RAPID: A Scalable and Controllable Physics-Informed Diffusion Framework for Real-Time Pedestrian Trajectory Generation

Zihan Yu
yuzh23@mails.tsinghua.edu.cn
Department of Electronic Engineering, BNRist, Tsinghua University
Beijing, USA

Huandong Wang*
wanghuandong@tsinghua.edu.cn
Department of Electronic Engineering, BNRist, Tsinghua University
Beijing, USA

Jingtao Ding
dingjt15@tsinghua.org.cn
Department of Electronic Engineering, BNRist, Tsinghua University
Beijing, USA

Yong Li*
liyong07@tsinghua.edu.cn
Department of Electronic Engineering, BNRist, Tsinghua University
Beijing, China

Abstract

Generating realistic and diverse pedestrian background flows is critical for training and validating autonomous driving systems. While recent diffusion-based models achieve state-of-the-art accuracy, they suffer from prohibitive inference latency, lack of physical consistency, and an inability to generalize across heterogeneous datasets, rendering them impractical for industrial Hardware-in-the-Loop testing. To address these challenges, we propose Real-time Adaptive Physics-Informed Diffusion (RAPID), a unified framework explicitly designed to balance high-fidelity generation with strict real-time constraints. First, we introduce a Canonical Representation Module that harmonizes diverse datasets via coordinate-invariant encoding and adaptive modality imputation, enabling unified training across varying scene scales. Second, we propose a Map Context Encoder that decouples computationally expensive map perception from the iterative denoising loop using cached latent embeddings. Third, a Physics-Informed Implicit Sampler integrates Social Force Model gradients as directional priors, encouraging physical consistency (e.g., collision avoidance). Extensive experiments on five heterogeneous benchmarks demonstrate that RAPID establishes a new state-of-the-art balance between fidelity and safety. Notably, it is the only framework capable of operating consistently below the 30 ms industrial threshold, maintaining almost constant inference latency regardless of crowd density. The system exhibits precise controllability over agent behaviors and has been successfully deployed in a production-grade autonomous-driving simulation platform as a core digital twin kernel for large-scale autonomous driving validation. The code is publicly available at https://github.com/tsinghua-fib-lab/RAPID.

*Corresponding authors.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818466

CCS Concepts

• **Computing methodologies** → **Real-time simulation**; *Neural networks*; Motion path planning; • **Applied computing** → Transportation.

Keywords

Diffusion Models, Pedestrian Trajectory Prediction, Physics-Informed Learning, Autonomous Driving Simulation, Real-time Inference

ACM Reference Format:
Zihan Yu, Huandong Wang, Jingtao Ding, and Yong Li. 2026. RAPID: A Scalable and Controllable Physics-Informed Diffusion Framework for Real-Time Pedestrian Trajectory Generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818466

1 Introduction

High-fidelity simulation has become the indispensable engine for developing and validating autonomous driving systems, enabling the safe and scalable training of driving agents before real-world deployment [20]. Within such simulation platforms, generating realistic pedestrian background flows is paramount. Unlike static obstacles, pedestrians are dynamic and interactive agents, whose complex behaviors ranging from road crossings to subtle interactions constitute the "long-tail" scenarios that autonomous vehicles must master [12, 34]. Therefore, an effective and efficient pedestrian simulation system, serving as a critical generator of training data, directly influences the capability of autonomous driving systems to perceive, predict, and plan in urban environments.

However, bridging the gap between current pedestrian simulation research and the rigorous demands of autonomous driving systems remains a formidable challenge. Traditionally, pedestrian modeling has been tailored for crowd evacuation [17], architecture design [10], or computer graphics [42], prioritizing microscopic precision and visual plausibility over computational efficiency and physical consistency required by autonomous driving systems. Recently, diffusion models have emerged as a dominant paradigm, achieving state-of-the-art accuracy by capturing the intrinsic stochasticity of human motion and incorporating semantic

map constraints [3, 6, 9, 14, 32, 38]. Despite their success, these methods face three critical limitations when deployed in industrial systems. First, existing datasets exhibit significant heterogeneity in data modalities (e.g., availability of maps or vehicle trajectories), scene scales, and coordinate definitions. Consequently, current methods typically train separate parameters for each dataset [25], failing to learn a unified model capable of generalizing across diverse scenarios from unstructured crowded plazas to complex mixed-traffic urban intersections. Second, the iterative denoising process of diffusion models usually incurs prohibitive inference latency, making them ill-suited for the millisecond-level real-time updates required by Hardware-in-the-Loop testing in industrial simulators [22]. Finally, lacking inherent physical constraints, purely data-driven approaches often generate "hallucinated" trajectories that violate physical consistency (e.g., collisions with obstacles), undermining the reliability and credibility of the simulation [13].

To address these challenges, we propose Real-time Adaptive Physics-Informed Diffusion (RAPID), a framework explicitly designed to balance high-fidelity generation with the strict real-time requirements of industrial autonomous driving simulation. First, to overcome data fragmentation and enable unified training across heterogeneous datasets, we introduce a canonical representation module, which establishes a unified feature space to harmonize heterogeneous scene data. Specifically, it employs a coordinate-invariant encoding scheme combining relative motion features with Fourier-based absolute positional encodings [40], ensuring the representation effectively handles vast differences in coordinate origins and spatial scales across scenes. Additionally, it incorporates an adaptive modality imputation mechanism that treats missing data fields (e.g., absent maps) as special latent states represented by learnable embeddings. Secondly, to tackle the inference latency bottleneck, we propose a map context encoder, which compresses scene information into fixed-size latent embeddings via learnable latent queries and agent-centric extraction. This design allows the computationally expensive map perception to be performed only once and cached at the start of the simulation, explicitly decoupling it from the iterative denoising loop and eliminating the overhead of repeatedly attending to high-resolution rasterized maps. Finally, we bridge the gap between data-driven fidelity and physical consistency via a physics-informed implicit sampler. Unlike standard diffusion models, our method integrates the data prior provided by the learned denoising model with physical gradients derived from the classical social force model [17], thus encouraging the generation of trajectories with better physics consistency.

Extensive experiments on five heterogeneous benchmarks (ETH / UCY [27, 31], SDD [33], GC [46], WayMo [39]) demonstrate that RAPID establishes a new state-of-the-art balance between prediction fidelity and physical consistency. In terms of efficiency, RAPID is the only framework capable of operating consistently below the 30 ms industrial real-time threshold. Crucially, due to the fully parallelized architecture, the inference latency remains almost constant regardless of crowd density, effectively solving the scalability bottleneck inherent in existing graph-based approaches. Beyond standard metrics, the model exhibits precise controllability over agent behaviors, accurately adhering to specified navigation intents (e.g., speed and destination) to support customized scenario generation. Furthermore, we validate that our unified training paradigm significantly boosts generalization on data-scarce domains. The proposed system has been successfully deployed in a production-grade autonomous-driving simulation platform as a core digital twin kernel, where it overcomes hardware heterogeneity on vendor-specific NPUs to support high-throughput virtual testing in production environments.

The main contributions of this work are summarized as follows:

- We propose a Canonical Representation Module that harmonizes heterogeneous data via coordinate-invariant encoding and adaptive modality imputation, enabling a single model to be jointly trained on diverse datasets and applied across scenarios ranging from crowded plazas to mixed-traffic intersections.
- We design a Map Context Encoder that compresses scene context into fixed-size embeddings via learnable queries, enabling one-time pre-computation that explicitly decouples heavy map perception from the iterative denoising loop.
- We introduce a Physics-Informed Implicit Sampler that integrates learned data priors with physical gradients, utilizing the gradients as directional priors to enforce physical consistency and significantly accelerate convergence.
- Extensive experiments and successful deployment on a production-grade autonomous-driving simulation platform demonstrate that RAPID achieves state-of-the-art prediction performance, scalable operational efficiency, and precise controllability.

2 Preliminaries and Problem Formulation

2.1 Problem Formulation

We focus on the task of generating realistic and physically compliant pedestrian trajectories in complex urban traffic scenarios. The problem is defined as predicting the future states of N pedestrians in a scene, conditioned on their historical states, the surrounding heterogeneous agents (e.g., vehicles), and the static scene context (e.g., maps).

Let $X = \{\mathbf{p}_t\}_{t=-T_{obs}+1}^{T_{pred}}$ denote the trajectories of N pedestrians, where $\mathbf{p}_t \in \mathbb{R}^{N\times 2}$ represents the 2D world coordinates of N pedestrian at time step t. We observe the history for $T_{obs} = 8$ time steps and aim to generate the future trajectory for $T_{pred} = 12$ time steps. All datasets were resampled to ensure $\Delta t = 0.4$ s. The goal is to learn a distribution $p_\theta(\mathbf{p}_{t>0}|C)$ that approximates the real data distribution while satisfying physical constraints based on the context C including historical trajectories of pedestrians $\{\mathbf{p}_{t\le 0}\}$ and vehicles $\{\mathbf{v}_t\}$ ($\mathbf{v}_t \in \mathbb{R}^{M\times 2}$ denotes the 2D world coordinates of M vehicles at time step t), agent intentions including destination $\mathbf{d} \in \mathbb{R}^{N\times 2}$ and desired speed $\mathbf{s}_{des} \in \mathbb{R}^{N\times 1}$, as well as a rasterized semantic map $M \in \mathbb{R}^{H\times W}$ encoding navigable areas and static obstacles.

2.2 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [18] are a class of generative models that learn to reverse a gradual noising process. The process involves two Markov chains: a forward diffusion process and a reverse denoising process.

Forward Process. Given a data sample $\mathbf{x}_0$ (in our case, the normalized future acceleration $\mathbf{a}_{t+1} \in \mathbb{R}^{N\times 2}$), Gaussian noise is incrementally added over K diffusion steps. The noisy state at step k,

denoted as $\mathbf{x}_k$, is sampled via:

$$q(\mathbf{x}_k|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_k; \sqrt{\bar{\alpha}_k}\mathbf{x}_0, (1-\bar{\alpha}_k)\mathbf{I}), \tag{1}$$

where $\bar{\alpha}_k = \prod_{s=1}^{k} \alpha_s$ is the cumulative product of noise schedule coefficients $\alpha_s \equiv \beta_s \in (0,1)$. We adopt a linear noise scheduler in this work, where the variance β_k increases linearly over K steps. When K is sufficiently large, $\mathbf{x}_K$ approximates standard Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Process. We adopt $\mathbf{x}_0$-prediction in this work, where a denosing network $f_\theta(\mathbf{x}_k, k, C)$ is trained to predict the denoised result $\hat{\mathbf{x}}_0$ at each step k, conditioned on the context C. The sampling transition is defined as:

$$\mathbf{x}_{k-1} = \frac{\bar{\alpha}_{k-1} - \bar{\alpha}_k}{1-\bar{\alpha}_k}\sqrt{\frac{1}{\bar{\alpha}_{k-1}}}\hat{\mathbf{x}}_0 + \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\sqrt{\frac{\bar{\alpha}_k}{\bar{\alpha}_{k-1}}}\mathbf{x}_k + \sigma_k \mathbf{z}, \tag{2}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is random noise, $\sigma_k = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\left(1 - \frac{\bar{\alpha}_k}{\bar{\alpha}_{k-1}}\right)$ controls the stochasticity of sampling.

Energy Guidance for Constrained Generation. The sampling process of diffusion models (e.g., DDPM and DDIM) can be conditionally steered to satisfy specific downstream constraints. Classifier Guidance [8] achieves this by biasing the sampling process toward a desired class label y using the gradient $\nabla_{\mathbf{x}_k} \log p_\phi(y|\mathbf{x}_k)$ from a pre-trained classifier. This paradigm can be generalized by replacing the discrete label with a continuous, differentiable potential function $\mathcal{U}(\mathbf{x}_k)$, allowing for arbitrary energy-based constraints [4]. Specifically, to steer the sampling, the estimated denoising result $\hat{\mathbf{x}}_0$ at each step k is updated as

$$\tilde{\mathbf{x}}_0 = \hat{\mathbf{x}}_0 + w\frac{1-\bar{\alpha}_k}{\sqrt{\bar{\alpha}_k}}\nabla_{\mathbf{x}_k}\mathcal{U}(\mathbf{x}_k), \tag{3}$$

where w is the guidance scale. In practice, accouding the one-step approximation discussed in [7], the gradient $\nabla_{\mathbf{x}_k}\mathcal{U}$ can be substituted with $\nabla_{\hat{\mathbf{x}}_0}\mathcal{U}$, enabling bypassing the computationally expensive backpropagation through the denoising network f_θ and significantly accelerating inference, enabling the integration of complex physical priors as zero-shot constraints at a low time cost.

3 Methodology

3.1 Method Overview

As illustrated in Figure 1, we propose Real-time Adaptive Physics-Informed Diffusion (RAPID), a unified framework designed to bridge the gap between high-fidelity generation and industrial-grade real-time constraints. The system consists of three synergistic modules: 1) The Canonical Representation Module (Section 3.2), which establishes a universal feature space to harmonize heterogeneous datasets. By employing a coordinate-invariant encoding scheme and adaptively imputing missing modalities, it provides a unified foundation for cross-domain learning across diverse coordinate systems and spatial scales. 2) The Map Context Encoder (Section 3.3), which explicitly decouples the computationally expensive map perception from the iterative inference loop. By compressing complex scene semantics into cached, fixed-size latent embeddings via learnable latent queries, it significantly reduces the memory footprint and eliminates the overhead of repetitive map processing. 3) The Physics-Informed Implicit Sampler (Section 3.4), which bridges the gap between data-driven fidelity and physical consistency. This module integrates physical gradients derived from the Social Force Model directly into the diffusion denoising process, thus enforcing physical consistency such as collision avoidance.

3.2 Canonical Representation Module

Training a universal foundation model requires overcoming significant heterogeneity across public datasets (e.g., ETH/UCY [27, 31], SDD [33], GC [46], WayMo [39]), which vary in coordinate definitions (pixels vs. meters), data modalities (presence of maps/vehicles), and spatial scales. To address this, we establish a unified encoding protocol harmonizing diverse data sources into a shared space. Specifically, we first align all input trajectories to a unified physical metric space. Since datasets like GC and SDD record data in perspective pixel coordinates while others use global meters, we employ a scene-specific homography transformation [16] $\mathbf{H} \in \mathbb{R}^{3\times3}$. We project raw pixel positions $\mathbf{p}_{pixel} = [x_{pixel}, y_{pixel}, 1]^T$ to canonical world coordinates $\mathbf{p}_{world} = [x_{world}, y_{world}, 1]^T$ via $\mathbf{p}_{world} \cong \mathbf{H} \cdot \mathbf{p}_{pixel}$, where $\cong$ denotes equality up to a scale factor. For datasets already in world coordinates, an identity transformation is applied. This ensures that all subsequent operations process physical displacements in meters.

Building on this geometrically aligned basis, we introduce the canonical representation module (Figure 1a), which operates on a comprehensive feature set comprising the current positions, velocities, historical trajectories, destinations, and intended speeds for pedestrians, positions and historical trajectories for vehicles, as well as the scenario map. To unify these inputs across datasets with varying availability and coordinates, the module resolves the remaining semantic and spatial heterogeneity through two key mechanisms:

Adaptive Modality Imputation. To handle missing data fields (e.g., the absence of maps or vehicle trajectories) within this feature set without introducing distribution shifts via zero-padding, we treat missing modalities as special latent states. Specifically, we employ a masked embedding strategy, where a learnable "null" embedding vector $\mathbf{e}_\emptyset \in \mathbb{R}^{D_f}$ is introduced for each of the features (excluding the current position and velocity necessarily required) to explicitly represent the absence of specific modality. During encoding, valid raw features $\mathbf{f}$ with a availability mask $m = 1$ are linearly projected into the latent space $\mathbb{R}^{D_f}$, whereas missing inputs whose availability mask $m = 0$ are substituted by their corresponding $\mathbf{e}_\emptyset$, that is,

$$\mathbf{h}_f = m \cdot \text{Encoder}(\mathbf{f}) + (1-m) \cdot \mathbf{e}_\emptyset. \tag{4}$$

This mechanism allows the module to adaptively learn a consistent default representation for absent contexts (e.g., automatically inferring a "no-vehicle" context state for ETH/UCY) while utilizing rich features when available.

Coordinate-Invariant State Encoding. To prevent the discrepancies in coordinate system origins and spatial scales across heterogeneous scene maps from impairing generalization performance, we adopt a dual-stream encoding strategy. First, to decouple local motion dynamics from global positioning, we encode historical trajectories and destinations as displacements relative to the agent's current position. A similar ego-centric normalization is applied to vehicle historical trajectories relative to their respective current positions. This ensures the motion representation is translation-invariant and robust across different coordinate origins. Secondly,

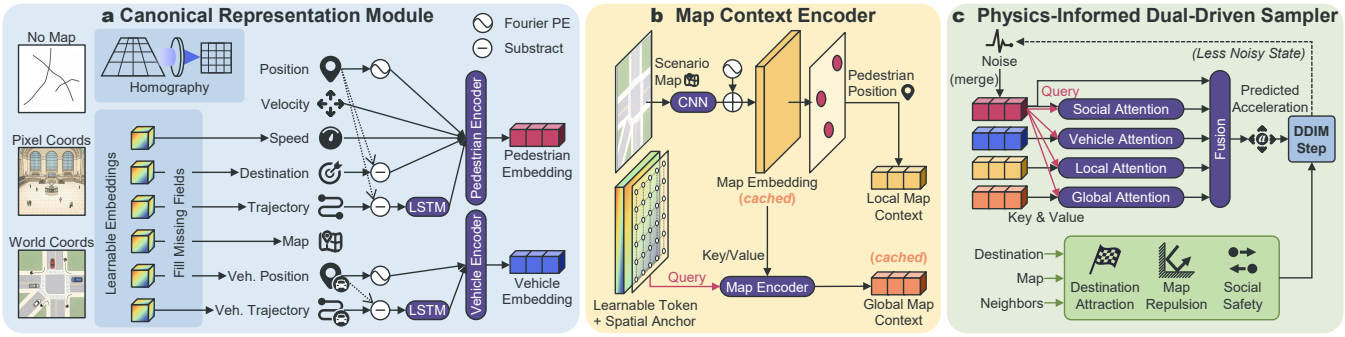

Figure 1: Overview of the Proposed RAPID Framework.

to enable the pedestrian to localize itself on the semantic map and establish precise spatial relationships, absolute position remains indispensable. Therefore, we encode the agent's current position $\mathbf{p} = [x, y]^T$ using 2-dimensional fourier positional encoding [40], which applies sinusoidal functions separately to each coordinate axis and projects the concatenated features to the latent dimension D_f via a learnable linear layer:

$$\mathrm{PE}(\mathbf{p}) = \mathrm{Linear}\left([\cdots, \sin(\omega_i \mathbf{p}), \cos(\omega_i \mathbf{p}), \cdots]\right). \tag{5}$$

Here, the frequency spectrum ω_i is linearly spaced from $2\pi \times 10^{-3}$ m^{-1} to $2\pi \times 2$ m^{-1}, where the lower frequency bound corresponds to a spatial wavelength of 1 km, ensuring global positional uniqueness across large-scale maps. Conversely, the upper frequency bound provides a resolution of 0.5 m, enabling the model to resolve fine-grained local dynamics at a scale comparable to individual pedestrians.

Finally, the processed latent features $\mathbf{h}_f$ belonging to the same category are aggregated. Specifically, all pedestrian-related features (e.g., historical dynamics, destination, and positional encodings) are summed and passed through a Pedestrian Encoder to yield a unified agent embedding. Similarly, vehicle-related features are aggregated and processed by a Vehicle Encoder. This categoric feature fusion ensures that the model captures the synergistic effects between different state variables within a unified agent-level representation.

3.3 Map Context Encoder for Efficiency

Scene maps provide critical environmental context, enabling agents to perceive navigable areas and generate physically plausible movements. To ensure universality across heterogeneous sources, we adopt a rasterized semantic map $\mathbf{M}$, which can be consistently derived from both vector-based data (e.g., WayMo) and raw top-down images (e.g., ETH/UCY). Crucially, to strictly maintain physical consistency across varying scene scales, we enforce a fixed physical resolution rather than resizing images to fixed dimensions. However, this strategy results in map tensors with variable spatial dimensions. In large-scale scenarios like WayMo that span hundreds of meters, the resulting feature maps become excessively large, incuring massive computational overhead and severely bottlenecking the denoising speed. To resolve this, we propose the Map Context Encoder (Figure 1b). This module compresses variable-sized semantic grids into compact, fixed-size latent embeddings. Crucially, this encoding is performed *once* prior to generation: the resulting embeddings are cached and efficiently queried throughout the iterative denoising steps, thereby decoupling the heavy map processing from the recurrent sampling process.

Global Context via Spatially-Anchored Compression. First, we encode the raw map $\mathbf{M}$ using a Convolutional Neural Network (CNN) backbone [28] to obtain a high-dimensional feature map $\mathbf{h}_{map}$, and inject Fourier positional encodings corresponding to the pixel's actual physical coordinates. To compress this variable-sized feature map into a fixed-size representation, inspired by Perceiver [23], we employ a set of learnable embeddings $\mathbf{Q}_{learn} \in \mathbb{R}^{K \times D}$ to query the map embeddings via cross-attention [41]. To explicitly encourage these learnable embeddings to capture distinct spatial regions, we introduce a grid of spatial anchors covering the map extent, calculate the physical coordinates for each grid point, and add their corresponding Fourier positional encodings to the learnable embeddings. Formally, the global map context is obtained by

$$\begin{aligned} Q &\equiv \mathbf{Q}_{learn} + \mathrm{PE}(\mathbf{p}_{anchor}), K \equiv V \equiv \mathbf{h}_{map} + \mathrm{PE}(\mathbf{p}_{map}), \\ \mathbf{z}_{global} &= \mathrm{Attention}(Q, K, V), \end{aligned} \tag{6}$$

where $\mathbf{p}_{anchor} \in \mathbb{R}^{K \times 2}$ represents the physical coordinates of $K = S^2$ grid points on the map, and $\mathbf{p}_{map} \in \mathbb{R}^{H \times W \times 2}$ represents the physical coordinates of each pixel on the map. This explicit spatial bias encourages each query to focus on a specific physical sub-region during the cross-attention process, ensuring that the compressed $\mathbf{z}_{global}$ preserves the global topological structure of the scene.

Local Context via Agent-Centric Extraction. While $\mathbf{z}_{global}$ captures the overall scene structure, collision avoidance requires precise knowledge of the immediate surroundings. Therefore, parallel to the global compression, we perform a localized readout. For each agent at position $\mathbf{p}$, we map its physical coordinates to the continuous index space of the feature map and extract the localized embedding $\mathbf{z}_{local}$ via bilinear interpolation. This agent-centric extraction complements the global perception, ensuring that each pedestrian captures fine-grained environmental details in its immediate vicinity directly.

Efficient Inference with Fixed-Size Context. $\mathbf{z}_{global}$ and $\mathbf{z}_{local}$ collectively describe both the global structure and local details of the map, which maintain a fixed token length and remain invariant to the increasing scale of the original scene. Crucially, we only need to perform the computationally intensive map encoding and compression once per scene, and utilize the cached lightweight embeddings throughout the subsequent diffusion denoising loop.

This strictly decouples the high-dimensional map processing from the repetitive sampling steps, significantly reducing GPU memory footprint and accelerating real-time inference.

3.4 Physics-Informed Implicit Sampler

As shown in Figure 1c, our denoising model predicts the denoised acceleration for each pedestrian at the next step $\hat{\mathbf{x}}_0 \equiv \hat{\mathbf{a}}_{t+1}$ based on the previously obtained embeddings of pedestrians and vehicles, as well as global and local maps contexts. Specifically, it uses the pedestrian embedding to query all four embeddings through cross attention, adds the results back to the pedestrian embedding, and predicts the output through a fully connected network.

While this data-driven prediction captures rich motion patterns, it does not explicitly guarantee physical feasibility. To address this, we introduce a physics-informed implicit sampler based on energy-guided denoising [4], which integrates classical Social Force Model (SFM) [17] priors directly into the DDIM sampling process [37]. Specifically, we define a differentiable physics energy that penalizes the discrepancy between the model-predicted acceleration $\hat{\mathbf{x}}_0$ and the SFM-predicted acceleration $\mathbf{a}$:

$$\tilde{\mathbf{x}}_0 = \underbrace{\hat{\mathbf{x}}_0}_{\text{Data Prior}} - \sum_i w_i \underbrace{\nabla_{\hat{\mathbf{x}}_0} \|\hat{\mathbf{x}}_0 - \mathbf{a}_i\|^2}_{\text{Physics Gradient}}, \tag{7}$$

where $i \in \{\text{destination}, \text{map}, \text{social}\}$ reflects the attractiveness of the destination, as well as the repulsive force of the map and other agents (see Appendix A for details). w_i is the guidance weight controlling the strength of the corresponding physical prior, which can be determined through standard hyperparameter optimization tools (e.g., Optuna [1]) automatically. It is worth noting that the proposed physics gradient guidance is applied only during inference (sampling), requiring no modification or retraining of the denoising model. Differernt from standard diffusion samplers like DDPM or DDIM that perform a stochastic exploration to find the data manifold, this sampler transforms this random walk into a physically-directed search by injecting SFM gradients, where the physical laws act as a strong directional prior, effectively orienting noisy trajectories toward feasible regions.

4 Performance Evaluation

We structure our experiments to answer the following research questions (RQs):

RQ1: How does our framework compare against leading baselines (GANs, GCNs, and Diffusion Models) in terms of trajectory fidelity (ADE/FDE) and physical compliance (Collision Rate)?

RQ2: Does the proposed unified training paradigm on heterogeneous datasets yield better generalization performance compared to training separate models for each domain?

RQ3: Can our method satisfy the strict real-time constraints of industrial autonomous driving systems simulation, and how does the inference latency scale with the number of pedestrians?

RQ4: Does the model exhibit high controllability in adhering to specified navigation intents (speed and destination) while maintaining reasonable diversity for pedestrian trajectory generation?

RQ5: What is the individual contribution of each design in the model to the overall performance?

4.1 Evaluation Settings

To comprehensively evaluate the universality and efficiency of the proposed RAPID framework for autonomous driving simulation, we conduct extensive experiments across five heterogeneous datasets. We employ a multi-dimensional suite of metrics spanning prediction accuracy, physical consistency, and computational efficiency, benchmarking our approach against four representative baselines.

4.1.1 Datasets and Experimental Protocols. To comprehensively evaluate the universality and efficiency of our framework, we conduct experiments on five open datasets: the small-scale **ETH/UCY** benchmark [27, 31] and three large-scale datasets, namely **SDD** [33], **GC** [46], and **WayMo** [39]. We adopt two distinct evaluation settings (see **Appendix B** for detailed dataset specifications, splitting rules, and hyperparameter configurations):

Standard Setting: We follow standard domain-specific protocols [25]. For ETH/UCY, we employ the leave-one-out strategy. For large-scale datasets, we utilize fixed train-test splits. Crucially, we strictly adhere to a temporal split (using the first 80% for training and the latter 20% for validation/testing) to prevent information leakage.

Unified Setting: To assess generalization, we train the model jointly on the target domain's training set combined with the training partitions of all other available datasets.

4.1.2 Evaluation Metrics. To comprehensively assess the performance of RAPID in the context of industrial autonomous driving simulation, we employ four quantitative metrics covering fidelity, safety, and efficiency:

Average Displacement Error (ADE): The average Euclidean distance between the ground truth and the predicted positions over all prediction time steps. Following standard practice [2, 15], we report the minimum ADE computed from 10 randomly sampled trajectories (Best-of-10), thus evaluating the model's capability to cover the true future mode within its predictive distribution.

Final Displacement Error (FDE): The Euclidean distance between the predicted and ground truth positions at the final time step (T_{pred}). Similar to ADE, this is calculated using the Best-of-10 strategy to assess the model's accuracy in predicting long-term destinations.

Collision Rate: The percentage of test samples where a collision occurs. A collision is identified if the Euclidean distance between any two agents falls below a safety threshold of $0.6m$ (approximating the typical human body diameter). Unlike ADE/FDE, this metric is averaged over all generated samples to strictly evaluate the physical consistency of the entire generated distribution. Furthermore, to distinguish model-induced collisions from intrinsic data noise, we exclude collision pairs that already exist in the ground truth trajectories.

Inference Time: The average wall-clock time required to generate the *actionable* prediction for the next simulation frame. In closed-loop simulation, agents typically update their states step-by-step. Therefore, for auto-regressive models (e.g., Ours and SPDiff) that predict recursively, we measure the latency of generating a single time step (i.e., 1 rollout step). Conversely, for non-autoregressive baselines (e.g., STGCNN, MID, TRACE) that generate the entire future trajectory (T_{pred} steps) in a holistic one-shot manner, we must measure the full generation time per inference call, as their

architecture does not support decoupling individual steps. To ensure a fair comparison, we measure the latency of all methods on the same server. (see Appendix B for details).

4.1.3 Baselines. To strictly evaluate RAPID's fidelity and efficiency, we select four representative baselines covering classic efficient architectures, state-of-the-art academic diffusion models, and high-fidelity industrial solutions:

Social-STGCNN [30]: A classic graph convolutional network (GCN) that models interactions via a spatio-temporal graph. We select it as the efficiency benchmark due to its lightweight architecture and rapid inference speed, representing the upper bound of efficiency for non-generative methods[1].

MID [14]: The pioneer denoise diffusion framework in this domain. It formulates trajectory prediction as a reverse process of motion indeterminacy diffusion. We include it as a representative of *pure data-driven* stochastic prediction without explicit physical constraints[2].

SPDiff [6]: A social physics-informed diffusion model that also incorporates the Social Force Model (SFM). Crucially, SPDiff integrates physics primarily by superimposing SFM predictions as conditions, which differs fundamentally from our method that leverages SFM gradients to *actively guide* the denoising sampling[3].

TRACE [32]: An industrial-grade controllable diffusion model developed by NVIDIA. Designed for high-fidelity character animation in computer graphics, it prioritizes visual realism and complexity.[4].

We adopted the official implementation and retrained the baseline models on our partitioned dataset. To ensure strict fairness, all baselines were allocated the same hyperparameter tuning budget as RAPID, including an identical grid over learning rates and batch sizes, as detailed in Appendix B.

4.2 Comparative Analysis (RQ1)

Table 1 reports the quantitative results on the standard ETH/UCY benchmark, while Table 2 summarizes the performance on large-scale heterogeneous datasets. As shown in the tables, our proposed framework establishes a new state-of-the-art balance between fidelity and safety. In terms of fidelity, the *Separate* version of our model achieves superior ADE/FDE compared to all baselines across most scenarios. Notably, on the challenging GC dataset, it reduces the ADE to 0.1891m, significantly outperforming the best baseline. Regarding safety, while the Separate model is already competitive, the incorporation of physics-informed guidance consistently reduces collision rates. For instance, on WayMo, guidance improves the collision rate to 4.15%, surpassing the physical validity-oriented baseline TRACE (4.32%).

Beyond the superiority of our framework, analyzing the performance variance among baselines reveals critical trade-offs in existing methodologies: First, MID generally outperforms the classic STGCNN in ADE/FDE, particularly on large-scale benchmarks. This confirms the advantage of data-driven diffusion models in capturing the high-fidelity, multi-modal nature of human motion. Secondly, SPDiff shows average performance on small datasets (ETH/UCY) but achieves significant gains on large-scale, dense datasets (GC/SDD), often outperforming MID. This suggests that incorporating physics constraints becomes increasingly beneficial for regularizing complex agent interactions as crowd density increases. Finally, TRACE, while suffering from higher displacement errors due to its focus on animation quality rather than precise trajectory matching, consistently achieves lower collision rates than most baselines. This validates its design priority on visual plausibility and physical validity over pure metric minimization. In contrast, our framework effectively synthesizes these strengths. By combining the high-fidelity generation of diffusion models (like MID) with explicit, gradient-based physical guidance (a more effective implementation of the principles in SPDiff and TRACE), RAPID achieves robust performance across both accuracy and safety metrics.

4.3 Unified Training Analysis (RQ2)

A core contribution of this work is the exploration of a unified generalist model capable of handling heterogeneous scenarios. We analyze the benefits and trade-offs of this Unified training strategy compared to the domain-specific Separate training. As shown in Table 1, the Unified strategy yields substantial gains on small-scale benchmarks. Most notably, on the Hotel and Univ scenes, the Unified model reduces the ADE by approximately 50% (e.g., Hotel: $0.4733 \rightarrow 0.2274$) compared to the Separate model. This confirms that leveraging diverse, large-scale data allows the model to learn robust pedestrian dynamics that effectively transfer to data-poor domains, mitigating the challenges associated with data scarcity.

On large-scale datasets (GC, WayMo), however, we observe an intriguing phenomenon: despite the Unified training set being a superset of the Separate one, the performance slightly degrades (e.g., GC ADE: $0.1891 \rightarrow 0.2245$). This can be attributed to the intrinsic behavioral heterogeneity across domains. Although our Canonical Representation aligns coordinates and viewpoints, differences in agent dynamics persist across different geographic regions, such as varying preferred walking speeds and divergent collision avoidance strategies. Consequently, the Unified model must capture these varying behaviors simultaneously, learning a "generalised" representation that inevitably trades off some domain-specific precision for universal applicability. We provide additional analysis of this trade-off, including the effects of sample-size imbalance and pedestrian-vehicle interaction patterns, in Appendix C. While the Separate model maximizes domain-specific fidelity for specific large datasets, the Unified model offers superior deployment scalability. In industrial systems, maintaining a single model that performs competitively across all scenarios (and significantly better on corner cases/small data) is often preferred over managing a siloed ensemble of specialized models, thereby significantly reducing the operational complexity and maintenance overhead of the simulation pipeline.

4.4 Efficiency and Scalability Analysis (RQ3)

In industrial autonomous driving simulation, particularly for Hardware-in-the-Loop (HIL) testing, the simulation frequency should align

[1] https://github.com/abduallahmohamed/Social-STGCNN
[2] https://github.com/Gutianpei/MID
[3] https://github.com/tsinghua-fib-lab/SPDiff
[4] https://github.com/nv-tlabs/trace

Table 1: Quantitative results on the ETH/UCY benchmark, including ADE (m) ↓ / FDE (m) ↓ / Collision Rate (%) ↓. *Separate* denotes training on the remaining 4 ETH/UCY scenes, while *Unified* additionally includes GC, SDD, and WayMo. *Physics* denotes the inclusion of physics-informed guidance.

Model	ETH			Hotel			Zara1			Zara2			Univ		
STGCNN	1.7230	3.4117	9.02	1.3674	2.5956	7.71	0.4262	0.7772	13.7	0.6420	1.1364	16.36	0.9371	1.9221	11.38
MID	0.9125	1.8737	8.21	0.4894	1.0098	7.34	0.2975	0.6484	14.62	0.3420	0.7753	12.65	0.9094	1.9913	9.65
SPDiff	1.3026	2.3449	15.44	1.2428	2.8150	8.02	0.3751	0.8644	18.19	0.3460	0.7928	15.13	1.0143	2.0178	11.89
TRACE	1.5767	2.8601	7.05	0.8812	1.1971	5.93	0.4318	0.7180	13.99	0.4024	0.7257	11.22	1.0007	1.6369	**8.41**
Ours (Seperate)	0.5288	1.0706	6.68	0.4733	1.1135	6.45	**0.1464**	**0.3073**	14.84	**0.1480**	**0.3091**	10.32	0.5566	1.0675	11.2
Ours (Seperate, Physics)	0.5862	1.1322	6.53	0.3826	0.8386	5.91	0.1959	0.3276	**13.1**	0.2036	0.4157	10.12	0.5206	0.9968	9.41
Ours (Unified)	**0.4826**	**0.9903**	5.83	**0.2274**	**0.4774**	5.38	0.1560	0.3138	15.09	0.1518	0.3175	15.34	**0.4186**	**0.8384**	10.77
Ours (Unified, Physics)	0.5345	1.0443	**5.66**	0.2412	0.5136	**5.35**	0.1771	0.3613	13.51	0.2180	0.4623	13.29	0.4428	0.8657	9.18

Table 2: Results on large-scale heterogeneous datasets, including ADE (m) ↓ / FDE (m) ↓ / Collision Rate (%) ↓. Meanings of Separate (*S.*) / Unified (*U.*) / Physics (*P.*) follow Table 1.

Model	GC			SDD			WayMo		
STGCNN	0.6559	1.1598	5.42	1.3056	2.3387	1.35	0.7737	1.4980	4.54
MID	0.4722	1.0158	4.49	1.3822	3.2407	1.2	0.8190	1.7119	5.06
SPDiff	0.4469	0.8958	6.14	0.6518	1.2988	1.55	0.5610	1.2262	5.77
TRACE	0.6504	0.9171	3.03	1.0316	1.5794	1.15	0.9357	1.2885	4.32
Ours (S.)	**0.1891**	**0.3537**	3.24	**0.3057**	**0.5863**	1.02	0.2940	0.5942	4.52
Ours (S., P.)	0.2227	0.3975	**2.83**	0.3537	0.6726	**0.85**	0.3067	0.6151	**4.15**
Ours (U.)	0.2245	0.4063	4.39	0.3331	0.6278	1.16	**0.2617**	**0.5398**	4.62
Ours (U., P.)	0.3800	0.6915	3.74	0.3888	0.7292	1.09	0.2741	0.5615	4.36

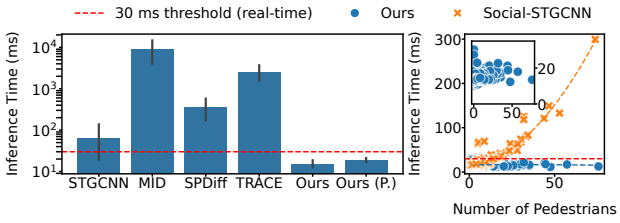

Figure 2: Scalability analysis of inference latency.

with the physical characteristics of onboard sensors, whose standard operating frequency is approximately 30 ms per frame. Consequently, the pedestrian generation module should complete inference within ~30 ms to effectively support the test loop.

In Figure 2 (Upper Left), we compare the average inference latency of our framework against state-of-the-art baselines. High-fidelity diffusion models (MID, TRACE) exhibit prohibitive latencies ($> 10^3$ ms) due to their iterative denoising processes, rendering them impractical for real-time applications. Crucially, RAPID is the only framework that consistently operates below the 30 ms threshold (red dashed line), achieving an average inference time of ~15 ms.

Furthermore, although Social-STGCNN is typically recognized for its efficiency, we observe that its average inference time fails to meet the real-time requirement in our evaluation. This performance degradation stems from its reliance on repeatedly reconstructing the interaction graph in large-scale scenarios, which creates a significant computational bottleneck as crowd density increases (in the upper right panel of Figure 3). In sharp contrast, RAPID maintains a remarkably flat latency profile. This scalability is attributed to our fully parallelized architecture, where all operations are executed via hardware-friendly parallel operators. By maximizing hardware parallelism, we effectively trade spatial memory for temporal efficiency, ensuring that inference time remains stable and real-time even as the number of agents grows.

4.5 Simulation Capability Analysis (RQ4)

In industrial autonomous driving simulation, a generative model must support both controllable behavior execution for targeted test cases and stochastic generation for realistic background flows. We evaluate these properties from the perspectives of intent controllability and effective diversity. The figures in this section provide visualizations of generated trajectories, and additional qualitative examples are provided in Appendix D.

Intent Controllability. We assess the model's ability to follow explicit motion intents by manipulating the desired speed v_{des} and destination $\mathbf{g}$ during inference. Figure 3 (Upper) shows that, without explicit speed guidance, the generated velocities remain strongly correlated with historical motion, indicating that the model preserves kinematic consistency by default. When target speeds are specified, the velocity distributions concentrate around the commanded values, demonstrating that the guidance can effectively enforce user-defined speeds regardless of prior motion. Similarly, as illustrated in Figure 3 (Lower), altering the destination condition causes trajectories to converge toward the specified targets, even when they deviate from the ground truth. This confirms that the proposed goal-driving guidance provides reliable spatial controllability, enabling the construction of customized and counterfactual simulation scenarios. Quantitatively, the R^2 between generated and target speeds reaches 0.8573, while the R^2 between generated and desired heading directions reaches 0.7714 at 12 steps and 0.8651 at 24 steps, further validating RAPID's controllability. Additional details are provided in Appendix E.

Effective Diversity under Physical Constraints. Beyond controllability, a simulator must generate diverse yet physically valid trajectories. Figure 4 reports ADE/FDE and APD as the number of samples increases. We observe that APD remains stable across different numbers, while prediction errors rapidly saturate with

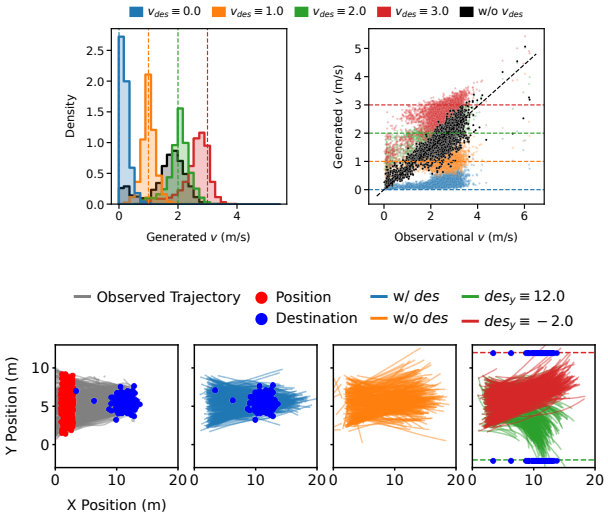

Figure 3: Analysis of speed and destination controllability.

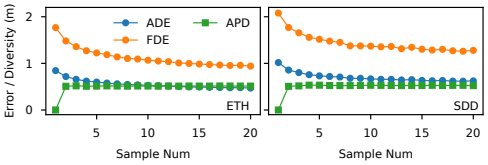

Figure 4: Analysis of generative stability and sample efficiency across different datasets (ETH and SDD).

only a few samples (≤ 5). This indicates that the proposed guidance constrains sampling to a physically feasible manifold, yielding meaningful diversity without introducing implausible outliers. As a result, a small number of samples is sufficient to balance diversity, accuracy, and inference efficiency in practical deployment.

4.6 Ablation Studies (RQ5)

To validate the individual contributions of each component within our unified framework, we conduct ablation studies on the GC, SDD, and WayMo datasets. The results are reported in Table 3.

Efficacy of Canonical Representation. We first evaluate the designs for harmonizing heterogeneous datasets. The removal of Relative Features results in a significant performance degradation across all benchmarks (e.g., SDD ADE deteriorates from 0.3331 to 0.4306), confirming that coordinate-invariant inputs are fundamental for generalized learning. Similarly, removing the Frequency Encoding leads to a distinct drop in these datasets. Furthermore, the Learnable Embedding is also proven effective, ensuring that the model adaptively handles missing modalities without distribution shifts.

Efficiency of Map Context Encoder. The Map Context Encoder is designed to decouple map perception from inference. Ablating the Spatial Anchor leads to increased prediction errors (GC ADE: 0.2848), indicating that explicit spatial priors assist the latent queries

Table 3: Ablation studies on model components.

Method	GC		SDD		WayMo	
	ADE ↓	FDE ↓	ADE	FDE	ADE	FDE
RAPID (Unified, Full)	**0.2245**	**0.4063**	**0.3331**	**0.6278**	0.2617	0.5398
w/o Frequency Encoding	0.2290	0.4163	0.3357	0.6379	0.2639	0.6613
w/o Relative Features	0.2781	0.5274	0.4306	0.8398	0.3379	0.7143
w/o Learnable Embedding	0.2405	0.4327	0.3337	0.6332	0.2770	0.5571
w/o Spatial Anchor	0.2848	0.5349	0.3440	0.6380	0.2484	0.5068
w/o Latent Query	0.3128	0.5738	0.3373	0.6325	**0.2427**	**0.4918**

in capturing topological structures. Most importantly, we investigate the impact of the Latent Query mechanism. As shown in Table 3, removing the latent query mechanism (i.e., using the full raw map features) does not yield a significant performance gain (e.g., SDD ADE remains comparable at 0.3373). This demonstrates that our compression strategy is highly efficient, preserving critical environmental information with negligible loss. However, the computational cost differs dramatically: without the latent query compression, the GPU memory consumption (at batch size 10) surges from 1.5GB to 5.0GB, and the inference latency nearly doubles from ~15ms to ~30ms. This confirms that the proposed Latent Query is a critical enabling technology, effectively pulling the inference speed back from the borderline of the industrial real-time threshold to a safe operational margin.

5 Deployment and Applications

Our pedestrian simulation model has been successfully integrated into a production-grade autonomous-driving cloud simulation platform operated by an industry partner, which provides large-scale simulation and data processing capabilities for industrial deployment. Within this ecosystem, RAPID serves as the core kernel for generating interactive, "Digital Twin" background traffic. During deployment, migrating the model from a standard NVIDIA GPU production environment to a vendor-specific NPU accelerator architecture introduced several infrastructure adaptation challenges inherent to heterogeneous computing. While the NPU's automatic migration adapter supports the majority of standard PyTorch operators, specific optimized attention components triggered regressions to CPU execution. This led to frequent Host-Device synchronization and excessive communication overhead, causing single-step inference latency to spike from the 15ms GPU baseline to over 50ms. To resolve this bottleneck, we implemented an Operator Recomposition strategy to optimize the computational graph. Specifically, we bypassed fused operators dependent on specific CUDA instruction sets and manually reconstructed the attention query mechanism using NPU-optimized Matrix Multiplication (MatMul) primitives. This adaptation enabled full-operator NPU acceleration across the entire framework. While this manual reconstruction is slightly less efficient than closed-source CUDA-optimized kernels, the resulting end-to-end latency was successfully reduced to 26ms. This performance remains comfortably below the 30ms real-time threshold required for industrial simulation, while maintaining a relative prediction accuracy of over 99% compared to the original Torch implementation. Consequently, RAPID enables high-throughput virtual testing and the orchestration of critical

"long-tail" corner cases without compromising the fidelity of the autonomous driving system's validation loop.

6 Related Work

Human trajectory modeling has transitioned from rule-based dynamics to data-driven generative frameworks. We categorize existing literature into three streams: classical knowledge-driven methods, deep learning/reinforcement learning schemes, and emerging diffusion paradigms.

Early research utilized mathematical rules and physical constraints. Cellular Automata (CA) [36] discretize scenes into grids for local transition-based simulation, while Velocity Obstacles (VO) [11] treat collision avoidance as geometric optimization. The Social Force Model (SFM) [17] remains seminal, modeling dynamics via attractive and repulsive forces. Although interpretable, these rigid rules struggle with real-world stochasticity and complex social nuances. To bridge this, hybrid models like Neural Social Physics (NSP) [44] and Physics-informed Crowd Simulator (PCS) [45] integrate physical constraints into neural architectures (e.g., as Neural Differential Equations or iterative learning frameworks). However, these methods often treat physics as hard constraints or post-filters, limiting their capacity to efficiently generate diverse, multi-modal distributions.

Deep learning has shifted focus toward data-driven modeling. Early works like Social-LSTM [2] and Social-GAN [15] used RNNs and GANs to capture social interactions, though the latter often faces mode collapse. To improve multi-modality and consistency, PECNet [29] and Social-VAE utilized CVAE frameworks, while Trajectron++ [35] and Social-STGCNN [30] leveraged graph-structured architectures to model heterogeneous dynamics and improve inference efficiency. Parallel to supervised methods, Reinforcement Learning (RL) approaches like CSRL [26] and HOPRL [19] treat trajectory generation as sequential decision-making. Notably, TECRL [43] incorporates a physics-informed reward based on the "least-effort theory." While RL excels at enforcing physical consistency via reward shaping, its hand-crafted objectives often struggle to capture the full fidelity of real-world data, leading to prediction inaccuracies.

Recently, Diffusion Models [18] have set new benchmarks in trajectory generation. MID [14] first framed prediction as a reverse diffusion process, followed by SingularTrajectory [3], which unified diverse motion modalities. In character animation, Trace & Pace [32] achieved high-fidelity control but with high computational overhead. To integrate social constraints, SPDiff [6] used SFM predictions as guidance conditions. Despite their success, these methods suffer from slow iterative sampling and lack explicit physical consistency guarantees. Unlike them, our RAPID framework bridges the industrial deployment gap via a physics-informed implicit sampler. By leveraging SFM gradients to actively guide denoising rather than mere conditioning, we ensure physical plausibility while optimizing for constant-latency, real-time inference in large-scale simulations.

Our canonical representation module uses coordinate-invariant encodings in local reference frames to support unified training across heterogeneous datasets. This design is inspired by the idea that removing redundant information through symmetry-based representations can substantially improve model performance, which has been widely validated in geometric GNNs through learned canonicalization functions and invariant/equivariant representations [5, 21, 24]. However, unlike the $SE(3)$ equivariance typically pursued in geometric GNNs, two-dimensional urban scenes impose practical constraints: while translation invariance can be handled by local coordinate-invariant motion features, rotation invariance would require costly map re-rasterization that violates the 30 ms real-time threshold, and reflection symmetry is often invalid under directional traffic rules and social norms.

7 Conclusion

This paper presents RAPID, a physics-informed diffusion framework that balances high-fidelity trajectory generation with industrial real-time constraints. By decoupling map perception and employing a Physics-Informed Implicit Sampler, RAPID ensures physical consistency while achieving state-of-the-art performance across five benchmarks. Notably, it is the only model to maintain sub-linear scalability with an average latency of 15 ms, well below the 30 ms real-time threshold. Successfully deployed on a production-grade autonomous-driving simulation platform, RAPID effectively generates physics-compliant background traffic and critical corner cases for the large-scale validation of autonomous driving systems.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2024YFC3307603, the National Natural Science Foundation of China under Grant No. 62476152. We thank Shuo Jin for his valuable support in CARLA deployment.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.

[3] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. 2024. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17890–17901.

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 843–852.

[5] Jiacheng Cen, Anyi Li, Ning Lin, Tingyang Xu, Yu Rong, Deli Zhao, Zihe Wang, and Wenbing Huang. 2026. Universally invariant learning in equivariant GNNs. *Advances in Neural Information Processing Systems* 38 (2026), 85715–85756.

[6] Hongyi Chen, Jingtao Ding, Yong Li, Yue Wang, and Xiao-Ping Zhang. 2024. Social physics informed diffusion model for crowd simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 474–482.

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. 2022. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*.

[8] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[9] Quancheng Du, Qiong Wu, Lingxi Li, Huansheng Ning, and Xiao Wang. 2025. Review and Perspectives on Pedestrian Trajectory Prediction for Safe Transportation. *IEEE Transactions on Intelligent Transportation Systems* 27, 1 (2025), 26–52.

[10] Tian Feng, Lap-Fai Yu, Sai-Kit Yeung, KangKang Yin, and Kun Zhou. 2016. Crowd-driven mid-scale layout design. *ACM Trans. Graph.* 35, 4 (2016), 132–1.
[11] Paolo Fiorini and Zvi Shiller. 1998. Motion planning in dynamic environments using velocity obstacles. *The international journal of robotics research* 17, 7 (1998), 760–772.
[12] Zheng Fu, Kun Jiang, Chuchu Xie, Yuhang Xu, Jin Huang, and Diange Yang. 2024. Summary and reflections on pedestrian trajectory prediction in the field of autonomous driving. *IEEE Transactions on Intelligent Vehicles* (2024).
[13] Mahsa Golchoubian, Moojan Ghafurian, Kerstin Dautenhahn, and Nasser Lashgarian Azad. 2023. Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review. *IEEE transactions on intelligent transportation systems* 24, 11 (2023), 11544–11567.
[14] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. 2022. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17113–17122.
[15] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2255–2264.
[16] Richard Hartley. 2003. *Multiple view geometry in computer vision*. Vol. 665. Cambridge university press.
[17] Dirk Helbing, Illés Farkas, and Tamas Vicsek. 2000. Simulating dynamical features of escape panic. *Nature* 407, 6803 (2000), 487–490.
[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
[19] Kaidong Hu, Brandon Haworth, Glen Berseth, Vladimir Pavlovic, Petros Faloutsos, and Mubbasir Kapadia. 2021. Heterogeneous crowd simulation using parametric reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics* 29, 4 (2021), 2036–2052.
[20] Xuemin Hu, Shen Li, Tingyu Huang, Bo Tang, Rouxing Huai, and Long Chen. 2023. How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2023), 593–612.
[21] Haowei Hua, Jingwen Yang, Wanyu Lin, and Pan Zhou. 2026. Revisiting the Canonicalization for Fast and Accurate Crystal Tensor Property Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 417–425.
[22] Huawei Technologies Co., Ltd. 2020. *Huawei MDC Intelligent Driving Computing Platform White Paper*. Technical Report. Huawei. Focusing on deterministic low-latency and high-performance computing for autonomous driving..
[23] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*. PMLR, 4651–4664.
[24] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. 2023. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*. PMLR, 15546–15566.
[25] Parth Kothari, Sven Kreiss, and Alexandre Alahi. 2021. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 7386–7400.
[26] Jaedong Lee, Jungdam Won, and Jehee Lee. 2018. Crowd simulation by deep reinforcement learning. In *Proceedings of the 11th ACM SIGGRAPH conference on motion, interaction and games*. 1–7.
[27] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. 2007. Crowds by example. In *Computer graphics forum*, Vol. 26. Wiley Online Library, 655–664.
[28] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* 33, 12 (2021), 6999–7019.
[29] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. 2020. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*. Springer, 759–776.
[30] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14424–14432.
[31] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*. IEEE, 261–268.
[32] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13756–13766.
[33] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*. Springer, 549–565.
[34] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research* 39, 8 (2020), 895–935.
[35] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*. Springer, 683–700.
[36] Siamak Sarmady, Fazilah Haron, and Abdullah Zawawi Talib. 2010. Simulating crowd movements using fine grid cellular automata. In *2010 12th International Conference on Computer Modelling and Simulation*. IEEE, 428–433.
[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
[38] Yiwen Song, Jingtao Ding, Jian Yuan, Qingmin Liao, and Yong Li. 2024. Controllable human trajectory generation using profile-guided latent diffusion. *ACM Transactions on Knowledge Discovery from Data* 19, 1 (2024), 1–25.
[39] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2446–2454.
[40] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems* 33 (2020), 7537–7547.
[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[42] Shanwen Yang, Tianrui Li, Xun Gong, Bo Peng, and Jie Hu. 2020. A review on crowd simulation and modeling. *Graphical Models* 111 (2020), 101081.
[43] Zihan Yu, Guozhen Zhang, Yong Li, and Depeng Jin. 2023. Understanding and modeling collision avoidance behavior for realistic crowd simulation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3052–3061.
[44] Jiangbei Yue, Dinesh Manocha, and He Wang. 2022. Human trajectory prediction via neural social physics. In *European conference on computer vision*. Springer, 376–394.
[45] Guozhen Zhang, Zihan Yu, Depeng Jin, and Yong Li. 2022. Physics-infused machine learning for crowd simulation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 2439–2449.
[46] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2871–2878.

A Details of Physics-Informed Guidance

In this section, we detail the specific formulations of the physical priors used in our physics-informed implicit sampler. Specifically, we leverage the classical Social Force Model (SFM) [17] to guide the diffusion sampling process. The SFM postulates that the motion of a pedestrian is governed by a sum of forces: a driving force reflecting the intention to reach a destination, and repulsive forces describing interactions with other agents and static obstacles.

In our framework, we treat these three components as distinct guidance targets $\mathbf{a}_i$, where $i \in \{\text{destination, map, social}\}$. The physical meaning and mathematical formulation of each term are defined as follows:

A.1 Destination Force (Attraction)

The destination force, $\mathbf{a}_{\text{dest}}$, represents the internal motivation of a pedestrian i to adapt their current velocity $\mathbf{v}_i$ towards a desired velocity vector. This term encourages the agent to move towards their goal g_i at a desired speed v_i^0. It is mathematically formalized as a relaxation process:

$$\mathbf{a}_{\text{dest}} = \frac{v_i^0 \mathbf{e}_i(t) - \mathbf{v}_i(t)}{\tau}, \tag{8}$$

where $\mathbf{e}_i(t) = \frac{\mathbf{g}_i - \mathbf{p}_i(t)}{\|\mathbf{g}_i - \mathbf{p}_i(t)\|}$ is the desired direction unit vector pointing from current position $\mathbf{p}_i$ to the destination $\mathbf{g}_i$, and τ is the relaxation time parameter characterizing how quickly the pedestrian adjusts their path. In our implementation, this term guides

the generated trajectory to ensure kinematic consistency with the agent's intended goal.

A.2 Social Force (Inter-Agent Repulsion)

The social force, $\mathbf{a}_{\text{social}}$, models the psychological tendency of pedestrians to maintain a certain safe distance from others to avoid collisions and uncomfortable proximity. For a pedestrian i, the repulsive force exerted by another pedestrian j is modeled as an exponential decay function of their distance:

$$\mathbf{a}_{\text{social}} = \sum_{j \neq i} A_{\text{soc}} \exp\left(\frac{-d_{ij}}{B_{\text{soc}}}\right) \mathbf{n}_{ij}, \tag{9}$$

where $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|$ is the Euclidean distance between the centers of pedestrians i and j. $\mathbf{n}_{ij}$ is the normalized vector pointing from j to i, representing the direction of the repulsive force. A_{soc} and B_{soc} are constants representing the interaction strength and range, respectively. This prior effectively penalizes generated states that result in inter-agent collisions or overcrowding.

A.3 Map Force (Obstacle Repulsion)

The map force, $\mathbf{a}_{\text{map}}$, ensures physical consistency with the static environment by preventing agents from penetrating walls or entering non-navigable areas. Similar to the social force, this interaction is modeled as a repulsive force from the nearest obstacle point O:

$$\mathbf{a}_{\text{map}} = \sum_{o \in O} A_{\text{obs}} \exp\left(\frac{-d_{io}}{B_{\text{obs}}}\right) \mathbf{n}_{io}, \tag{10}$$

where d_{io} denotes the distance to the nearest static obstacle, and $\mathbf{n}_{io}$ is the direction vector pointing away from the obstacle. A_{obs} and B_{obs} control the magnitude and decay range of the obstacle repulsion. By integrating this term into the diffusion sampling, our model actively steers trajectories away from unnavigable regions defined in the semantic map.

In practice, searching for the nearest obstacle point O in a high-resolution rasterized map for every agent at every step is computationally prohibitive. To satisfy real-time constraints, we implement Eq. (10) using a convolution-like approximation with a pre-computed force template. Specifically, we generate a fixed-size local repulsion kernel $\mathbf{K} \in \mathbb{R}^{(2R+1)\times(2R+1)\times 2}$ centered at the origin, where each entry $\mathbf{K}_{u,v}$ represents the repulsive vector from a hypothetical obstacle at local offset (u, v) calculated via the exponential decay formula Eq. (10). This kernel is cached to avoid redundant computation. During inference, for each pedestrian i at position $\mathbf{p}_i$, we extract a local occupancy patch $\mathcal{M}^i_{local}$ of size $(2R+1)\times(2R+1)$ from the global semantic map. The effective map force is then computed by aggregating the pre-computed forces masked by the actual obstacle occupancy:

$$\mathbf{a}_{\text{map}} \approx \sum_{u=-R}^{R} \sum_{v=-R}^{R} \mathcal{M}^i_{local}(u, v) \cdot \mathbf{K}_{u,v}. \tag{11}$$

In the inference phase, these three acceleration components $\mathbf{a}_{\text{dest}}$, $\mathbf{a}_{\text{social}}$, and $\mathbf{a}_{\text{map}}$ serve as the targets $\mathbf{a}_i$ in Eq. (7), calculating the gradients to iteratively refine the noisy state $\hat{\mathbf{x}}_0$ towards physically feasible regions.

B Detailed Experimental Settings

Dataset Preprocessing and Splits. We provide detailed protocols for the datasets used in our evaluation:

- **ETH/UCY:** This benchmark comprises 5 scenes with relatively sparse crowds. Following [25], we use the standard leave-one-out strategy, training on 4 scenes and testing on the remaining one.
- **SDD:** We follow the standard practice [25] where the last scene of each scene class is reserved for testing, while the others are utilized for training and validation.
- **WayMo:** To focus on complex interactions pertinent to autonomous driving, we filter the dataset to select the top 500 scenes with the highest pedestrian density.
- **GC:** We utilize the standard train-test split provided by the dataset.

For GC and the selected WayMo scenes, we strictly implement a temporal split, allocating the first 80% of the timeline for training/validation and the subsequent 20% for testing.

Unified Training Details. In the **Unified** setting, the training data is a superset of the standard setting. For instance, when testing on a specific ETH/UCY scene, the training data comprises the remaining 4 ETH/UCY scenes combined with the complete training sets of SDD, GC, and WayMo. This setup evaluates the model's ability to learn robust representations from heterogeneous data sources.

Implementation and Hyperparameters. The training process is governed by an early stopping mechanism with a patience of 200 epochs, terminating when the validation loss ceases to improve. We determine optimal hyperparameters via grid search, exploring learning rates in $\{10^{-5}, 10^{-4}, 10^{-3}\}$ and batch sizes in $\{32, 64, 128\}$. The same grid-search budget is allocated to every baseline and RAPID variant, and all models are selected according to their validation performance under the same data splits. All experiments were conducted on a high-performance computing server equipped with dual AMD EPYC 7742 64-Core Processors and **8 × NVIDIA A100-SXM4 GPUs (80GB VRAM)**. The system is supported by 1TB of RAM and operates on Ubuntu 22.04 LTS with a CUDA 12.2 environment.

C Analysis of Unified Training Trade-offs

The Unified setting is designed to learn a generalist model from heterogeneous trajectory datasets, but it does not consistently improve performance across all training domains, mainly due to two factors: First, the unified training set is highly imbalanced in sample size. When GC, SDD, and WayMo are trained jointly, the much larger WayMo subset can dominate gradient updates, making the unified model fit mixed-traffic patterns more strongly while underfitting the smaller GC and SDD domains. This imbalance helps explain why the Unified setting improves or remains highly competitive on WayMo, but can slightly degrade ADE/FDE on GC and SDD compared with training separate domain-specific models. Second, the datasets differ in their intrinsic interaction patterns. GC and SDD mainly contain dense pedestrian-pedestrian interactions in subway or campus-like open spaces, whereas WayMo contains sparser pedestrian crowds but substantially richer pedestrian-vehicle interactions. As a result, dense GC/SDD samples can provide useful hard cases for improving interaction reasoning in WayMo, while

the pedestrian-vehicle behaviors emphasized by WayMo provide less direct benefit to GC and SDD. Since RAPID targets autonomous driving simulation, where mixed pedestrian-vehicle interaction is central, we regard this trade-off as acceptable for deployment, especially given the reduced operational complexity of maintaining a single unified model.

To further verify these explanations, we conduct an additional unified-training experiment using only GC and SDD while excluding WayMo, following the same evaluation protocol as Table 2. As shown in Table 4, once the dominant WayMo subset is removed, unified training improves both GC and SDD over their separate counterparts, indicating that the degradation in the full unified setting mainly comes from sample-size imbalance rather than an inherent limitation of joint training.

Table 4: Supplementary unified-training results on GC and SDD. We report ADE (m) ↓ / FDE (m) ↓ / Collision Rate (%) ↓.

Model	GC	SDD
Ours (Separate)	0.1891 / 0.3537 / 3.24	0.3057 / 0.5863 / 1.02
Ours (Unified, GC & SDD)	0.1844 / 0.3540 / 1.48	0.3017 / 0.5803 / 0.37

D Qualitative Trajectory Visualizations

Figure 5 provides representative qualitative examples comparing RAPID with the classical Social Force Model (SFM) in the same pedestrian-vehicle interaction case (Blue dots denote pedestrians, orange squares denote vehicles), where RAPID generates smoother and more anticipatory pedestrian responses to nearby vehicles, while SFM tends to react only when vehicles become close, producing less natural avoidance behavior. More generated trajectory demos, including dynamic videos, are available in our public GitHub repository.

E Controllability Analysis

We systematically evaluate RAPID's speed and heading controllability by comparing the alignment between commanded inputs and generated outputs. Figure 6 reports the quantitative results under three control dimensions.

The left subplot evaluates speed control by comparing generated speeds with the commanded speeds, where the R^2 reaches 0.8573, showing a strong linear correlation and precise velocity-following capability. The middle and right subplots evaluate heading control by comparing generated heading angles with the desired heading directions. The R^2 reaches 0.7714 at 12 steps and further increases to 0.8651 at 24 steps, suggesting that destination guidance becomes more stable over longer rollouts and validating that the physics-guided sampling mechanism can achieve both precise speed control and robust destination following.

F Failure Analysis

Although RAPID can generate realistic pedestrian-pedestrian and pedestrian-vehicle interactions in common traffic scenarios, it may still fail under extreme crowd density or complex obstacle layouts. In ultra-dense scenes, the SFM-based repulsive guidance can conflict with the data-driven motion prior, leading to unsmooth trajectories or occasional collisions. This occurs partly because pedestrian embeddings do not explicitly encode all neighboring states; instead, they query neighboring agents through attention over positional encodings, which can mix interaction information when many pedestrians are nearby.

Another failure mode appears when pedestrians are close to complex obstacles and need to temporarily move away from their destinations to find a feasible path. In such cases, the compressed map representation may not preserve sufficiently detailed long-range path information, making it difficult for the model to plan around distant openings. These issues could be mitigated by adding richer neighbor states to pedestrian embeddings or increasing the number of learnable map tokens, but both changes would increase computational cost and inference latency. Since such extreme density and obstacle configurations are uncommon in mixed-traffic autonomous driving simulation, RAPID prioritizes real-time performance while maintaining robust behavior in typical pedestrian-vehicle scenarios.

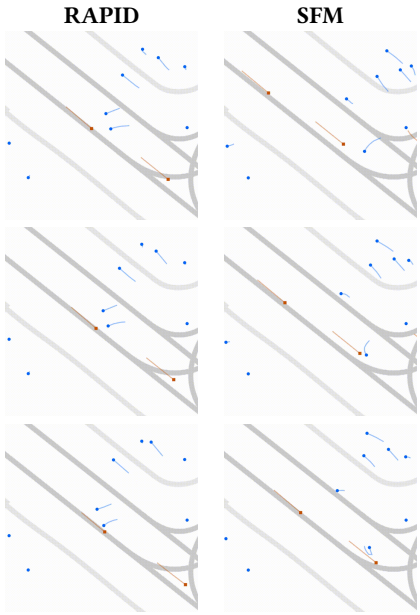

Figure 5: Qualitative comparison in a pedestrian-vehicle interaction scenario.

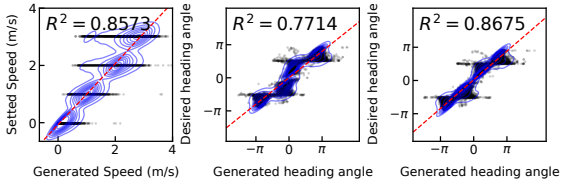

Figure 6: Quantitative analysis of controllability.