PLFlow: Prior-Conditioned Latent Flow Matching for Map-Compliant Forecasting

Yiming Xu[1], Hao Cheng[2], and Monika Sester[1]

[1] Leibniz Universität Hannover, Germany
[2] University of Twente, The Netherlands

Abstract. Trajectory forecasting requires a compact set of diverse futures that remain compliant with observable road geometry. Flow matching provides powerful conditional generation with simulation-free training, but is challenging to apply directly to trajectory forecasting. Three practical issues arise: (i) forecasting naturally starts from a history-conditioned distribution rather than a standard Gaussian, (ii) multi-step generation has high latency and discretized ODE sampling is brittle under intermediate-state drift, and (iii) unordered samples do not directly provide a compact, ordered set of probabilistic modes. We propose PLFlow (Prior-Conditioned Latent Flow Matching), which formulates forecasting as *history→future* conditional transport in a shared trajectory-VAE latent space with a history-induced Gaussian prior and training-time Gaussian augmentation for robust integration. To enable real-time inference, we pool and re-cluster teacher rollouts into K representative latent centers and distill them into a one-step student through permutation-invariant matching. The resulting student outputs an ordered set of probabilistic trajectories while the stochastic flow process remains in the teacher. Experiments on Argoverse 1 and 2 demonstrate competitive forecasting accuracy and improved road-geometry compliance. Controlled comparisons indicate that repeated state-dependent transport contributes to lower map-violation rates, whereas repeated endpoint prediction does not.

Keywords: Trajectory forecasting · Flow matching · Map consistency

1 Introduction

Trajectory forecasting predicts multiple plausible futures from motion history, HD maps, and nearby agents [1, 13, 37]. Most systems return a compact set of trajectories with confidence scores [5,18,20,23,49,50]. The output is consumed by downstream planners under a small, fixed hypothesis budget—commonly $K = 6$ on these benchmarks. This interface demands coverage of mutually exclusive maneuvers, meaningful confidence ranking, and feasible secondary modes. Because minADE and minFDE score only the closest prediction, they can obscure invalid alternatives; full-set road-geometry measures provide a complementary view. Displacement accuracy alone is insufficient for downstream planning: secondary hypotheses may leave the drivable area or conflict with lane direction

even when the best mode is accurate. Forecasting should therefore preserve multimodality and road-geometry fidelity across the full set—an important structural prerequisite, not evidence of safe or defensive behavior.

Conditional diffusion and flow models offer expressive multimodal generation under contextual guidance [12,21,24,33]. Flow matching is particularly appealing because it learns a vector field with a simulation-free training objective [22]. Its direct use in trajectory forecasting nevertheless presents three challenges. First, forecasting provides structured motion history, suggesting an informative source rather than a task-agnostic Gaussian. Second, iterative ODE evaluation is costly, and coarse integration can visit poorly supervised intermediate states [2]. Third, stochastic rollouts do not directly provide the fixed number of ordered modes and probabilities required by forecasting interfaces.

We propose **PLFlow** (**P**rior-Conditioned **L**atent **Flow** Matching), which learns conditional transport between observed and future motion in a shared trajectory-VAE manifold [17]. Embedding both segments with the same encoder and decoder makes source and target comparable and reframes generic noise-to-data synthesis as motion continuation. A mixture of flow heads represents distinct transport branches, while scene features are reused at every vector-field evaluation so that each branch can respond to map geometry and agent interactions throughout transport. Its source is induced by the history posterior rather than a standard Gaussian, while the current latent state, transport time, and scene context condition the vector field throughout generation. This state-dependent conditioning allows scene context to redirect intermediate latents as transport evolves. A lightweight training-time state-coverage strategy improves robustness to coarse integration without discarding the history anchor.

For efficient fixed-K prediction, a multi-step teacher generates diverse future latents, which are re-clustered into representative modes and distilled into a one-pass student. The student directly outputs ordered trajectories and probabilities, while stochastic transport and intra-mode variation remain in the teacher. This separation provides both a high-quality generative operating point and a deterministic deployment interface.

We evaluate PLFlow on Argoverse 1 (AV1) and Argoverse 2 (AV2) [1, 37] using standard forecasting and map-consistency metrics. PLFlow achieves competitive displacement accuracy and improves off-road, off-yaw, and lane-miss rates over strong regression baselines. To isolate the effect of conditional transport, we compare endpoint prediction and vector-field integration under the same source, scene context, architecture, and augmentation. With vector-field integration, increasing the number of evaluations from one to three improves all four road-geometry metrics: off-road rate, off-yaw rate, top-mode lane-miss rate, and best-of-six lane-miss rate. Repeated endpoint prediction shows no consistent improvement. This matched comparison provides evidence that repeated state-dependent transport contributes materially to map consistency.

Our main contributions are as follows:

- We formulate trajectory forecasting as conditional transport from a history-induced prior to the future distribution in a shared motion manifold.

- We adapt latent flow to robust coarse-step generation and distill its stochastic teacher into an efficient fixed-K probabilistic predictor.
- We demonstrate competitive forecasting accuracy and improved map consistency on AV1 and AV2, supported by controlled analysis of state-dependent transport.

2 Related Work

Trajectory Prediction. Current mainstream trajectory forecasting combines structured scene encoding with a fixed set of K hypotheses and associated probabilities. Vectorized and graph-based representations preserve map topology and enable explicit agent–map and agent–agent reasoning via Transformers [5,20,23,28,30,31,41,49,50]. A dominant paradigm is winner-takes-all (WTA) multi-hypothesis learning, including DETR-style query decoding, where only the best-matching hypothesis is regressed and mode probabilities are learned with an auxiliary classification objective [18, 23, 30, 47, 49, 50]. Several works further refine hypotheses through recurrent unrolling or trajectory-conditioned interaction [14, 38, 48]. We adopt QCNet's query-centric graph encoder [49]; its scene features are computed once and reused across agents and teacher/student heads. Complementary off-road, off-yaw, and lane-miss measures [8,10,29] test whether all predicted hypotheses respect road geometry beyond displacement accuracy.
Generative Modeling for Trajectory Forecasting. Diffusion and score-based models provide expressive conditional generation but typically require iterative sampling [12, 33]. These paradigms have been explored in pedestrian and autonomous-driving forecasting [16, 36, 40, 45]. In particular, CDT-P [40] uses map and endpoint tokens to control a diffusion-based AV2 predictor, providing a directly comparable conditional generative baseline. Flow matching and rectified flow learn conditional velocity fields and support ODE-based generation with simulation-free training [21, 24]. Existing flow formulations for time-series forecasting commonly follow a generic "noise→data" paradigm. This design underuses the fact that forecasting already provides a structured history and that history and future can be represented within the same latent space. Unlike these formulations, PLFlow starts from a history-conditioned prior and models *history→future* transport.

Recent flow-based predictors include TrajFlow [42], a single-pass confidence-ranked forecaster, and MoFlow [4], which distills flow matching for human trajectory forecasting. TrajFlow is evaluated on Waymo and MoFlow on pedestrian datasets, so neither reports directly comparable AV1/AV2 results under our protocol. More importantly, our focus is forecasting-specific state-dependent transport in dense HD-map scenes and its relation to road-geometry compliance.
Few/One-step Acceleration and Distillation. Iterative sampling motivates progressive and distribution-matching distillation [25, 27, 43], consistency-style training [11, 32], and recent one-step objectives for rectified flow [3, 7, 19, 51]. Under strong conditioning, one-step parameterizations can behave similarly to direct endpoint regression and make predictions less dependent on intermediate states. We therefore separate the two roles: the multi-step teacher retains

stochastic, state-dependent flow generation, whereas the distilled student preserves its representative mode-level targets in a standard fixed-K interface. We do not claim that the student retains the teacher's full intra-mode uncertainty or flow semantics.

3 Method

3.1 Problem Setup

We consider an autonomous-driving scene observed at discrete timesteps $t \in \{1, \dots, T_h\}$. At each timestep, upstream perception provides the states of surrounding agents together with an HD vector map. Let $\mathcal{A}_t$ denote the detected dynamic agents at time t, and let $A = |\mathcal{A}_{T_h}|$ be the number of agents considered at the current time. For an agent $i \in \mathcal{A}_t$, its state $\mathbf{s}_i^t$ includes its 2D position $\mathbf{p}_i^t \in \mathbb{R}^2$, heading, kinematic cues such as velocity or displacement, validity, and semantic category. The HD map is represented as M lane polylines or polygons $\mathcal{M} = \{L_j\}_{j=1}^{M}$. Each map element contains sampled points with local geometry and semantics such as direction, boundary type, lane type, and intersection membership.

Given the target history $\mathbf{X}_i \in \mathbb{R}^{T_h \times 2}$, neighboring-agent histories, and relevant map elements, our goal is to predict K plausible future trajectories over T_f steps and their associated probabilities,

$$\left\{\hat{\mathbf{Y}}_i^{(k)}, \hat{\pi}_i^{(k)}\right\}_{k=1}^{K}, \qquad \hat{\mathbf{Y}}_i^{(k)} \in \mathbb{R}^{T_f \times 2}, \quad \sum_{k=1}^{K} \hat{\pi}_i^{(k)} = 1. \tag{1}$$

We omit validity masks from subsequent notation when they are clear from context.

3.2 Architecture

Overview. Fig. 1 summarizes PLFlow. Given agent histories and local HD-map context, a scene encoder produces a context embedding $\mathbf{c}_i$ for each target agent. A shared trajectory VAE embeds both history and future segments into the *same* latent space, yielding history statistics $(\boldsymbol{\mu}_{x,i}, \boldsymbol{\sigma}_{x,i})$ and a future latent mean $\boldsymbol{\mu}_{y,i}$. On this shared manifold, we train a K-head conditional rectified-flow teacher that transports samples from a history-conditioned start distribution toward future latents. Teacher rollouts are then summarized into a fixed set of representative modes and distilled into a one-step student that predicts K ordered latent displacements and probabilities.

Scene Encoder Backbone. A strong scene encoder must fuse map topology, multi-agent interactions, and temporal motion cues into representations that are both expressive and efficient. We adopt QCNet's query-centric encoder [49] as our backbone. The map is encoded as polylines or polygons, while agent

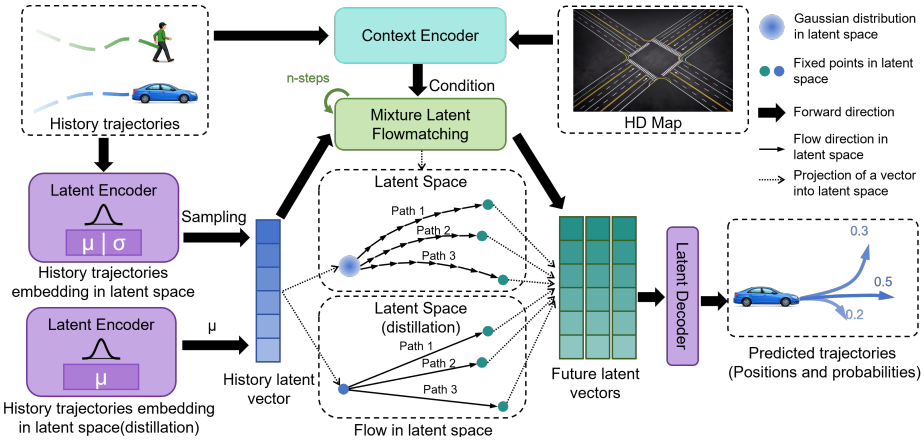

Fig. 1: Overview of PLFlow. The teacher performs scene-conditioned latent transport from a history-induced start distribution. Its rollouts are pooled and re-clustered into a fixed-K target set, which is aligned to a one-step student by permutation-invariant matching. The stochastic flow process resides in the teacher; the student provides the deterministic fixed-K deployment interface.

history is represented by A agents over T_h timesteps. QCNet encodes agents and map elements in local coordinate frames and aligns them through deterministic spatial transforms anchored at each agent, producing features stable under global translation and rotation. For each target, learnable queries aggregate nearby agents and relevant map elements into $\mathbf{c}_i \in \mathbb{R}^{d_c}$. The same scene features are reused by all flow heads, ODE steps, and the distilled student.

Trajectory Latent Manifold via a Shared VAE. We learn a compact trajectory latent space using a shared sequence VAE so that history and future trajectories are embedded into the same manifold. Let $\mathbf{X}_i \in \mathbb{R}^{T_h \times 2}$ and $\mathbf{Y}_i \in \mathbb{R}^{T_f \times 2}$ denote agent-centric history and future positions. To reuse one VAE despite different horizons, both segments are represented at length $L = T_f$ with masks. The future is encoded chronologically; the history excludes the current frame, is reversed from the most recent state, and is zero-padded to length L. An encoder E_ϕ maps a segment $\tilde{\mathbf{Z}}$ to a diagonal Gaussian,

$$(\boldsymbol{\mu}_z, \log \boldsymbol{\sigma}_z^2) = E_\phi(\tilde{\mathbf{Z}}), \qquad \mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The shared increment decoder D_ψ predicts per-step displacements and reconstructs positions by cumulative summation. We optimize

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \beta D_{\text{KL}}\Big(q_\phi(\mathbf{z} \mid \tilde{\mathbf{Z}}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\Big), \tag{3}$$

using masked ℓ_2 reconstruction and a weak $\beta = 10^{-4}$ regularizer. Applying the shared encoder gives $(\boldsymbol{\mu}_{x,i}, \boldsymbol{\sigma}_{x,i})$ for history and $(\boldsymbol{\mu}_{y,i}, \boldsymbol{\sigma}_{y,i})$ for the future. Their means are directly comparable points on the same learned motion manifold, and the downstream flow target is $\mathbf{x}_{1,i} \triangleq \boldsymbol{\mu}_{y,i}$.

The encoder is a one-layer GRU with 128 hidden units and 64 learnable Fourier frequency bands; the decoder is an MLP and the latent dimension is $d_z = 32$. After pretraining, the VAE is frozen for teacher and student training. This keeps the decoder identical across transport variants and lets the ablations isolate latent alignment and generation behavior.

History-conditioned Prior. Standard rectified flow commonly starts from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Forecasting already provides a structured motion history, so PLFlow instead derives the source from the history posterior,

$$p_0(\mathbf{x} \mid \mathbf{X}_i) = \mathcal{N}\big(\boldsymbol{\mu}_{x,i}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x,i})\big), \qquad \mathbf{x}_{0,i} = \boldsymbol{\mu}_{x,i} + \boldsymbol{\sigma}_{x,i} \odot \boldsymbol{\epsilon}. \tag{4}$$

This source is stochastic but remains anchored in a behaviorally meaningful neighborhood of the observed dynamics. The same noise realization is shared across mode heads so that mode diversity is learned by the conditional mixture rather than introduced by unrelated initial samples.

Conditional Mixture Latent Flow Matching. For $t \sim \mathcal{U}(0, 0.95)$, the linear rectified-flow path and target velocity are

$$\mathbf{x}_{t,i} = (1-t)\mathbf{x}_{0,i} + t\mathbf{x}_{1,i}, \qquad \mathbf{u}^\star_i = \mathbf{x}_{1,i} - \mathbf{x}_{0,i}. \tag{5}$$

A K-head decoder predicts $\{\mathbf{v}^{(k)}_\theta(\mathbf{x}_{t,i}, t, \mathbf{c}_i)\}_{k=1}^K$ and logits $\{\ell^{(k)}_{\theta,i}\}_{k=1}^K$. Each head uses seven scene-conditioned refinement layers. At every layer, a latent-mode token is modulated by a time embedding and attends to the current latent state, map polygons, target history, and neighboring-agent features. Consequently, the update depends jointly on the current latent state, transport time, and scene condition.

As in common multi-hypothesis forecasters, the teacher uses WTA regression. We select a winning mode $k^\star$ using the final-layer prediction at the deterministic history anchor $\boldsymbol{\mu}_{x,i}$ $(t = 0)$ and reuse that assignment across sampled times and auxiliary layers, reducing mode switching during training. For a selected head, the endpoint estimate is

$$\hat{\mathbf{x}}^{(k)}_{1,i} = \mathbf{x}_{t,i} + (1-t)\mathbf{v}^{(k)}_\theta(\mathbf{x}_{t,i}, t, \mathbf{c}_i). \tag{6}$$

Here, the vector field is conditioned on scene context $\mathbf{c}_i$ in addition to the history-conditioned source in Eq. (4). The factor $(1-t)$ accounts for the remaining integration interval from t to 1.

Detached Self-Conditioning. Training only on exact linear interpolants does not expose the field to states produced by its own endpoint errors. We therefore add a training-only corrective pass. After obtaining $\hat{\mathbf{x}}^{(k)}_{1,i}$ at one sampled time t, we detach the estimate, use it as the endpoint of a new interpolation, independently sample $t' \sim \mathcal{U}(0, 0.95)$, and form

$$\tilde{\mathbf{x}}^{(k)}_{1,i} = \mathrm{sg}\Big[\hat{\mathbf{x}}^{(k)}_{1,i}\Big], \qquad \tilde{\mathbf{x}}^{(k)}_{t',i} = (1-t')\mathbf{x}_{0,i} + t'\tilde{\mathbf{x}}^{(k)}_{1,i}. \tag{7}$$

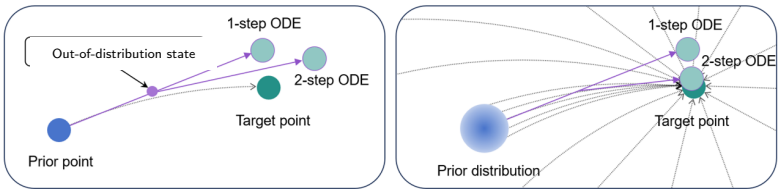

Fig. 2: Robust conditional transport. Mean-to-mean training covers a narrow set of states, so discretization error can move the solver into poorly supervised regions (left). History-conditioned sampling and occasional standard-Gaussian replacement expand coverage while retaining scene conditioning, enabling the vector field to correct perturbed intermediate states (right).

Although the interpolated state is induced by the predicted endpoint, the second field query is supervised toward the true future endpoint,

$$\tilde{\mathbf{u}}_i^{(k)\star} = \frac{\mathbf{x}_{1,i} - \tilde{\mathbf{x}}_{t',i}^{(k)}}{1 - t'}. \tag{8}$$

This exposes the field to a model-induced state and trains a corrective direction back to the true future rather than merely repeating endpoint prediction. Self-conditioning is disabled at inference and excluded from latency measurements.

The teacher loss combines latent velocity regression, decoded trajectory regression, and mixture probability learning:

$$\mathcal{L}_{\text{reg}} = \left\| \mathbf{v}_\theta^{(k^\star)} - \mathbf{u}_i^\star \right\|_2^2 + \lambda_{\text{traj}} \left\| D_\psi(\hat{\mathbf{x}}_{1,i}^{(k^\star)}) - \mathbf{Y}_i \right\|_2^2, \tag{9}$$

$$\mathcal{L}_\pi = -\log \sum_{k=1}^{K} \pi_{\theta,i}^{(k)} \exp\left(-\left\| \mathbf{v}_\theta^{(k)} - \mathbf{u}_i^\star \right\|_2^2 \right), \qquad \boldsymbol{\pi}_{\theta,i} = \text{softmax}(\boldsymbol{\ell}_{\theta,i}). \tag{10}$$

We set $\lambda_{\text{traj}} = 1$. The regression terms are applied to both the ordinary path state in Eq. (5) and the self-conditioned state in Eq. (7). All seven decoder layers receive regression supervision with geometrically decayed auxiliary weights; $\mathcal{L}_\pi$ is applied to the final layer.

Gaussian Sampling and State-space Augmentation. Although Eq. (5) provides simulation-free training, a coarse Euler solver can enter latent regions that are not well covered by the sampled paths. We expand training coverage by replacing the history-conditioned source in Eq. (4) with $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with probability 0.1. The target future and scene condition remain unchanged. This deliberately presents the vector field with off-anchor states and encourages it to recover a map-compatible direction from $\mathbf{c}_i$ rather than relying only on a narrow history neighborhood. At inference, generation always begins from the history-conditioned source. Fig. 2 illustrates the distinction.

Teacher Sampling and Fixed-K Targets. The teacher solves $d\mathbf{x}/dt = \mathbf{v}_\theta^{(k)}(\mathbf{x}, t, \mathbf{c}_i)$ with R Euler steps. For each of its K heads, it draws S history-

conditioned starts and obtains endpoints

$$\mathbf{x}_{r+1,i}^{(k,s)} = \mathbf{x}_{r,i}^{(k,s)} + \frac{1}{R}\mathbf{v}_\theta^{(k)}\left(\mathbf{x}_{r,i}^{(k,s)}, \frac{r}{R}, \mathbf{c}_i\right). \tag{11}$$

We use $R = 3$ and $S = 10$. We pool all KS endpoints and run K-means again on the complete set to obtain $K = 6$ centers $\{\bar{\mathbf{x}}_{1,i}^{(j)}\}_{j=1}^{K}$, rather than retaining the original teacher-head partition. Consequently, samples from different heads can merge when they describe the same maneuver, while samples from one head can separate when they describe different maneuvers. This converts the unordered stochastic rollout set into a compact set of representative latent targets.

One-step Student Distillation. The deterministic student predicts latent displacements $\{\hat{\mathbf{d}}_i^{(k)}\}_{k=1}^{K}$ relative to the history mean and probabilities $\{\hat{\pi}_i^{(k)}\}_{k=1}^{K}$ in one pass. Using k to index student modes and j to index the re-clustered teacher centers, the student endpoints and teacher targets are

$$\hat{\mathbf{x}}_{1,i}^{(k)} = \boldsymbol{\mu}_{x,i} + \hat{\mathbf{d}}_i^{(k)}, \qquad \bar{\mathbf{d}}_i^{(j)} = \bar{\mathbf{x}}_{1,i}^{(j)} - \boldsymbol{\mu}_{x,i}. \tag{12}$$

Because cluster indices are arbitrary, we align the target set to the student set with a one-to-one Hungarian assignment,

$$\rho_i = \arg\min_{\rho \in \mathfrak{S}_K} \sum_{k=1}^{K} \left\| \hat{\mathbf{d}}_i^{(k)} - \bar{\mathbf{d}}_i^{(\rho(k))} \right\|_2^2, \tag{13}$$

where $\mathfrak{S}_K$ denotes all permutations of K elements. This produces dense supervision for every student mode without assuming that the teacher's original heads have persistent semantic identities.

Mode scores are trained from ground truth, not from teacher-cluster frequencies. Let $e_i^{(k)}$ be the masked ℓ_2 trajectory error between $D_\psi(\boldsymbol{\mu}_{x,i} + \hat{\mathbf{d}}_i^{(k)})$ and $\mathbf{Y}_i$. We form the detached soft target $q_i^{(k)} = \exp(-e_i^{(k)}) / \sum_j \exp(-e_i^{(j)})$, without a temperature parameter, and optimize $\mathcal{L}_{\text{cls}} = -\sum_k q_i^{(k)} \log \hat{\pi}_i^{(k)}$. The GT-best mode receives the largest target mass, while near-best modes retain smaller nonzero targets. This resembles QCNet-style soft mode classification, but uses ℓ_2 trajectory error rather than a probabilistic NLL. The complete objective is

$$\mathcal{L}_{\text{student}} = \sum_{k=1}^{K} \left\| \hat{\mathbf{d}}_i^{(k)} - \bar{\mathbf{d}}_i^{(\rho_i(k))} \right\|_2^2 + \min_k \left\| D_\psi(\boldsymbol{\mu}_{x,i} + \hat{\mathbf{d}}_i^{(k)}) - \mathbf{Y}_i \right\|_2^2 + \mathcal{L}_{\text{cls}}. \tag{14}$$

The trajectory-space term has unit weight. The student retains the teacher's representative maneuver set but collapses each cluster to one latent center; it therefore does not preserve intra-cluster variance. At test time, $\hat{\mathbf{Y}}_i^{(k)} = D_\psi(\boldsymbol{\mu}_{x,i} + \hat{\mathbf{d}}_i^{(k)})$.

3.3 Training and Inference

We train PLFlow in three stages. (1) **Shared VAE:** we pretrain (E_ϕ, D_ψ) on unified history and future segments, then freeze it as the common latent interface and trajectory synthesizer. (2) **Mixture latent-flow teacher:** we optimize the K-head vector field using history-conditioned starts, random-time rectified-flow supervision, detached self-conditioning, WTA regression, and mixture probability learning. (3) **One-step student:** we generate teacher rollouts with three Euler steps and ten starts per head, pool and re-cluster all endpoints into six centers, and train the student after Hungarian alignment. The flow teacher and student are each trained for 50 epochs with batch size 16.

At teacher inference, Eq. (11) generates stochastic rollouts from the history-conditioned prior. At deployment, we instead run the student once, form $\hat{\mathbf{x}}_{1,i}^{(k)} = \boldsymbol{\mu}_{x,i} + \hat{\mathbf{d}}_i^{(k)}$, and decode $\hat{\mathbf{Y}}_i^{(k)} = D_\psi(\hat{\mathbf{x}}_{1,i}^{(k)})$. Thus, iterative flow generation is used to construct and evaluate the teacher, while the final student provides six trajectories and probabilities without ODE integration.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate on the AV1 and AV2 datasets [1, 37]. Both are sampled at 10 Hz: AV1 provides 2 s of history and a 3 s prediction horizon, whereas AV2 provides 5 s of history and a 6 s horizon. For $K = 6$, we report minADE_6, minFDE_6, MR_6, and confidence-aware b-minFDE_6. On AV2 vehicle focal agents, we further evaluate off-road rate, off-yaw rate, and lane miss rate (LMR). Off-road rate marks an agent if any predicted waypoint leaves the valid lane polygons [10]; off-yaw rate measures predicted segments whose heading differs from the local lane direction by more than 45°, excluding intersections [8]; and LMR tests whether predicted and ground-truth lane positions exceed a speed-dependent graph-distance threshold [29]. LMR_1 evaluates the top-ranked mode, while LMR_6 accepts the best of six modes. All compliance metrics are percentages, and lower is better.

Formally, let $\mathcal{M}_i$ be the union of valid lane polygons around agent i and let $\hat{\mathbf{y}}_{i,t}^{(k)}$ be waypoint t of mode k. The per-agent off-road indicator is

$$\text{OR}_i = \mathbb{1}\left[\exists\, k, t : \hat{\mathbf{y}}_{i,t}^{(k)} \notin \mathcal{M}_i\right]. \tag{15}$$

This stringent definition penalizes a prediction set if any of its modes leaves the drivable area. For off-yaw, let $\mathcal{V}_i^{(k)}$ denote valid, non-intersection segments and $\Delta\theta_{i,t}^{(k)}$ the difference between predicted heading and local lane direction:

$$\text{OY}_i = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|\mathcal{V}_i^{(k)}|} \sum_{t \in \mathcal{V}_i^{(k)}} \mathbb{1}\left[|\Delta\theta_{i,t}^{(k)}| > 45^\circ\right]. \tag{16}$$

Table 1: AV2 test results. The flow teacher prioritizes predictive quality; the distilled variant provides the deployment operating point. Best results are bold.

Method	$\text{minFDE}_1\downarrow$	$\text{minADE}_1\downarrow$	$\text{minFDE}_6\downarrow$	$\text{minADE}_6\downarrow$	$\text{MR}_6\downarrow$	$\text{b-minFDE}_6\downarrow$
FRM [26]	5.93	2.37	1.81	0.89	0.29	2.47
HDGT [15]	5.37	2.08	1.60	0.84	0.21	2.24
MTR [30]	4.39	1.74	1.44	0.73	0.15	1.98
HPTR [46]	4.61	1.84	1.43	0.73	0.19	2.03
HeteroGCN [6]	4.40	1.72	1.34	0.69	0.18	1.90
ProphNet [35]	4.74	1.80	1.33	0.68	0.18	1.88
HiMAP [41]	4.61	1.81	1.33	0.68	0.17	1.96
QCNet [49]	4.30	1.69	1.29	0.65	0.16	1.91
CDT-P [40]	—	—	1.25	0.80	0.16	—
SmartRefine [48]	4.17	1.65	1.23	0.63	0.15	1.86
DeMo [44]	**3.74**	**1.49**	1.17	**0.61**	**0.13**	1.84
DONUT [18]	4.06	1.66	1.16	0.63	0.14	**1.79**
PLFlow Teacher (3-step)	3.97	1.53	**1.15**	**0.61**	0.14	1.81
PLFlow Teacher-μ (3-step)	4.31	1.67	1.26	0.64	0.16	1.91
PLFlow Distilled	4.18	1.63	1.24	0.63	0.15	1.88

Finally, if $m_i^{(k)} \in \{0, 1\}$ indicates a lane miss for mode k, then $\text{LMR}_1 = m_i^{(k_{\text{top}})}$ and $\text{LMR}_6 = \min_{k \le 6} m_i^{(k)}$. Together, the metrics distinguish global drivable-area validity, local directional compatibility, and coverage of the realized lane-level maneuver.

Implementation. The scene encoder and flow decoder use hidden dimension 128, eight attention heads of dimension 16, and dropout 0.1. The shared VAE uses latent dimension 32 and is trained for 64 epochs with batch size 48, AdamW, learning rate 5×10^{-4}, weight decay 10^{-4}, and a 2,000-step KL warm-up. The flow teacher and distilled student are each trained for 50 epochs with batch size 16, AdamW, learning rate 5×10^{-4}, weight decay 10^{-4}, and cosine annealing. The decoded trajectory-space loss has weight 1. Auxiliary layer supervision has total weight 0.3 with decay factor 0.8. Map-to-mode and agent-to-mode edges are independently dropped with probability 0.05 during training and retained at evaluation. During teacher training, $t \sim \mathcal{U}(0, 0.95)$, and the history-conditioned source is replaced by a standard Gaussian with probability 0.1. The final teacher uses three Euler steps and ten samples per head, pools all endpoints, and reclusters them into six output modes. The distilled predictor returns six modes in one pass.

4.2 Forecasting Accuracy

Table 1 shows that the stochastic three-step teacher achieves the best minFDE_6 (1.15) and ties the best minADE_6 (0.61). Its advantage over Teacher-μ confirms that sampling from the history-conditioned neighborhood provides useful hypothesis diversity. The distilled student increases minFDE_6 from 1.15 to 1.24, but remains competitive with strong fixed-K predictors and improves over the conditional diffusion baseline CDT-P in both minFDE_6 and minADE_6. The teacher and student therefore define complementary operating points: stochastic multi-step forecasting for accuracy and deterministic one-pass prediction for deployment.

Table 2: Controlled latency comparison on AV2. All methods use one RTX 3090, batch size 1, and the same scenario-level protocol.

Method	Latency↓	minFDE_6↓	MR_6↓
QCNet [49]	83.82 ms	1.29	0.16
PLFlow Teacher (3-step)	167.22 ms	**1.15**	**0.14**
PLFlow Distilled	**76.92 ms**	1.24	0.15

Table 3: AV1 validation results. Best results are bold.

Method	minFDE_6↓	minADE_6↓	MR_6↓
DenseTNT [9]	1.05	0.73	0.10
FRM [26]	0.99	0.68	—
HiVT [50]	0.96	0.66	0.09
HiMAP [41]	0.94	0.66	0.09
LAformer [23]	0.92	0.64	—
HPNet [34]	**0.87**	0.64	**0.07**
QCNet [49]	0.89	0.64	0.08
DeMo [44]	0.90	**0.59**	**0.07**
DGFNet [39]	0.89	0.63	—
PLFlow Teacher (3-step)	0.88	0.60	**0.07**
PLFlow Teacher-μ (3-step)	0.90	0.63	0.08
PLFlow Distilled	0.90	0.62	0.08

Table 2 reports controlled end-to-end latency on a single RTX 3090. The student runs in 76.92 ms, compared with 83.82 ms for QCNet and 167.22 ms for the three-step teacher, recovering most of the teacher's accuracy at less than half its latency.

Table 3 shows the same ordering on AV1: the stochastic teacher is competitive with specialized predictors and outperforms deterministic initialization. The smaller teacher–student gap is consistent with easier mode compression over AV1's shorter 3 s horizon.

4.3 Conditional Fidelity Beyond Displacement Accuracy

Conditional generation does not by itself guarantee map-consistent trajectories. We therefore measure structural fidelity over the complete prediction set and then isolate the contribution of state-dependent transport through a controlled endpoint-versus-vector-field comparison.

Table 4 evaluates adherence to the HD-map condition. Because off-road and off-yaw score all six modes, their improvements cannot be attributed solely to the GT-best trajectory. The three-step teacher performs best on all four metrics; relative to DeMo, it reduces off-road by 9.7%, off-yaw by 8.8%, LMR_1 by 12.9%, and LMR_6 by 7.4%. The distilled student also outperforms all regression baselines, showing that dense teacher supervision transfers much of the mode-level road structure without iterative inference. These gains are larger than the corresponding differences in minFDE_6, highlighting structural fidelity as PLFlow's principal empirical advantage.

Table 4: Road-geometry compliance on AV2 validation vehicles (%). The three-step teacher is best on all four measures; the distilled predictor preserves most gains in one pass.

Method	Off-road↓	Off-yaw↓	LMR_1↓	LMR_6↓
HiMAP [41]	6.44	0.75	61.01	22.39
QCNet [49]	6.00	0.71	59.77	21.61
DeMo [44]	5.46	0.68	59.70	20.61
PLFlow Teacher (3-step)	**4.93**	**0.62**	**51.99**	**19.08**
PLFlow Teacher-μ (3-step)	5.11	0.64	55.74	19.86
PLFlow Distilled	5.09	0.66	54.29	19.24

Table 5: Endpoint prediction versus state-dependent transport on AV2 validation. All variants use the same history-conditioned source, scene condition, and state-space augmentation. Road-geometry compliance metrics are percentages.

Target	Steps	$minFDE_6$	$minADE_6$	MR_6	Off-road	Off-yaw	LMR_1	LMR_6
Endpoint $\mathbf{x}_1$	1	1.25	0.73	0.16	5.37	0.67	57.27	20.56
Endpoint $\mathbf{x}_1$	3	1.27	0.73	0.17	5.34	0.67	57.74	20.32
Vector field $\mathbf{v}$	1	1.24	0.73	0.16	5.21	0.66	58.48	20.84
Vector field $\mathbf{v}$	3	**1.22**	**0.72**	**0.15**	**4.93**	**0.62**	**51.99**	**19.08**

4.4 State-Dependent Transport and Road-Geometry Compliance

Table 5 isolates the learning target while holding the source, scene context, augmentation, and architecture fixed. Repeating endpoint prediction from one to three evaluations leaves off-road and off-yaw nearly unchanged and yields no consistent LMR improvement. In contrast, three-step vector-field integration improves every metric over one-step integration: off-road decreases from 5.21% to 4.93%, off-yaw from 0.66% to 0.62%, LMR_1 from 58.48% to 51.99%, and LMR_6 from 20.84% to 19.08%. At each state, $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$ re-evaluates the current latent with the scene condition, allowing conditional redirection toward map-compatible regions. Because the endpoint control does not benefit from repeated evaluation, additional computation alone does not explain the gain. Within this controlled setting, the pattern provides evidence that state-dependent conditional transport materially contributes to map consistency. This does not mean that flow matching is intrinsically map-aware: map information enters only through $\mathbf{c}$, and the observed benefit requires both informative conditioning and repeated conditional refinement.

4.5 Forecasting-Specific Flow Adaptations

Source and state coverage. Table 6 separates the effects of source design and training-time state coverage. Mean-to-mean training performs well at one step but degrades with additional integration, indicating insufficient coverage of intermediate states. A standard-Gaussian source benefits from multiple steps but remains weaker because it discards the history anchor. With the shared deterministic evaluation anchor, the history-conditioned training source performs

Table 6: Source design and robustness to Euler integration on AV2 validation. All variants are evaluated from the deterministic history anchor $\boldsymbol{\mu}_x$ to isolate training-time coverage.

Training source	Steps	$\text{minFDE}_6\downarrow$	$\text{minADE}_6\downarrow$	$\text{MR}_6\downarrow$
Deterministic history mean	1	1.26	0.73	0.17
	2	1.28	0.73	0.17
	3	1.31	0.74	0.18
	4	1.34	0.75	0.18
Standard Gaussian	1	1.34	0.75	0.18
	2	1.30	0.74	0.17
	3	1.29	0.73	0.17
	4	1.29	0.73	0.17
History-conditioned prior	1	1.24	0.73	0.16
	2	1.24	0.73	0.16
	3	1.25	0.73	0.17
	4	1.26	0.73	0.17
History prior + augmentation	1	1.24	0.73	0.16
	2	1.23	0.73	**0.15**
	3	**1.22**	**0.72**	**0.15**
	4	1.23	**0.72**	**0.15**

best; augmentation enables further gains from additional evaluations. Their combination preserves motion-specific structure while covering the states required for robust iterative correction.

Table 7: Shared versus separate history/future VAEs on AV2 validation. Downstream metrics use one-step mean transport; compliance values are percentages.

VAE	Reconstruction		Forecasting		Map compliance			
	FDE↓	ADE↓	$\text{minFDE}_6\downarrow$	$\text{minADE}_6\downarrow$	Off-road↓	Off-yaw↓	$\text{LMR}_1\downarrow$	$\text{LMR}_6\downarrow$
Shared VAE	**0.064**	**0.031**	**1.24**	**0.73**	**5.19**	**0.66**	**57.29**	**19.92**
Separate VAEs	**0.064**	0.032	1.31	0.74	5.97	0.70	60.17	22.31

Aligned latent geometry. Table 7 compares a shared trajectory VAE with separately trained history and future VAEs. Despite nearly identical reconstruction error, the shared representation improves minFDE_6, minADE_6, and all four compliance metrics, including minFDE_6 from 1.31 to 1.24 and LMR_6 from 22.31% to 19.92%. The gain therefore reflects better transport geometry rather than decoder fidelity: a common coordinate system avoids the alignment burden introduced by independent latent spaces.

4.6 Qualitative Results

Fig. 3 shows that one-step predictions can deviate from lane curvature, whereas three-step transport better follows map geometry. The distilled student preserves the dominant mode structure without ODE integration.

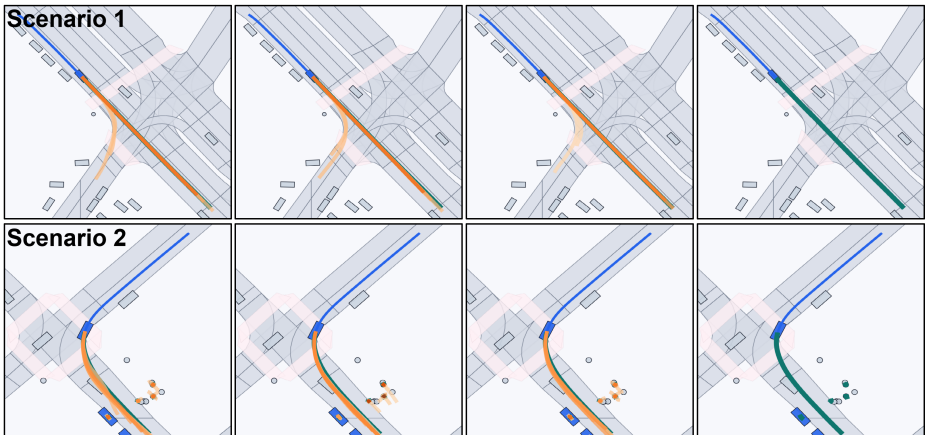

Fig. 3: Qualitative results. (A–C) show PLFlow predictions from teacher sampling with one-step ODE, teacher sampling with three-step ODE, and the distilled one-step student; (D) shows ground truth. History is blue, predicted futures are orange (darker indicates higher probability), and ground truth is green.

4.7 Limitations and Safety Scope

PLFlow has three principal limitations. Its VAE, flow teacher, and student form a multi-stage pipeline with greater engineering cost than direct fixed-K regression. The latent formulation represents an entire future as one implicit code, which may accumulate approximation error over long horizons. The student retains mode-level diversity but represents each teacher cluster by one center, so stochastic intra-cluster uncertainty and flow semantics reside in the teacher. A compatible extension is to predict a covariance for each student mode and supervise it from teacher-rollout covariance, but we do not evaluate that extension here. Finally, road-geometry compliance does not guarantee safe or defensive driving; we do not measure collision risk, closed-loop planning, or reactions to latent hazards. Our claim is therefore limited to improved fidelity to observable road geometry.

5 Conclusion

We presented PLFlow, which adapts conditional latent flow matching to forecasting through a history-induced source, shared motion manifold, robust state coverage, and fixed-K distillation. Its clearest benefit appears beyond displacement accuracy: on AV2, it consistently reduces off-road, off-yaw, and lane-miss rates relative to strong regression baselines. Additional vector-field evaluations improve road-geometry compliance whereas repeated endpoint prediction does not, providing controlled evidence for state-dependent transport as a major contributor in our setting.

References

1. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8748–8757 (2019)
2. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
3. Dao, Q., Phung, H., Dao, T.T., Metaxas, D.N., Tran, A.: Self-corrected flow distillation for consistent one-step and few-step image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2654–2662 (2025)
4. Fu, Y., Yan, Q., Wang, L., Li, K., Liao, R.: Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17282–17293 (2025)
5. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11525–11533 (2020)
6. Gao, X., Jia, X., Li, Y., Xiong, H.: Dynamic scenario representation learning for motion forecasting with heterogeneous graph convolutional recurrent networks. IEEE Robotics and Automation Letters **8**(5), 2946–2953 (2023)
7. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
8. Greer, R., Deo, N., Trivedi, M.: Trajectory prediction in autonomous driving with a lane heading auxiliary loss. IEEE Robotics and Automation Letters **6**(3), 4907–4914 (2021)
9. Gu, J., Sun, C., Zhao, H.: Densetnt: End-to-end trajectory prediction from dense goal sets. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15303–15312 (2021)
10. Hallgarten, M., Kisa, I., Stoll, M., Zell, A.: Stay on track: A frenet wrapper to overcome off-road trajectories in vehicle motion prediction. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 795–802. IEEE (2024)
11. Heek, J., Hoogeboom, E., Salimans, T.: Multistep consistency models. arXiv preprint arXiv:2403.06807 (2024)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Huang, Y., Du, J., Yang, Z., Zhou, Z., Zhang, L., Chen, H.: A survey on trajectory-prediction methods for autonomous driving. IEEE transactions on intelligent vehicles **7**(3), 652–674 (2022)
14. Huang, Y., Cheng, Y., Wang, K.: Trajectory mamba: Efficient attention-mamba forecasting model based on selective ssm. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12058–12067 (2025)
15. Jia, X., Wu, P., Chen, L., Liu, Y., Li, H., Yan, J.: Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. IEEE transactions on pattern analysis and machine intelligence **45**(11), 13860–13875 (2023)
16. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: 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)

17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)
18. Knoche, M., de Geus, D., Leibe, B.: Donut: A decoder-only model for trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28903–28912 (2025)
19. Lee, S., Lin, Z., Fanti, G.: Improving the training of rectified flows. Advances in neural information processing systems **37**, 63082–63109 (2024)
20. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., Urtasun, R.: Learning lane graph representations for motion forecasting. In: European Conference on Computer Vision. pp. 541–556. Springer (2020)
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
22. Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R.T., Lopez-Paz, D., Ben-Hamu, H., Gat, I.: Flow matching guide and code. arXiv preprint arXiv:2412.06264 (2024)
23. Liu, M., Cheng, H., Chen, L., Broszio, H., Li, J., Zhao, R., Sester, M., Yang, M.Y.: 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)
24. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
25. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
26. Park, D., Ryu, H., Yang, Y., Cho, J., Kim, J., Yoon, K.J.: Leveraging future relationship reasoning for vehicle trajectory prediction. In: The Eleventh International Conference on Learning Representations (2023)
27. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
28. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: European Conference on Computer Vision. pp. 683–700. Springer (2020)
29. Schmidt, J., Monninger, T., Jordan, J., Dietmayer, K.: Lmr: Lane distance-based metric for trajectory prediction. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–6. IEEE (2023)
30. Shi, S., Jiang, L., Dai, D., Schiele, B.: Motion transformer with global intention localization and local movement refinement. Advances in Neural Information Processing Systems **35**, 6531–6543 (2022)
31. Song, N., Zhang, B., Zhu, X., Zhang, L.: Motion forecasting in continuous driving. Advances in Neural Information Processing Systems **37**, 78147–78168 (2024)
32. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
33. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
34. Tang, X., Kan, M., Shan, S., Ji, Z., Bai, J., Chen, X.: Hpnet: Dynamic trajectory forecasting with historical prediction attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15261–15270 (2024)

35. Wang, X., Su, T., Da, F., Yang, X.: Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21995–22003 (2023)
36. Wang, Y., Tang, C., Sun, L., Rossi, S., Xie, Y., Peng, C., Hannagan, T., Sabatini, S., Poerio, N., Tomizuka, M., et al.: Optimizing diffusion models for joint trajectory prediction and controllable generation. In: European Conference on Computer Vision. pp. 324–341. Springer (2024)
37. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
38. Xiao, L., Pan, Z., Wang, Z., Cao, Z., Li, W.: Srefiner: Soft-braid attention for multi-agent trajectory refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 960–969 (2025)
39. Xin, G., Chu, D., Lu, L., Deng, Z., Lu, Y., Wu, X.: Multi-agent trajectory prediction with difficulty-guided feature enhancement network. IEEE Robotics and Automation Letters (2025)
40. Xu, Y., Cheng, H., Sester, M.: Controllable diverse sampling for diffusion based motion behavior forecasting. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 2397–2404. IEEE (2024)
41. Xu, Y., Yang, Y., Cheng, H., Sester, M.: Himap: History-aware map-occupancy prediction with fallback. arXiv preprint arXiv:2602.17231 (2026)
42. Yan, Q., Zhang, B., Zhang, Y., Yang, D., White, J., Chen, D., Liu, J., Liu, L., Zhuang, B., Shi, S., et al.: Trajflow: Multi-modal motion prediction via flow matching. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3923–3928. IEEE (2025)
43. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
44. Zhang, B., Song, N., Zhang, L.: Decoupling motion forecasting into directional intentions and dynamic states. Advances in Neural Information Processing Systems **37**, 106582–106606 (2024)
45. Zhang, H., Li, G., Guan, Z., Wu, J., Bu, S., Chai, J.: A spatio-temporal flow matching framework for pedestrian trajectory prediction. In: NeurIPS 2024 Workshop on Behavioral Machine Learning
46. Zhang, Z., Liniger, A., Sakaridis, C., Yu, F., Gool, L.V.: Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. Advances in Neural Information Processing Systems **36**, 57481–57499 (2023)
47. Zhao, H., Gao, J., Lan, T., Sun, C., Sapp, B., Varadarajan, B., Shen, Y., Shen, Y., Chai, Y., Schmid, C., et al.: Tnt: Target-driven trajectory prediction. In: Conference on robot learning. pp. 895–904. PMLR (2021)
48. Zhou, Y., Shao, H., Wang, L., Waslander, S.L., Li, H., Liu, Y.: Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15281–15290 (2024)
49. Zhou, Z., Wang, J., Li, Y.H., Huang, Y.K.: Query-centric trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17863–17873 (2023)
50. Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.: Hivt: Hierarchical vector transformer for multi-agent motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8823–8833 (2022)

51. Zhu, Y., Liu, X., Liu, Q.: Slimflow: Training smaller one-step diffusion models with rectified flow. In: European Conference on Computer Vision. pp. 342–359. Springer (2024)