PARETOTAIL: DISTRIBUTIONAL PARETO ADAPTATION FOR LONG-TAIL MOTION FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Motion forecasting in naturalistic driving is inherently long-tailed, leaving rare yet safety-critical behaviors severely underrepresented. Existing approaches address this problem either by strengthening tail-sample learning or by generating additional rare behaviors, both increasing the influence of underrepresented behaviors during training. However, we observe a long-tail Pareto problem: stronger tail learning can improve rare-behavior forecasting while destabilizing predictions over the naturalistic distribution. We present ParetoTail, a framework that combines controllable tail generation with distributional Pareto adaptation. ParetoTail represents driving behavior in a continuous space of longitudinal motion, lateral motion, and interaction, and uses a conditional diffusion model to generate diverse rare trajectories in sparsely covered regions. To incorporate these behaviors without unrestricted distributional drift, we model a deficit-weighted tail correction field together with a naturalistic prediction geometry, whose interaction defines a compact tail-efficient adaptation space. Within this space, ParetoTail seeks a distributionally Pareto-dominating update when tail and naturalistic objectives admit co-improvement; otherwise, it maximizes tail improvement under explicit naturalistic-performance and prediction-drift constraints. Experiments demonstrate that ParetoTail improves forecasting on challenging long-tail behaviors while preserving strong performance on common naturalistic scenarios, with generated trajectories remaining controllable, diverse, and physically plausible.

1 INTRODUCTION

The long-tail challenge remains one of the major obstacles to the reliable deployment of autonomous driving systems. Due to the inherently long-tailed nature of real-world driving, naturalistic datasets densely cover routine behaviors but contain only limited observations of rare, complex, and safety-critical situations. Consequently, data-driven motion forecasting models can achieve strong average performance while remaining unreliable in precisely those scenarios where accurate prediction is most critical for safe planning and decision making (Caesar et al., 2020; Wang et al., 2023). This creates a fundamental mismatch between improving overall forecasting performance and improving prediction on underrepresented behaviors: advances driven by increasingly large naturalistic datasets or stronger forecasting models are naturally dominated by well-represented behaviors, while the sparse tail continues to receive insufficient supervision. Improving prediction under such underrepresented behaviors has therefore become an important problem for robust autonomous driving.

As illustrated in Fig. 1, progress in motion forecasting has largely followed two directions in the common–tail performance space. General-purpose forecasting improves overall predictive capability through larger-scale naturalistic data and stronger prediction models, but its gains are dominated by well-represented common behaviors. Long-tail forecasting, in contrast, explicitly strengthens learning on underrepresented behaviors. Algorithm-centric approaches, including loss reweighting, contrastive representation learning, and meta-learning, modify the learning procedure to emphasize existing tail samples (Wang et al., 2023; Rao et al., 2025), while data-centric approaches use targeted collection or controllable generation to increase the exposure of rare behaviors (Park et al., 2025). Despite their different forms, both long-tail directions essentially increase the effective learning influence of tail behaviors. However, stronger tail learning can improve rare-behavior prediction while simultaneously disturbing previously learned common behaviors. We refer to this tension between tail improvement and common-behavior preservation as the long-tail Pareto problem.

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.

The key difficulty is that increasing tail influence determines how strongly rare behaviors affect learning, but not how the resulting model should adapt. Different parameter updates can yield similar tail improvements while inducing substantially different changes to predictions over the broader naturalistic distribution. Consequently, the common–tail trade-off observed under a particular reweighting, augmentation, or training strategy does not necessarily represent an intrinsic Pareto limit; it may instead arise from the restricted adaptation path explored by that strategy. Existing methods may occasionally reach favorable joint improvements, but such outcomes are implicit rather than explicitly identified or controlled. This raises a more fundamental question: can we determine whether a Pareto-improving adaptation direction exists, and, when it does not, control the unavoidable trade-off explicitly?

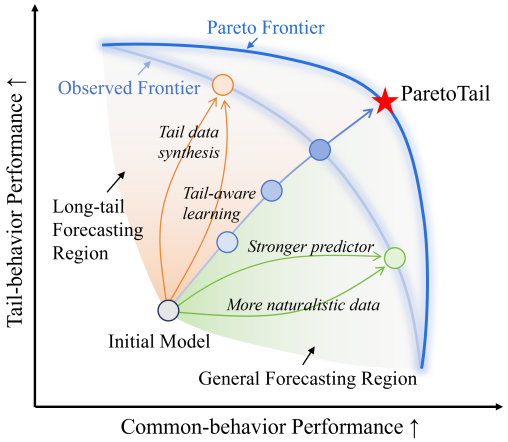


Figure 1: Common–tail performance landscape in motion forecasting. ParetoTail explicitly searches for Pareto-efficient adaptation directions beyond the trade-offs typically explored by general-purpose and long-tail forecasting methods.

To address this problem, we propose **ParetoTail**, consisting of a controllable long-tail trajectory generator and a Pareto-aware forecasting adaptation framework. The generator synthesizes diverse rare trajectories characterized by multi-dimensional motion and interaction attributes, providing observations in sparsely covered behavioral regions. ParetoTail then models how these underrepresented behaviors seek to modify the forecasting model through a continuous tail-correction field, while a naturalistic prediction geometry characterizes how candidate updates affect predictions over the original driving distribution. Their interaction identifies adaptation directions that provide strong and broadly distributed tail benefits with limited disturbance to naturalistic forecasting. Based on this structure, ParetoTail first searches for a distributionally dominating update that improves both objectives when such a direction exists; when conflict is unavoidable, it instead maximizes tail improvement under explicit naturalistic-performance and prediction-drift constraints.

In summary, our contributions are as follows:

- We develop a controllable long-tail trajectory generator that synthesizes diverse and physically plausible rare behaviors through multi-dimensional motion and interaction conditions, enriching sparsely observed regions of naturalistic driving data.
- We propose ParetoTail, which explicitly models the long-tail Pareto problem through a continuous tail-correction structure and a naturalistic prediction geometry, and adaptively seeks either a distributionally dominating update or a controlled Pareto compromise according to the local common–tail interaction.
- We conduct extensive experiments to validate both the generated long-tail trajectories and the proposed adaptation framework. ParetoTail substantially improves forecasting performance on challenging long-tail scenarios while maintaining strong performance on common naturalistic behaviors.

2 Related Work

Long-Tail Trajectory Generation. The long-tail challenge fundamentally stems from insufficient observations of rare driving behaviors, making data generation a direct approach to improving behavioral coverage. Early traffic generation and simulation methods relied heavily on handcrafted rules or learned generative models with limited control, and often struggled to jointly achieve realistic, diverse, and interaction-consistent behaviors (Suo et al., 2021; Tan et al., 2021). Recent generative models, including diffusion-based trajectory models and learned world models, provide substantially stronger multimodality and controllability, enabling joint multi-agent generation under behavioral, physical, or scenario-specific conditions (Zhong et al., 2022; Pronovost et al., 2023;

Jiang et al., 2023; Zhang et al., 2026). These advances have made targeted generation of rare behaviors increasingly practical, and recent work has begun to directly use controllable generation for long-tail trajectory augmentation (Lu et al., 2024; Park et al., 2025).

Long-Tail Trajectory Learning. Another line of research improves long-tail prediction by strengthening learning of underrepresented trajectories. Early approaches relied on difficulty-aware sampling, reweighting, or contrastive learning to emphasize challenging cases (Makansi et al., 2021; Wang et al., 2023). Subsequent methods recognize the heterogeneity of tail behaviors, learning structured representations to distinguish and organize diverse rare motion patterns (Lan et al., 2024; Yang et al., 2024). More recent work shifts from static rebalancing toward adaptive tail modeling, allowing the learner to identify underrepresented behaviors from evolving training signals and scene context and adjust its learning strategy accordingly (Zhang et al., 2024; Rao et al., 2025; 2026a;b; Wang et al., 2026). Overall, this line of work has evolved from coarse tail rebalancing toward increasingly structured and adaptive modeling of heterogeneous long-tail behaviors.

Taken together, generation-based and learning-based approaches have improved how underrepresented behaviors are exposed, identified, and emphasized during training. Despite sophisticated tail characterization and adaptation mechanisms, their primary focus remains on strengthening the learning influence of rare behaviors. Prior studies have observed that such rebalancing can induce head–tail trade-offs or conflicting learning directions (Li et al., 2024; Yang et al., 2024; Rao et al., 2026a). What remains less explicitly modeled is how a tail-induced adaptation direction changes predictions over the broader naturalistic distribution. ParetoTail focuses on this interaction through the joint modeling of tail corrections and naturalistic prediction geometry.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given the observed motion histories of all agents in a driving scene, we denote the historical context as $X = \{s_i^t\}$, where each state s_i^t contains the position, velocity, and heading of agent i at time t, and let M denote the corresponding static map context. Our objective is to forecast the future trajectory Y of a target agent by learning the conditional distribution $p_\theta(Y \mid X, M)$. To capture the inherent multimodality of future motion, the predictor approximates this distribution with K candidate future trajectories and their associated confidence scores, denoted by $\{(\hat{Y}^{(k)}, \pi^{(k)})\}_{k=1}^{K}$, where $\pi^{(k)} \geq 0$.

3.2 OVERALL FRAMEWORK

The overall framework of **ParetoTail** is illustrated in Fig. 2. It consists of two complementary stages. First, an independently trained controllable generator synthesizes diverse trajectories in sparsely observed regions of the continuous behavioral space, enriching the available long-tail observations beyond the naturalistic training data. Second, rather than directly mixing these generated samples into conventional training, ParetoTail models the distributional correction induced by underrepresented behaviors and characterizes how candidate parameter updates affect predictions over the original naturalistic distribution. By jointly exploiting the directional structure of tail corrections and the naturalistic prediction geometry, the framework identifies a compact tail-efficient adaptation space and performs Pareto-aware optimization within it. When a direction providing broad tail improvement together with naturalistic improvement is available, ParetoTail seeks a distributionally dominating update; otherwise, it switches to a controlled adaptation regime that maximizes tail benefit while explicitly limiting naturalistic degradation and prediction drift.

3.3 CONTINUOUS BEHAVIORAL REPRESENTATION

Trajectory rarity is inherently continuous and may arise from different physical sources. We characterize driving behavior along three semantically distinct aspects: longitudinal motion, lateral motion, and interaction. The first two follow the route-aligned decomposition widely used in autonomous driving, separating motion along a reference route from deviation across it (Werling et al., 2010), and thereby capture the target agent's intrinsic kinematics. However, rare behavior can also arise from how the target responds to surrounding traffic, which cannot be described by single-agent mo-

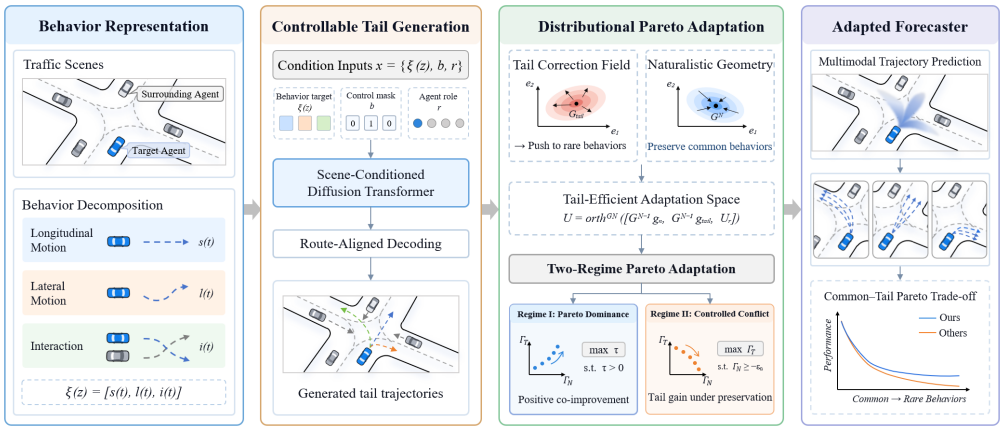


Figure 2: Overview of the proposed ParetoTail framework. The framework integrates continuous behavior representation, controllable tail generation, and distributional Pareto adaptation for long-tail motion forecasting while preserving performance on common naturalistic driving scenarios.

tion alone and is explicitly modeled in modern multi-agent forecasting (Yuan et al., 2021; Sun et al., 2022). We therefore introduce interaction as a complementary relational aspect. Together, the three components provide a physically interpretable decomposition of driving behavior into along-route motion, cross-route motion, and inter-agent interaction.

For each trajectory z, we collect the corresponding fine-grained attributes as $\xi(z) = [\xi_{\text{lon}}(z), \xi_{\text{lat}}(z), \xi_{\text{int}}(z)]$. Motion-related attributes are computed in a route-aligned coordinate system, and all attributes are normalized using statistics from the naturalistic training data. The fine-grained descriptor retains explicit physical semantics and is used for controllable generation and sample verification. Detailed attribute definitions are provided in Appendix A.

For distributional modeling, directly operating on all fine-grained attributes would lead to a sparse, high-dimensional behavioral space. We therefore preserve the three semantic axes while summarizing the attributes within each group into a single continuous coordinate. For $g \in \{\text{lon}, \text{lat}, \text{int}\}$,

$$a_g(z) = \left(\sum_{k \in I_g} w_k \, \tilde{\xi}_k(z)^\gamma \right)^{1/\gamma}, \qquad \sum_{k \in I_g} w_k = 1, \tag{1}$$

where I_g indexes the attributes in group g, and γ controls the emphasis on strongly expressed attributes. The resulting descriptor $a(z) = [a_{\text{lon}}(z), a_{\text{lat}}(z), a_{\text{int}}(z)]$ defines the continuous behavioral space used for coverage estimation and ParetoTail adaptation.

3.4 Controllable Tail Generator

Given a naturalistic scene (X, M) and a requested fine-grained behavioral target $\xi^\star$, we generate scene-consistent futures that realize the specified rare behavior. To preserve interaction consistency, the target agent and its relevant neighbors are generated jointly. We denote the generation condition by $\eta = (\xi^\star, b, r)$, where b specifies the actively controlled attributes and r encodes the semantic roles of participating agents. The conditional generator is formulated as

$$\widetilde{Y} \sim q_\psi(Y \mid X, M, \eta), \tag{2}$$

where $\widetilde{Y}$ contains the joint future trajectories of all generated agents. During training, $\xi^\star$ is extracted from ground-truth futures using the behavioral attributes in Sec. 3.3; at inference time, it can be explicitly specified to target sparsely observed behaviors.

We instantiate q_ψ as a scene-conditioned diffusion model over route-aligned control sequences. Each generated agent is associated with a topology-compatible route, and its future is represented by longitudinal acceleration and lateral velocity rather than global Cartesian coordinates. Scene context,

behavioral conditions, agent roles, and route information are encoded into actor–time tokens, which are jointly denoised by a Transformer across all participating agents and future steps. A differentiable route-based decoder then integrates the denoised controls into Cartesian trajectories. This representation provides direct control over physically meaningful motion variables while allowing the target maneuver and surrounding responses to be generated coherently.

The generator is trained with a standard conditional diffusion objective together with behavioral and physical constraints,

$$L_{\text{gen}} = L_{\text{diff}} + \lambda_b L_{\text{beh}} + \lambda_r L_{\text{road}} + \lambda_d L_{\text{dyn}} + \lambda_c L_{\text{col}} + \lambda_i L_{\text{int}}, \tag{3}$$

where the auxiliary terms enforce behavioral fidelity, road and dynamic feasibility, collision validity, and interaction consistency. At inference time, multiple diffusion samples are generated and refined using the same differentiable constraints, followed by a hard validity check on the realized behavioral attributes and physical feasibility. Accepted trajectories form an offline synthetic tail pool in the same input–supervision format as naturalistic data. The generator is trained independently and remains fixed during ParetoTail adaptation. Full architectural, optimization, and sampling details are provided in Appendix B.

3.5 Distributional Pareto Adaptation

Given observations of underrepresented behaviors, the key question is not only how strongly they should influence learning, but in which parameter directions the predictor should adapt. ParetoTail addresses this by coupling a distributional tail-correction field with a naturalistic prediction geometry. Their interaction defines a compact tail-efficient adaptation space, within which we distinguish locally compatible common–tail improvement from unavoidable conflict.

Distributional Tail Correction. Let $L_N(\theta) = \mathbb{E}_{z\sim P_N}[\ell_\theta(z)]$ denote the forecasting objective over the naturalistic distribution, with gradient $g_N = \nabla_\theta L_N(\theta)$. For a behavioral location a, we define the behavior-conditioned objective $L(\theta; a) = \mathbb{E}[\ell_\theta(z) \mid a(z) = a]$ and its gradient $g(a) = \nabla_\theta L(\theta; a)$. The behavior-specific correction relative to naturalistic training is

$$c(a) = g(a) - g_N. \tag{4}$$

Varying a continuously therefore defines a correction field describing how different behavioral regions seek to modify the predictor.

Because naturalistic data are highly non-uniform over the behavioral space, we weight this field according to the local coverage deficit. Let $p_N(a)$ denote the naturalistic behavioral density and $p_{\text{tar}}(a)$ the target behavioral distribution induced by the requested generation conditions over the valid space $\mathcal{A}$. We define

$$d(a) = \frac{[p_{\text{tar}}(a) - p_N(a)]_+}{\int_{\mathcal{A}} [p_{\text{tar}}(u) - p_N(u)]_+ \, \mathrm{d}u}. \tag{5}$$

Importantly, p_{tar} describes the desired generation coverage rather than the empirical distribution of accepted synthetic samples, thereby separating intended tail coverage from generator-specific sampling bias. We denote the deficit-weighted mean direction and covariance of the correction field by g_D and Σ_T, respectively; their estimation is detailed in Appendix C.

Naturalistic Prediction Geometry. Tail corrections characterize how under-covered behaviors encourage the model to change, but parameter updates of comparable magnitude can induce substantially different changes in the forecasting function. We therefore evaluate candidate updates by their prediction-space effect over the naturalistic distribution. Let $D_x(F_{\theta'}(x), F_\theta(x))$ denote a differentiable discrepancy between two multimodal forecasts for scene x, accounting for trajectory displacement and mode-confidence variation. Its local pullback to parameter space defines the naturalistic prediction geometry

$$G_N = \mathbb{E}_{x\sim P_N^X}\left[J_F(x)^\top W(x) J_F(x)\right] + \lambda_G I, \tag{6}$$

where P_N^X denotes the context marginal of the naturalistic distribution, $J_F(x) = \partial F_\theta(x)/\partial\theta$, and $W(x)$ is the local output-space metric induced by D_x. For a small parameter update s, the quadratic form $s^\top G_N s$ approximates the resulting prediction disturbance over naturalistic behaviors. Thus, G_N measures parameter directions according to how strongly they perturb the forecasting function, rather than by Euclidean parameter distance. The concrete forecast discrepancy and matrix-free computation of G_N are provided in Appendix C.

Tail-Efficient Adaptation Space. We seek directions that retain substantial variation of the deficit-weighted tail corrections, as characterized by Σ_T, while inducing limited naturalistic prediction disturbance measured by G_N. This leads to the generalized eigenproblem

$$\Sigma_T u_r = \lambda_r G_N u_r, \qquad u_r^\top G_N u_s = \delta_{rs}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots . \tag{7}$$

The leading R generalized eigenvectors $U_T = [u_1, \ldots, u_R]$ capture heterogeneous tail-correction modes that are most efficient under the naturalistic prediction geometry. We further include the geometry-aware naturalistic and mean tail learning directions,

$$U = \mathrm{orth}_{G_N}\left(\left[G_N^{-1} g_N,\ G_N^{-1} g_D,\ U_T\right]\right), \qquad U^\top G_N U = I. \tag{8}$$

The resulting low-dimensional space jointly represents naturalistic learning, average tail adaptation, and heterogeneous tail corrections. Its variational characterization is provided in Appendix C.

Two-Regime Pareto Adaptation. The observed common–tail trade-off may reflect either a restricted adaptation path or an intrinsic local conflict. ParetoTail distinguishes these cases within the tail-efficient space. Let $s = Uz$ denote a candidate descent update, with $\theta^+ = \theta - s$. We define the local improvement surrogates

$$\begin{aligned} \Gamma_N(s) &= g_N^\top s - \frac{\beta_N}{2}\|s\|_{G_N}^2, \\ \Gamma_T(s) &= g_D^\top s - \kappa\sqrt{s^\top \Sigma_T s} - \frac{\beta_T}{2}\|s\|_{G_N}^2, \end{aligned} \tag{9}$$

where $\|s\|_{G_N}^2 = s^\top G_N s$, $\beta_N, \beta_T \geq 0$ are local curvature coefficients accounting for finite-step effects, and $\kappa \geq 0$ controls the penalty on behavioral variation of the tail improvement. Γ_N measures local naturalistic improvement, whereas Γ_T favors large mean tail improvement while discouraging directions whose benefits are concentrated on only a subset of under-covered behaviors.

Regime I: Distributional Pareto dominance. Let $\mathcal{B}_r = \{z : \|z\|_2 \leq r\}$ denote a local trust region with radius r. We first test whether a locally compatible update exists by solving

$$\max_{z \in \mathcal{B}_r,\ \tau} \ \tau \quad \text{s.t.} \quad \Gamma_N(Uz) \geq \omega_N \tau,\ \Gamma_T(Uz) \geq \omega_T \tau, \tag{10}$$

where $\omega_N, \omega_T > 0$ control the relative scaling of the two improvement margins. Because $U^\top G_N U = I$, $\|Uz\|_{G_N}^2 = \|z\|_2^2$, so $\mathcal{B}_r$ directly limits the local naturalistic prediction disturbance. When $\tau^\star > 0$, the resulting update has positive margins for both objectives and is locally distributionally Pareto-dominating under the surrogate model.

Regime II: Controlled Pareto conflict. If $\tau^\star = 0$, no positive co-improvement margin is available within the current local surrogate model and adaptation space. We then maximize tail improvement subject to an explicit naturalistic preservation budget,

$$\max_{z \in \mathcal{B}_r} \Gamma_T(Uz) \quad \text{s.t.} \quad \Gamma_N(Uz) \geq -\epsilon_N, \tag{11}$$

where $\epsilon_N \geq 0$ specifies the maximum tolerated local degradation of the naturalistic objective. Thus, ParetoTail exploits co-improvement directions when available and otherwise maximizes tail benefit under explicit naturalistic-performance and prediction-drift control.

4 Experiments

4.1 Experimental Setup

We evaluate ParetoTail on three widely used motion forecasting datasets: nuScenes (Caesar et al., 2020), NGSIM (Deo & Trivedi, 2018), and HighD (Krajewski et al., 2018). nuScenes serves as our primary benchmark for evaluating overall forecasting accuracy, long-tail performance, and the Common–Tail trade-off, while NGSIM and HighD provide complementary evaluation in highway driving scenarios. For nuScenes, we report minimum average displacement error (minADE), minimum final displacement error (minFDE), and miss rate (MR) under different numbers of predicted modes. For NGSIM and HighD, we follow their standard 1–5 s prediction protocols and report horizon-wise RMSE. All controlled comparisons use the same forecasting architecture, with only the training or adaptation strategy varied. Detailed dataset splits, evaluation protocols, model configurations, training procedures, and comparison settings are provided in Appendix D.1.

4.2 LONG-TAIL GENERATION EVALUATION

We first evaluate whether the proposed generator can produce valid, controllable, and diverse long-tail trajectories. As shown in Table 1, 99.5% of the 3,000 generated trajectories pass the automatic joint-validity check, while all trajectories pass the center-distance collision diagnostic. A complementary manual inspection of 100 randomly sampled trajectories further shows that 95 satisfy the expected map and interaction semantics. The generator also achieves an 80.0% strict requested-condition hit rate across ten event categories, with the lowest category-level rate remaining at 72.0%, indicating consistent controllability across different tail behaviors. Meanwhile, the generated set exhibits broad behavioral coverage, with a normalized event entropy of 0.90 and a median nearest-neighbor future RMS of 1.950 m. Together, these results show that the generator can reliably realize requested rare behaviors while maintaining high validity and nontrivial trajectory diversity.

Table 1: Generation quality of 3,000 synthetic trajectories across ten requested event categories. The manual map and interaction audit is conducted on 100 randomly sampled trajectories.

Diagnostic	Value	Diagnostic	Value
Automatic joint validity	99.5%	Normalized event entropy	0.90
Center-distance collision-free	100.0%	Median nearest-neighbor future RMS	1.95 m
Strict requested-condition hit	80.0%	Distinct target futures	3,000/3,000
Lowest event-category hit	72.0%	Manual map/interaction check	95/100

4.3 MOTION FORECASTING PERFORMANCE

Overall Performance. Table 2 compares ParetoTail with published motion forecasting methods on the full nuScenes validation set. ParetoTail achieves 1.17 m minADE_5 and 2.34 m minFDE_5, while obtaining the best reported minADE_{10} and MR_5 among the compared methods. Its minFDE_1 of 6.55 m is also competitive with the best reported result of 6.45 m. These results show that improving underrepresented behaviors does not come at the expense of overall forecasting performance. Since the baseline results are taken from the corresponding literature rather than matched reruns, we use this comparison to establish the overall competitiveness of ParetoTail, while controlled comparisons under the same forecasting architecture are presented in Sec. 4.4.

Table 2: Overall nuScenes validation performance. ADE/FDE are in metres and MR_5 is a proportion. **Bold** and underlined values denote the best and second-best reported results.

Model	Venue	minADE_{10}	minADE_5	minFDE_5	minFDE_1	MR_5
NEST (Wang et al., 2025)	AAAI	1.05	<u>1.18</u>	<u>2.39</u>	6.87	<u>0.50</u>
PGP (Deo et al., 2022)	CoRL	<u>1.03</u>	1.30	2.52	7.17	0.61
Autobot (Girgis et al., 2022)	ICLR	<u>1.03</u>	1.37	3.40	8.19	0.62
MultiPath (Chai et al., 2020)	CoRL	1.14	1.44	4.83	7.69	0.76
AgentFormer (Yuan et al., 2021)	ICCV	1.31	1.59	3.14	**6.45**	0.68
Trajectron++ (Salzmann et al., 2020)	ECCV	1.51	1.88	5.63	9.52	0.70
ParetoTail	–	**1.00**	**1.17**	**2.34**	<u>6.55</u>	**0.48**

Long-Tail Performance. We next examine forecasting performance on difficult and underrepresented behaviors. Table 3 evaluates the upper tail of each model's own minFDE_5 error distribution. ParetoTail achieves the lowest minADE_5 and minFDE_5 across all five worst-case subsets. On the worst 1% of predictions, ParetoTail obtains 6.50/15.00 m minADE_5/minFDE_5, compared with 7.55/18.78 m for Q-EANet, the strongest competing result in this subset. The advantage remains consistent as the evaluated subset expands, with ParetoTail reaching 4.10/10.30 m on the worst 5%, compared with 4.55/11.49 m for Q-EANet. Since each predictor independently defines its own worst-case subset, this experiment characterizes the severity of each model's error tail rather than providing a paired comparison on identical scenes.

Table 3: Self-ranked worst-case performance on nuScenes in minADE_5/minFDE_5 (m). Each model independently ranks its validation instances by minFDE_5. Lower is better; bold and underlined values denote the best and second-best results.

Model	Venue	Top 1%	Top 2%	Top 3%	Top 4%	Top 5%	All
Q-EANet (Chen et al., 2024)	IET ITS	7.55/18.78	6.15/15.58	5.44/13.76	4.94/12.49	4.55/11.49	1.20/2.45
LAformer (Liu et al., 2024)	CVPR	8.19/19.03	6.73/15.81	5.89/13.90	5.33/12.60	4.90/11.61	1.19/2.42
PGP (Deo et al., 2022)	CoRL	8.86/21.92	7.21/17.90	6.24/15.68	5.52/13.77	5.02/12.44	1.28/2.52
UniTraj (MTR) (Feng et al., 2024)	ECCV	7.84/21.69	6.44/18.06	5.69/15.95	5.18/14.49	4.78/13.37	**1.15**/2.61
ParetoTail	–	**6.50/15.00**	**5.65/13.90**	**5.15/12.20**	**4.45/11.10**	**4.10/10.30**	1.17/**2.34**

To directly assess the Common–Tail trade-off, Table 4 evaluates all controlled variants on the same fixed density-defined partition. ParetoTail reduces the Tail minADE_5 to 2.27 m, compared with 2.80 m for Natural-only and 2.69 m for Naive, while simultaneously lowering the corresponding Common error to 1.11 m from 1.16 m and 1.18 m. This joint improvement shows that stronger forecasting on underrepresented behaviors does not require sacrificing performance on common naturalistic behaviors. Additional experiments further demonstrate consistent generalization to highway scenarios, with average RMSEs of 1.35 m on NGSIM and 0.17 m on HighD, while retaining an inference time of approximately 14 ms per batch on an RTX 3090. Detailed cross-dataset and efficiency results are provided in Appendices D.2 and D.4, respectively.

Table 4: nuScenes minADE_5 (m) on the full validation set and the fixed Common and Tail subsets. **Bold** and underlined values denote the best and second-best reported results.

Model	Overall	Fixed Common	Fixed Tail
Natural-only	1.25	1.16	2.80
Grouped	1.27	1.19	2.74
Naive	1.25	1.18	2.69
ParetoTail	**1.17**	**1.11**	**2.27**

4.4 Ablation Studies

We compare different training strategies using the same forecasting architecture in Table 5. Behavioral grouping alone does not improve the Natural-only predictor, while directly incorporating generated tail samples with Naive training does not yield consistent gains. In contrast, ParetoTail improves all reported metrics, reducing minADE_5/minFDE_5 from 1.25/2.46 m with Naive training to 1.17/2.34 m, while minFDE_1 decreases from 7.01 to 6.55 m and MR_5 from 0.49 to 0.48. The improvement is also pronounced on difficult cases: for the self-ranked worst 1% and 5% subsets, ParetoTail reduces minADE_5/minFDE_5 from 8.24/20.72 m and 5.22/13.12 m to 6.50/15.00 m and 4.10/10.30 m, respectively. These results indicate that the gains cannot be attributed to generated tail samples alone, but depend on how they are incorporated during adaptation. We further ablate the individual components in Appendix D.3; removing tail generation, coverage weighting, the correction field, the naturalistic prediction geometry, or the two-regime switch consistently degrades Tail performance, while removing the naturalistic budget improves Tail accuracy at the expense of Common and overall performance, consistent with its role in controlling tail-oriented adaptation.

Table 5: Training controls on nuScenes. ADE/FDE denote minADE/minFDE (m), and MR_5 is a proportion. All variants use the same forecasting architecture. Results are mean $\pm$ SD over three seeds; lower is better and bold denotes the best result.

Model	minADE_{10}	minADE_5	minFDE_5	minFDE_1	MR_5
Natural-only	1.01 ($\pm$0.01)	1.25 ($\pm$0.01)	2.46 ($\pm$0.03)	6.98 ($\pm$0.07)	0.49 ($\pm$0.01)
Grouped	1.01 ($\pm$0.01)	1.27 ($\pm$0.01)	2.52 ($\pm$0.03)	7.11 ($\pm$0.08)	0.49 ($\pm$0.01)
Naive	1.01 ($\pm$0.01)	1.25 ($\pm$0.01)	2.46 ($\pm$0.03)	7.01 ($\pm$0.08)	0.49 ($\pm$0.01)
ParetoTail	**1.00** ($\pm$0.01)	**1.17** ($\pm$0.01)	**2.34** ($\pm$0.02)	**6.55** ($\pm$0.06)	**0.48** ($\pm$0.01)

4.5 Pareto Adaptation Analysis

We examine whether the local optimization signals of ParetoTail translate into the intended realized behavior during adaptation. Regime I is activated in 40.00% of the updates, and 90.00% of these updates lead to realized decreases in both the naturalistic and tail losses. The predicted improvement margin is also positively associated with the realized joint gain, with a Spearman correlation of 0.50.

When no compatible co-improvement direction is identified and ParetoTail switches to Regime II, the naturalistic budget is violated in only 1.00% of the corresponding updates. These results indicate that the two-regime mechanism generally behaves as intended, exploiting compatible updates when available while effectively controlling naturalistic degradation under conflicting objectives.

4.6 Qualitative Forecasting Analysis

Figure 3 presents representative nuScenes comparisons between Q-EANet (Chen et al., 2024) and ParetoTail. In the successful cases of Fig. 3(a,b), the ground-truth future involves pronounced turning behavior on a curved road and at an intersection, respectively. ParetoTail produces prediction modes that more closely follow the corresponding turning geometry and cover the observed future, illustrating its ability to better represent challenging motion patterns that are less frequently observed in naturalistic training data. Additional successful examples are provided in Appendix D.5.

Figures 3(c,d) illustrate two remaining failure modes. In (c), ParetoTail captures the overall turning tendency but distributes its modes across several plausible continuations, leaving uncertainty in the precise route and turning profile. In (d), one mode follows an alternative feasible branch under ambiguous road topology. Both cases primarily reflect scene-level multimodal ambiguity rather than an inability to represent the rare behavior itself. ParetoTail addresses how underrepresented behaviors should influence distributional adaptation, but does not explicitly resolve latent route intent when multiple futures remain plausible. Stronger intent and route-disambiguation modeling therefore provides a complementary direction for future improvement.

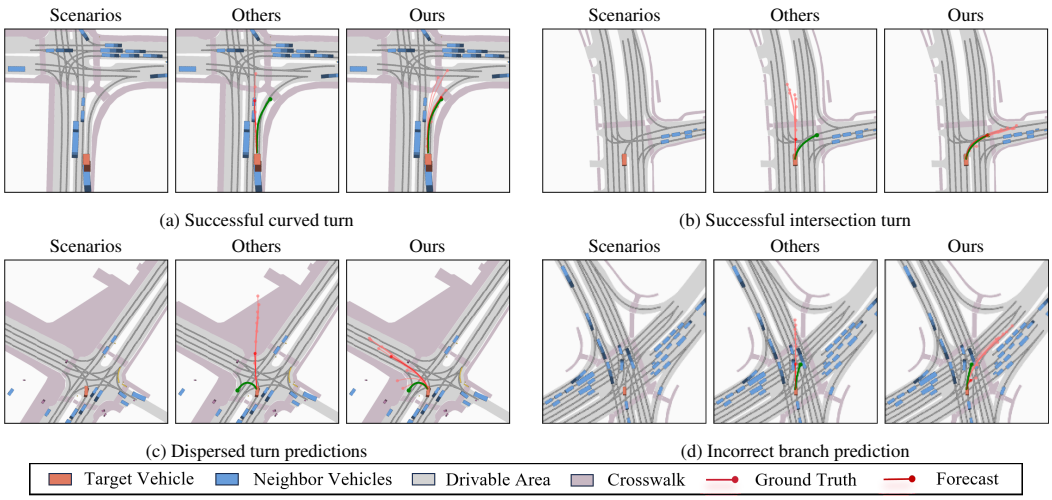


Figure 3: Representative qualitative comparisons on nuScenes. Panels (a,b) show successful turn-following cases, while panels (c,d) illustrate residual failure modes.

5 Conclusion

We present ParetoTail, a framework for long-tail motion forecasting that explicitly addresses the tension between improving underrepresented behaviors and preserving performance over the broader naturalistic distribution. ParetoTail combines controllable tail trajectory generation with distributional Pareto adaptation, using a deficit-weighted tail correction field and naturalistic prediction geometry to construct a compact tail-efficient adaptation space. Within this space, the proposed two-regime optimization exploits common–tail co-improvement when compatible directions exist and otherwise constrains naturalistic degradation while improving tail performance. Experiments across nuScenes, NGSIM, and HighD show that ParetoTail improves challenging long-tail forecasting while maintaining strong overall and common-behavior performance, with generated trajectories exhibiting high validity, controllability, and diversity. The adaptation further incurs no additional inference-time model complexity. Remaining qualitative failures are primarily associated with scene-level route and intent ambiguity, suggesting that stronger intent disambiguation could complement distributional tail adaptation in future work.

AI USE STATEMENT

Generative AI tools were used to assist with early-stage organization and drafting of this manuscript. The authors are responsible for verifying all mathematical derivations, citations, experimental claims, and implementation details, and for revising or removing any AI-assisted content that is not supported by the final study.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pp. 11621–11631, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning*, pp. 86–99. PMLR, 2020.

Jiuyu Chen, Zhongli Wang, Jian Wang, and Baigen Cai. Q-eanet: Implicit social modeling for trajectory prediction via experience-anchored queries. *IET Intelligent Transport Systems*, 18(6): 1004–1015, 2024.

Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1468–1476, 2018.

Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pp. 203–212. PMLR, 2022.

Lan Feng, Mohammadhossein Bahari, Kaouther Messaoud Ben Amor, Éloi Zablocki, Matthieu Cord, and Alexandre Alahi. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *European Conference on Computer Vision*, pp. 106–123. Springer, 2024.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2022.

Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9644–9653, 2023.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pp. 2118–2125. IEEE, 2018.

Zhengxing Lan, Yilong Ren, Haiyang Yu, Lingshan Liu, Zhenning Li, Yinhai Wang, and Zhiyong Cui. Hi-scl: Fighting long-tailed challenges in trajectory prediction with hierarchical wave-semantic contrastive learning. *Transportation Research Part C: Emerging Technologies*, 165: 104735, 2024.

Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-tailed learning as multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3190–3198, 2024.

Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2039–2049, 2024.

Jack Lu, Kelvin Wong, Chris Zhang, Simon Suo, and Raquel Urtasun. Scenecontrol: Diffusion for controllable traffic scene generation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16908–16914. IEEE, 2024.

Osama Makansi, Özgün Cicek, Yassine Marrakchi, and Thomas Brox. On exposing the challenging long tail in future prediction of traffic actors. pp. 13147–13157, 2021.

Seokha Moon, Hyun Woo, Hongbeen Park, Haeji Jung, Reza Mahjourian, Hyung-gun Chi, Hyerin Lim, Sangpil Kim, and Jinkyu Kim. Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In *European Conference on Computer Vision*, pp. 361–379. Springer, 2024.

Daehee Park, Monu Surana, Pranav Desai, Ashish Mehta, Reuben John, and Kuk-Jin Yoon. Generative active learning for long-tail trajectory prediction via controllable diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27839–27850, 2025.

Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. *Advances in Neural Information Processing Systems*, 36:68873–68894, 2023.

Bin Rao, Haicheng Liao, Yanchen Guan, Chengyue Wang, Bonan Wang, Jiaxun Zhang, and Zhenning Li. Amd: Adaptive momentum and decoupled contrastive learning framework for robust long-tail trajectory prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28849–28858. IEEE, 2025.

Bin Rao, Haicheng Liao, Chengyue Wang, Keqiang Li, Zhenning Li, and Hai Yang. Sail: Scene-aware adaptive iterative learning for long-tail trajectory prediction in autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 190:105757, 2026a.

Bin Rao, Chengyue Wang, Haicheng Liao, Qianfang Wang, Yanchen Guan, Jiaxun Zhang, Xingcheng Liu, Meixin Zhu, Kanye Ye Wang, and Zhenning Li. Differentiable semantic meta-learning framework for long-tail motion forecasting in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 962–970, 2026b.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2020.

Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6533–6542. IEEE, 2022.

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.

Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 892–901, 2021.

Chengyue Wang, Haicheng Liao, Bonan Wang, Yanchen Guan, Bin Rao, Ziyuan Pu, Zhiyong Cui, Cheng-Zhong Xu, and Zhenning Li. Nest: A neuromodulated small-world hypergraph trajectory prediction model for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 808–816, 2025.

Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. pp. 1400–1409, 2023.

Yunxi Wang, Xukai Zhai, Mingze Miao, and Zhishuai Yin. Ulptp-id: A unified learning framework under imbalanced data distributions for pedestrian trajectory prediction. In *Intelligent Transportation Engineering: Proceedings of the 10th International Conference (ICITE 2025), Beijing, China, 24-26 October 2025*, pp. 688–701. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2026.

Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a frenet frame. In *2010 IEEE international conference on robotics and automation*, pp. 987–993. IEEE, 2010.

Biao Yang, Kai Yan, Chuan Hu, Hongyu Hu, Zhitao Yu, and Rongrong Ni. Dynamic subclass-balancing contrastive learning for long-tail pedestrian trajectory prediction with progressive refinement. *IEEE Transactions on Automation Science and Engineering*, 2024.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9813–9823, 2021.

Cong Zhang, Bangyang Wei, Yang Liu, and Samuel Labi. World model-based long-tail and scenario-specific generation for autonomous driving. *Journal of Intelligent and Connected Vehicles*, 2026.

Junrui Zhang, Mozhgan Pourkeshavarz, and Amir Rasouli. Tract: A training dynamics aware contrastive learning framework for long-tail trajectory prediction. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 3282–3288. IEEE, 2024.

Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. *arXiv preprint arXiv:2210.17366*, 2022.

Appendix

A Behavioral Attribute Definitions

We provide the detailed definitions of the fine-grained behavioral attributes introduced in Sec. 3.3. Let $p_i^t \in \mathbb{R}^2$ denote the future position of agent i at step t, with sampling interval Δt. The corresponding velocity is approximated by

$$q_i^t = \frac{p_i^t - p_i^{t-1}}{\Delta t}. \tag{12}$$

For motion-related attributes, we use a route-aligned coordinate system defined by the tangent and normal directions (t_i^t, n_i^t) of the reference route. The target agent is indexed by $i = 0$.

Longitudinal. Longitudinal behavior describes how strongly the target changes its motion along the driving direction. We project the velocity onto the route tangent and compute its temporal derivatives as

$$\begin{aligned} v_{i,t}^{\parallel} &= (q_i^t)^\top t_i^t, \\ a_{i,t}^{\parallel} &= \frac{v_{i,t}^{\parallel} - v_{i,t-1}^{\parallel}}{\Delta t}, \\ j_{i,t}^{\parallel} &= \frac{a_{i,t}^{\parallel} - a_{i,t-1}^{\parallel}}{\Delta t}. \end{aligned} \tag{13}$$

We characterize longitudinal behavior using

$$\begin{aligned} \xi_{\mathrm{lon}}(z) &= \left[\Delta v^{\parallel},\ a_{\max}^{\parallel},\ b_{\max}^{\parallel},\ j_{\max}^{\parallel}\right], \\ \Delta v^{\parallel} &= \max_t v_{0,t}^{\parallel} - \min_t v_{0,t}^{\parallel}, \\ a_{\max}^{\parallel} &= \max_t \left(a_{0,t}^{\parallel}\right)_+, \\ b_{\max}^{\parallel} &= \max_t \left(-a_{0,t}^{\parallel}\right)_+, \\ j_{\max}^{\parallel} &= \max_t \left|j_{0,t}^{\parallel}\right|. \end{aligned} \tag{14}$$

Here, $\Delta v^{\parallel}$ measures the overall variation in longitudinal speed, while $a_{\max}^{\parallel}$ and $b_{\max}^{\parallel}$ capture strong acceleration and braking events, respectively. The maximum jerk $j_{\max}^{\parallel}$ complements these magnitude-based quantities by capturing abrupt changes in longitudinal control. Together, they distinguish routine speed evolution from aggressive or rapidly varying longitudinal maneuvers.

Lateral. Lateral behavior describes deviation from the nominal driving direction and the geometric complexity of the maneuver. We define the lateral velocity and motion heading as

$$\begin{aligned} v_{i,t}^{\perp} &= (q_i^t)^\top n_i^t, \\ \psi_{i,t} &= \mathrm{atan2}\left(q_{i,t}^y, q_{i,t}^x\right). \end{aligned} \tag{15}$$

Let $d_{i,t}$ denote the signed lateral displacement from the reference route. For trajectory segments with nonzero displacement, we approximate the local curvature by

$$\kappa_{i,t} = \frac{\left|\mathrm{wrap}\left(\psi_{i,t} - \psi_{i,t-1}\right)\right|}{\left\|p_i^t - p_i^{t-1}\right\|_2}, \tag{16}$$

where $\mathrm{wrap}(\cdot)$ maps angular differences to $[-\pi, \pi)$.

The lateral descriptor is defined as

$$\begin{aligned}
\xi_{\mathrm{lat}}(z) &= \left[\Delta d,\ v^{\perp}_{\max},\ \Delta\psi_{\max},\ \kappa_{\max}\right], \\
\Delta d &= \max_t d_{0,t} - \min_t d_{0,t}, \\
v^{\perp}_{\max} &= \max_t \left|v^{\perp}_{0,t}\right|, \\
\Delta\psi_{\max} &= \max_t \left|\mathrm{wrap}\left(\psi_{0,t} - \psi_{0,0}\right)\right|, \\
\kappa_{\max} &= \max_t \kappa_{0,t}.
\end{aligned} \tag{17}$$

The lateral excursion Δd reflects the overall magnitude of deviation from the reference route, while $v^{\perp}_{\max}$ measures the intensity of lateral motion. The heading variation $\Delta\psi_{\max}$ captures large directional changes, and $\kappa_{\max}$ emphasizes locally sharp trajectory geometry. These attributes jointly characterize behaviors ranging from mild lateral drift to aggressive lane changes or turning maneuvers.

Interaction. Interaction behavior characterizes how the target evolves relative to surrounding agents. For a pair of agents (i, j), we define their relative displacement and Euclidean separation as

$$\begin{aligned}
r^t_{ij} &= p^t_j - p^t_i, \\
d^t_{ij} &= \left\|r^t_{ij}\right\|_2.
\end{aligned} \tag{18}$$

For agent pairs with positive separation, the nonnegative closing speed is

$$c^t_{ij} = \left[-\frac{(r^t_{ij})^{\top}(q^t_j - q^t_i)}{d^t_{ij}}\right]_+, \tag{19}$$

which measures the component of relative velocity that reduces the inter-agent distance. We further define an inverse time-to-collision surrogate as

$$\mathrm{ITTC}^t_{ij} = \frac{c^t_{ij}}{d^t_{ij}}. \tag{20}$$

The interaction descriptor is

$$\begin{aligned}
\xi_{\mathrm{int}}(z) &= \left[d^{-1}_{\min},\ c_{\max},\ \mathrm{ITTC}_{\max}\right], \\
d_{\min} &= \min_{t,j\neq 0} d^t_{0j}, \\
c_{\max} &= \max_{t,j\neq 0} c^t_{0j}, \\
\mathrm{ITTC}_{\max} &= \max_{t,j\neq 0} \mathrm{ITTC}^t_{0j}.
\end{aligned} \tag{21}$$

The inverse minimum distance captures spatial proximity, while $c_{\max}$ distinguishes approaching interactions from nearby but non-closing agents. $\mathrm{ITTC}_{\max}$ combines proximity and closing intensity, providing a time-sensitive measure of potentially critical interactions. Together, the three attributes describe complementary spatial, kinematic, and temporal aspects of agent interaction.

The complete fine-grained descriptor is

$$\xi(z) = \left[\xi_{\mathrm{lon}}(z),\ \xi_{\mathrm{lat}}(z),\ \xi_{\mathrm{int}}(z)\right]. \tag{22}$$

All attributes are defined such that larger values correspond to stronger behavioral extremity. To make attributes with different physical units comparable while reducing sensitivity to extreme outliers, we normalize each attribute using statistics from the naturalistic training split. Let b^{99}_k denote the 99th percentile of the k-th attribute; the normalized value is

$$\tilde{\xi}_k(z) = \mathrm{clip}\left(\frac{\xi_k(z)}{b^{99}_k}, 0, 1\right). \tag{23}$$

The normalization statistics are fixed after being estimated from the naturalistic training data and are shared by generated and evaluation trajectories. The normalized attributes are subsequently aggregated into the compact behavioral descriptor $a(z)$ using Eq. 1.

B CONTROLLABLE TAIL GENERATOR

This section provides the architectural, optimization, and sampling details of the controllable tail generator introduced in Sec. 3.4. Given a naturalistic scene and a requested behavioral condition, the generator produces joint multi-agent futures that realize the specified motion characteristics while remaining compatible with the observed scene, road topology, and inter-agent relationships.

Conditioning and Route Compatibility. Let $\xi^\star$ denote the requested fine-grained behavioral attributes defined in Appendix A. We represent the generation condition as

$$\eta = (\xi^\star, b, r)\,, \tag{24}$$

where b indicates the actively controlled attributes and r specifies the semantic roles of participating agents, such as the controlled target and its interacting neighbors. During training, $\xi^\star$ is extracted from the ground-truth future using the same behavioral measurements as in Appendix A; at inference time, the active attributes can be directly specified to target desired long-tail behaviors. We encode the condition as

$$h_\eta = E_\eta(\xi^\star, b, r). \tag{25}$$

Because a requested maneuver must remain compatible with the scene topology, each generated agent is associated with a feasible route selected from map-derived candidates. Let H_i denote the observed history of agent i and $R_{i,m}$ its m-th feasible route candidate. We score each candidate by

$$s_{i,m} = \frac{E_R(R_{i,m})^\top E_Q(H_i, h_\eta)}{\sqrt{d_h}}, \tag{26}$$

where E_R and E_Q denote the route and query encoders and d_h is the embedding dimension. Topologically invalid candidates are masked before normalization, allowing the observed motion and requested behavior to jointly determine a compatible route for future generation.

Joint Route-Aligned Diffusion. Rather than diffusing global Cartesian positions, we represent the joint future by a route-aligned control tensor

$$U^0 \in \mathbb{R}^{A\times T_f\times 2}, \tag{27}$$

where A is the number of jointly generated agents and T_f is the future horizon. The two channels correspond to longitudinal acceleration and lateral velocity. This representation aligns the generative variables with the behavioral decomposition in Sec. 3.3 and provides direct control over physically interpretable motion quantities.

The standard forward diffusion process is

$$U^k = \sqrt{\bar\alpha_k}\,U^0 + \sqrt{1-\bar\alpha_k}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{28}$$

where k denotes the diffusion step and $\bar\alpha_k$ follows the cumulative noise schedule.

For each agent i and future step t, the denoiser constructs an actor–time token

$$\begin{aligned} h^k_{i,t} =& E_U(U^k_{i,t}) + E_H(H_i) + E_R(R_i) + E_{\text{role}}(r_i) \\ &+ h_\eta + E_k(k) + E_t(t), \end{aligned} \tag{29}$$

which combines the noisy control state, observed history, selected route, semantic role, behavioral condition, diffusion step, and future time. All valid actor–time tokens are jointly processed by a Transformer, enabling the requested target maneuver and the responses of surrounding agents to be coordinated within the same diffusion sample.

Let $\mathcal{V}$ denote the set of valid generated agents. The denoising objective is

$$L_{\text{diff}} = \mathbb{E}_{k,U^0,\epsilon}\left[\frac{1}{|\mathcal{V}|T_f}\sum_{i\in\mathcal{V}}\sum_{t=1}^{T_f}\left\|\epsilon_{i,t} - \epsilon_{\psi,i,t}(U^k, k, X, M, \eta)\right\|_2^2\right]. \tag{30}$$

Differentiable Decoding and Validity Objectives. The denoised variables are first mapped to bounded physical controls,

$$\begin{aligned} a_{i,t}^{\parallel} &= a_{\max} \tanh(U_{i,t,1}^{0}), \\ v_{i,t}^{\perp} &= v_{\perp,\max} \tanh(U_{i,t,2}^{0}), \end{aligned} \tag{31}$$

where $a_{\max}$ and $v_{\perp,\max}$ define admissible ranges of longitudinal acceleration and lateral velocity. Let $s_{i,t}$ and $d_{i,t}$ denote longitudinal progress and lateral offset along the selected route. The controls are integrated according to

$$\begin{aligned} v_{i,t}^{\parallel} &= \mathrm{clip}\left(v_{i,t-1}^{\parallel} + a_{i,t}^{\parallel}\Delta t, 0, v_{\max}\right), \\ s_{i,t} &= s_{i,t-1} + v_{i,t}^{\parallel}\Delta t, \\ d_{i,t} &= d_{i,t-1} + v_{i,t}^{\perp}\Delta t. \end{aligned} \tag{32}$$

Given the selected route centerline $\gamma_i(s)$ and its normal direction $n_i(s)$, the Cartesian future is recovered as

$$\widetilde{p}_i^{t} = \gamma_i(s_{i,t}) + d_{i,t}\, n_i(s_{i,t}). \tag{33}$$

The decoder is differentiable, allowing behavioral and physical objectives to directly refine the generated controls.

To enforce the requested behavioral condition, let $\mathcal{I}_b = \{k : b_k = 1\}$ denote the set of actively controlled attributes and let $g_k(\widetilde{Y})$ be the corresponding differentiable measurement on the generated future. We define

$$L_{\mathrm{beh}} = \frac{1}{|\mathcal{I}_b|} \sum_{k \in \mathcal{I}_b} \mathrm{Huber}\left(\widetilde{g}_k(\widetilde{Y}) - \widetilde{\xi}_k^{\star}\right), \tag{34}$$

where the behavioral measurements are expressed on the normalized scale defined in Appendix A. Smooth approximations are used for max- or min-based measurements when gradients are required.

Physical feasibility is encouraged through road and dynamic constraints. Let $w_i(s_{i,t})$ denote the admissible lateral half-width of the selected route. We use

$$\begin{aligned} L_{\mathrm{road}} &= \mathbb{E}_{i,t}\left[|d_{i,t}| - w_i(s_{i,t})\right]_+^2, \\ L_{\mathrm{dyn}} &= \mathbb{E}_{i,t}\left[\left(\frac{|j_{i,t}| - j_{\max}}{j_{\max}}\right)_+^2 + \left(\frac{|\kappa_{i,t}| - \kappa_{\max}}{\kappa_{\max}}\right)_+^2\right], \end{aligned} \tag{35}$$

where $j_{i,t}$ and $\kappa_{i,t}$ denote the realized jerk and curvature. The first term keeps generated trajectories within route support, while the second penalizes excessive higher-order or geometric motion that is not directly bounded by the control representation.

For multi-agent validity, let $d_{ij}^{t,\mathrm{body}}$ denote the signed separation between the physical footprints of agents i and j, with negative values indicating overlap. Collision and relational consistency are enforced by

$$\begin{aligned} L_{\mathrm{col}} &= \mathbb{E}_t\left[\sum_{i<j}\left(-d_{ij}^{t,\mathrm{body}}\right)_+^2\right], \\ L_{\mathrm{int}} &= \frac{1}{N_{\mathrm{rel}}}\sum_{l=1}^{N_{\mathrm{rel}}} \mathrm{Huber}\left(\rho_l(\widetilde{Y}, r) - \rho_l^{\star}\right), \end{aligned} \tag{36}$$

where ρ_l denotes a differentiable relational measurement associated with the requested agent roles. While behavioral attributes already control quantities such as proximity and closing intensity, L_{int} additionally enforces relational properties such as interaction ordering, agent roles, and temporal consistency. Importantly, the collision term penalizes physical overlap rather than close proximity, allowing near-conflict behaviors to remain valid long-tail samples.

The complete training objective is

$$L_{\mathrm{gen}} = L_{\mathrm{diff}} + \lambda_b L_{\mathrm{beh}} + \lambda_r L_{\mathrm{road}} + \lambda_d L_{\mathrm{dyn}} + \lambda_c L_{\mathrm{col}} + \lambda_i L_{\mathrm{int}}. \tag{37}$$

Sampling and Acceptance. At inference time, conditional diffusion provides the primary control over the generated future. To further improve condition fidelity, we refine each denoised sample using the same differentiable behavioral and validity objectives. Given an initial denoised control sequence $\bar{U}$, the refinement objective is

$$\begin{aligned} J_{\text{guide}}(U) =& \lambda_b L_{\text{beh}}(U) + \lambda_r L_{\text{road}}(U) + \lambda_d L_{\text{dyn}}(U) \\ &+ \lambda_c L_{\text{col}}(U) + \lambda_i L_{\text{int}}(U) + \lambda_{\text{tr}} \|U - \bar{U}\|_2^2, \end{aligned} \tag{38}$$

where the final term limits refinement away from the sample proposed by the learned diffusion model and thereby preserves diversity across different noise seeds.

Each refined trajectory is subsequently subjected to a deterministic validity audit. We recompute the fine-grained descriptor $\xi(\widetilde{Y})$ and retain a sample only if the requested behavioral attributes fall within their prescribed tolerances, the generated trajectories remain compatible with the selected routes and drivable regions, dynamic quantities satisfy feasibility limits, no physical overlap occurs, and the requested interaction roles and temporal relationships are preserved. Multiple diffusion seeds are sampled for each behavioral condition to obtain diverse valid futures.

The accepted trajectories form an offline synthetic tail pool. Before forecasting adaptation, all generator-specific variables, including behavioral conditions, role labels, and route-selection metadata, are discarded. Generated and naturalistic samples therefore share the same history–map input and future-trajectory supervision format for the downstream forecasting model. The generator is trained independently and remains fixed throughout ParetoTail adaptation.

C Distributional Pareto Adaptation Details

This section provides the finite-sample estimation, prediction-space geometry, subspace construction, and numerical optimization details of the distributional Pareto adaptation introduced in Sec. 3.5. All gradients, geometric quantities, and parameter updates are defined over the full parameter space θ of the forecasting model.

Behavioral Coverage and Correction Estimation. Let $\mathcal{D}_N$ denote the naturalistic training set and $\mathcal{D}_G$ the accepted synthetic tail pool. For each trajectory z, the compact descriptor $a(z) \in [0,1]^3$ is computed as in Sec. 3.3.

We distinguish the desired generation coverage from the distribution actually realized by the generator. Let $\{\xi_j^\star\}_{j=1}^{N_{\text{tar}}}$ denote the fine-grained behavioral conditions requested from the generator. Each condition is mapped to its compact behavioral coordinate $a_j^\star$ using the same normalization and aggregation as real trajectories. The target behavioral density is estimated from these requested coordinates as

$$\widehat{p}_{\text{tar}}(a) = \frac{1}{N_{\text{tar}}} \sum_{j=1}^{N_{\text{tar}}} K_h(a - a_j^\star), \tag{39}$$

where $K_h(\cdot)$ is a Gaussian kernel with bandwidth h. In contrast, the naturalistic density is estimated by

$$\widehat{p}_N(a) = \frac{1}{|\mathcal{D}_N|} \sum_{z_j \in \mathcal{D}_N} K_h(a - a(z_j)). \tag{40}$$

Using the same kernel and bandwidth for both distributions ensures that their difference reflects coverage rather than estimator mismatch.

The empirical distribution of accepted synthetic trajectories,

$$\widehat{p}_G(a) = \frac{1}{|\mathcal{D}_G|} \sum_{z_j \in \mathcal{D}_G} K_h(a - a(z_j)), \tag{41}$$

is not used to define the coverage deficit. It instead characterizes the behavioral support actually supplied by the generator. This separation prevents generator bias, rejection rates, or mode preference from redefining which behavioral regions are considered under-covered.

To discretize the continuous correction field, we construct a set of behavioral anchors $\mathcal{A}_M = \{a_m\}_{m=1}^M$ over the valid behavioral region. In practice, we first evaluate the unnormalized coverage deficit on a dense three-dimensional grid,

$$\widetilde{d}_m = [\widehat{p}_{\mathrm{tar}}(a_m) - \widehat{p}_N(a_m)]_+ , \tag{42}$$

and retain anchors with non-negligible deficit and sufficient local sample support. Let

$$n_m^{\mathrm{eff}} = \sum_{z_j \in \mathcal{D}_N \cup \mathcal{D}_G} K_h(a(z_j) - a_m) \tag{43}$$

denote the effective support around anchor a_m. Anchors with $n_m^{\mathrm{eff}} < n_{\min}$ are discarded because their behavior-conditioned gradients cannot be estimated reliably. The remaining deficit weights are normalized as

$$d_m = \frac{\widetilde{d}_m}{\sum_{l \in \mathcal{M}} \widetilde{d}_l}, \qquad m \in \mathcal{M}, \tag{44}$$

where $\mathcal{M}$ indexes the retained anchors.

For each retained anchor, we estimate the behavior-conditioned gradient through a locally weighted forecasting objective rather than individual per-sample gradients. Let $\mathcal{D}_A = \mathcal{D}_N \cup \mathcal{D}_G$ denote the adaptation pool and define

$$L_m(\theta) = \frac{\sum_{z_j \in \mathcal{D}_A} K_h(a(z_j) - a_m)\, \ell_\theta(z_j)}{\sum_{z_j \in \mathcal{D}_A} K_h(a(z_j) - a_m)}, \qquad g_m = \nabla_\theta L_m(\theta). \tag{45}$$

The reference naturalistic gradient is estimated from a naturalistic minibatch $\mathcal{B}_N$,

$$g_N = \nabla_\theta \left[\frac{1}{|\mathcal{B}_N|} \sum_{z_j \in \mathcal{B}_N} \ell_\theta(z_j) \right], \qquad c_m = g_m - g_N. \tag{46}$$

The resulting finite set $\{c_m\}_{m \in \mathcal{M}}$ approximates the continuous correction field $c(a)$.

The deficit-weighted mean correction is

$$\bar{c} = \sum_{m \in \mathcal{M}} d_m c_m, \qquad g_D = g_N + \bar{c}. \tag{47}$$

Rather than explicitly materializing the full covariance matrix, we define

$$C = \left[\sqrt{d_1}(c_1 - \bar{c}), \ldots, \sqrt{d_M}(c_M - \bar{c}) \right], \qquad \Sigma_T = CC^\top. \tag{48}$$

where the columns correspond only to retained anchors. This low-rank factorization represents the deficit-weighted covariance of the tail corrections without constructing a dense parameter-space matrix.

Naturalistic Prediction Geometry. The naturalistic geometry measures parameter updates through their effect on the forecasting function rather than their Euclidean magnitude. Let the multimodal predictor output

$$F_\theta(x) = \left\{ \left(\widehat{Y}_\theta^{(k)}(x), \pi_\theta^{(k)}(x) \right) \right\}_{k=1}^K . \tag{49}$$

We define the local forecast discrepancy as

$$D_x(F_{\theta'}, F_\theta) = \lambda_{\mathrm{traj}} D_{\mathrm{traj}}(F_{\theta'}, F_\theta) + \lambda_{\mathrm{conf}} D_{\mathrm{conf}}(F_{\theta'}, F_\theta). \tag{50}$$

The trajectory component measures displacement between corresponding modes,

$$D_{\mathrm{traj}} = \sum_{k=1}^K \pi_\theta^{(k)} \frac{1}{T_f} \sum_{t=1}^{T_f} \left\| \widehat{Y}_{\theta'}^{(\sigma_x(k)),t} - \widehat{Y}_\theta^{(k),t} \right\|_2^2, \tag{51}$$

where σ_x denotes the mode correspondence. For predictors with persistent mode queries, the native mode ordering is used. Otherwise, σ_x is obtained by minimum-cost trajectory matching to the reference prediction and is treated as fixed when evaluating the local geometry.

The confidence component preserves the multimodal probability structure,

$$D_{\text{conf}} = D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\theta'}^{\sigma_x}\right), \tag{52}$$

where the probabilities of the perturbed forecast are reordered using the same correspondence.

Let

$$W(x) = \frac{1}{2}\,\nabla^2_{F'} D_x\left(F', F_\theta(x)\right)\big|_{F'=F_\theta(x)} \tag{53}$$

denote the local output-space metric induced by this discrepancy. For a sufficiently small parameter update s, first-order expansion of the predictor and second-order expansion of the discrepancy yield

$$\begin{aligned} F_{\theta-s}(x) &= F_\theta(x) - J_F(x)s + o(\|s\|), \\ D_x\left(F_{\theta-s}(x), F_\theta(x)\right) &= s^\top J_F(x)^\top W(x) J_F(x) s + o(\|s\|^2). \end{aligned} \tag{54}$$

Averaging this quantity over naturalistic contexts gives the geometry in Eq. 6.

The full matrix G_N is never explicitly constructed. Given a vector v, we evaluate

$$G_N v = \frac{1}{|\mathcal{B}_N|} \sum_{x \in \mathcal{B}_N} J_F(x)^\top W(x) J_F(x) v + \lambda_G v, \tag{55}$$

using Jacobian–vector and vector–Jacobian products. The damping coefficient $\lambda_G > 0$ ensures numerical stability and makes the local metric positive definite. Linear systems of the form $G_N x = v$ are solved by conjugate gradients, avoiding both explicit matrix construction and inversion.

Tail-Efficient Subspace. The low-rank form $\Sigma_T = CC^\top$ allows the generalized eigenproblem in Eq. 7 to be solved without operating directly in the full parameter dimension. We first solve

$$G_N Q = C, \qquad Q = G_N^{-1} C, \tag{56}$$

where each column of Q is obtained through conjugate gradients using the matrix–vector product in Eq. 55. We then construct the small matrix

$$B = C^\top G_N^{-1} C = C^\top Q \in \mathbb{R}^{M \times M}. \tag{57}$$

Let

$$B v_r = \lambda_r v_r, \qquad v_r^\top v_s = \delta_{rs}, \tag{58}$$

and retain the leading R nonzero eigenpairs. The corresponding parameter-space generalized eigenvectors are recovered as

$$u_r = \frac{G_N^{-1} C v_r}{\sqrt{\lambda_r}}. \tag{59}$$

These vectors satisfy

$$\Sigma_T u_r = \lambda_r G_N u_r, \qquad u_r^\top G_N u_s = \delta_{rs}. \tag{60}$$

Thus, the large generalized eigenproblem is reduced to an eigendecomposition whose dimension is determined by the number of behavioral anchors rather than by the number of model parameters.

The generalized eigenspace also admits a variational interpretation. For any R-dimensional basis V satisfying $V^\top G_N V = I$, define

$$\mathcal{J}(V) = \operatorname{tr}\left(V^\top \Sigma_T V\right). \tag{61}$$

Proposition 1. The leading generalized eigenvectors $U_T = [u_1, \ldots, u_R]$ maximize the retained deficit-weighted correction variation:

$$U_T \in \arg\max_{V^\top G_N V = I} \operatorname{tr}\left(V^\top \Sigma_T V\right), \qquad \max \mathcal{J}(V) = \sum_{r=1}^{R} \lambda_r. \tag{62}$$

Proof. Let $B_G = G_N^{-1/2} \Sigma_T G_N^{-1/2}$ and define $Q_V = G_N^{1/2} V$. Then $Q_V^\top Q_V = I$ and

$$\operatorname{tr}\left(V^\top \Sigma_T V\right) = \operatorname{tr}\left(Q_V^\top B_G Q_V\right). \tag{63}$$

By the Ky Fan maximum principle, this trace is maximized by the leading R eigenvectors of B_G, with optimum equal to the sum of the corresponding eigenvalues. Mapping these vectors back through $G_N^{-1/2}$ yields the leading generalized eigenvectors of (Σ_T, G_N).

The final adaptation space additionally preserves the geometry-aware naturalistic and mean tail learning directions. We solve

$$G_N q_N = g_N, \qquad G_N q_D = g_D, \tag{64}$$

and construct

$$U = \mathrm{orth}_{G_N}\left([q_N, q_D, U_T]\right), \qquad U^\top G_N U = I. \tag{65}$$

All linear solves and orthonormalization are therefore performed with respect to the prediction geometry while still acting on the full forecasting-model parameter space.

Local Pareto Optimization. For an update $\theta^+ = \theta - s$, we locally model the naturalistic objective by

$$L_N(\theta - s) \le L_N(\theta) - g_N^\top s + \frac{\beta_N}{2}\|s\|_{G_N}^2, \tag{66}$$

which yields the naturalistic improvement surrogate $\Gamma_N(s)$ in Eq. 9.

For the tail objective, consider a random behavioral location $A \sim d(a)$. The first-order behavior-conditioned improvement has mean

$$\mathbb{E}_{A\sim d}\left[g(A)^\top s\right] = g_D^\top s \tag{67}$$

and variance

$$\mathrm{Var}_{A\sim d}\left[g(A)^\top s\right] = s^\top \Sigma_T s. \tag{68}$$

The conservative first-order margin

$$g_D^\top s - \kappa\sqrt{s^\top \Sigma_T s} \tag{69}$$

therefore favors updates that provide strong average tail improvement without concentrating the gain on only a small subset of under-covered behaviors. Adding the finite-step curvature penalty gives $\Gamma_T(s)$ in Eq. 9.

Because the final basis satisfies $U^\top G_N U = I$, all optimization is performed in the low-dimensional coordinate z. Define

$$\begin{aligned} \widetilde{g}_N &= U^\top g_N, \\ \widetilde{g}_D &= U^\top g_D, \\ \widetilde{\Sigma}_T &= U^\top \Sigma_T U. \end{aligned} \tag{70}$$

The two surrogates become

$$\begin{aligned} \Gamma_N(Uz) &= \widetilde{g}_N^\top z - \frac{\beta_N}{2}\|z\|_2^2, \\ \Gamma_T(Uz) &= \widetilde{g}_D^\top z - \kappa\sqrt{z^\top \widetilde{\Sigma}_T z} - \frac{\beta_T}{2}\|z\|_2^2. \end{aligned} \tag{71}$$

We use the local adaptation region

$$\mathcal{B}_r = \{z : \|z\|_2 \le r\}, \tag{72}$$

which is equivalent to $\|Uz\|_{G_N} \le r$. The two programs in Sec. 3.5 are therefore solved entirely in this compact coordinate space. Since Γ_N and Γ_T are concave in z, their superlevel constraints together with $\mathcal{B}_r$ define convex feasible regions; both regimes can consequently be solved efficiently using a standard conic optimizer.

The curvature coefficients β_N and β_T account for the discrepancy between first-order gains and finite parameter updates. In implementation, they are calibrated on held-out naturalistic and tail minibatches using small probe updates within the current adaptation space. For a probe direction u with $\|u\|_{G_N} = 1$ and step size δ, we estimate the local curvature through the symmetric second difference

$$\widehat{\beta}_L(u) = \frac{L(\theta + \delta u) - 2L(\theta) + L(\theta - \delta u)}{\delta^2}, \tag{73}$$

and use a conservative upper quantile across probe directions for the corresponding coefficient. The remaining parameters r, κ, ϵ_N, ω_N, and ω_T are selected on the validation split and kept fixed during evaluation.

D Additional Experimental Details

D.1 Datasets, Metrics, and Implementation Details

For nuScenes (Caesar et al., 2020), the naturalistic training set contains 32,186 instances and evaluation is conducted on 9,041 validation instances. To construct the fixed Common–Tail partition used in our distributional analysis, we fit the behavioral-density rule exclusively on the naturalistic training set, which identifies 1,610 natural-tail instances and 30,576 common instances. The fitted rule is then frozen and applied to the validation set, resulting in a fixed Tail subset of 453 instances and a disjoint Common subset of 8,588 instances. We additionally evaluate on NGSIM (Deo & Trivedi, 2018) and HighD (Krajewski et al., 2018) under their respective 1–5 s forecasting protocols.

For nuScenes, we report minADE_{10}, minADE_5, minFDE_5, minFDE_1, and MR_5 where applicable, with lower values indicating better performance. For the worst-case evaluation, each model independently ranks all validation instances according to its own minFDE_5, and we evaluate the highest-error 1%–5% subsets. These model-specific subsets are distinct from the fixed density-defined Common and Tail subsets, which remain identical across methods. For NGSIM and HighD, we report RMSE at prediction horizons from 1 to 5 s together with the average across horizons.

All controlled nuScenes experiments use the same predictor with hidden dimension 64, 10 output modes, and dropout 0.1. The predictor is trained for 120 epochs with batch size 32 using AdamW, a learning rate of 5×10^{-4}. The controllable generator uses hidden dimension 64, two Transformer layers, four attention heads, and 64 diffusion steps, and is trained for 80 epochs with batch size 64 using AdamW, a learning rate of 2×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. ParetoTail uses a frozen offline pool of 3,000 generated trajectories and performs 1,000 adaptation updates. The tail-efficient subspace retains $R = 2$ generalized eigen-directions, with a direction-correlation threshold of 10^{-9} and a regime-switch threshold of 10^{-6}. We set the geometry damping to $\lambda_G = 10^{-3}$ and solve the geometry linear systems using conjugate gradients with relative tolerance 10^{-5} and at most 100 iterations. The two low-dimensional optimization problems are solved using SLSQP with at most 500 iterations and function tolerance 10^{-10}.

We compare four training configurations. *Natural-only* uses only the original naturalistic training data; *Grouped* additionally applies the density-grouped training strategy; *Naive* directly incorporates the generated tail pool without Pareto adaptation; and *ParetoTail* applies the complete proposed framework using the same naturalistic data and synthetic pool. Grouped and Naive are evaluated at the end of their natural-data training procedure, whereas ParetoTail is evaluated after 1,000 adaptation updates; therefore, their checkpoint-selection procedures are not identical.

D.2 Cross-Dataset Performance

We additionally evaluate ParetoTail on NGSIM and HighD to assess forecasting performance beyond the urban nuScenes setting. As shown in Table A.1, ParetoTail achieves the lowest average RMSE of 1.35 m on NGSIM and 0.17 m on HighD. The improvement is particularly evident at longer prediction horizons, with 5 s RMSEs of 2.45 m and 0.34 m on NGSIM and HighD, respectively, compared with 2.82 m and 0.43 m for CITF. These results demonstrate consistent forecasting performance across different driving environments and prediction horizons.

Table A.1: NGSIM and HighD RMSE (m) at 1–5 s and the average across horizons. Lower is better; **bold** and underlined values denote the best and second-best reported results.

Model	NGSIM						HighD					
	1 s	2 s	3 s	4 s	5 s	Avg	1 s	2 s	3 s	4 s	5 s	Avg
CF-LSTM	0.55	1.10	1.78	2.73	3.82	2.00	0.18	0.42	1.07	1.72	2.44	1.17
WSiP	0.56	1.23	2.05	3.08	4.34	2.25	0.20	0.60	1.21	2.07	3.14	1.44
STDAN	0.42	1.01	1.69	2.56	3.67	1.87	0.19	0.27	0.48	0.91	1.66	0.70
HiT	0.38	0.90	<u>1.42</u>	2.08	2.87	1.53	0.08	0.13	0.22	0.39	0.61	0.29
CITF	**0.30**	<u>0.81</u>	<u>1.42</u>	<u>2.04</u>	<u>2.82</u>	<u>1.48</u>	**0.04**	<u>0.09</u>	<u>0.18</u>	<u>0.30</u>	<u>0.43</u>	<u>0.21</u>
ParetoTail	<u>0.32</u>	**0.78**	**1.37**	**1.85**	**2.45**	**1.35**	<u>0.05</u>	**0.08**	**0.16**	**0.23**	**0.34**	**0.17**

D.3 Component Ablation

Table A.2 provides the complete component-wise ablation of ParetoTail. The full model achieves the lowest overall minADE_5 of 1.17 m while maintaining strong performance on both the fixed Common and Tail subsets. Removing tail generation increases the Tail error from 2.27 to 2.62 m, while removing coverage weighting, the correction field, or the naturalistic prediction geometry raises it to 2.60, 2.47, and 2.50 m, respectively. Disabling the two-regime switch also degrades both Common and Tail performance to 1.14 and 2.39 m, indicating the benefit of explicitly separating compatible and conflicting adaptation regimes. Removing the naturalistic budget yields the lowest Tail error of 2.15 m, but simultaneously increases the Overall and Common errors from 1.17/1.11 m to 1.20/1.15 m. This behavior is consistent with the intended role of the budget: it limits aggressive tail-oriented updates to preserve forecasting performance over the broader naturalistic distribution. Overall, the complete model provides a more favorable balance between tail improvement and naturalistic preservation.

Table A.2: Component ablation on nuScenes, reported as minADE_5 (m) on the full validation set and the fixed Common and Tail subsets. Rows A–F each remove one component from the complete model G: TG, tail generation; CW, coverage weighting; CF, correction field; NG, naturalistic prediction geometry; RS, two-regime switch; and NB, naturalistic budget. Lower is better; bold and underlined values denote the best and second-best results based on unrounded values.

Model	Components						Performance (minADE_5)		
	TG	CW	CF	NG	RS	NB	Overall	Fixed Common	Fixed Tail
A	×	✓	✓	✓	✓	✓	1.20	1.13	2.62
B	✓	×	✓	✓	✓	✓	<u>1.18</u>	**1.11**	2.60
C	✓	✓	×	✓	✓	✓	1.19	1.12	2.47
D	✓	✓	✓	×	✓	✓	1.20	1.13	2.50
E	✓	✓	✓	✓	×	✓	1.20	1.14	2.39
F	✓	✓	✓	✓	✓	×	1.20	1.15	**2.15**
G (Full)	✓	✓	✓	✓	✓	✓	**1.17**	<u>1.11</u>	<u>2.27</u>

D.4 Efficiency Analysis

ParetoTail introduces the correction-field and naturalistic-geometry computations only during adaptation, while inference uses the resulting forecasting model without additional modules. As shown in Table A.3, ParetoTail requires approximately 14 ms per batch on an RTX 3090, matching the inference efficiency of the Natural-only predictor while improving minADE_5, minFDE_1, and MR_5 from 1.25 m, 6.98 m, and 0.49 to 1.17 m, 6.55 m, and 0.48, respectively. Compared with the reported baselines, ParetoTail also maintains favorable inference efficiency while achieving strong forecasting accuracy.

Table A.3: Inference time and nuScenes forecasting accuracy on an RTX 3090. ADE/FDE are in metres and MR_5 is a proportion; lower is better.

Model	Time (ms)	minADE_5 (m)	minFDE_1 (m)	MR_5
Trajectron++ (Salzmann et al., 2020)	38	1.88	9.52	0.70
VisionTrap (Moon et al., 2024)	53	1.36	8.72	0.61
MultiPath (Chai et al., 2020)	87	1.44	7.69	0.74
LAformer (Liu et al., 2024)	115	1.19	6.89	0.54
PGP (Deo et al., 2022)	215	1.30	7.17	0.61
Natural-only	∼14	1.25	6.98	0.49
ParetoTail	∼14	**1.17**	**6.55**	**0.48**

D.5 Additional Qualitative Results

Figure A.1 provides additional qualitative comparisons between Q-EANet (Chen et al., 2024) and ParetoTail across six nuScenes scenarios. Across different road geometries and interaction contexts, ParetoTail produces multimodal forecasts that cover the observed future while remaining consistent with the local road structure.

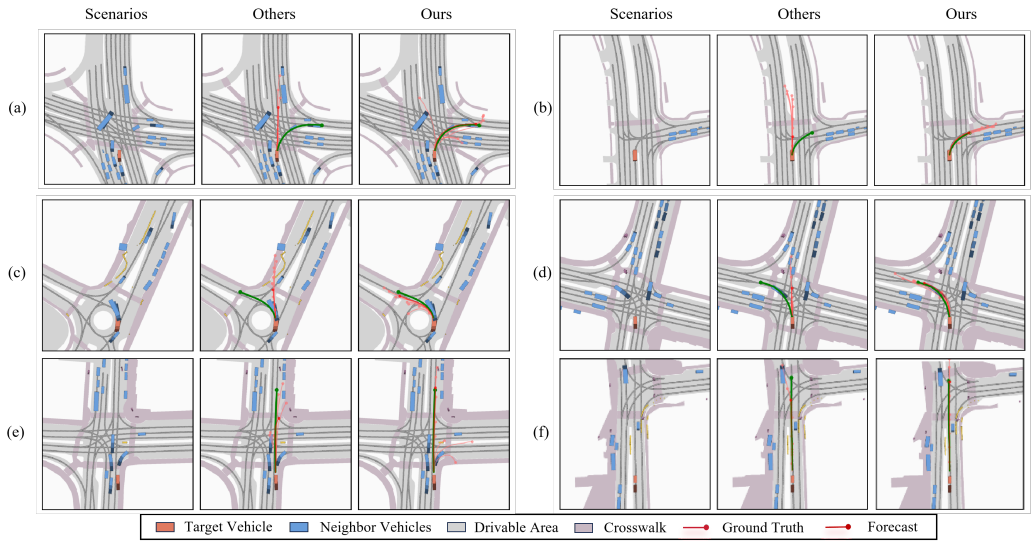


Figure A.1: Additional qualitative comparisons between Q-EANet (Chen et al., 2024) and ParetoTail on six nuScenes scenarios. In each group (a–f), the three columns show the scene context, Q-EANet, and ParetoTail (Ours), from left to right.