agent-time positions critical in any same-scenario rollout, while refined credit compares aligned rewards across the group. The complete joint rollouts remain intact. Experiments on the Waymo Open Motion Dataset (WOMD, Ettinger et al. (2021)) demonstrate reduced collision rates with low average displacement error (ADE), yielding a favorable safety–fidelity trade-off.

Our main contributions are:

- We adapt Dense Learning to tokenized multi-agent post-training with policy-dependent collision mining and episode-consistent physical interaction masks, addressing scenario and agent–time sparsity while retaining joint transitions.
- We develop critic-free aligned agent–time credit that compares local physical rewards across same-scenario rollouts without broadcasting episode-level credit.
- We evaluate the safety–fidelity trade-off, credit granularity, and robustness across detector, reward, temporal, group-size, and seed settings.

2 Related Work

2.1 Data-Driven Traffic Simulation

Learning-based traffic simulation spans motion forecasting and closed-loop behavior generation, with WOSAC (Montali et al. (2023)) providing a standardized evaluation protocol. MTR (Shi et al. (2022)), MTR++ (Shi et al. (2024)), and MVTA (Wang et al. (2023)) predict multimodal trajectories via regression tasks. VBD (Huang et al. (2026)) and SceneDiffuser (Jiang et al. (2024); Tan et al. (2025)) model trajectory or scene distributions using generative models. TrafficSim (Suo et al. (2021)), SimNet (Bergamini et al. (2021)), and TrafficBots (Zhang et al. (2023b)) learn multi-agent, reactive, and configurable behavior, respectively.

MotionLM (Seff et al. (2023)) and Trajeglish (Philion et al. (2023)) formulate multi-agent motion generation as next-token prediction. SMART (Wu et al. (2024)) scales this formulation with joint map and motion tokenization, BehaviorGPT (Zhou et al. (2024)) uses next-patch prediction for efficiency, and CAT-K (Zhang et al. (2025)) addresses covariate shift through closed-loop supervised fine-tuning. These tokenized policies provide a natural interface for post-training with closed-loop feedback.

2.2 RL Post-Training

Proximal Policy Optimization (PPO, Schulman et al. (2017)) stabilizes policy updates through clipping, while GRPO (Shao et al. (2024)) supplies critic-free group-relative advantages. Our method combines these mechanisms. In traffic simulation, RTR (Zhang et al. (2023a)) combines imitation with factorized PPO, TrafficRLHF (Cao et al. (2024)) learns realism preferences, Peng et al. (Peng et al. (2024)) fine-tune autoregressive behavior with policy gradients, and RIFT (Chen et al. (2025)) uses group-relative fine-tuning for realism and controllability. Related ego-policy research studies RL fine-tuning of imitation, language-model, and diffusion planners (Lu et al. (2023); Tang et al. (2025); Li et al. (2026)), as well as expert-guided and safety-constrained optimization (Peng et al. (2022); Wen et al. (2020); Isele et al. (2018)).

For tokenized simulators, SMART-R1 (Pei et al. (2025)) performs metric-oriented post-training, R1Sim (Wang et al. (2026)) combines entropy-guided sampling with token-level safety-aware group-relative optimization and reference-policy regularization, and RLFTSim (Ahmadi et al. (2026)) derives a reward from WOSAC realism. These methods establish the value of token-level feedback and group-relative optimization, but do not define the policy-dependent collision-scenario mining and physical interaction-to-agent mask studied here. Our focus is the frequency and localization of sparse safety feedback, with ADE providing per-rollout fidelity supervision. Global realism objectives can instead benefit from rollout-level scalar rewards.

2.3 Dense Learning

D2RL (Feng et al. (2023)) addresses the curse of rarity by removing non-informative states and reconnecting critical states for efficient safety validation. Subsequent work studies adaptive scenario selection (Yang et al. (2025)) and implicit traffic distributions (Ren et al. (2026)), while Dense Learning (Feng et al. (2026)) broadens the informative-sample view to failures and successes that contribute to the target policy gradient.

Dense Learning concentrates optimization on informative samples, whereas reward densification increases feedback frequency. Our contribution is a task-specific adaptation that preserves complete joint rollouts while selecting safety-relevant scenarios and agent-time updates.

3 Methodology

3.1 Preliminaries

Tokenized multi-agent policy. Our model backbone is built upon SMART (Wu et al. (2024)), which discretizes continuous motion into a token vocabulary $\mathcal{V}$ and autoregressively predicts motion tokens for all simulated agents through agent motion policy π_θ, where θ is the trainable parameters. At timestep t, let $\mathbf{s}_t = (s_{1,t}, \ldots, s_{N,t}) \in \mathcal{S}$ denote the joint history-augmented state of N agents and let $\mathcal{C}$ denote the static map context for each scenario. Each agent selects a motion token $a_{i,t} \in \mathcal{V}$, forming the joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t}) \in \mathcal{A} = \mathcal{V}^N$. SMART factorizes the same-step joint policy across agents:

$$\pi_\theta\left(\mathbf{a}_t|\mathbf{s}_t, \mathcal{C}\right) = \prod_{i=1}^{N} \pi_\theta^i\left(a_{i,t}|\mathbf{s}_t, \mathcal{C}\right) \tag{1}$$

where π_θ^i denotes the categorical token policy of agent i and still conditions on the joint scene through temporal, map-to-agent, and agent-to-agent attention.

We initialize π_θ from a supervised pretrained SMART checkpoint. During post-training, the map encoder and motion-token embeddings are frozen, while the agent interaction layers and token prediction head are updated.

RL formulation and reward design. Conditioned on $\mathcal{C}$, we formulate simulation as the finite-horizon Markov Decision Process (MDP): $\mathcal{M}_\mathcal{C} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are defined above, $\mathcal{P}$ is the state transition function that decodes sampled tokens and advances the joint rollout, $\mathcal{R}$ is the per-agent-step reward function defined below, ρ_0 is the empirical distribution of logged traffic histories, and $\gamma = 1$ because the implemented estimator uses undiscounted finite-horizon process rewards. The final method uses the immediate process reward at each generated step, corresponding to a one-step horizon rather than a return-to-go estimator.

We retain a physical reward for each valid agent-step: $r_{i,t} = w_{\text{ade}} r^{\text{ade}}_{i,t} + w_{\text{col}} r^{\text{col}}_{i,t}$, where w_{ade} and w_{col} weight trajectory fidelity and collision avoidance, respectively. The ADE component is

$$r^{\text{ade}}_{i,t} = -\left\|\hat{\mathbf{p}}_{i,t} - \mathbf{p}^{\text{gt}}_{i,t}\right\|_2 \tag{2}$$

where $\hat{\mathbf{p}}$ and $\mathbf{p}^{\text{gt}}$ denote the simulated and logged positions of an agent, respectively. This component provides a per-step trajectory-fidelity signal whose temporal average corresponds to ADE. The collision component is

$$r^{\text{col}}_{i,t} = -\mathbb{I}\left(\exists j \neq i : \text{gap}\left(\hat{\mathbf{b}}_{i,t}, \hat{\mathbf{b}}_{j,t}\right) < 0\right) \tag{3}$$

where $\hat{\mathbf{b}}$ denotes the simulated bounding box of an agent, approximated using three circles derived from its physical dimensions. The signed distance $\text{gap}(\cdot,\cdot)$ is negative when the three-circle approximations overlap, and $\mathbb{I}(\cdot)$ is the indicator function. The ADE term provides continuous feedback throughout a rollout, whereas the sparse collision term directly reflects the primary safety objective of reducing simulated collisions.

3.2 Refined Group-Relative Credit Assignment

For each scenario $\xi = (\mathbf{s}_0, \mathcal{C}) \sim \mathcal{D}$, the behavior policy $\pi_{\theta_{\text{old}}}$ generates G closed-loop rollouts with shared initial history and map context but independently sampled motion tokens. Each rollout yields physical agent-step rewards $\{r_{e,i,t}\}^{G,N,T}_{e=1,i=1,t=1}$, where e indexes rollouts and T is the simulation horizon. In existing work, COMA (Foerster et al. (2018)) uses counterfactual multi-agent baselines, while RUDDER (Arjona-Medina et al. (2019)) decomposes delayed returns. We instead compare aligned physical rewards without a learned critic or return decomposition. Unlike outcome-supervised GRPO (Shao et al. (2024)), which broadcasts a scalar rollout advantage across tokens, our credit retains both agent and timestep resolution. For a certain (i,t), we normalize rewards across the G same-scenario rollouts:

$$\hat{A}_{e,i,t} = \tilde{r}_{e,i,t} = \frac{r_{e,i,t} - \mu_{i,t}}{\max(\sigma_{i,t}, \epsilon)} \tag{4}$$

where $\tilde{r}$ denotes the normalized reward, $\epsilon = 10^{-6}$ is a numerical floor, and $\mu_{i,t}$ and $\sigma_{i,t}$ are respectively the group mean and standard deviation:

$$\mu_{i,t} = \frac{1}{G}\sum_{e=1}^{G} r_{e,i,t}, \quad \sigma_{i,t} = \sqrt{\frac{1}{G}\sum_{e=1}^{G}\left(r_{e,i,t} - \mu_{i,t}\right)^2} \tag{5}$$

Eq. 5 assumes G valid samples at each aligned position. Validity-aware statistics are given in Appendix A.3. This aligned normalization compares alternative outcomes at the same agent-time position and preserves information about when and for which agent the sampled rollouts differ. It serves as a local process-reward surrogate; unbiased full-horizon policy-gradient estimation is outside its scope.

For tokens sampled from $\pi_{\theta_{\text{old}}}$, the PPO-style clipped objective is

$$\mathcal{J}^{\text{clip}}_{e,i,t} = \min\left[\frac{\pi^i_\theta(a_{e,i,t}|\mathbf{s}_{e,t},\mathcal{C})}{\pi^i_{\theta_{\text{old}}}(a_{e,i,t}|\mathbf{s}_{e,t},\mathcal{C})}\hat{A}_{e,i,t}, \text{clip}\left(\frac{\pi^i_\theta(a_{e,i,t}|\mathbf{s}_{e,t},\mathcal{C})}{\pi^i_{\theta_{\text{old}}}(a_{e,i,t}|\mathbf{s}_{e,t},\mathcal{C})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{e,i,t}\right] \tag{6}$$

where ε is the clipping threshold. The clipped loss is

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{e=1}^{G}\frac{1}{N_e}\sum_{i=1}^{N}\sum_{t=1}^{T} m_{e,i,t}\mathcal{J}^{\text{clip}}_{e,i,t}\right] \tag{7}$$

where $m_{e,i,t} \in \{0,1\}$ is the validity mask and $N_e = \sum_{i,t} m_{e,i,t}$ This objective is the basis for hierarchical densification in the next section.

3.3 Hierarchical Safety-Oriented Dense RL

Safety-critical interactions are sparse both across traffic scenarios and within individual rollouts. Uniform sampling therefore allocates most policy updates to ordinary behavior, while the few samples that expose meaningful collision-avoidance decisions are rarely observed. Dense Learning (Feng et al. (2026)) concentrates training on informative events and states.

The ideal informative-event argument retains a set of rollouts whose complement has zero expected contribution to the target policy gradient. Conditioning on that set then preserves the expected gradient direction up to a positive scale (Appendix A.2). We therefore combine episode-level sampling that prioritizes scenarios whose probes collide with state-level masking at aligned agent-time positions across same-scenario rollouts. Together, the two levels address rarity across scenarios and sparsity within each selected rollout.

Episode-level densification. Identifying the ideal set is intractable for a pretrained joint simulator, so we use collision occurrence as a policy-dependent operational proxy for scenario selection. At the beginning of epoch k, we sample a candidate set $\mathcal{B}_k \subset \mathcal{D}$ and generate one stochastic probe rollout $\tau_\xi \sim \pi_{\theta_{\text{old}}}$ for each initial scenario $\xi = (\mathbf{s}_0, \mathcal{C}) \in \mathcal{B}_k$. We define the practical informative-event indicator as

$$u(\tau_\xi) = \mathbb{I}_c(\tau_\xi) \tag{8}$$

where $\mathbb{I}_c(\cdot)$ indicates whether the rollout contains at least one collision. We use collision-only episode mining because the near-miss criteria select nearly 90% of candidate scenarios and therefore provide little episode-level densification. The same physical near-miss measures remain active for state-level masking. This single observed outcome serves as a selection proxy for scenario mining without estimating the scenario's collision probability. The probe-positive subset is

$$\mathcal{B}_k^+ = \{\xi \in \mathcal{B}_k | u(\tau_\xi) = 1\} \tag{9}$$

The fixed-size epoch training subset $\widetilde{\mathcal{B}}_k$ samples from $\mathcal{B}_k^+$, filling any shortfall with randomly sampled probed noncollision scenarios from $\mathcal{B}_k \setminus \mathcal{B}_k^+$ (see Alg. 1). We discard the probes and generate G fresh training rollouts per scenario in $\widetilde{\mathcal{B}}_k$ using $\pi_{\theta_{\text{old}}}$. Training can therefore include probe-negative scenarios. We employ retrospective data densification (Feng et al. (2026)) by regenerating probe rollouts with the evolving policy and refreshing the informative labels of candidate scenarios at every training epoch. Refreshing reduces label staleness as the policy learns to resolve previously risky interactions.

State-densified multi-agent GRPO. In dense learning theory, a state $\mathbf{s}$ is informative if taking some action $\mathbf{a}$ changes the expected future return:

$$\mathbf{s} \in \mathcal{S}_c \iff V^\pi(\mathbf{s}) \neq Q^\pi(\mathbf{s}, \mathbf{a}),\ \exists\, \mathbf{a} \in \mathcal{A} \tag{10}$$

where $\mathcal{S}_c$ denotes the informative-state set, $\pi = \pi_\theta$ denotes the current parameterized policy, and V^π and Q^π denote its state-value and state-action value functions, respectively. Thus, a state is non-informative when $Q^\pi(\mathbf{s}, \mathbf{a}) = V^\pi(\mathbf{s})$ for all $\mathbf{a} \in \mathcal{A}$. Such a state $\mathbf{s} \notin \mathcal{S}_c$ has zero expected policy-gradient contribution:

$$\mathbb{E}_{\mathbf{a}\sim\pi}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s}) Q^\pi(\mathbf{s}, \mathbf{a})\right] = V^\pi(\mathbf{s}) \nabla_\theta \sum_{\mathbf{a}\in\mathcal{A}} \pi_\theta(\mathbf{a}|\mathbf{s}) = 0 \tag{11}$$

Thus, omitting these joint-state contributions preserves the expected policy gradient. Omitted agent-token terms must individually have zero expected contribution. The joint-state result alone does not justify an arbitrary token mask. Variance reduction requires additional covariance assumptions (Appendix A.2).

However, directly evaluating Eq. 10 is intractable in practice. The joint action space grows as $|\mathcal{V}|^N$, and whether an action is safety-critical depends on the coupled future responses of multiple agents. We use a tractable detector of closed-loop physical interactions as an approximation to the theoretical informative-state set. For rollout e, $c_{e,i,t} \in \{0, 1\}$ marks collision or near-miss events at evaluated vehicle agent-time positions. This physical mask is vehicle-centered: other object types can participate in interactions and collision evaluation, but receive no state-mask entries. We assign $c_{e,i,t}$ to the token whose decoded motion produces the inspected physical state. The default detector uses no pre-collision window. Geometry and thresholds appear in Appendix A.1 and Figure 4, and precursor-window alternatives in Appendix A.7.

To preserve the same-scenario counterfactual comparison, we take the union of the critical positions across the G rollouts before applying the validity mask:

$$\bar{c}_{e,i,t} = \mathbb{I}\left(\sum_{e'=1}^{G} c_{e',i,t} > 0\right) \tag{12}$$

Thus, a position detected as critical in one rollout is retained in every rollout where that aligned agent-time position is valid. The union does not label the other rollouts as collisions. For each rollout, we combine this episode-consistent critical indicator with the validity mask defined in the previous section:

$$m^{\text{dense}}_{e,i,t} = m_{e,i,t}\bar{c}_{e,i,t},\ N^{\text{dense}}_{e} = \sum_{i=1}^{N}\sum_{t=1}^{T} m^{\text{dense}}_{e,i,t} \tag{13}$$

Importantly, the group-relative advantages $\hat{A}_{e,i,t}$ are computed across all G rollouts before applying this mask. This preserves comparisons between alternative closed-loop outcomes generated from the same initial scenario and retains the agent-step credit assignment introduced in the previous section. The mask is applied only when accumulating the clipped policy-gradient objective:

$$\mathcal{L}_{\text{DenseRL}} = -\mathbb{E}\left[\frac{1}{|\mathcal{E}^+|}\sum_{e\in\mathcal{E}^+}\frac{1}{N^{\text{dense}}_{e}}\sum_{i=1}^{N}\sum_{t=1}^{T} m^{\text{dense}}_{e,i,t}\mathcal{J}^{\text{clip}}_{e,i,t}\right] \tag{14}$$

where $\mathcal{E}^+ = \{e|N^{\text{dense}}_{e} > 0\}$ contains rollouts with at least one valid critical agent-step. A group with no valid critical token does not contribute to the RL loss.

Unlike the state reconnection used in D2RL, SMART generates coupled multi-agent trajectories: independently reconnecting an agent-step would disrupt joint transition consistency. We therefore preserve complete rollouts and apply masking only to the surrogate loss. The ideal gradient results require exact informative-event and state selection (Appendix A.2). Together, the physical proxy, immediate-reward group normalization, random mask normalization, and PPO clipping define a practical biased surrogate for full-horizon policy-gradient optimization.

4 Experiments

We organize our experiments around the following questions:

- **Q1: Overall effectiveness.** Does Dense RL reduce collisions relative to supervised and RL baselines while maintaining per-rollout fidelity?
- **Q2: Credit assignment.** Does aligned agent–time credit improve safety and optimization behavior over advantages broadcast across agents or timesteps?
- **Q3: Hierarchical densification.** Do episode-level scenario densification and state-level selective updates provide complementary gains?
- **Q4: Design sensitivity.** How rollout budget, detector settings, and credit or mask designs affect the safety–fidelity trade-off?

4.1 Experimental Setup

Dataset and evaluation protocol. We evaluate on WOMD under the closed-loop WOSAC protocol (Montali et al. (2023)), initializing all RL variants from the same pretrained SMART-7M checkpoint and motion-token vocabulary. A fixed 640-scenario validation set selects checkpoints, and a

Table 1: Testing results and token-reconstruction references. Bold marks column-best model metrics (references excluded), and shading marks Dense RL. Three-seed results report mean ± sample standard deviation.

Method	CR(%) ↓	OR(%) ↓	ADE(m) ↓
Logged ground truth	2.325	11.286	0.0000
Ground-truth tokens	5.378	11.809	0.1567
SMART (Wu et al. (2024))	4.080	11.869	2.8230
CAT-K (Zhang et al. (2025))	3.664	11.599	2.7100
Rollout-level credit	29.325	34.105	11.1556
Uniform refined	3.394 ± 0.064	11.511 ± 0.057	$\mathbf{2.0805 \pm 0.0358}$
Dense RL	$\mathbf{2.914 \pm 0.024}$	$\mathbf{11.346 \pm 0.024}$	2.1055 ± 0.0166

fixed 882-scenario testing set reports performance with 32 rollouts per scenario under a common sampling and evaluator protocol. Both sets are sampled separately from the WOMD validation split using different random seeds. For the main comparison and new sensitivity studies, all metrics use each run's minimum-validation-CR checkpoint.

Metrics. Our primary metrics are *collision rate* (**CR**), *offroad rate* (**OR**), and **ADE**, all computed from the released WOSAC evaluator. Within each scenario, CR and OR average the evaluator's collision and offroad event indicators over evaluated agents and sampled rollouts. An indicator records whether the event occurs at any valid future step. ADE averages trajectory displacement error rather than selecting the best sampled trajectory. We then average scenario-level metrics equally over the testing set. Together, they measure direct closed-loop safety and the fidelity of a sampled rollout.

CR/OR and ADE answer different questions from the collision/offroad indication likelihood and minimum ADE (minADE) shown on the public WOSAC leaderboard. Collision/offroad indication likelihood evaluates distributional agreement of event indicators, whereas CR/OR report their absolute simulated frequencies. minADE selects the closest trajectory from multiple samples and therefore measures candidate-set coverage, whereas ADE measures the fidelity of a sampled rollout. This distinction is particularly important in safety-critical interactions, where the feasible safe behavior may be concentrated around a small number of modes and the tolerance for an unsafe sample is low. A diverse policy can obtain a favorable minADE while still producing a high ADE and unsafe individual rollouts; our objective emphasizes the reliability of each sampled simulation. A detailed discussion of the differences among these metrics, together with the rationale for their use in our safety-oriented setting, is provided in Appendices A.13 and A.14.

Baselines and implementation. We compare SMART and CAT-K with three group-relative credit variants: *Rollout-level credit*, *Agent-only credit*, and uniformly sampled aligned agent-time GRPO (*Uniform refined*). Uniform refined disables both densification levels. The other two credit variants broadcast advantages across both agents and time, or across time alone, respectively. All three variants use ADE/collision rewards and uniform scenario sampling, with configurations and fixed-rollout aggregation audits in Appendices A.3 and A.6. WOSAC leaderboard methods without reproducible checkpoints and direct CR/OR/ADE are not compared numerically. Unless stated otherwise, $G = 32$, $w_{\text{ade}} = 1.0$, and $w_{\text{col}} = 0.1$. Dense RL uses auto collision-scenario scanning and the basic state mask without auxiliary losses. Dense RL and Uniform refined report three-seed means and sample standard deviations (seeds 42/14/51). Other variants use seed 42 only.

4.2 Main Results

Table 1 shows that Dense RL lowers mean CR from 4.080% for SMART and 3.394% for Uniform refined to 2.914%, with comparable or lower OR despite using no offroad reward. Mean ADE is 2.1055 m versus 2.0805 m for Uniform refined, a difference of approximately 0.025 m. RMM is 0.7694, below SMART's 0.7852 but above Uniform refined's 0.7655, showing that similar ADE can coexist with changes in distributional realism. Appendices A.10–A.11 report the complete WOSAC components and vehicle-level distributions of speed, acceleration, angular motion, object distance, time headway (THW), and time to collision (TTC).

Rollout-level credit degrades all three direct metrics, consistent with the aggregation audit and with broadcasting sparse safety feedback to unrelated decisions (Appendix A.6). The observed failure is specific to broadcasting the present sparse safety/fidelity reward; rollout-level rewards remain appropriate for objectives defined at trajectory level.

The logged trajectories already exhibit a CR of 2.325% on the common testing set, providing an empirical reference that may include unavoidable collisions. Our post-training objective targets simulated collision behavior under the learned policy. Ground-truth token reconstruction further increases CR to 5.378%. This observation has two implications. First, the logged trajectories contain a substantial number of near-miss interactions, for which small intermediate-state errors can change whether a collision occurs. Second, it exposes the precision limit imposed by the fixed motion-token vocabulary. Although vocabulary refinement lies outside the scope of post-training, closed-loop RL can learn to avoid tokens that turn such near-miss interactions into collisions. After subtracting logged CR from each method's mean, Dense RL reduces excess CR by 44.9% relative to Uniform refined. The raw reduction is 14.1%, or 0.480 percentage points (Appendix A.8).

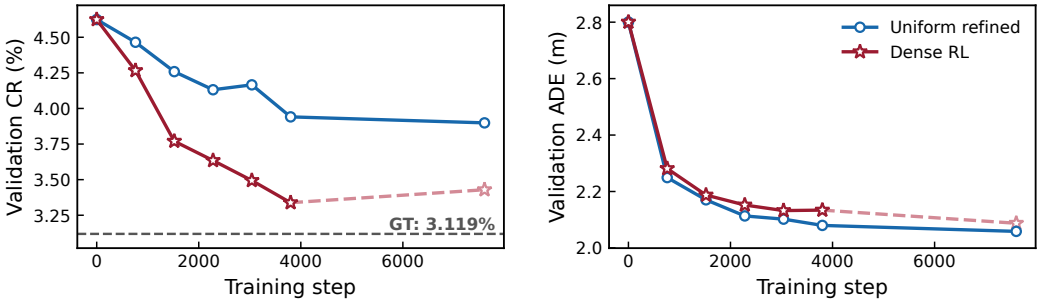

Figure 2: Seed-42 training curves on the common validation set. Each run uses its minimum-validation-CR checkpoint, so Dense RL's unselected epoch-9-to-19 continuation is pale and dashed. The gray dashed line denotes logged CR (3.119%).

At seed 42, Dense RL reaches 3.338% validation CR at step 3,800, versus 3.899% at step 7,600 for Uniform refined (Figure 2). We select each run's checkpoint with the lowest CR on this fixed validation set, yielding epoch 9 for Dense RL and epoch 19 for Uniform refined. Checkpoint selection depends solely on validation CR. Each step is one grouped minibatch with two PPO passes, and the curves omit episode-mining probes. Including probes, Dense RL uses 8.15 million rollouts to its selected checkpoint versus 15.56 million for Uniform refined. The complete 20-epoch totals are 16.30 and 15.56 million, respectively, with validation rollouts excluded. Appendix A.4 reports optimization/probe counts and elapsed job times separately.

4.3 Ablation Studies

Table 2 compares reward terms, credit granularity, and densification. *Only ADE reward* offers limited collision reduction, whereas *Only collision reward* causes severe trajectory drift. Their combination better balances safety and fidelity. Agent-only credit improves on Rollout-level credit but remains worse than aligned agent-time credit, supporting both resolutions for this reward design. Episode mining increases exposure to rare interactions and state masking focuses updates on critical agent-time positions. Together they yield the lowest CR among the four variants, with slightly higher ADE than Episode-only or Uniform refined. Refreshing every one or two epochs outperforms a fixed initial subset, supporting policy adaptation without a uniquely optimal frequency (Appendix A.5.3).

Some auxiliary full-token GRPO, KL, or behavior-cloning variants improve distributional realism, but none improves both CR and ADE over the default (Appendix A.9). Their exclusion reflects this safety–fidelity trade-off.

Design sensitivity and diagnostic checks. Results remain broadly stable across the tested detector thresholds and collision weights (Appendix A.5). Larger groups improve CR with additional rollouts, but $G = 64$ does not improve on $G = 32$ at a fixed per-epoch optimization-rollout budget. Longer temporal returns degrade both Dense RL and Uniform refined training, supporting imme-

diate process credit in our setup (Appendix A.6). Parameter-gradient replay shows that mask expansion changes update directions and coverage, yet a two-second collision precursor window does not lower testing CR (Appendix A.7). We therefore treat offline gradient similarity as a diagnostic rather than as evidence of closed-loop improvement.

Table 2: Core comparisons at seed 42 on the common testing set. Dense RL and Uniform refined reuse the main runs at this seed. These single-seed diagnostic ablations are interpreted descriptively. Configuration and selection details appear in Appendix A.3.

Variant	CR(%) ↓	OR(%) ↓	ADE(m) ↓
Dense RL	**2.933**	**11.374**	2.1222
Only ADE reward	3.387	11.640	2.1074
Only collision reward	2.986	45.049	23.0570
Rollout-level credit	29.325	34.105	11.1556
Agent-only credit	7.830	15.954	2.7128
Uniform refined	3.323	11.521	2.0612
Episode-only	3.164	11.444	**2.0509**
State-only	3.313	11.674	2.0917

4.4 Qualitative Results

Figure 3 illustrates a safety-critical cut-in interaction from the testing set. As one agent cuts into an adjacent lane, Uniform refined produces incompatible motions for the interacting agents and results in a collision. In contrast, our Dense RL policy captures their interaction during the cut-in and adapts their relative motion, allowing the maneuver to proceed without collision.

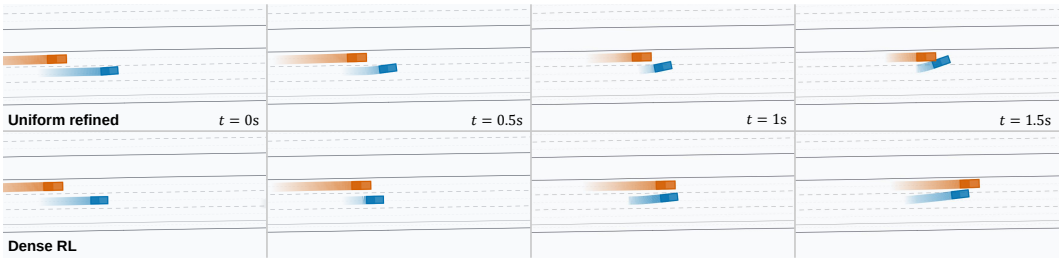

Figure 3: Cut-in interaction from Uniform refined and Dense RL runs, matched by initial scenario and sampled rollout index.

5 Conclusion

This work presented a hierarchical safety-oriented adaptation of Dense RL for post-training tokenized multi-agent traffic simulators under sparse safety-critical supervision. Refined credit retains per-agent, per-step physical rewards and compares aligned agent-time positions across same-scenario rollouts, providing critic-free local process credit without broadcasting one rollout score to all tokens. Retrospective episode densification and critical-state masking concentrate optimization on informative scenarios and interaction steps while preserving complete joint rollouts. Experiments on SMART and WOMD reduce closed-loop collisions while maintaining low per-rollout ADE and comparable offroad rate across three seeds.

The evidence is limited to one backbone and dataset. Failures are not stratified by interaction topology or logged versus model-induced causes. Future work should test additional backbones and datasets, richer physical criteria, and more expressive motion tokens.

AI USE STATEMENT

We used generative AI tools to assist with literature organization, language editing, LaTeX and figure preparation, code review, methodological and experimental-design discussion, theoretical derivation and consistency checking, and interpretation and organization of experimental results. Generative AI was not used to generate experimental measurements or replace the training and evaluation pipelines. All AI-assisted methodological suggestions, derivations, code changes, and interpretations were independently reviewed and verified by the authors, who take full responsibility for the final content and reported results.

ETHICS STATEMENT

This work uses the de-identified Waymo Open Motion Dataset for research on traffic simulation and does not involve human subjects, interventions, or the release of personal information. The safety-oriented evaluation is intended to improve simulation-based testing of autonomous-driving systems. We report both safety and fidelity metrics because optimizing collision reduction alone can produce unrealistic behavior, and we discuss this trade-off explicitly in the paper and appendix.

REPRODUCIBILITY STATEMENT

The main paper specifies the tokenized simulator, reward, credit aggregation, densification procedure, and evaluation protocol. The appendix provides the physical detector, algorithm, assumptions and propositions, resolved hyperparameters, execution environment, checkpoint rule, complete metric tables, and additional controlled variants. Anonymous supplementary materials contain the implementation and scripts needed to reproduce training, fixed-rollout audits, evaluation, and figure generation. The released configuration records the random seeds and checkpoint paths used for the reported results.

REFERENCES

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

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Luca Bergamini, Yawei Ye, Oliver Scheel, Long Chen, Chih Hu, Luca Del Pero, Błażej Osiński, Hugo Grimmett, and Peter Ondruska. Simnet: Learning reactive self-driving simulations from real-world observations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5119–5125. IEEE, 2021.

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

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

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

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

Shuo Feng, Haojie Zhu, Haowei Sun, Xintao Yan, Linxuan He, Jingxuan Yang, Guangzhen Su, Boqi Li, Shu Li, Ling Wang, et al. Breaking through safety performance stagnation in autonomous vehicles with dense learning. *Nature Communications*, 17(1):3163, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Jaime Fernández Fisac, and Chen Lv. Versatile behavior diffusion for generalized traffic agent simulation. *IEEE Transactions on Intelligent Transportation Systems*, 2026.

David Isele, Alireza Nakhaei, and Kikuo Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–6. IEEE, 2018.

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

Hongchen Li, Tianyu Li, Jiazhi Yang, Haochen Tian, Caojun Wang, Lei Shi, Mingyang Shang, Zengrong Lin, Gaoqiang Wu, Zhihui Hao, et al. Plannerrft: Reinforcing diffusion planners through closed-loop and sample-efficient fine-tuning. *arXiv preprint arXiv:2601.12901*, 2026.

Henry X Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *nature communications*, 15 (1):4808, 2024.

Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7553–7560. IEEE, 2023.

Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36:59151–59171, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Muleilan Pei, Shaoshuai Shi, and Shaojie Shen. Advancing multi-agent traffic simulation via r1-style reinforcement fine-tuning. *arXiv preprint arXiv:2509.23993*, 2025.

Zhenghao Peng, Quanyi Li, Chunxiao Liu, and Bolei Zhou. Safe driving via expert guided policy optimization. In *Conference on Robot Learning*, pp. 1554–1563. PMLR, 2022.

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

Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. *arXiv preprint arXiv:2312.04535*, 2023.

Kun Ren, Yizhou Xu, Jingxuan Yang, Yi Zhang, Jianming Hu, and Shuo Feng. Intelligent safety testing of autonomous vehicles with implicit traffic distributions by dense reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2026.

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

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024.

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.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Scenediffuser++: City-scale traffic simulation via a generative world model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1570–1580, 2025.

Xiaolong Tang, Meina Kan, Shiguang Shan, and Xilin Chen. Plan-r1: Safe and feasible trajectory planning as language modeling. *arXiv preprint arXiv:2505.17659*, 2025.

Yu Wang, Tiebiao Zhao, and Fan Yi. Multiverse transformer: 1st place solution for waymo open sim agents challenge 2023. *arXiv preprint arXiv:2306.11868*, 2023.

Ziyan Wang, Peng Chen, Ding Li, Chiwei Li, Qichao Zhang, Zhongpu Xia, and Guizhen Yu. Learning rollout from sampling: An r1-style tokenized traffic simulation model. *IEEE Robotics and Automation Letters*, 11(5):6336–6343, 2026.

Lu Wen, Jingliang Duan, Shengbo Eben Li, Shaobing Xu, and Huei Peng. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–7. IEEE, 2020.

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

Jingxuan Yang, Ruoxuan Bai, Haoyuan Ji, Yi Zhang, Jianming Hu, and Shuo Feng. Adaptive testing environment generation for connected and automated vehicles with dense reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Chris Zhang, James Tu, Lunjun Zhang, Kelvin Wong, Simon Suo, and Raquel Urtasun. Learning realistic traffic agents in closed-loop. In *7th Annual Conference on Robot Learning*, 2023a.

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*, pp. 1522–1529, 2023b.

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

Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun J Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. *Advances in Neural Information Processing Systems*, 37:79597–79617, 2024.

A Appendix

A.1 Physical Interaction Detection and Agent-Step Assignment

This section specifies how continuous multi-agent rollouts are converted into the collision indicators and critical agent-step masks used by reinforcement learning (RL) and state-level densification. The final method uses collisions alone for episode-level mining and both collisions and near misses for state-level masking. In preliminary screening, their union selected approximately 90% of the probed candidate scenarios. This high acceptance rate provided insufficient contrast with the original scenario distribution and therefore produced little effective episode densification. Accordingly, the final experiments use collision occurrence as the episode-level mining criterion, while clearance, time-to-collision (TTC), and time headway (THW) remain active for the finer-grained state-level detector. Episode mining changes which initial scenarios are trained, whereas state densification determines which aligned token positions contribute to the policy update. Figure 4 summarizes the geometry and the two distinct uses of these physical criteria.

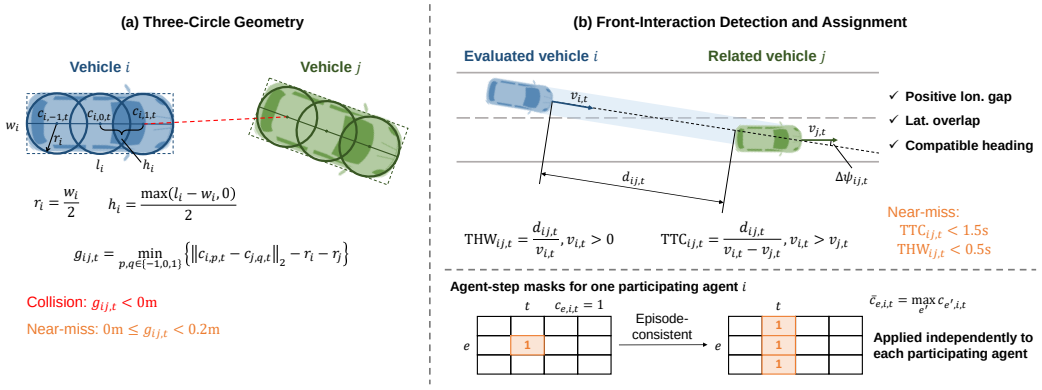

Figure 4: Physical interaction detection and assignment. (a) Three-circle geometry: collision uses $g_{ij,t} < 0$, and near-distance uses $0 \le g_{ij,t} < 0.2$ m. (b) Front-interaction criteria and episode-consistent assignment to evaluated vehicle agent-steps.

A.1.1 Three-Circle Interaction Geometry

Each oriented agent box is approximated by three circles derived from its length ℓ_i, width w_i, center $\mathbf{p}_{i,t}$, and heading $\psi_{i,t}$. The common radius and half-offset are

$$r_i = \frac{w_i}{2}, \; h_i = \frac{\max(\ell_i - w_i, 0)}{2} \tag{15}$$

For $p \in \{-1, 0, 1\}$, the circle center is

$$\mathbf{c}_{i,p,t} = \mathbf{p}_{i,t} + p \cdot h_i \cdot [\cos\psi_{i,t}, \sin\psi_{i,t}]^\top \tag{16}$$

The signed inter-agent gap is the minimum clearance over the nine circle pairs:

$$g_{ij,t} = \min_{p,q\in\{-1,0,1\}} (\|\mathbf{c}_{i,p,t} - \mathbf{c}_{j,q,t}\|_2 - r_i - r_j) \tag{17}$$

As illustrated for two vehicles in Figure 4(a), negative $g_{ij,t}$ indicates overlap and hence a collision between agents i and j, whereas $0 \le g_{ij,t} < 0.2$ m denotes a near-distance interaction. Only the former is used by collision-based episode mining, while both conditions can activate the state-level mask. Invalid agents, self-pairs, and cross-scenario pairs are excluded.

A.1.2 Near-Miss Detection and Pair Assignment

For state-level densification, a valid pair is critical if its signed gap is below 0.2 m or if it forms a valid front interaction satisfying the TTC or THW criterion. Front interactions require positive longitudinal separation, lateral overlap, and approximately aligned headings. The implementation excludes pairs with heading difference above 75° and then accepts geometrically stronger overlap or

Table 3: Final physical detection settings.

Setting	Value
Episode mining event	Collision
Collision signed-gap criterion	$g_{ij,t} < 0$ m
State near-distance threshold	0.2 m
TTC thresholds	1.5 s
THW thresholds	0.5 s
Simulation interval	0.5 s

near-parallel pairs with heading difference at most 10°. Only the nearest valid front agent is retained for each evaluated vehicle. The implementation does not restrict the other participant's object type. These geometric conditions and the associated longitudinal quantities are visualized in the upper part of Figure 4(b). For longitudinal separation $d_{ij,t} > 0$, ego speed $v_{i,t}$, and front-agent speed $v_{j,t}$,

$$\mathrm{THW}_{ij,t} = \frac{d_{ij,t}}{v_{i,t}}, \ \mathrm{TTC}_{ij,t} = \frac{d_{ij,t}}{v_{i,t} - v_{j,t}} \tag{18}$$

THW is finite only when $v_{i,t} > 0$, and TTC only for positive closing speed. Undefined values do not trigger a critical event. The active state-level rule is

$$[g_{ij,t} < 0.2 \text{ m}] \vee [\mathrm{TTC}_{ij,t} < 1.5 \text{ s}] \vee [\mathrm{THW}_{ij,t} < 0.5 \text{ s}] \tag{19}$$

Pairwise events are projected to the evaluated vehicle agents participating in the interaction. Pedestrians and cyclists can be interaction partners but receive no entries in this physical state mask. Collision-based episode mining and official all-agent evaluation are not vehicle-only. The resulting $c_{e,i,t}$ is aligned with the token decision whose decoded motion produces the inspected state, and is intersected with evaluated-role, validity, and generated-step masks. No pre-collision temporal window is used. Applying $c_{e,i,t}$ independently would select different agent-time locations in different rollouts: for example, a position would enter the loss only in the rollout that collides, while the safe outcomes at the same aligned position would be removed. This would discard the within-group contrast on which the group-relative advantage is based and make loss selection depend directly on the sampled outcome. We therefore take the union over the rollout dimension and broadcast it back to every rollout, as shown in the lower part of Figure 4(b):

$$\bar{c}_{e,i,t} = \max_{e' \in \{1,\ldots,G\}} c_{e',i,t}, \quad m^{\mathrm{dense}}_{e,i,t} = m_{e,i,t} \bar{c}_{e,i,t} \tag{20}$$

Here, $m_{e,i,t}$ is the effective evaluated-role, validity, and generated-step mask. Thus, if any rollout identifies (i, t) as critical, every rollout in which that agent-time position is valid contributes its corresponding token loss. The union does not label the other rollouts as collisions. It only retains their aligned safe or less-critical outcomes as counterfactual comparisons. Conversely, an agent-time position that is noncritical in every rollout is removed from loss accumulation. Complete joint transitions are retained. The mask affects only loss accumulation.

A.1.3 End-to-End Densification Procedure

Alg. 1 combines the complete pipeline described in the main text with the physical assignment above. It makes explicit that probe rollouts are discarded during training, aligned advantages are estimated before state masking, gradients are accumulated over the rollout group, and an optimizer step is taken only after all selected rollout losses have been backpropagated.

Algorithm 1 Hierarchical Dense RL post-training

Require: Dataset $\mathcal{D}$, policy π_θ, group size G, training epochs K_{train}, PPO epochs K_{PPO}, clip threshold ε, numerical floor $\epsilon = 10^{-6}$

for epoch $k = 1, \ldots, K_{\text{train}}$ **do**
  Sample candidate scenarios $\mathcal{B}_k \subset \mathcal{D}$
  Generate one probe rollout for each $\xi \in \mathcal{B}_k$ using the current policy
  Collect $\mathcal{B}_k^+ = \{\xi : \text{the probe contains a collision}\}$
  Form the epoch training subset $\widetilde{\mathcal{B}}_k$ from $\mathcal{B}_k^+$, filling any remaining slots with probed noncollision scenarios
  Discard all probe rollouts
  for each minibatch of initial scenarios from $\widetilde{\mathcal{B}}_k$ **do**
    Set $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ and precompute the static map features
    Under no gradient, generate G fresh complete joint rollouts $\{\tau_e\}_{e=1}^{G} \sim \pi_{\theta_{\text{old}}}$
    Store sampled tokens, rollout states, validity masks, and old token log-probabilities
    Compute $r_{e,i,t} = -w_{\text{ade}}\|\hat{\mathbf{p}}_{e,i,t} - \mathbf{p}_{i,t}^{\text{gt}}\|_2 - w_{\text{col}}\mathbb{I}(\exists j \neq i : g_{e,ij,t} < 0)$
    For every aligned (i, t), compute mask-aware $\mu_{i,t}$ and $\sigma_{i,t}$ across the G rollouts
    Compute $\hat{A}_{e,i,t} = (r_{e,i,t} - \mu_{i,t})/\max(\sigma_{i,t}, \epsilon)$ before state masking
    Detect Eq. 19, project pairwise events to evaluated vehicles, and construct $c_{e,i,t}$
    Apply Eq. 20 and set $m_{e,i,t}^{\text{dense}} = m_{e,i,t}\bar{c}_{e,i,t}$
    Set $\mathcal{E}^+ = \{e : N_e^{\text{dense}} = \sum_{i,t} m_{e,i,t}^{\text{dense}} > 0\}$
    if $\mathcal{E}^+ = \varnothing$ **then**
      Skip the RL update for this minibatch
    else
      for PPO epoch $h = 1, \ldots, K_{\text{PPO}}$ **do**
        Clear accumulated gradients
        for each rollout e with $N_e^{\text{dense}} = \sum_{i,t} m_{e,i,t}^{\text{dense}} > 0$ **do**
          Re-evaluate the stored sampled tokens under π_θ and obtain their new log-probabilities
          Compute $\rho_{e,i,t}(\theta) = \exp(\{\log \pi_\theta^i(a_{e,i,t}) - \log \pi_{\theta_{\text{old}}}^i(a_{e,i,t})\})$
          Compute $\mathcal{J}_{e,i,t}^{\text{clip}} = \min\Big(\rho_{e,i,t}\hat{A}_{e,i,t}, \text{clip}(\rho_{e,i,t}, 1-\varepsilon, 1+\varepsilon)\hat{A}_{e,i,t}\Big)$
          Compute $\mathcal{L}_e = -\left(N_e^{\text{dense}}\right)^{-1} \sum_{i,t} m_{e,i,t}^{\text{dense}} \mathcal{J}_{e,i,t}^{\text{clip}}$
          Backpropagate $\mathcal{L}_e/|\mathcal{E}^+|$ and accumulate gradients without updating parameters
        end for
        Clip the accumulated gradient norm and take one optimizer step
      end for
      Step the learning-rate scheduler once
    end if
  end for
  Refresh retrospective scenario mining with the updated policy in epoch $k + 1$
end for

A.2 Gradient Properties of Densified Group-Relative Optimization

We formalize which components of the densified estimator preserve a policy-gradient direction and which are practical stabilization approximations.

A.2.1 Definitions and Assumptions

Let $P_\theta(\tau)$ denote the distribution of a complete joint rollout under a fixed initial-scenario distribution, and let $R(\tau)$ be its target return. For the uniformly sampled data objective, this distribution averages uniformly over the training scenarios before sampling policy trajectories. The score-function contribution is

$$\mathbf{g}_\theta(\tau) = R(\tau)\nabla_\theta \log P_\theta(\tau),\ \nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim P_\theta}[\mathbf{g}_\theta(\tau)] \tag{21}$$

Let Φ be the theoretical informative-event set, $\mathcal{S}_c$ the informative-state set, and $\rho_\Phi = P_\theta(\Phi) > 0$. The analysis uses the following assumptions.

1. Rollouts are sampled on-policy, or distribution shift is corrected by exact importance weighting.
2. Excluded trajectories have zero expected target-gradient contribution: $\mathbb{E}_{P_\theta}[\mathbf{g}_\theta(\tau)\mathbb{I}(\tau \notin \Phi)] = \mathbf{0}$.
3. For each non-informative joint state $\mathbf{s} \notin \mathcal{S}_c$, $Q^\pi(\mathbf{s}, \mathbf{a}) = V^\pi(\mathbf{s})$ for every feasible joint action $\mathbf{a}$.

4. Given an initial scenario, the G rollout samples are conditionally independent and identically distributed.
5. The baseline for rollout e is independent of its current sampled action conditional on that rollout's pre-action history, initial scenario, and aligned agent-time index.
6. Selections and masks generated by $\pi_{\theta_{\text{old}}}$ are fixed within one update. No gradient is propagated through discrete detection or selection.

Collision-based episode selection approximates Φ, while the collision and near-miss state detector approximates $\mathcal{S}_c$ (Sec. A.1). These assumptions define an ideal target-gradient argument. The physical proxies do not establish Assump. 2 or 3, and the normalized, masked, and clipped process-credit loss is not covered by these guarantees.

A.2.2 Episode-Level Direction Preservation

Proposition 1 (informative-event conditioning). Define

$$q_\theta(\tau) = \frac{P_\theta(\tau)\mathbb{I}(\tau \in \Phi)}{\rho_\Phi} \tag{22}$$

Under Assump. 2,

$$\mathbb{E}_{q_\theta}[\mathbf{g}_\theta(\tau)] = \frac{1}{\rho_\Phi}\nabla_\theta J(\theta) \tag{23}$$

Thus, exact informative-event conditioning preserves the expected gradient direction, with positive scale $1/\rho_\Phi$.

Proof. Substitution of Eq. 22 yields

$$\mathbb{E}_{q_\theta}[\mathbf{g}_\theta] = \frac{1}{\rho_\Phi}\mathbb{E}_{P_\theta}[\mathbf{g}_\theta(\tau)\mathbb{I}(\tau \in \Phi)] \tag{24}$$

$$= \frac{1}{\rho_\Phi}\left(\mathbb{E}_{P_\theta}[\mathbf{g}_\theta(\tau)] - \mathbb{E}_{P_\theta}[\mathbf{g}_\theta(\tau)\mathbb{I}(\tau \notin \Phi)]\right) \tag{25}$$

$$= \frac{1}{\rho_\Phi}\nabla_\theta J(\theta) \tag{26}$$

where the last equality follows from Assump. 2. *End of proof.*

The implementation selects initial scenarios rather than exact trajectories: one stochastic probe supplies a collision label, and $\widetilde{\mathcal{B}}_k$ prioritizes probe-positive scenarios while filling any shortfall with probed noncollision scenarios. Fresh training rollouts use this completed subset, which can include probe-negative scenarios and is distinct from exact informative-trajectory conditioning. Proposition 1 characterizes ideal informative-event sampling. Practical retrospective scenario mining should therefore be interpreted as optimizing a policy-dependent rare-event objective rather than as an exactly unbiased estimator of the uniformly sampled data objective.

A.2.3 State-Level Gradient Preservation

Proposition 2 (zero contribution of non-informative states). Under Assump. 3, for $\mathbf{s} \notin \mathcal{S}_c$,

$$\mathbb{E}_{\mathbf{a}\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})Q^\pi(\mathbf{s},\mathbf{a})] = \mathbf{0} \tag{27}$$

Proof. Since $Q^\pi(\mathbf{s},\mathbf{a}) = V^\pi(\mathbf{s})$ for all $\mathbf{a}$,

$$\mathbb{E}_{\mathbf{a}\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})Q^\pi(\mathbf{s},\mathbf{a})] \tag{28}$$

$$= V^\pi(\mathbf{s})\sum_{\mathbf{a}} \pi_\theta(\mathbf{a}|\mathbf{s})\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s}) \tag{29}$$

$$= V^\pi(\mathbf{s})\nabla_\theta \sum_{\mathbf{a}} \pi_\theta(\mathbf{a}|\mathbf{s}) \tag{30}$$

$$= V^\pi(\mathbf{s})\nabla_\theta \mathbf{1} = \mathbf{0} \tag{31}$$

End of proof.

For the factorized same-step policy, $\log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathcal{C}) = \sum_i \log \pi^i_\theta(a_{i,t}|\mathbf{s}_t, \mathcal{C})$. Although every factor conditions on the joint scene, the score decomposes into agent-token terms. If the critical mask exactly retains all informative agent-step terms, omitting the remaining zero-expectation terms preserves the gradient expectation. This statement requires the corresponding zero-expectation condition for the omitted agent-token terms. A zero joint-state contribution alone does not justify removing an arbitrary subset of agent factors. The implementation retains the complete state trajectory and masks only these loss terms. It never reconnects transitions or alters future states after generation.

A.2.4 GROUP-RELATIVE BASELINES

At aligned location (i, t), define the leave-one-out baseline

$$b_{-e,i,t} = \frac{1}{G-1} \sum_{e'=1, e' \neq e}^{G} r_{e',i,t} \tag{32}$$

Proposition 3 (baseline subtraction). Under Assump. 4 and 5, replacing $r_{e,i,t}$ by $r_{e,i,t} - b_{-e,i,t}$ does not change the expected score-function gradient at that aligned location.

Proof. Condition on the initial scenario and the focal rollout's pre-action history. Under Assump. 4 and 5, $b_{-e,i,t}$ is conditionally independent of $a_{e,i,t}$. All expectations below are conditional on this history, so

$$\mathbb{E}\left[b_{-e,i,t} \nabla_\theta \log \pi^i_\theta(a_{e,i,t}|\mathbf{s}_{e,t}, \mathcal{C})\right] \tag{33}$$

$$= \mathbb{E}[b_{-e,i,t}] \sum_a \pi^i_\theta(a|\mathbf{s}_{e,t}, \mathcal{C}) \nabla_\theta \log \pi^i_\theta(a|\mathbf{s}_{e,t}, \mathcal{C}) \tag{34}$$

$$= \mathbb{E}[b_{-e,i,t}] \nabla_\theta \sum_a \pi^i_\theta(a|\mathbf{s}_{e,t}, \mathcal{C}) \tag{35}$$

$$= \mathbb{E}[b_{-e,i,t}] \nabla_\theta \mathbf{1} = \mathbf{0} \tag{36}$$

End of proof.

When all G entries are valid, the full-group mean is $\bar{r}_{i,t} = \frac{1}{G} \sum_e r_{e,i,t}$. Before standard-deviation normalization,

$$r_{e,i,t} - \bar{r}_{i,t} = \frac{G-1}{G} \left(r_{e,i,t} - b_{-e,i,t}\right) \tag{37}$$

For a complete valid group, the centered estimator has the same expected direction as its leave-one-out counterpart up to the fixed positive factor $\frac{G-1}{G}$. With mask-dependent group size $n_{i,t}$, the algebra instead uses $(n_{i,t}-1)/n_{i,t}$, and a random validity pattern need not preserve a common scale across samples. This baseline argument concerns baseline subtraction only; return-to-go remains a separate credit construction. A correctly defined complete return multiplied by all trajectory score terms remains a valid likelihood-ratio estimator, so the empirical limitations of our rollout-level credit variant are task-specific.

Composition with critical-state masking. The aligned advantages are computed from the complete valid rollout group before critical-state selection, as specified in the main paper and Algorithm 1. During optimization, the episode-consistent mask in Eq. 20 is treated as fixed data generated by $\pi_{\theta_{\text{old}}}$. Thus, the gradient of the densified surrogate is

$$\nabla_\theta \mathcal{L}_{\text{DenseRL}} = -\mathbb{E}\left[\frac{1}{|\mathcal{E}^+|} \sum_{e \in \mathcal{E}^+} \frac{1}{N^{\text{dense}}_e} \sum_{i,t} \text{sg}\big(m^{\text{dense}}_{e,i,t}\big) \nabla_\theta \mathcal{J}^{\text{clip}}_{e,i,t}\right] \tag{38}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Stop-gradient specifies differentiation through the saved surrogate; statistical dependence between an outcome-dependent mask and sampled actions remains. Thus, noncritical positions may enter the group statistics used to construct the aligned advantages, but only valid critical positions contribute to loss accumulation and backpropagation. Eq. 38 specifies where masking enters the implemented gradient. Its direction-preservation interpretation remains subject to the idealized conditions of Propositions 2 and 3.

A.2.5 Scope of the Practical GRPO Estimator

Propositions 1–3 establish direction preservation or unbiased baseline subtraction under the stated idealized assumptions. Several implementation choices deliberately move beyond that estimator.

- Division by a random finite-group standard deviation couples each sample to all rewards and is generally biased relative to the unnormalized score-function gradient.
- Normalization by the random critical-token count N_e^{dense} reweights rollout contributions and is likewise part of the practical surrogate.
- Although the importance ratio corrects action-probability shift at states sampled by the behavior policy, PPO clipping truncates this ratio and retains the behavior-policy state distribution. The resulting clipped surrogate is a biased stabilization once π_θ departs from $\pi_{\theta_{\mathrm{old}}}$. Any variance-reduction claim therefore requires assumptions beyond clipping.
- A physical proxy can omit informative decisions or include states that do not satisfy Assump. 3. Episode mining also deliberately changes the initial-scenario distribution without importance correction.
- The practical agent-step advantage is constructed from the immediate physical reward at the aligned decoded state and therefore targets localized process credit instead of return-to-go or action-value estimation.

The propositions separate three ideal arguments: informative-event conditioning preserves direction up to a positive scale, omission of zero-expectation state contributions preserves the gradient expectation, and an action-independent leave-one-out baseline leaves the expected score contribution unchanged. The implemented objective relaxes these assumptions through scenario mining, physical token selection, immediate-reward normalization, and PPO clipping, yielding an empirically evaluated process-credit surrogate. Densification increases the observed frequency of proxy-informative samples and motivates improved sample allocation. A strict variance or convergence guarantee would additionally require detector-accuracy, covariance, and moment assumptions.

A.3 Implementation and Reproducibility Details

A.3.1 Model, Optimization, and Dense RL Settings

The pretrained SMART-7M model contains 7,001,024 parameters, of which 5,344,384 are trainable in our experiment. We update the last six agent-interaction layers and token prediction head while freezing the map encoder and motion-token embeddings. Table 4 records the resolved settings of the final Dense RL method.

The final configuration uses collision-based auto scanning, the basic state mask, immediate aligned credit, and no auxiliary objective. The scanner stops when its target subset is filled or the maximum probe budget is reached; actual probe counts can be below this cap. At $G = 32$, each epoch contains 380 grouped minibatches with 64 initial scenarios and 32 joint rollouts per scenario. Each minibatch is reused for two PPO passes; the plots count one grouped minibatch as one training step.

Component ablations. The credit ablations use uniform scenario sampling with both densification levels disabled and share the ADE/collision reward weights, group size, and PPO settings. Rollout-level credit aggregates across agents and timesteps, Agent-only credit aggregates across timesteps, and Uniform refined retains aligned agent-time credit. Appendix A.6 verifies these aggregation rules on fixed rollouts.

Validity-aware group statistics. At each aligned agent-time position, $v_{e,i,t}$ indicates that both the generated position and its logged target are valid. Writing $n_{i,t} = \max(\sum_{e=1}^{G} v_{e,i,t}, 1)$, the implementation computes

$$\mu_{i,t} = \frac{1}{n_{i,t}} \sum_{e=1}^{G} v_{e,i,t} r_{e,i,t}, \quad \sigma_{i,t} = \sqrt{\frac{1}{n_{i,t}} \sum_{e=1}^{G} v_{e,i,t} \left(r_{e,i,t} - \mu_{i,t}\right)^2}$$

Table 4: Resolved hyperparameters for the final method.

Category	Parameter	Value
Model	Backbone / vocabulary	SMART-7M / 555 motion tokens
Model	Trainable / total parameters	5,344,384 / 7,001,024
Model	Updated modules	All 6 layers and prediction head
Model	Hidden dimension / attention heads	128 / 8
Model	Map / agent interaction layers	3 / 6
Model	Historical / future raw steps	11 / 80
Rollout	Frequency / generated horizon	2 Hz / 16 token steps
Rollout	Group size G	32
Rollout	Sampling	Top-k with $k = 48$, temperature 1.0
Rollout	Closed-loop validation rollouts	32
Reward	ADE / collision	1.0 / 0.1
GRPO	PPO epochs / clipping threshold	2 / 0.2
Dense RL	Probe scanner / maximum at default G	Auto / 121,600 scenarios per epoch
Dense RL	State detector	Basic distance–TTC–THW detector
Optimization	Optimizer / scheduler	AdamW / cosine decay
Optimization	Initial learning rate / minimum ratio	4×10^{-5} / 0.05
Optimization	Warmup steps	0
Optimization	Gradient-norm clipping	5.0
Optimization	Per-GPU batch size / training-data fraction	16 / 0.05
Optimization	Dataloader workers	16
Optimization	Maximum epochs / precision	20 / 32-bit
Evaluation	WOSAC batches / workers per rank	10 / 4
Auxiliary terms	Full-token GRPO / KL / BC coefficients	10 / 0.01 / 0.1
Auxiliary terms	Entropy loss coefficients	0.0

These statistics reduce to Eq. 5 when all G samples are valid. The numerical floor in Eq. 4 handles zero variance, and positions with no valid samples contribute no policy loss. Validity is distinct from critical-state selection: group statistics are computed before applying the critical-state mask.

A.3.2 Data, Randomness, and Checkpoint Selection

Training uses the Waymo Open Motion Dataset (WOMD) training split. Testing uses top-48 token sampling and temperature 1.0 for all methods. The common monitoring set contains 640 scenarios from the WOMD validation split, and the fixed testing set contains 882 scenarios selected with a different sampling seed. All reported testing runs use the same 882 IDs, evaluator version, and 32-rollout protocol.

The selection rule for the main comparison and new sensitivity studies is minimum monitoring CR within each training run, without metric-specific reselection. Selected checkpoint epochs for the main three-seed comparison are reported in Table 12; sensitivity runs follow the same minimum-monitoring-CR rule.

The rollout-level and agent-only baselines report their validation-selected post-training checkpoints; the pretrained model is listed separately.

The main Dense RL and Uniform refined comparisons use training seeds 42, 14, and 51. Other variants use seed 42 unless stated otherwise. The release will propagate each recorded seed to Python, NumPy, PyTorch, CUDA, distributed samplers, and dataloader workers.

A.3.3 Software, Hardware, and Release Artifacts

Table 5 summarizes the hardware, operating system, software stack, evaluator version, and numerical precision used in our experiments.

Table 5: Execution environment.

Component	Version or specification
GPU model / count / memory	NVIDIA H20 / 4 / 96G
CPU model / host memory	Intel 8457C / 1.84T
Operating system	Ubuntu
Python / CUDA	3.11.9 / 12.6
PyTorch / Lightning / Hydra	2.4.1 / 2.6.5 / 1.3.2
Waymo and WOSAC evaluator	1.6.7 (2025)
Numerical precision	32-bit

A.4 Computational Cost

Hierarchical densification adds policy probes for retrospective episode mining in addition to grouped training rollouts. Table 6 complements the training-iteration curves with rollout counts and elapsed job times for the main seed-42 runs. The main run uses four H20 GPUs, 32-bit precision, and approximately 5.34 million trainable parameters.

One rollout denotes one complete joint multi-agent trajectory for one initial scenario. Training rollouts are the fresh grouped samples used to compute rewards and policy gradients, whereas probe rollouts provide single-sample collision labels for scenario selection and are discarded before optimization. The reported totals exclude validation and testing rollouts. With a global batch size of 64 scenarios, optimization-rollout counts equal the recorded grouped-minibatch iterations times $64G$. Two PPO passes reuse each generated group. They increase loss evaluations and optimizer calls without increasing this rollout count. Probe totals use the actual number of scanned scenarios. Elapsed time spans training through the corresponding validation record and therefore includes validation and runtime overhead.

To the selected checkpoint, Dense RL uses 8.146 million optimization-plus-probe rollouts (7.7824 million optimization and 0.3636 million probes), compared with 15.565 million optimization rollouts for Uniform refined. Across the complete 20-epoch runs, Dense RL uses 16.296 million total rollouts and 197.9 h, versus 15.565 million rollouts and 165.1 h for Uniform refined; the selected Dense RL checkpoint is reached after 97.9 h. These job-level times include validation and runtime overhead, and selected-checkpoint costs differ because validation minima occur at different epochs. The full ordinary $G = 64$ run uses twice as many optimization rollouts as $G = 32$, while the full fixed-budget $G = 64$ run matches the latter's 15.565 million optimization rollouts.

Table 6: Recorded compute for seed-42 Dense RL and Uniform refined. Rollout counts are in millions and exclude evaluation. Hours include in-job validation and are archival costs, not standardized timings. Epochs are zero-based, and PPO reuse does not add rollouts.

Run	Epoch	Iterations	Optimization	Probe	Total	Hours
Through selected checkpoint						
Dense RL	9	3,800	7.7824	0.3636	8.1460	97.9
Uniform refined	19	7,600	15.5648	0	15.5648	165.1
Complete training						
Dense RL	19	7,600	15.5648	0.7316	16.2964	197.9
Uniform refined	19	7,600	15.5648	0	15.5648	165.1

A.5 Design Sensitivity under Different Sampling Budgets

We study group size, detector thresholds, collision-reward weight, and mining refresh frequency at training seed 42. Each point reports the testing metrics of the checkpoint selected within its own run. The default point is reused across studies and is not an independent repetition. Table 7 gives the exact values. Figures 5–8 visualize CR, OR, and ADE on separate aligned axes because their units and scales differ. RMM is retained in the table as a complementary distributional metric.

Table 7: Single-seed design sensitivity on the common 882-scenario testing set. The default row is shared by $G = 32$ in both budget protocols, the default detector, $w_{\text{col}} = 0.1$, and every-epoch refresh. "Fixed budget" refers to optimization rollouts per epoch and the configured training horizon, not equal cost at separately selected checkpoints.

Variant	CR(%) ↓	OR(%) ↓	ADE(m) ↓	RMM ↑
Default	2.933	11.374	2.1222	0.7687
Ordinary budget, $G = 8$	3.247	11.656	2.2034	**0.7711**
Ordinary budget, $G = 16$	3.085	11.515	2.1515	0.7697
Ordinary budget, $G = 64$	**2.836**	**11.188**	**2.0727**	0.7686
Fixed budget, $G = 8$	3.103	11.582	2.1999	0.7702
Fixed budget, $G = 16$	3.047	11.532	2.1336	0.7696
Fixed budget, $G = 64$	2.999	11.403	2.1388	0.7703
Detector tighter	2.929	11.334	2.1142	0.7696
Detector looser	2.850	11.319	2.0961	0.7691
$w_{\text{col}} = 0.05$	2.973	11.327	2.0903	0.7681
$w_{\text{col}} = 0.20$	3.097	11.624	2.1423	0.7698
Refresh every 2 epochs	2.964	11.293	2.0925	0.7692
Fixed initial subset	3.147	11.548	2.2022	0.7680

A.5.1 Group Size: Sampling Depth versus Scenario Coverage

Under the ordinary protocol, $G \in \{8, 16, 32, 64\}$ changes the number of same-scenario rollouts while retaining the per-GPU batch size, training-data fraction, and 20-epoch cap. Consequently, increasing G increases the number of optimization rollouts and policy-loss evaluations. CR decreases monotonically from 3.247% to 2.836% in these runs, but this trend combines the effects of group size and additional sampling.

The fixed-budget protocol instead sets the training-data fraction to $0.05 \times 32/G$, namely 0.20, 0.10, 0.05, and 0.025. It retains the per-GPU batch size of 16 and two PPO passes, while scaling the number of grouped minibatches per epoch inversely with G. This yields approximately 778,240 optimization rollouts per epoch. Larger groups therefore allocate more samples to fewer initial scenarios, while scenario coverage and the number of optimizer updates change accordingly. This controls the rollout budget but not total FLOPs, mining probes, validation cost, or all aspects of the optimization trajectory.

Under this protocol, CR is 3.103%, 3.047%, 2.933%, and 2.999% for increasing G. At a fixed optimization-rollout budget, the ordinary-protocol gain at $G = 64$ is no longer monotonic. The paired CR difference between fixed-budget $G = 64$ and $G = 32$ is $+0.067$ percentage points (95% scenario-bootstrap CI $[-0.092, 0.211]$, Appendix A.8), indicating no clear CR improvement from increasing the group beyond 32 at this budget. Because checkpoints are selected separately by validation CR, their accumulated rollouts at selection are not equal even when the per-epoch budgets and maximum training horizon agree.

A.5.2 Detector Thresholds and Collision Weight

The detector study jointly changes the near-distance, TTC, and THW thresholds. The tight, default, and loose settings are $(0.1\,\text{m}, 1.0\,\text{s}, 0.3\,\text{s})$, $(0.2\,\text{m}, 1.5\,\text{s}, 0.5\,\text{s})$, and $(0.3\,\text{m}, 2.0\,\text{s}, 0.7\,\text{s})$. Collision-based episode mining and the zero signed-gap collision threshold remain unchanged. CR remains between 2.850% and 2.933%, with similar OR and ADE, supporting robustness within this tested joint neighborhood. Because the thresholds are varied jointly, this experiment characterizes neighborhood robustness rather than individual-threshold effects or global optimality.

With $w_{\text{ade}} = 1$, collision weights of 0.05, 0.1, and 0.2 yield CR of 2.973%, 2.933%, and 3.097%, respectively. Doubling the default weight worsens CR, OR, and ADE, whereas halving it gives slightly lower ADE with slightly higher CR. The safety penalty influences learning through normalized, coupled advantages and policy evolution. A larger coefficient need not produce a lower final event rate. The default is a supported operating choice, not a globally optimized weight.

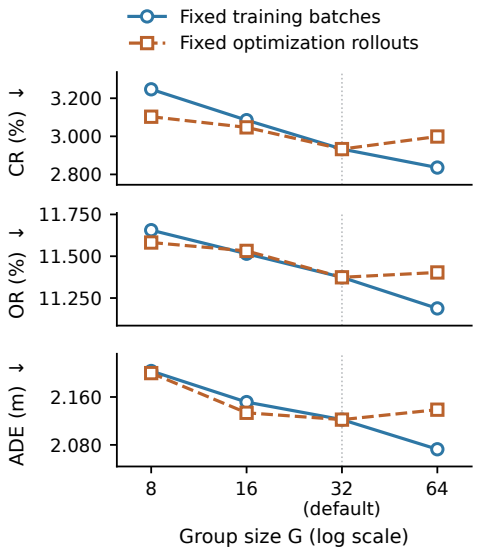

Figure 5: Group-size sensitivity under ordinary and fixed optimization-rollout budgets. Both curves reuse the same $G = 32$ default. The log-scaled group-size axis orders the tested settings.

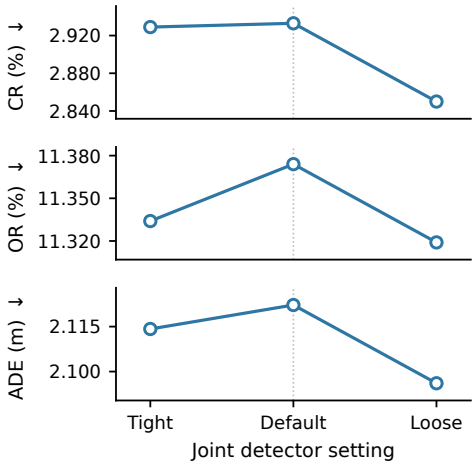

Figure 6: Joint detector-threshold sensitivity. Tight/default/loose are ordered configurations, not a numerical distance axis.

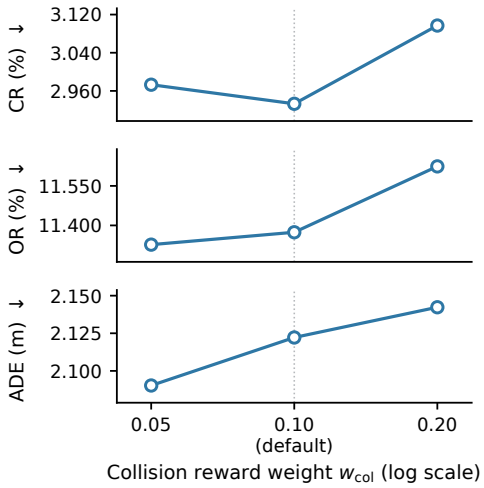

Figure 7: Collision-reward-weight sensitivity with the ADE weight fixed to one.

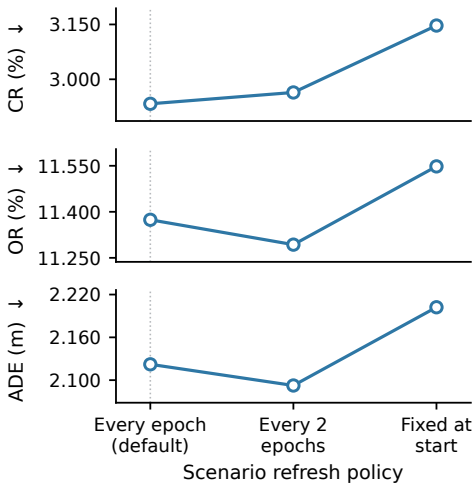

Figure 8: Scenario-refresh sensitivity. The three settings are categorical and equally spaced for readability.

A.5.3 Refreshing the Mined Scenario Subset

We compare mining at every epoch, every two epochs, and only once at the beginning of training. The recorded configurations differ in refresh frequency while retaining the same reward, group size, detector, and scanner settings. Periodic refresh gives CR of 2.933% and 2.964%, compared with 3.147% for the fixed subset. The fixed subset also increases ADE to 2.2022 m, whereas refreshing every two epochs gives 2.0925 m. These observations are consistent with the initial mined subset becoming less informative as the policy changes. These observations support periodic policy-dependent mining. Because the two ablation runs terminate before the configured 20-epoch cap, the comparison uses their validation-selected checkpoints and reflects checkpoint-level performance rather than equal-duration convergence.

A.6 Credit Aggregation and Temporal-Return Diagnostics

A.6.1 Same-Rollout Aggregation Analysis

We export complete rollout groups from the selected Dense RL, Rollout-level credit, and Agent-only credit policies on the same 882 scenario IDs. Each export uses $G = 32$ with 32 independently sampled rollouts per scenario, top-48 sampling, and temperature 1.0. For each policy, all alternative credit rules are recomputed on the same rollout rewards, validity masks, and token log-probabilities without updating the policy. This controls the aggregation rule within each policy export; trajectories differ across policies.

Aligned credit normalizes rewards across rollouts separately at each agent-time position. Agent-only credit first aggregates over valid timesteps, then normalizes across rollouts for each agent and broadcasts the result in time. Rollout credit aggregates over valid agents and timesteps before group normalization and broadcasts one scalar to all valid tokens in that rollout. We inspect the gradient of the corresponding surrogate with respect to saved token log-probabilities. These token-score gradients describe credit allocation before multiplication by the model Jacobian and therefore differ from parameter gradients or Adam updates.

Table 8: Same-export token-score-gradient alignment with aligned agent-time credit. Cosines are averaged over scenarios with defined nonzero pairs. The mask fraction is reported only to characterize the basic detector on each policy's trajectories.

Exported policy	Agent-only cosine	Rollout cosine	Basic token fraction
Dense RL	0.660	0.165	19.52%
Rollout-level credit	0.756	0.163	68.26%
Agent-only credit	0.673	0.215	29.56%

Rollout broadcasting yields a substantially different direction on all three policies, while retaining agent identity is closer to aligned credit. On the Dense RL export, the mean score-gradient norms are 0.00865, 0.00907, and 0.01154 for aligned, agent-only, and rollout credit, respectively. The nonzero score-gradient norms make a vanished learning signal an unlikely explanation for the poor rollout baseline. Aggregation instead mixes rewards from heterogeneous agents and times and assigns the resulting preference to decisions that may not have produced it, a plausible task-specific mismatch for a sparse collision penalty combined with frequent displacement costs. Rollout-level optimization can remain effective when the objective is defined at trajectory level, as in realism-oriented traffic post-training (Pei et al. (2025)). This same-data analysis verifies the intended aggregation semantics and provides mechanistic evidence about credit construction, while other factors may also contribute to the closed-loop performance gaps.

A.6.2 Finite-Horizon Returns and Closed-Loop Training

For a continuing discounted task, the policy-gradient advantage $A^\pi = Q^\pi - V^\pi$ is defined using the complete future return, $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$ with $0 \le \gamma < 1$, rather than an immediate reward alone (Sutton & Barto (2018)). For a finite episode, the corresponding return ends at termination. Here the generated trajectory contains 16 decisions, so the theoretical reference is the full remaining return within that trajectory. Short-horizon estimators can incorporate a bootstrapped continuation value,

but our critic-free comparison uses no such value estimate. We test whether propagating more of the available future agent-specific rewards improves the localized group-relative surrogate.

For $H \in \{1, 4, 16\}$ we replace the immediate reward used for group normalization by

$$R^{(H)}_{e,i,t} = \sum_{k=0}^{\min(H-1,T-t)} r_{e,i,t+k}, \quad \gamma = 1 \tag{39}$$

where invalid future entries are excluded. Group normalization remains aligned at (i, t) and precedes any dense loss mask. At 2 Hz, $H = 4$ covers a two-second window, while $H = 16$ includes all available future rewards in the generated horizon. Uniform training disables both densification levels, so its $H = 4$ variant tests whether the same failure occurs without critical-state selection.

Table 9: Temporal-return training results at seed 42. Rows use their validation-selected checkpoints. Uniform $H = 16$ was not trained.

Sampling	H	CR(%)	OR(%)	ADE(m)	RMM
Dense RL	1	**2.933**	**11.374**	2.1222	0.7687
Dense RL	4	3.724	12.180	2.2839	**0.7693**
Dense RL	16	4.325	12.740	2.4193	0.7670
Uniform refined	1	3.323	11.521	**2.0612**	0.7649
Uniform refined	4	4.971	12.851	2.3148	0.7644

Among the tested horizons, $H = 1$ gives the lowest CR, OR, and ADE for each sampling scheme. Longer returns worsen all three metrics at the selected checkpoints. The Uniform refined $H = 4$ run also deteriorates, with validation CR increasing from 5.606% at epoch 0 to 13.774% at epoch 3, and ADE from 2.317 to 3.207 m. The same degradation under Uniform refined indicates that the failure extends beyond dense masking. Table 9 reports the validation-selected checkpoints, including early minima from runs that later deteriorate.

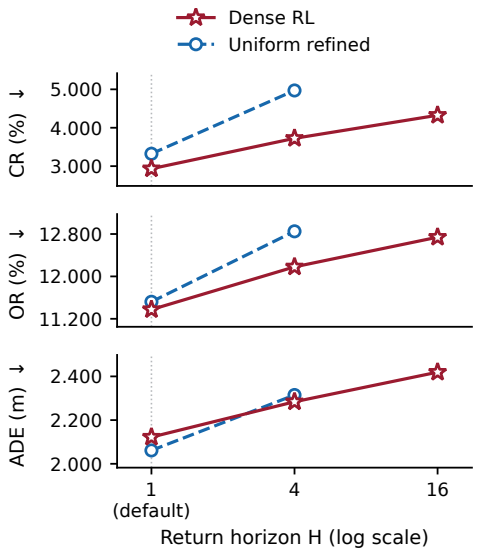

Figure 9: Testing sensitivity to the return horizon. Each point uses the separately selected checkpoint in Table 9; accumulated training cost can therefore differ across points.

A.6.3 What the Training and Fixed-Rollout Diagnostics Explain

Across horizons, per-position normalized advantage maxima remain near 5.57, mean importance ratios stay between 1.000 and 1.002, and mean clipping fractions remain below 0.8%. In contrast, pre-clipping parameter-gradient spikes and rising policy entropy coincide with worsening metrics: the Dense RL $H = 4$ maximum gradient norm rises from 0.61 to 3.25 at epoch 2, while Dense RL

$H = 16$ and Uniform refined $H = 4$ reach 5.70 and 1.97. Gradient clipping still bounds the applied gradient, and these correlations remain descriptive rather than causal.

On the fixed dense export, increasing H from 1 to 4 and 16 changes the mean token-score-gradient cosine with $H = 1$ from 1 to 0.825 and 0.666. The diagnostic rollout-ranking-change rate rises from zero to 0.0040 and 0.0381, while collision-precursor negative-credit coverage rises from 0.390 to 0.503 and 0.506. Thus, longer returns alter both where credit is assigned and the direction of the surrogate, despite similar normalized advantage magnitudes. Overlapping return windows repeatedly include the same future rewards, and later stochastic actions by multiple agents affect those rewards. Group normalization controls marginal scale but does not remove this cross-token dependence. Frequent ADE costs can dominate return-based rankings, while sustained overlaps can repeat collision penalties across adjacent windows. Keeping $H = 1$ preserves the direct correspondence between a local agent-time reward and its token, limiting the repeated attribution of later joint-rollout divergence to earlier decisions. Together, these diagnostics are consistent with credit-construction and policy-drift effects. We therefore use $H = 1$ as the empirical design choice for this critic-free setup. Learned critics, generalized advantage estimation, and tuned PPO for long-horizon control remain outside the evaluated configurations.

A.7 State-Mask Coverage, Parameter Gradients, and Closed-Loop Validation

A.7.1 Masks and Precursor Coverage

We use 882 rollout groups from the selected seed-42 Dense RL checkpoint to compare four masks while holding rewards, aligned $H = 1$ advantages, and generated trajectories fixed. The basic mask uses the physical detector and group union described in Appendix A.1. The critical-window mask extends every critical event backward by four token decisions (two seconds). The collision-window mask instead retains collision events and their four-step precursor windows, without retaining unrelated near misses. The union-window mask takes the union of the basic mask and the collision window. All masks are restricted to valid generated tokens and made consistent across the rollout group.

The basic mask covers 40.7%, 37.6%, and 33.7% of valid agent-steps within 0.5, 1, and 2 seconds before collisions, respectively, over 680 eligible scenarios. The event step itself is excluded from these precursor measurements. The expanded masks cover the defined collision windows by construction, while temporal proximity remains an imperfect causal proxy. The collision-only window retains only 44.5% of the near-miss tokens, whereas the union window preserves the complete basic supervision. Table 10 quantifies the resulting expansion.

Table 10: Coverage and pre-clipping parameter-gradient magnitudes on fixed rollouts. Token fractions use the common effective valid-token denominator. Norm statistics include all 882 scenarios, including zero gradients. Norm-ratio quantiles use scenarios where the basic gradient is nonzero.

Mask	Tokens	Fraction	Active scenarios	Mean norm	Median ratio [P10, P90]
Basic (default)	1,626,272	19.52%	748	3.832	1.000 [1.000, 1.000]
Critical window	2,509,632	30.12%	771	3.644	0.980 [0.557, 2.341]
Collision window	2,238,112	26.86%	696	3.042	0.915 [0.000, 2.315]
Union window	2,364,736	28.38%	771	3.639	1.000 [0.617, 1.750]

A.7.2 Replay in Trainable-Parameter Space

For each scenario and mask, we replay the stored tokens through the selected checkpoint and backpropagate the aligned surrogate through the trainable interaction layers and prediction head. The gradient is accumulated over the 32 rollouts with the same loss normalization and recorded before gradient clipping, without taking an optimizer step. This replay yields trainable-parameter gradients while varying only the mask; the aggregation rule and return horizon remain fixed.

The union window adds 45.4% more effective tokens and activates 23 additional scenarios, yet its median gradient-norm ratio to basic is approximately one. Token-count normalization and vector cancellation prevent a direct proportionality between selected token count and gradient magnitude.

The collision-only window also activates these 23 scenarios but removes 75 previously active scenarios, consistent with its loss of near-miss supervision.

Table 11: Per-scenario parameter-gradient cosine with the basic mask. Undefined zero-gradient pairs are omitted rather than assigned cosine zero. Each row has its own valid-pair denominator.

Candidate mask	Pairs	Mean	Median	P10–P90	Negative pairs
Union window	748	0.779	0.889	0.317–1.000	18 (2.41%)
Critical window	748	0.693	0.789	0.187–0.996	31 (4.14%)
Collision window	673	0.641	0.783	0.019–0.992	48 (7.13%)

On the same 673 scenarios where all masks have nonzero gradients, the corresponding mean cosines are 0.754, 0.705, and 0.641. The same ordering persists on the common 673-scenario subset, reducing sensitivity to valid-pair composition. Among the 639 basic/union pairs where tokens are actually added, the mean cosine is 0.741, showing that mask expansion can materially change the aggregate update direction.

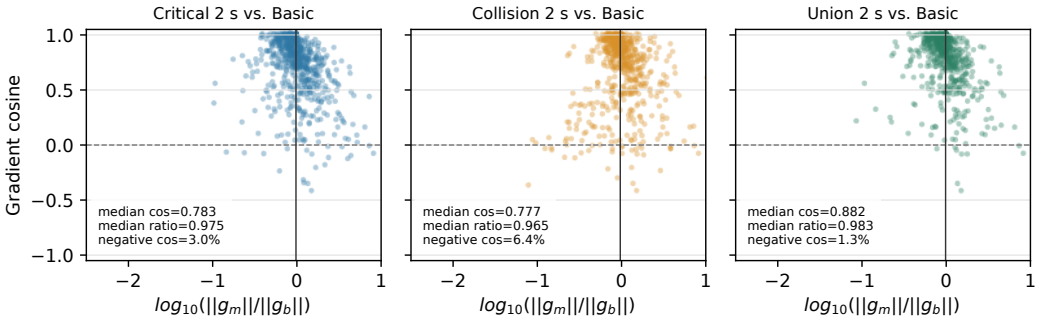

Figure 10: How precursor expansion changes parameter-gradient direction and magnitude. Each point is a scenario. The horizontal coordinate is $\log_{10}(\|g_{\text{candidate}}\|/\|g_{\text{basic}}\|)$ and the vertical coordinate is their cosine. The plotted subset requires $\|g_{\text{basic}}\| > 10^{-4}$ and a positive candidate norm to avoid unstable ratios and undefined logarithms. Panel summaries refer to that plotted subset, whereas Table 11 includes all defined pairs. The point $(0, 1)$ denotes unchanged magnitude and direction.

Gradient distribution. Across active scenarios, roughly 34%, 32–33%, and 25% of squared gradient norm fall in temporal, agent-to-agent, and map-to-agent attention, respectively, with 7–9% in the prediction head. The mask changes therefore affect the interaction network as well as the output head. These module-level magnitudes characterize where gradients occur but are insufficient to infer behavioral causes, aggregate update direction, generalization, or future collision reduction.

A.7.3 Closed-Loop Check of the Union Window

The union mask is the targeted training counterpart because it adds collision precursors while retaining all basic supervision. Its selected checkpoint gives CR/OR/ADE/RMM of $2.959\%/11.406\%/2.1034\,\text{m}/0.7691$, compared with $2.933\%/11.374\%/2.1222\,\text{m}/0.7687$ for basic. The paired CR difference is $+0.026$ percentage points with 95% CI $[-0.114, 0.151]$. The ADE difference is -0.0188 m with CI $[-0.0466, 0.0084]$. The intervals overlap zero for both CR and ADE, indicating no clear improvement at this uncertainty level. Broader precursor coverage and changed gradient geometry are therefore insufficient by themselves to improve CR, so we retain the basic mask as the default and treat the broader windows as diagnostic alternatives.

A.8 Seed Variation and Paired Evaluation Uncertainty

Table 12 reports all three training seeds for the main comparison. Means and sample standard deviations weight training seeds equally. Checkpoints are selected independently within each run.

All three Dense RL runs have lower CR than all three Uniform refined runs, while Uniform refined has slightly lower mean ADE.

Table 12: Main comparison by training seed. Checkpoint epochs are zero-based. The mean rows in Table 1 summarize these same runs.

Method	Seed	Epoch	CR(%)	OR(%)	ADE(m)
Uniform refined	42	19	3.323	11.521	2.0612
Uniform refined	14	5	3.447	11.450	2.1218
Uniform refined	51	12	3.413	11.563	**2.0586**
Dense RL	42	9	2.933	11.374	2.1222
Dense RL	14	11	**2.887**	11.334	2.0890
Dense RL	51	11	2.922	**11.331**	2.1053

Defining excess CR as $\mathrm{CR} - \mathrm{CR}_{\mathrm{GT}}$, we compute its relative reduction using seed means:

$$\frac{\overline{\mathrm{CR}}_{\text{Uniform refined}} - \overline{\mathrm{CR}}_{\text{Dense RL}}}{\overline{\mathrm{CR}}_{\text{Uniform refined}} - \mathrm{CR}_{\mathrm{GT}}} = 44.9\% \tag{40}$$

The corresponding raw-CR reduction is 14.1%, or 0.480 percentage points. The logged-trajectory CR serves as an empirical reference. Token reconstruction matches endpoints, but discrepancies along decoded segments can change geometric collision indicators despite small average displacement errors; the higher reconstruction CR therefore reflects tokenization sensitivity while a learned policy can still avoid unsafe token choices.

For evaluation uncertainty, we pair the scenario-level metric records of two fixed checkpoints and resample the 882 scenario IDs with replacement 10,000 times. Each resample uses the same IDs on both sides, averages within-scenario metric differences, and keeps all 32 rollouts within a sampled scenario as one cluster. The 2.5th and 97.5th percentiles give the intervals in Table 13. These fixed-checkpoint intervals capture scenario-level evaluation uncertainty conditional on the saved policies and rollout samples; training-seed, checkpoint-selection, repeated-simulation, and multiple-testing uncertainty are outside their scope.

Table 13: Paired scenario-bootstrap differences at seed 42 (left minus right). CR and OR differences are percentage points. ADE is in meters. Lower differences favor the left policy except for RMM. The default $G = 32$ and basic rows denote the same Dense RL run.

Comparison	Metric	Difference	95% interval
Dense RL − Uniform refined	CR	-0.390	$[-0.603, -0.186]$
	OR	-0.148	$[-0.341, 0.041]$
	ADE	$+0.0610$	$[0.0295, 0.0944]$
	RMM	$+0.00382$	$[0.00146, 0.00624]$
Fixed ($G = 64$) − ($G = 32$)	CR	$+0.067$	$[-0.092, 0.211]$
	OR	$+0.029$	$[-0.096, 0.153]$
	ADE	$+0.0166$	$[-0.0118, 0.0451]$
	RMM	$+0.00164$	$[0.00003, 0.00330]$
Union window − basic	CR	$+0.026$	$[-0.114, 0.151]$
	OR	$+0.032$	$[-0.095, 0.155]$
	ADE	-0.0188	$[-0.0466, 0.0084]$
	RMM	$+0.00044$	$[-0.00110, 0.00206]$

The interval for Dense RL versus Uniform refined supports lower CR and higher RMM for this seed-42 pair, along with a measurable ADE cost. The corresponding three-seed mean ADE difference is smaller (0.0250 m), showing that fixed-checkpoint bootstrap uncertainty differs from training-seed uncertainty. The fixed-budget and union-window intervals contain zero for CR, OR, and ADE, providing no clear evidence of improvement at this evaluation scale.

A.9 Auxiliary Objectives and Fidelity Controls

We test whether auxiliary supervision improves the safety–fidelity trade-off of localized RL updates. Starting from the hierarchical objective, the auxiliary variants combine full-token GRPO, reference-policy Kullback–Leibler (KL) divergence regularization, and teacher-forced behavior cloning (BC):

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{DenseRL}} + \alpha\mathcal{L}_{\text{GRPO}} + \beta\mathcal{L}_{\text{KL}} + \lambda\mathcal{L}_{\text{BC}} \tag{41}$$

where α, β, and λ weight the full-token GRPO, KL, and BC objectives, respectively. *Full-token GRPO* uses the same aligned agent-time advantages as the main method but activates the unmasked objective in Eq. 7 over all valid agent-steps. It is distinct from the rollout-level credit baseline in Table 1. For the *KL* loss, π_{ref} is a frozen copy of the pretrained policy, and the full-vocabulary divergence is evaluated on closed-loop states:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}\left[\frac{1}{G}\sum_{e=1}^{G}\frac{1}{N_e}\sum_{i,t} m_{e,i,t} D_{\text{KL}}\left[\pi_\theta^i \| \pi_{\text{ref}}^i\right]\right] \tag{42}$$

For categorical motion tokens, this divergence is computed exactly over the complete vocabulary:

$$D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}] = \sum_{a\in\mathcal{V}} \pi_\theta(a|\mathbf{s},\mathcal{C}) \log \frac{\pi_\theta(a|\mathbf{s},\mathcal{C})}{\pi_{\text{ref}}(a|\mathbf{s},\mathcal{C})} \tag{43}$$

The *Teacher-forced BC* variant applies the standard next-token cross-entropy to logged trajectories just like SMART pretraining, inspired by PPO-ptx (Ouyang et al. (2022)):

$$\mathcal{L}_{\text{BC}} = -\mathbb{E}\left[\frac{1}{N_{\text{BC}}}\sum_{i,t} m_{i,t}^{\text{BC}} \log \pi_\theta^i\left(a_{i,t}^{\text{gt}} | \mathbf{s}_t^{\text{gt}}, \mathcal{C}\right)\right] \tag{44}$$

where $a_{i,t}^{\text{gt}}$ is the logged motion token, $m_{i,t}^{\text{BC}}$ is its validity mask, and $N_{\text{BC}} = \sum_{i,t} m_{i,t}^{\text{BC}}$. In tables and figures, the prefix "+" denotes adding the indicated auxiliary loss to the densified objective. KL is reference-policy regularization and BC is teacher-forced behavior cloning. Individual variants activate the corresponding coefficient in Eq. 41, while combined variants activate multiple coefficients. Their active coefficients are $\alpha = 10$, $\beta = 0.01$, and $\lambda = 0.1$. Inactive terms have coefficient zero.

Table 14: Auxiliary-objective comparisons at seed 42. Full-token GRPO retains aligned credit and is distinct from rollout-level broadcasting.

Variant	CR(%) ↓	OR(%) ↓	ADE(m) ↓	RMM ↑
Dense RL (no auxiliary loss)	**2.933**	**11.374**	2.1222	0.7687
+ Full-token GRPO	3.198	11.422	2.0724	0.7653
+ KL	3.227	11.563	2.1779	**0.7743**
+ BC	4.209	12.077	2.2517	0.7742
+ Full-token GRPO + KL	3.141	11.446	**2.0612**	0.7690
+ Full-token GRPO + BC	3.570	11.773	2.1197	0.7688

None of these variants has both lower CR and lower ADE than the default. Full-token GRPO and its KL combination obtain lower ADE at higher CR, whereas KL and BC improve RMM to approximately 0.774. Teacher-forced BC alone has the largest CR increase among the auxiliary variants. These results justify using the unaugmented safety-oriented objective as the default in the tested setting, while preserving the alternative value of regularization for distributional realism.

A.10 Complete WOSAC Metrics Results

Official WOSAC metrics, including the Realism Meta Metric (RMM), provide a complementary assessment of distributional realism. We evaluate them on the same selected checkpoints and saved rollouts used for direct CR, OR, and ADE. Every entry is computed over the same testing set with Waymo Open Dataset evaluator v1.6.7 and the 2025 Sim Agents Challenge configuration. Tables 15 and 16 report the aggregate metrics and their detailed components, respectively.

Table 15: Official WOSAC summary metrics for the main methods and ablations. Rows marked "3 seeds" report the corresponding three-seed mean. The remaining rows are single runs.

Method or variant	RMM ↑	Kinematic ↑	Interactive ↑	Map ↑	minADE ↓
SMART	0.7852	0.4970	0.8107	0.9172	1.2767
CAT-K	**0.7881**	**0.5010**	**0.8137**	**0.9192**	**1.2735**
Rollout-level credit	0.5845	0.3236	0.6158	0.6935	7.0655
Uniform refined (3 seeds)	0.7655	0.4657	0.7929	0.9017	1.6690
Dense RL (3 seeds)	0.7694	0.4682	0.7990	0.9035	1.6439
Only ADE reward	0.7657	0.4665	0.7921	0.9026	1.6490
Only collision reward	0.5233	0.2626	0.7121	0.4297	20.3284
Agent-only credit	0.7488	0.4612	0.7766	0.8774	1.6505
Episode-only	0.7672	0.4676	0.7955	0.9019	1.6578
State-only	0.7658	0.4678	0.7920	0.9023	1.6620
+ Full-token GRPO	0.7653	0.4658	0.7927	0.9013	1.6658
+ KL	0.7743	0.4780	0.8026	0.9074	1.6252
+ BC	0.7742	0.4842	0.8004	0.9063	1.4443
+ Full-token GRPO + KL	0.7690	0.4740	0.7960	0.9028	1.6433
+ Full-token GRPO + BC	0.7688	0.4730	0.7954	0.9037	1.5751

Table 16: Detailed official WOSAC components for the main methods. Rows marked "3 seeds" report the corresponding three-seed mean. Higher likelihoods and lower simulated event rates are better.

Method	Speed	Accel.	Ang. speed	Ang. accel.
SMART	0.3915	0.4107	0.5252	0.6603
CAT-K	**0.3986**	**0.4143**	**0.5286**	**0.6625**
Rollout-level credit	0.1646	0.2735	0.3554	0.5007
Uniform refined (3 seeds)	0.3431	0.3862	0.4942	0.6391
Dense RL (3 seeds)	0.3448	0.3930	0.4974	0.6375

Method	Obj. dist.	Collision ind.	TTC
SMART	**0.3999**	0.9647	0.8365
CAT-K	**0.3999**	**0.9692**	**0.8386**
Rollout-level credit	0.2891	0.6748	0.7949
Uniform refined (3 seeds)	0.3568	0.9585	0.8151
Dense RL (3 seeds)	0.3584	0.9684	0.8163

Method	Road edge	Offroad ind.	TL viol.	Sim. CR	Sim. OR	Sim. TL
SMART	0.6922	0.9532	0.9625	0.0408	0.1187	0.0261
CAT-K	**0.6947**	**0.9554**	**0.9631**	0.0366	0.1160	0.0241
Rollout-level credit	0.5547	0.6737	0.9315	0.2933	0.3410	**0.0167**
Uniform refined (3 seeds)	0.6676	0.9375	0.9569	0.0339	0.1151	0.0192
Dense RL (3 seeds)	0.6676	0.9395	0.9589	**0.0291**	**0.1135**	0.0193

The supervised SMART and CAT-K models obtain the highest RMM values, 0.7852 and 0.7881, respectively. Our safety-oriented post-training lowers the mean RMM to 0.7694. Among its constituent buckets, the largest decrease relative to SMART occurs in kinematic likelihood (0.4970 to 0.4682), while interactive and map-based likelihoods decrease from 0.8107 and 0.9172 to 0.7990 and 0.9035. These three bucket changes account for the reduction in aggregate distributional realism measured by RMM. Separately, minADE increases from 1.2767 to 1.6439 m, indicating reduced best-of-set trajectory coverage. MinADE is reported as a complementary metric and does not contribute to RMM.

Relative to Uniform refined, hierarchical densification improves mean RMM from 0.7655 to 0.7694 and raises all three constituent bucket scores. It separately reduces minADE from 1.6690 to 1.6439 m. It simultaneously lowers mean simulated CR from 0.0339 to 0.0291 and mean simulated OR from 0.1151 to 0.1135. The detailed components show corresponding gains in collision indica-

tion likelihood (0.9585 to 0.9684), offroad indication likelihood (0.9375 to 0.9395), and traffic-light violation likelihood (0.9569 to 0.9589), while the kinematic components remain mixed. Thus, hierarchical densification recovers part of the official WOSAC performance lost during RL post-training while further improving direct closed-loop safety.

The ablations reinforce this trade-off. Reference-policy KL and teacher-forced BC raise RMM to approximately 0.774, but their direct CR increases to 3.227% and 4.209%, respectively. Collision-only training has 2.986% CR while degrading RMM to 0.5233 and increasing minADE to 20.3284 m. Optimizing safety and per-rollout fidelity therefore requires considering the official distributional metrics together with direct CR, OR, and ADE rather than relying on either family alone.

A.11 Vehicle-Only Safety and Fidelity Metrics

WOMD contains vehicles, pedestrians, and cyclists, all of which can enter the official evaluated-agent set. We additionally report a vehicle-anchor analysis with the trajectory evaluator used for the saved rollouts. Following the WOMD convention, an agent is a vehicle when its object type is 1. It selects vehicle anchors from the logged `tracks_to_predict` role and applies the trajectory evaluator's validity and generated-step masks. Vehicle CR counts a collision whenever an evaluated vehicle overlaps any valid participant, including a pedestrian or cyclist. It is not restricted to vehicle–vehicle pairs. Vehicle OR and ADE use the corresponding valid vehicle trajectories. The trajectory evaluator pools eligible agent records and uses its own geometric collision and offroad routines, whereas the official all-agent metrics average WOSAC scenario records equally. The vehicle columns are complementary measurements because both the evaluated population and the evaluator reductions and geometries differ from the official WOSAC metrics.

Vehicle-conditioned metrics use the same checkpoints selected by all-agent monitoring CR. For CR and OR, Table 17 reports the fraction of evaluated agents that exhibit the corresponding event at least once. ADE is averaged over the complete evaluated rollout.

We additionally compare the generated and logged vehicle distributions through histogram-based KL divergence in Table 18. The statistics cover vehicle kinematics, map-relative position, and pairwise interaction geometry. Lower values indicate a distribution closer to the logged trajectories. These divergences quantify distributional agreement rather than near-miss event frequency.

Dense RL achieves a mean vehicle CR of 1.464% and OR of 5.424% across three seeds, while reducing vehicle ADE from 3.7037 m for SMART to 2.7608 m. Relative to the Uniform refined three-seed means of 2.161% CR, 5.845% OR, and 2.7477 m ADE, Dense RL training lowers vehicle CR and OR with a 0.0131 m increase in ADE. The vehicle-anchor results support the same broad safety–fidelity trade-off under this complementary evaluator. Rollout-level credit and Agent-only credit have substantially worse vehicle metrics than Dense RL training, while collision-only optimization incurs severe offroad behavior and trajectory error.

Compared with SMART, Dense RL substantially reduces the divergence in linear speed (0.0120 to 0.0033), nearest-object distance (0.0053 to 0.0026), and TTC (0.0094 to 0.0085), while increasing the acceleration and angular-motion divergences. Relative to the Uniform refined three-seed means, it has lower nearest-object and TTC divergence and slightly lower road-edge divergence, but larger angular-speed, angular-acceleration, and THW divergence. The safety improvement is accompanied by closer agreement in some interaction-geometry distributions and shifts in short-term dynamics. These aggregate divergences do not establish that sharper evasive responses caused the shifts. The extreme KL values for rollout-level credit and collision-only optimization are consistent with their large ADE and OR, whereas the auxiliary KL and BC variants better preserve several kinematic distributions without matching the vehicle CR of the proposed method.

Table 17: Official all-agent metrics and complementary vehicle-anchor metrics from the trajectory evaluator. CR and OR are percentages and ADE is in meters. The two evaluator reductions and geometries differ as described in Appendix A.11. Rows marked "3 seeds" report means. All-agent sample standard deviations appear in Table 1. Other rows are single runs.

Method or variant	All CR ↓	Veh. CR ↓	All OR ↓	Veh. OR ↓	All ADE ↓	Veh. ADE ↓
SMART	4.080	2.1561	11.869	5.6129	2.8230	3.7037
CAT-K	3.664	2.0352	11.599	**5.3030**	2.7100	3.6097
Rollout-level credit	29.325	27.9450	34.105	29.7808	11.1556	13.9242
Uniform refined (3 seeds)	3.394	2.161	11.511	5.845	2.0805	2.7477
Dense RL (3 seeds)	**2.914**	**1.464**	**11.346**	5.424	2.1055	2.7608
Only ADE reward	3.387	2.2347	11.640	6.1089	2.1074	2.7808
Only collision reward	2.986	2.0184	45.049	40.7977	23.0570	29.0731
Agent-only credit	7.830	6.411	15.954	10.563	2.7128	3.4971
Episode-only	3.164	1.8913	11.444	5.7392	**2.0509**	**2.7091**
State-only	3.313	2.1369	11.674	6.1282	2.0917	2.7567
+ Full-token GRPO	3.198	1.9236	11.422	5.8003	2.0724	2.7306
+ KL	3.227	1.6144	11.563	5.4295	2.1779	2.8564
+ BC	4.209	2.8015	12.077	6.1740	2.2517	2.9414
+ Full-token GRPO + KL	3.141	1.9605	11.446	6.0239	2.0612	2.7217
+ Full-token GRPO + BC	3.570	2.2865	11.773	5.9691	2.1197	2.7985

Table 18: Histogram KL divergence between simulated and logged vehicle distributions. Lower is closer. THW and TTC denote time headway and time to collision. Rows marked "3 seeds" average the per-run divergences, rather than pooling the underlying histograms. Other rows are single runs.

Method or variant	Speed	Accel.	Ang. spd.	Ang. acc.	Road dist.	Obj. dist.	THW	TTC
SMART	0.0120	0.1165	0.0156	0.1207	0.0016	0.0053	**0.0026**	0.0094
CAT-K	0.0050	**0.1082**	**0.0112**	**0.1007**	**0.0010**	0.0054	0.0054	**0.0057**
Rollout-level credit	0.6877	0.1536	0.1834	0.1245	0.1257	0.2223	0.1445	0.5652
Uniform refined (3 seeds)	0.0028	0.1318	0.0221	0.1379	0.0025	0.0035	0.0081	0.0106
Dense RL (3 seeds)	0.0033	0.1312	0.0294	0.1703	0.0025	**0.0026**	0.0096	0.0085
Only ADE reward	0.0035	0.1343	0.0227	0.1463	0.0027	0.0032	0.0080	0.0115
Only collision reward	1.1183	0.3836	0.6514	0.5126	0.1433	0.3309	0.3252	0.0661
Agent-only credit	**0.0025**	0.1532	0.0658	0.2531	0.0053	0.0143	0.0122	0.0414
Episode-only	0.0029	0.1305	0.0210	0.1339	0.0025	0.0036	0.0063	0.0095
State-only	0.0033	0.1319	0.0223	0.1476	0.0025	0.0032	0.0071	0.0110
+ Full-token GRPO	0.0033	0.1335	0.0205	0.1395	0.0023	0.0032	0.0061	0.0103
+ KL	0.0043	0.1259	0.0147	0.1323	0.0022	0.0027	0.0073	0.0067
+ BC	0.0039	0.1277	0.0151	0.1231	0.0019	0.0042	0.0061	0.0117
+ Full-token GRPO + KL	0.0030	0.1244	0.0156	0.1265	0.0023	0.0029	0.0060	0.0098
+ Full-token GRPO + BC	0.0031	0.1305	0.0188	0.1362	0.0025	0.0038	0.0064	0.0114

A.12 Empirical Safety–Fidelity Trade-Off

Figure 11 complements Tables 2 and 14 by jointly visualizing collision rate and per-rollout ADE for the supervised baselines and the original component and auxiliary variants. Each policy point is a single checkpoint. In particular, the Dense RL and Uniform refined points use seed 42 rather than their three-seed means. Lower values on both axes are preferred. The blue points are non-dominated among the non-degenerate comparison variants, and the shaded area visualizes the region dominated by this alternative frontier. The main method is highlighted separately. The axes are capped at ADE > 3.0 m and CR $> 3.75\%$ to preserve resolution around the practical operating range. Out-of-range methods are placed at the corresponding boundaries, while their exact values remain in the cited tables and Table 1.

The main method extends the comparison frontier in the safety-oriented direction: it achieves the lowest CR among the non-degenerate variants, while Episode-only and *+ Full-token GRPO + KL* attain slightly lower ADE at higher collision rates. Relative to Episode-only, the complete method reduces seed-42 CR from 3.164% to 2.933%, with ADE changing from 2.0509 to 2.1222 m. The

collision-only reward has CR of 2.986%, which is also higher than the default, and its ADE of 23.0570 m and OR of 45.049% indicate severely degraded behavior. The plotted frontier is descriptive for these checkpoints and hyperparameter settings. The separate group-size study, for example, includes a lower-CR $G = 64$ run obtained with a larger rollout budget.

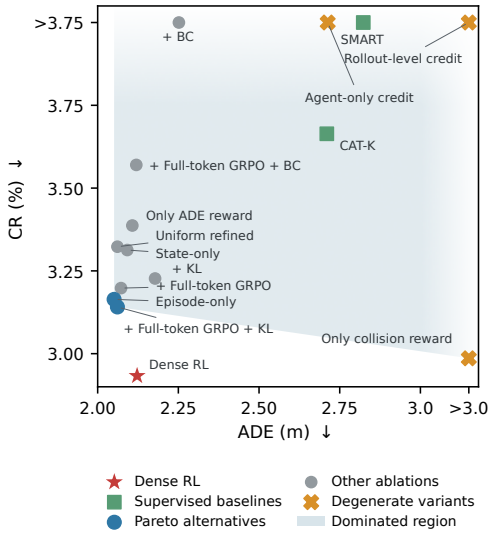

Figure 11: Empirical safety–fidelity trade-off for the supervised baselines and ablations. Lower CR and ADE are preferred. Capped axes retain severely degraded variants without compressing the practical operating range.

A.13 Collision Rate versus Collision Indication Likelihood

In the main experiments, we use collision rate (CR) instead of WOSAC collision indication likelihood as the safety metric. This is because direct CR and WOSAC collision indication likelihood answer different questions. CR is the empirical frequency of geometric collision indicators in generated closed-loop rollouts under the evaluated validity mask, directly measuring the event targeted by safety-oriented post-training. Collision indication likelihood instead measures how closely the distribution of simulated collision indicators matches the logged-data distribution. High likelihood reflects agreement with the logged collision-indicator distribution and can coexist with a higher absolute collision frequency. Agent count, valid duration, and repeated collision states can also change the relative ranking of checkpoints.

Figure 12 retrospectively scans the available paired validation records across experiment families. We restrict the display to likelihood ≥ 0.95 and CR $\leq 10\%$ to inspect the high-likelihood operating region. This retains 1,012 checkpoint records from 80 runs out of 1,732 records from 117 runs. Within this restricted range, Pearson correlation is $r = -0.688$ (run-cluster-bootstrap 95% CI $[-0.790, -0.573]$) and Spearman correlation is $\rho = -0.681$ (95% CI $[-0.796, -0.540]$). The intervals use 10,000 bootstrap samples that resample complete training runs as clusters, preserving dependence among checkpoints from the same optimization trajectory.

The association remains negative, but the moderate rank correlation, visible branch structure, and deviations from the fitted line show that likelihood does not uniquely determine CR once models have entered the high-likelihood regime. The linear fit accounts for approximately 47% of pooled pointwise variance ($r^2 = 0.473$), leaving substantial CR variation unresolved. This distinction follows from the metric definitions: likelihood rewards agreement with the logged collision-indicator distribution, whereas CR always prefers fewer simulated collisions. We therefore use collision indication likelihood as a realism diagnostic and direct CR for safety-oriented checkpoint comparison. Final model comparisons therefore use the fixed testing set and retain both direct rates and the complete official metrics in Sec. A.10. Because the historical monitoring runs used different validation subsets and between 2 and 10 batches per checkpoint, we treat the pooled association as descriptive.

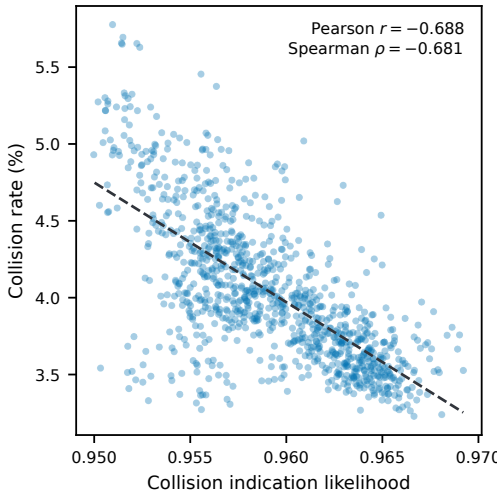

Figure 12: Collision likelihood versus direct CR for historical checkpoints with likelihood ≥ 0.95 and CR $\leq 10\%$. The dashed fit visualizes the association only.

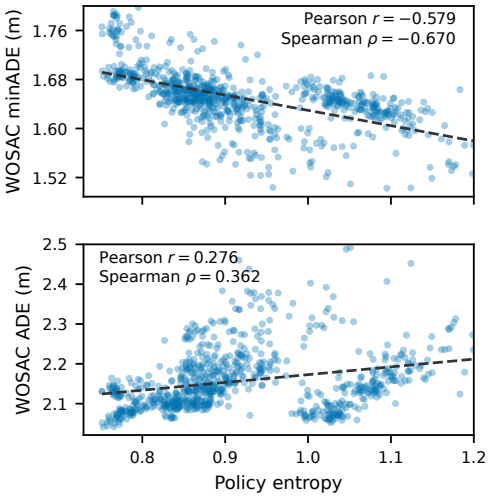

Figure 13: Entropy versus minADE and ADE for historical checkpoints with entropy ≤ 1.2, minADE ≤ 1.8 m, and ADE ≤ 2.5 m. Dashed fits visualize the associations only.

A.14 ADE VERSUS MINADE AND POLICY ENTROPY

In the main experiments, we use ADE instead of minimum average displacement error (minADE) as the trajectory-fidelity metric. ADE averages displacement error over all sampled rollouts and describes fidelity expected from one stochastic simulation, while minADE selects the closest trajectory among multiple samples and instead asks whether the candidate set contains at least one trajectory close to the log. This distinction matters in safety-critical interactions, where safe actions may occupy a narrow mode and evaluation should reflect the quality of each rollout in addition to best-of-set coverage. A diverse pretrained policy can therefore achieve low minADE while retaining higher per-rollout ADE. Token entropy characterizes uncertainty or diversity of the categorical policy, whereas fidelity is assessed by trajectory metrics.

Figure 13 retrospectively scans every paired validation record in experiment families. We jointly restrict entropy to $[0, 1.2]$, minADE to $[0, 1.8]$ m, and ADE to $[0, 2.5]$ m so both panels use identical checkpoints. This retains 755 checkpoints from 53 runs out of 1,267 records from 92 runs. In this retrospective operating range, entropy and minADE are negatively associated: Pearson $r = -0.579$ (run-cluster-bootstrap 95% CI $[-0.658, -0.485]$) and Spearman $\rho = -0.670$ (95% CI $[-0.756, -0.553]$). Entropy and ADE have a weaker positive association: Pearson $r = 0.276$ (95% CI $[0.140, 0.419]$) and Spearman $\rho = 0.362$ (95% CI $[0.176, 0.556]$). All intervals use 10,000 bootstrap samples that resample complete training runs, preserving dependence among checkpoints from the same optimization trajectory.

The contrasting associations show why ADE and minADE are not interchangeable measures of fidelity. Higher entropy broadens candidate coverage and increases the chance that at least one trajectory lies close to the log, even when a typical rollout remains inaccurate. ADE averages over rollouts and is less strongly associated with entropy in these records. Because our safety-oriented setting requires each sampled rollout to be safe and plausible, we use ADE as the primary fidelity metric and minADE as a complementary coverage diagnostic. Entropy is used as an interpretive diversity diagnostic; optimization and simulator quality are assessed by separate objectives and metrics. These historical associations are descriptive, while final comparisons use fixed-set CR/OR/ADE.