EC-IGM: EXPLICIT EXPECTATION COMPENSATION THROUGH INFINITESIMAL GENERATOR MATCHING FOR DRIFT-FREE TRAJECTORY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose EC-IGM, a trajectory generation approach via matching the infinitesimal generator of a latent stochastic process with a deterministic kinematic vector field, aiming to address kinematic drift during long-term prediction. Conventional generation methods focus on matching the distribution of generated samples to real data. However, these methods overlook the expectation bias induced by stochastic noise during nonlinear mapping, which causes the generated trajectories to progressively deviate from actual vehicle kinematics over extended prediction horizons. In contrast, our EC-IGM leverages the infinitesimal generator to explicitly compensate for the expectation bias induced by stochastic noise in nonlinear mapping, thereby aligning the expected evolution of the latent stochastic process with the deterministic kinematic vector field. This alignment ensures consistency between the generated trajectories and actual vehicle kinematics during long-term prediction. Furthermore, to circumvent the extreme local curvature introduced by modeling non-linear mappings, we decouple kinematic uncertainty from multimodal behavioral intentions by transforming the complex joint distribution into an independent modeling of stochastic kinematic noise, achieving sampling-free stochastic kinematic evolution with $\mathcal{O}(KL^3)$ computational complexity. Compared with state-of-the-art approaches, the EC-IGM achieves a kinematic compliance score of 98.9 during long-term prediction, representing a 12% improvement. Extensive experiments on the NAVSIM and Bench2Drive benchmarks demonstrate that our approach achieves state-of-the-art performance, setting new records with 93.2 PDMS and 92.28 DrivingScore, while significantly outperforming existing methods in maintaining kinematic consistency during long-term prediction.

1 INTRODUCTION

In end-to-end autonomous driving Chen et al. (2024b); Zhang et al. (2022), generative models Rowe et al. (2025); Wu et al. (2025); Zheng et al. (2025b); Hu et al. (2025); Wei et al. (2024) are a core paradigm for mapping high-dimensional perception to planning Zheng et al. (2024a); Xing et al. (2025); Yang et al. (2024); Park et al. (2025). Within this framework, generating trajectories that strictly conform to actual vehicle kinematic properties is fundamental to precise planning Hu et al. (2026); Wang et al. (2024b). However, existing methods exhibit significant limitations in maintaining consistency between trajectories and vehicle kinematics during long-term prediction. This issue arises from relying on sampling to approximate distribution evolution under nonlinear mapping, which inevitably introduces expectation bias induced by stochastic noise and results in pronounced kinematic drift. Specifically, these methods overlook the expectation bias accumulation during long-term prediction, implicitly assuming that sampled distribution expectations spontaneously conform to the actual vehicle kinematics. In reality, according to Jensen's inequality $E[f(x)] \geq f(E[x])$ Särkkä & Svensson (2013), nonlinear mapping disrupts the symmetry of the stochastic noise distribution, causing the transformed expectation to deviate from the original distribution expectation. The expectation bias arising from the coupling of nonlinear curvature and state variance is not corrected by explicit compensation. Therefore, the expected trajectory gradually drifts away from the actual vehicle kinematics during long-term prediction. Furthermore, existing methods couple kinematic uncertainty with multimodal intentions, requiring extensive sampling at high computational cost to approximate the true distribution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

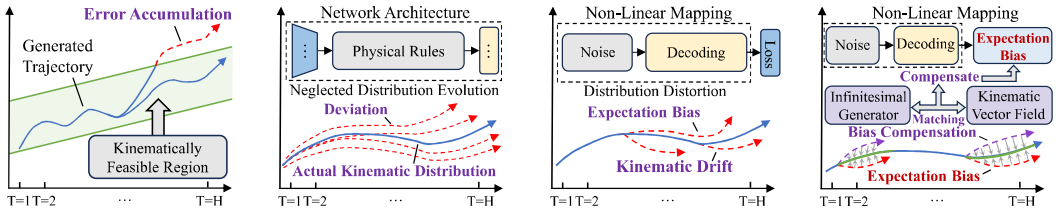

Figure 1: Comparison of different constraint paradigms in the generation process. The proposed EC-IGM leverages the infinitesimal generator to explicitly match the deterministic kinematic vector field, which compensates the expectation bias induced by stochastic noise under nonlinear mappings.

Existing generative approaches for kinematic consistency are categorized into post-projection methods Djeumou et al. (2024); Christopher et al. (2024), hard-constraint embedding methods Wang et al. (2025); Ma et al. (2024), and soft-regularization guidance methods Zheng et al. (2024b); Xiao et al. (2025). Post-projection methods force trajectories into kinematically feasible regions via mathematical projections (Fig. 1(a)). However, this paradigm fails to suppress error accumulation from nonlinear mapping over long horizons, causing progressive kinematic deviation. Hard-constraint embedding methods integrate physical rules into architectures or the inference process (Fig. 1(b)). This paradigm focuses solely on constraint satisfaction while neglecting distribution evolution, still resulting in kinematic drift in long-term predictions. Soft-regularization guidance methods implicitly guide the kinematic compliance of the distribution through physical penalty terms (Fig. 1(c)). The inherent flaw of this paradigm is the failure to explicitly correct expectation bias induced by stochastic noise during nonlinear transformation, leading to kinematic drift in long-term predictions.

These observations motivate us to propose an approach to resolve kinematic drift in long-term prediction. Accordingly, we propose EC-IGM, a novel paradigm that explicitly aligns the infinitesimal generator of the latent stochastic process with a deterministic kinematic vector field. Unlike sampling-based approximations, our approach leverages the infinitesimal generator to explicitly compensate for expectation bias induced by stochastic noise in nonlinear mappings. This ensures the expected evolution of the distribution after nonlinear decoding conforms to actual vehicle kinematics (Fig. 1(d)). Furthermore, to circumvent the extreme local curvature introduced by modeling non-linear mappings, we decouple kinematic uncertainty from driving intentions by transforming the complex joint trajectory distribution into independent modeling of stochastic kinematic noise, achieving sampling-free stochastic kinematic evolution with $\mathcal{O}(KL^3)$ computational complexity.

We validate the performance of our method on the Navsim and Bench2Drive datasets, achieving 93.2 PDMS and 92.28 DrivingScore, respectively. Crucially, we focused on evaluating the consistency between trajectories and actual vehicle kinematics over long-term prediction. Compared with state-of-the-art methods, the kinematic compliance score of our EC-IGM reaches 98.9, improving upon the baseline by 12%. Notably, even across the extended 8-second prediction horizon, our method strictly bounds the expectation bias of actual vehicle kinematics. These findings confirm that our EC-IGM effectively resolves the kinematic drift issue in long-term predictions.

2 Related Work

End-to-end Autonomous Driving. Trajectory generation methods for end-to-end autonomous driving have evolved from early deterministic regression Wu et al. (2022); Prakash et al. (2021) toward stochastic generative paradigms Wang et al. (2024a); Liao et al. (2025). The discrete query paradigm Hu et al. (2023) enhances the ability to model uncertainty. However, approximating continuous distributions with finite queries neglects cumulative stochastic perturbations, inevitably causing trajectories to deviate from actual vehicle kinematic boundaries. To address this, continuous approaches like diffusion models capture the probability distribution of trajectories Mao et al. (2023). Although this paradigm excels at representing trajectory diversity, the gradually accumulated errors during the iterative sampling process still lead to kinematic drift. Flow matching paradigms Fu et al. (2025) reduce sampling iterations via deterministic probability flows, but fundamentally overlook the expectation bias induced by stochastic noise under nonlinear mappings, leaving kinematic drift

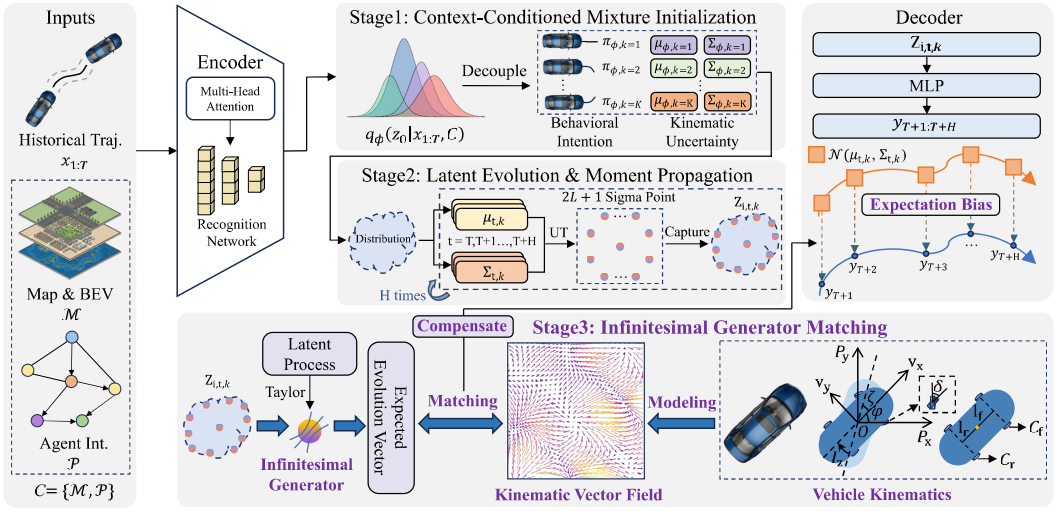

Figure 2: The EC-IGM encodes scene inputs into feature sequences. Stage 1 decomposes the variational posterior $q_\phi(\mathbf{z}_0 \mid \mathbf{x}_{1:T}, \mathcal{C})$ to explicitly decouple kinematic uncertainty from behavioral intention. Stage 2 employs $2L+1$ deterministic sigma points for nonlinear moment propagation, extracting the predicted mean $\dot{\mu}$ and covariance $\dot{\Sigma}$ sequences. Stage 3 compensates for the expectation bias by matching the latent infinitesimal generator with the deterministic kinematic vector field.

unresolved in long-term predictions. By compensating for this nonlinear expectation bias, our approach ensures the expected stochastic evolution strictly adheres to actual vehicle kinematics.

Physics-constrained Guidance Strategies. Early post-projection techniques forcibly corrected trajectories at inference Cheng et al. (2019). However, this paradigm disrupts original distributions and causes mode collapse. To mitigate this, Chen *et al.* Chen et al. (2024a) introduced implicit regularization via physical loss penalties, yet their approach suffers from kinematic expectation bias due to nonlinear mappings. Inspired by classifier guidance, Zheng *et al.* Yao et al. (2026) explored real-time correction of sampled trajectories based on the gradients of physics constraint functions. Nevertheless, this paradigm fails to address the expectation bias arising from the coupling of stochastic noise and nonlinear kinematics, leaving severe kinematic drift unresolved. To overcome these limitations, we achieve precise alignment between the infinitesimal generator and the deterministic kinematic vector field during evolution, explicitly correcting the kinematic expectation bias caused by stochastic noise in nonlinear mappings.

Stochastic Differential Equations. Stochastic differential equations (SDEs) have become a powerful framework for modeling complex data distributions. Song *et al.* Song et al. (2021) pioneered a diffusion process paradigm based on forward and reverse SDEs, which has since been widely adopted in autonomous driving. The DiffusionPlanner Zheng et al. (2025a) formulated the trajectory generation process as SDE path sampling. However, their iterative sampling relies on first-order approximations, where nonlinear decoding inevitably induces expectation bias that violates actual vehicle kinematics. While recent works inject physical priors into the SDE drift term White et al. (2023); Park (2025), they treat constraints as mere external perturbations, failing to address the expectation bias of Itô correction terms under nonlinear mappings Xia et al. (2025). To resolve this, we leverage the infinitesimal generator for second-order expectation compensation, ensuring the expected evolution strictly adheres to vehicle kinematics across long-term predictions.

3 METHOD

We formulate trajectory generation as stochastic generator matching rather than matching sampled trajectories. The infinitesimal generator of the learned latent stochastic kinematics is aligned with a deterministic kinematic vector field. Conditioned on historical trajectories $\mathbf{x}_{1:T}$, map context $\mathcal{M}$ (BEV rasterizations), and interactions $\mathcal{P}$, our framework evolves a latent state $\mathbf{z}_t$ via a physics-constrained SDE, decoded into future states $\mathbf{y}_{T+1:T+H}$. The overall framework of our framework is

depicted in Fig. 2. Alignment is achieved through unscented moment propagation and Itô calculus-based Oksendal (2013) kinematic bias correction. This pipeline enables fully differentiable training with $\mathcal{O}(KL^3)$ complexity. The latent dimension $L \in [8, 32]$ balances capacity with constraints, while $K = 6$ initial states capture distinct behavioral modes.

3.1 Decoupled Initialization of Intent and Uncertainty

The latent initial state $\mathbf{z}_0$ captures multimodal intent anchored to scene topology $\mathcal{C} = \{\mathcal{M}, \mathcal{P}\}$. We parameterize the variational posterior $q_\phi(\mathbf{z}_0 \mid \mathbf{x}_{1:T+H}, \mathcal{C})$ as a Gaussian mixture over $k \in \{1, \dots, K\}$, where each component k corresponds to the latent initial state of a distinct behavioral mode:

$$q_\phi(\mathbf{z}_0 \mid \mathbf{x}_{1:T+H}, \mathcal{C}) = \sum_{k=1}^{K} \pi_{\phi,k} \mathcal{N}(\mu_{\phi,k}, \Sigma_{\phi,k}), \tag{1}$$

where the mixing weights $\pi_{\phi,k}$ represent the categorical probabilities, and each component $\mathcal{N}(\mu_{\phi,k}, \Sigma_{\phi,k})$ characterizes mode-specific kinematic uncertainty. This formulation decouples intent-level multimodality from local kinematic uncertainty, avoiding highly nonlinear drift terms to separate modes. Multimodality arises from intent-conditioned components evolved under scene-aware SDE kinematics. To obtain these parameters, a recognition network processes historical trajectories. It employs a cross-attention mechanism to fuse BEV features from $\mathcal{M}$ with graph-based interaction embeddings from $\mathcal{P}$. The encoder applies multi-head attention over tokenized BEV grids and agent graphs to extract and encode kinematic features into low-dimensional latent mixture parameters.

3.2 Deterministic Propagation of Latent Kinematic Moments

We model the latent kinematics for each mode k as an Itô SDE Oksendal (2013). This enables the continuous-time evolution of the multimodal posterior:

$$d\mathbf{z}_{t,k} = \mathbf{f}_\theta(\mathbf{z}_{t,k}, t, \mathcal{C})\, dt + \sigma_\theta(\mathbf{z}_{t,k}, t, \mathcal{C})\, d\mathbf{W}_t, \tag{2}$$

where $\mathbf{f}_\theta$ is a lightweight MLP parameterizing the drift, conditioned on the scene context $\mathcal{C}$. The diffusion matrix σ_θ captures multi-agent correlations, while $\mathbf{W}_t$ denotes a shared Wiener process modeling shared environmental stochasticity. The modes differentiate through their distinct initial conditions and drift terms as they evolve under physics-generator alignment.

We employ the Unscented Transform (UT) Julier & Uhlmann (2004) as a deterministic approximation to the moment evolution induced by the latent SDE. This enables deterministic and parallel evolution of per-mode moments $\mu_{t,k}$ and $\Sigma_{t,k}$ with second-order accuracy for the nonlinear drift $\mathbf{f}_\theta$, while preserving the $\mathcal{O}(KL^3)$ complexity. At any time t, the $2L + 1$ sigma points $\mathcal{Z}_{i,t,k}$ are constructed deterministically from the instantaneous Gaussian $\mathcal{N}(\mu_{t,k}, \Sigma_{t,k})$ of dimension L:

$$\mathcal{Z}_{0,t,k} = \mu_{t,k}, \tag{3}$$

$$\mathcal{Z}_{i,t,k} = \mu_{t,k} + \left(\sqrt{(L+\lambda)\Sigma_{t,k}}\right)_i, \quad i = 1, \dots, L, \tag{4}$$

$$\mathcal{Z}_{i+L,t,k} = \mu_{t,k} - \left(\sqrt{(L+\lambda)\Sigma_{t,k}}\right)_i, \quad i = 1, \dots, L. \tag{5}$$

The point $\mathcal{Z}_{0,t,k}$ is the mean, while $\mathcal{Z}_{i,t,k}$ and $\mathcal{Z}_{i+L,t,k}$ are symmetric pairs obtained from the matrix square root of the covariance. These points capture the uncertainty in the latent space. The term $\sqrt{(L+\lambda)\Sigma_{t,k}}$ is the matrix square root, and $\lambda = \alpha^2(L + \kappa) - L$ is the standard UT scaling factor parameterized by α and κ. The weights for the mean $\omega^{(\mu)}$ and covariance $\omega^{(\Sigma)}$ are defined as:

$$\omega_0^{(\mu)} = \frac{\lambda}{L+\lambda}, \quad \omega_0^{(\Sigma)} = \frac{\lambda}{L+\lambda} + (1 - \alpha^2 + \beta), \quad \omega_i^{(\mu)} = \omega_i^{(\Sigma)} = \frac{1}{2(L+\lambda)} \quad (i \geq 1), \tag{6}$$

with $\beta = 2$ typically used for optimal Gaussian approximation.

Using these weighted sigma points and the deviations $\Delta_{i,t,k} = \mathcal{Z}_{i,t,k} - \mu_{t,k}$, the continuous-time moment dynamics are:

$$\dot{\mu}_{t,k} = \sum_{i=0}^{2L} \omega_i^{(\mu)} \mathbf{f}_\theta(\mathcal{Z}_{i,t,k}), \tag{7}$$

$$\dot{\Sigma}_{t,k} = \sum_{i=0}^{2L} \omega_i^{(\Sigma)} \left[\Delta_{i,t,k} \mathbf{f}_\theta(\mathcal{Z}_{i,t,k})^\top + \mathbf{f}_\theta(\mathcal{Z}_{i,t,k}) \Delta_{i,t,k}^\top \right] + \sum_{i=0}^{2L} \omega_i^{(\Sigma)} \sigma_\theta(\mathcal{Z}_{i,t,k}) \sigma_\theta(\mathcal{Z}_{i,t,k})^\top. \quad (8)$$

The diffusion term aligns with the Fokker-Planck equation Frank (2005), capturing Wiener process uncertainty. Backpropagation utilizes discrete-time integration via automatic differentiation, maintaining $\mathcal{O}(KL^3)$ reverse complexity.

3.3 Infinitesimal Generator Matching

To quantify the expectation bias inherent in the nonlinear mapping, we perform a second-order Taylor expansion of the decoder function $g(\mathbf{z}_{t,k})$ with respect to the latent stochastic process $\mathbf{z}_{t,k} \sim \mathcal{N}(\mu_{t,k}, \Sigma_{t,k})$ via Itô's Lemma Oksendal (2013). For a latent process $d\mathbf{z}_t = \mathbf{f}_\theta \, dt + \sigma_\theta \, d\mathbf{W}_t$, the infinitesimal evolution of the observable state $\mathbf{y}_t = g(\mathbf{z}_t)$ is defined as:

$$dg^{(j)}(\mathbf{z}_{t,k}) = \left(\nabla_{\mathbf{z}} g^{(j)\top} \mathbf{f}_\theta + \frac{1}{2} \sum_{d_1,d_2=1}^{L} \left(\sigma_\theta \sigma_\theta^\top \right)_{d_1 d_2} \frac{\partial^2 g^{(j)}}{\partial z_{d_1} \partial z_{d_2}} \right) dt + \nabla_{\mathbf{z}} g^{(j)\top} \sigma_\theta \, d\mathbf{W}_t. \quad (9)$$

Consequently, the observable drift is not simply the push-forward of the latent drift $\nabla_{\mathbf{z}} g^\top \mathbf{f}_\theta$. The curvature of the decoder, captured by its Hessian, interacts with the stochastic diffusion $\sigma_\theta \sigma_\theta^\top$ to introduce a systematic bias. Unlike deterministic calculus, the quadratic variation of the Wiener process $(d\mathbf{W}_t)^2 = dt$ promotes this second-order term to a first-order contribution in the drift.

Physical constraints are enforced in the observable state space by aligning the stochastic generator with the deterministic vehicle kinematics vector field, ensuring that decoded trajectories satisfy vehicle kinematics regardless of the latent representation.

In stochastic calculus, the infinitesimal generator $\mathcal{A}$ characterizes the expected time evolution of any observable under latent stochasticity. To account for the second-order stochastic effects, we use Itô calculus to define $\mathcal{A}$ and correct the expectation bias for each output component j and mode k:

$$\mathcal{A}[g^{(j)}(\mathbf{z}_{t,k})] = \nabla_{\mathbf{z}} g^{(j)}(\mathbf{z}_{t,k})^\top \mathbf{f}_\theta(\mathbf{z}_{t,k}) + \frac{1}{2} \sum_{d_1,d_2=1}^{L} \left(\sigma_\theta(\mathbf{z}_{t,k}) \sigma_\theta(\mathbf{z}_{t,k})^\top \right)_{d_1 d_2} \frac{\partial^2 g^{(j)}(\mathbf{z}_{t,k})}{\partial z_{d_1} \partial z_{d_2}}, \quad (10)$$

computed via vector-Jacobian products in automatic differentiation at $\mathcal{O}(L^2)$ cost per component. The decoder g is implemented as a shallow MLP to ensure stable second-order differentiation.

We align the Itô-corrected drift with the deterministic kinematic vector field by predicting mode-specific controls $\mathbf{u}_{t,k} = [a, \delta]^\top$ via an independent head $h_\psi(\mathbf{z}_{t,k}, \mathcal{C})$. This ensures that the expected evolution of the decoded states strictly follows physically valid trajectories. The soft constraint jointly optimizes g and h_ψ through gradients of the loss, balancing physical feasibility and data-driven fitting without compounding errors from direct model integration.

The system parameters are extracted from scene metadata or estimated via an auxiliary network. We adopt a standard right-handed body frame. The observable state vector is $\mathbf{y}_{t,k} = [p_x, p_y, v_x, v_y, \varphi, \zeta]^\top$, where p_x and p_y are the position coordinates in the world frame, v_x and v_y are the longitudinal and lateral velocity components, φ is the heading angle and ζ is the yaw rate. Let $\mathcal{F}_{\text{veh}}(\mathbf{y}_{t,k}, \mathbf{u}_{t,k})$ denote the right-hand side of the deterministic kinematic vector field governing vehicle motion Kong et al. (2015). To ensure numerical stability, this formulation employs a regularized longitudinal velocity $\tilde{v}_x = \max(|v_x|, \epsilon)$ for tire force projections and incorporates a kinematic fallback for low-speed regimes where $|\mathbf{v}| < v_{\min}$, with $\mathbf{v} = [v_x, v_y]^\top$. Here, m is the vehicle mass, I_z is the yaw moment of inertia about the vertical axis, C_f and C_r are the front and rear tire cornering stiffnesses, and l_f and l_r are the distances from the center of gravity to the front and rear axles, respectively:

$$\dot{\mathbf{y}}_{t,k} = \begin{bmatrix} \dot{p_x} \\ \dot{p_y} \\ \dot{v_x} \\ \dot{v_y} \\ \dot{\varphi} \\ \dot{\zeta} \end{bmatrix} = \begin{bmatrix} v_x \cos\varphi - v_y \sin\varphi \\ v_x \sin\varphi + v_y \cos\varphi \\ a + v_y \zeta - \frac{C_f}{m} \left(\frac{v_y + l_f \zeta}{\tilde{v}_x} - \delta \right) \sin\delta \\ -v_x \zeta + \frac{1}{m} \left[C_f \left(\frac{v_y + l_f \zeta}{\tilde{v}_x} - \delta \right) \cos\delta + C_r \left(\frac{v_y - l_r \zeta}{\tilde{v}_x} \right) \right] \\ \zeta \\ \frac{1}{I_z} \left[l_f C_f \left(\frac{v_y + l_f \zeta}{\tilde{v}_x} - \delta \right) \cos\delta - l_r C_r \left(\frac{v_y - l_r \zeta}{\tilde{v}_x} \right) \right] \end{bmatrix}. \quad (11)$$

Table 1: Comparison on NAVSIM navtest split. All methods employ ResNet-34 backbone.

Method	**KC↑**	**Latency↑**	NC↑	DAC↑	TTC↑	Comf↑	EP↑	**PDMS↑**
UniAD Hu et al. (2023)	86.2	558 ms	97.8	91.9	92.9	**100**	78.8	83.4
PARA-Drive Weng et al. (2024)	87.5	137 ms	97.9	92.4	93.0	<u>99.8</u>	79.3	84.0
LTF Chitta et al. (2023)	85.1	152 ms	97.4	92.8	92.4	**100**	79.0	83.8
Transfuser Chitta et al. (2023)	85.9	108 ms	97.7	92.8	92.8	**100**	79.2	84.0
DRAMA Yuan et al. (2024)	88.4	116 ms	98.0	93.1	<u>94.8</u>	**100**	80.1	85.5
VAD Jiang et al. (2023)	90.1	302 ms	97.2	89.1	91.6	**100**	76.0	80.9
Hydra-MDP-V_{8192} Li et al. (2024b)	91.5	274 ms	98.3	96.0	94.6	**100**	78.7	86.5
DiffusionDrive Liao et al. (2025)	92.8	45 ms	98.2	96.2	94.7	**100**	82.2	88.1
GoalFlow Xing et al. (2025)	94.2	<u>32 ms</u>	98.4	<u>98.3</u>	94.6	**100**	85.0	90.3
Diffusion-Phy Hu et al. (2026)	<u>95.1</u>	104 ms	<u>98.6</u>	<u>98.3</u>	94.7	**100**	<u>87.8</u>	<u>91.5</u>
EC-IGM	**98.6**	**24 ms**	**98.9**	**98.8**	**95.5**	**100**	**88.2**	**93.2**

The generator-matching loss minimizes the expected discrepancy between $\mathcal{A}$ and $\mathcal{F}_{\text{veh}}$ under the latent distribution $q_\phi(\mathbf{z}_t)$, approximated deterministically via the Unscented Transform over collocation points t_c obtained from adaptive Runge-Kutta integration and summed over modes:

$$\mathcal{L}_{\text{gen}} = \sum_k \pi_k \sum_{t_c} \sum_{i=0}^{2L} \omega_i^{(\mu)} \left\| \mathcal{A}[g(\mathcal{Z}_{i,t_c,k})] - \mathcal{F}_{\text{veh}}(g(\mathcal{Z}_{i,t_c,k}), h_\psi(\mathcal{Z}_{i,t_c,k}, \mathcal{C})) \right\|_2^2 . \tag{12}$$

Compared with discrete-time predictors, our continuous-time alignment ensures that heteroscedastic uncertainty stays kinematically grounded across the entire prediction horizon. Alignment at individual instants approximates global consistency with an error of $\mathcal{O}(\Delta t^2)$ per Itô expansion.

3.4 Kinematics-Aligned Training Objective

Our framework optimizes a composite objective end-to-end:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \gamma D_{\text{KL}}(q_\phi(\mathbf{z}_0 \mid \mathbf{x}_{1:T+H}, \mathcal{C} \parallel p(\mathbf{z}_0 \mid \mathcal{C})) + \rho_{\text{gen}} \mathcal{L}_{\text{gen}}, \tag{13}$$

where the prior $p(\mathbf{z}_0 \mid \mathcal{C})$ is a neural conditional GMM that matches the structure of q_ϕ and is generated by a context encoder. The coefficient γ promotes modal diversity and prevents posterior collapse, while ρ_{gen} anchors latent geometry to kinematics. The reconstruction loss employs winner-takes-all (WTA) over the best-matching mode $\hat{k}$, with soft-WTA in early training to mitigate dead modes. The best-matching mode $\hat{k}$ is identified by minimizing the Mahalanobis distance between the ground-truth trajectory and the UT-propagated Gaussian moments $(\mu^y_{t,\hat{k}}, \Sigma^y_{t,\hat{k}})$. The reconstruction term is approximated via UT-propagated moments as Gaussian log-likelihoods in observation space:

$$\mathcal{L}_{\text{recon}} = -\log \pi_{\hat{k}} - \sum_{t=T+1}^{T+H} \log \mathcal{N}(\mathbf{x}_t \mid \mu^y_{t,\hat{k}}, \Sigma^y_{t,\hat{k}} + \Sigma_{\text{obs}}), \tag{14}$$

where $\mu^y_{t,\hat{k}} = \sum_i \omega_i^{(\mu)} g(\mathcal{Z}_{i,t,\hat{k}})$ and $\Sigma^y_{t,\hat{k}} = \sum_i \omega_i^{(\Sigma)} [g(\mathcal{Z}_{i,t,\hat{k}}) - \mu^y_{t,\hat{k}}][g(\mathcal{Z}_{i,t,\hat{k}}) - \mu^y_{t,\hat{k}}]^\top$ are the decoded mean and covariance from sigma points, together with observation noise Σ_{obs}.

4 Experiments

4.1 Datasets and Metrics

Datasets. EC-IGM was evaluated on the NAVSIM Dauner et al. (2024) and Bench2Drive Jia et al. (2024) datasets. For NAVSIM, derived from the nuPlan Karnchanachari et al. (2024) dataset to emphasize dynamic intent changes, the official *navtrain* (103k) and *navtest* (12k) splits were utilized. For Bench2Drive, constructed from the real-world data of Think2Drive Li et al. (2024a), the model was trained on 10k segments and assessed via the official benchmark, utilizing the CARLA Dosovitskiy et al. (2017) simulator to replicate real-world scenarios.

Table 2: Closed-loop results on Bench2Drive. Latency is average inference time on RTX 4090.

Method	Latency↓	KC↑			Avg. L2↓	Closed-loop↑				Ability (%) ↑					
		3s	5s	8s		Eff	Comf	SR	**DS**	MG	OT	EB	GW	TS	**Mean**
PDM-Closed Dauner et al. (2023)	**4 ms**	35.2	24.7	15.6	3.52	48.37	21.98	0.00	18.26	0.00	0.00	0.00	0.00	4.49	0.89
UniAD-Tiny Hu et al. (2023)	445 ms	88.5	82.7	71.2	0.80	123.92	**47.04**	13.18	40.73	8.89	9.33	20.00	20.00	15.43	14.73
UniAD-Base Hu et al. (2023)	558 ms	90.1	85.4	75.8	0.73	129.21	43.58	16.36	45.81	14.10	17.78	21.67	10.00	14.21	15.55
VAD Jiang et al. (2023)	359 ms	89.2	83.3	72.4	0.91	157.94	46.01	15.00	42.35	8.11	24.44	18.64	20.00	19.15	18.07
DriveTransformer Jia et al. (2025)	212 ms	96.5	93.7	86.1	0.62	100.64	20.78	35.01	63.46	17.57	35.00	48.36	40.00	52.10	38.60
Diffusion-Phy Hu et al. (2026)	39 ms	98.5	95.2	92.7	0.92	162.74	30.41	36.51	67.49	32.40	21.25	54.19	62.73	50.47	44.65
Expert-VADv2 Jiang et al. (2026)	258 ms	92.3	87.5	78.4	0.58	168.45	38.22	58.34	78.18	40.44	48.33	78.42	40.00	65.78	54.59
GaussianFusion Liu et al. (2025)	32 ms	93.8	89.1	82.6	0.78	178.62	44.25	54.40	79.10	36.60	64.40	66.50	53.3	60.80	56.30
SimLingo Renz et al. (2025)	215 ms	91.5	82.3	71.8	0.73	**259.23**	33.67	67.27	85.07	54.01	57.04	88.33	53.33	82.45	67.03
R2SE Liu et al. (2026a)	125 ms	88.4	80.2	68.5	0.92	243.89	23.26	69.54	86.28	53.33	61.25	90.00	50.00	84.21	67.76
HiP-AD Tang et al. (2025)	139 ms	91.2	85.6	76.9	0.69	203.12	19.36	69.09	86.77	50.00	84.44	83.33	40.00	72.10	65.98
BridgeDrive Liu et al. (2026b)	72 ms	95.8	92.1	84.3	0.48	236.49	20.98	**74.99**	87.99	69.92	66.67	90.00	50.00	89.47	73.15
EC-IGM	24 ms	**99.7**	**99.4**	**98.9**	**0.41**	256.47	35.21	74.79	**92.28**	**81.33**	**85.12**	**94.21**	**66.46**	**90.33**	**83.49**

Evaluation Metrics. For NAVSIM, we adopted the PDMS Dauner et al. (2024), aggregating sub-metrics: NC, DAC, TTC, Comf, and EP. For Bench2Drive, four closed-loop metrics were used: Eff, Comf, SR, and DS. Furthermore, the evaluation encompasses five driving skills: MG, OT, EB, GW, and TS. Full names are listed in the appendix. In addition, we introduce **Kinematic Compliance (KC)** based on Bansal *et al.*'s method Bansal et al. (2019), measuring the proportion of trajectories within a model-agnostic feasible space $\mathcal{S}_{\text{phys}}$ over horizon H. While our framework optimizes $\mathcal{F}_{\text{veh}}$, $\mathcal{S}_{\text{phys}}$ serves as a broad envelope derived from real-world traffic statistics. We define $\mathcal{S}_{\text{phys}}$ by empirical bounds: acceleration $|a| \leq 8$ m/s^2, jerk ≤ 10 m/s^3, and curvature $\kappa \leq 0.2$ m^{-1}:

$$\text{KC} = \frac{1}{N}\sum_{i=1}^{N}[\forall t \in [T, T+H], \mathbf{y}_{i,t} \in \mathcal{S}_{\text{phys}}] \times 100 \tag{15}$$

4.2 Implementation Details

All models employ a ResNet-34 backbone. We trained the EC-IGM from scratch on two RTX A6000 GPUs using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 5×10^{-4}), a cosine annealing schedule, and a 1000-step linear warmup. NAVSIM training required 100 epochs, a batch size of 512, and a learning rate of 6×10^{-4}, whereas Bench2Drive used 30 epochs, a batch size of 64, and a learning rate of 1×10^{-4}. We set the latent SDE dimension to $L = 16$ and configured the unscented transform with $\alpha = 1 \times 10^{-3}$, $\beta = 2$, and $\kappa = 0$. Dataset metadata directly provides vehicle physical parameters.

4.3 Quantitative Analysis

Comparisons on NAVSIM. As shown in Table 1, our EC-IGM achieves 93.2 PDMS and 98.6 KC, outperforming state-of-the-art baselines on NAVSIM. It outperforms GoalFlow by 9.2 PDMS and 4.4 KC, since explicit expectation bias compensation ensures the generated trajectories align with actual vehicle kinematics. Relative to the DiffusionDrive, which accelerates inference via truncated two-step denoising, our EC-IGM attains a superior latency of 24 ms. This efficiency derives from one-pass deterministic moment propagation via UT, thereby eliminating multi-step sampling. Compared to Diffusion-Phy, which constrains output sample distributions via first-order physical guidance, our method compensates for biases induced by second-order terms, yielding a 3.5 KC improvement.

Comparisons on Bench2Drive. Table 2 shows that our EC-IGM achieves state-of-the-art Success Rate and Driving Score. Our model achieves the second-lowest latency, surpassed only by PDM-Closed, which lacks complex perception processing. Additionally, incorporating a deterministic kinematic vector field derived from actual kinematic evolution secures 98.9 KC. Our EC-IGM also attains the highest mean performance across five driving skills and Driving Score, demonstrating strong interaction-aware capabilities in challenging scenarios.

Kinematic expectation bias in long-term prediction. As shown in Fig. 3(a), while standard models develop kinematic deviations exceeding 2 meters over 8 seconds, our method strictly bounds this

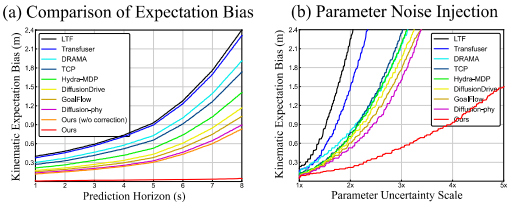

Figure 3: Kinematic Expectation Bias.

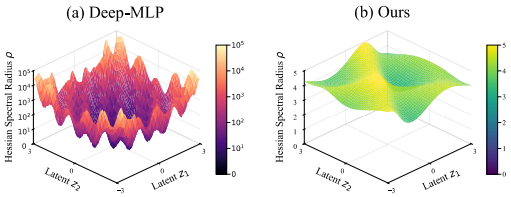

Figure 4: Latent Space Topology.

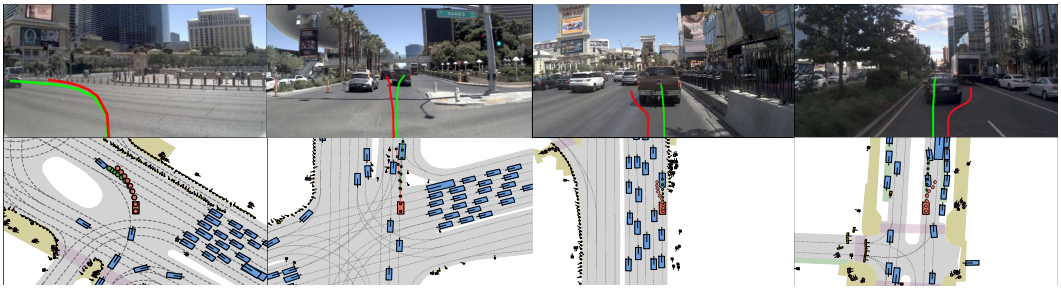

Figure 5: Qualitative results of generated trajectories on the NAVSIM dataset.

deviation within 0.1 meters. Without compensation, our model exhibits substantial drift similar to traditional approaches. This confirms that explicit deviation compensation via the infinitesimal generator is the primary mechanism for achieving long-term kinematic consistency.

Uncertainties of Physical Parameters. We evaluated system robustness against parameter uncertainty in Fig. 3(b). Injecting scaled noise up to $5\times$ into kinematic parameters causes an exponential collapse in the kinematic compliance of baselines. In contrast, our EC-IGM maintains a significantly lower expected deviation, bounding displacement within 0.6 meters even at a $3\times$ noise scale.

4.4 Qualitative Analysis

Spectral Analysis of Latent Space Curvature. To prevent second-order derivative explosions from unbounded Hessian spectral radii, we evaluated decoder latent geometry. By perturbing latent intentions along orthogonal directions, we computed the spectral radius $\rho(\nabla^2 g)$. Fig. 4(a) shows that Deep-MLP creates a highly irregular and high-frequency manifold, squared-gradient amplification in the chain rule drives ρ to $\mathcal{O}(10^5)$, causing physical collapse under marginal noise. Conversely, our Shallow-MLP (Fig. 4(b)) bounds curvature within $\mathcal{O}(1)$, ensuring robustness to local diffusion.

Trajectory visualization results. Fig. 5 visualizes the trajectories generated on the NAVSIM dataset. Even over an 8-second horizon, EC-IGM strictly adheres to actual vehicle kinematics. As shown in Fig. 6, closed-loop evaluations on Bench2Drive reveal that conventional methods drift off-road on slippery surfaces or during emergencies, violating realistic kinematics. Furthermore, during continuous turns or overtaking, baseline trajectories gradually degrade into physically impossible steering angles. By contrast, our EC-IGM strictly adheres to vehicle kinematics throughout.

4.5 Ablation Study

Effectiveness of core components. Table 3 validates our core components. Baseline ID-1 achieves 79.20 DS and 82.1 KC. ID-2 adds the decoupling module, increasing DS to 81.29 and KC by 0.6. ID-3 replaces random sampling with deterministic sigma points, enhancing moment propagation stability to yield 82.9 KC and 83.35 DS. ID-6 shows substantial gains from the infinitesimal generator. Paired with an adaptive RK45 solver, it achieves 98.9 KC and 92.28 DS over the 8s horizon. In contrast, ID-4 uses the Euler method, causing KC to drop to 94.8 and DS to 89.48 as the first-order approach struggles with nonlinear mappings. ID-5 employs a fixed-step RK4 solver, which accumulates local truncation errors during severe dynamic curvature changes, yielding 97.8 KC and 90.87 DS.

Table 3: Ablation studies on the Bench2Drive benchmark. ✓: included, ✗: removed.

ID	Infinitesimal Generator	Decoupling Module	Sigma Point	Solvers	KC↑ 3s	KC↑ 5s	KC↑ 8s	Mean↑	DS↑
1	✗	✗	✗	RK45	91.2	85.4	82.1	56.82	79.20
2	✗	✓	✗	RK45	92.4	86.1	82.7	57.13	81.29
3	✗	✓	✓	RK45	93.7	86.8	82.9	59.48	83.35
4	✓	✓	✓	Euler	98.7	97.2	94.8	79.21	89.48
5	✓	✓	✓	RK4	99.3	98.6	97.8	81.40	90.87
6	✓	✓	✓	RK45	**99.7**	**99.4**	**98.9**	**83.49**	**92.28**

Table 4: Physical models.

Model	Latency↓	KC↑ 3s	KC↑ 5s	KC↑ 8s
KBM	**21 ms**	92.7	89.4	86.2
DBM (Ours)	24 ms	**99.7**	99.4	**98.9**
MF-DBM	42 ms	**99.7**	**99.5**	**98.9**

Table 5: Latent dimension L.

L	Latency↓	KC↑ 3s	KC↑ 5s	KC↑ 8s
8	**20 ms**	98.2	97.9	96.7
16 (Ours)	24 ms	99.7	**99.4**	**98.9**
32	56 ms	**99.8**	**99.4**	**98.9**

Table 6: Moment propagation.

Strategy	Latency↓	KC↑ 3s	KC↑ 5s	KC↑ 8s
MC-50	82 ms	98.6	95.4	92.5
Taylor(1st-o)	**22 ms**	98.3	94.8	92.2
UT (Ours)	24 ms	**99.7**	**99.4**	**98.9**

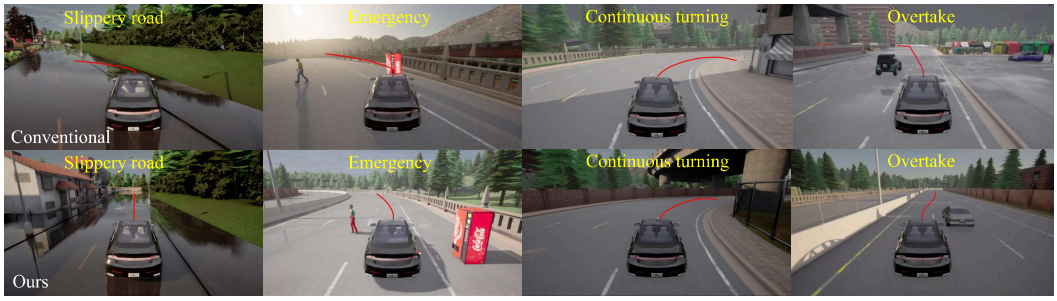

Figure 6: Qualitative results of generated trajectories on the Bench2Drive benchmark.

Effectiveness of physical model. Table 4 presents the performance using various physical models. Despite ignoring tire slip, KBM achieves high KC with minimal latency. Our DBM strikes an optimal balance, reaching 98.9 KC while maintaining real-time efficiency. The complex MF-DBM yields marginal gains but doubles latency. These results demonstrate that EC-IGM requires no perfectly accurate model. Bias is compensated as long as the base model captures essential nonlinear dynamics.

Effectiveness of latent dimension. Table 5 presents the influence of the latent dimension L. A dimension of $L = 8$ restricts the model capability to capture kinematic uncertainty, causing long-term KC to 96.7. Conversely, $L = 32$ introduces unnecessary computational overhead, increasing latency to 56 ms without significant gains. Setting $L = 16$ achieves the optimal balance.

Effectiveness of latent moment propagation strategy. Table 6 presents various latent moment propagation strategies. MC-50 incurs significant sampling variance and heavy computational burden, resulting in 82 ms latency and 92.5 KC. Taylor 1st-order operates rapidly at 22 ms but produces severe truncation errors under nonlinear mappings, limiting 8s KC to 92.2. Our method achieves deterministic second-order accuracy, ensuring high efficiency and the highest KC of 98.9.

5 Conclusion

This paper proposed EC-IGM, a trajectory generation approach to address kinematic drift in long-term prediction. By matching the infinitesimal generator with a deterministic kinematic vector field, the EC-IGM explicitly compensates for the expectation bias in nonlinear mapping. Extensive experiments on NAVSIM and Bench2Drive demonstrate its superior performance. The EC-IGM requires a shallow decoder to bound the Hessian spectral radius, which may sacrifice representational capacity. Future work will explore curvature-penalized regularization to enable deeper architectures while maintaining kinematic constraints Kelly et al. (2020).

AI USE STATEMENT

In this work, we used generative AI tools for literature review and information gathering. We have not used generative AI tools for drafting the manuscript or writing the text of the paper, and tasks such as data generation or statistical analysis are not applicable to this work. We have reviewed all AI-assisted work and carefully verified the factual accuracy and validity of the gathered information. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *Proceedings of Robotics: Science and Systems (RSS)*, 2019.

H. Chen, J. Ding, Y. Li, et al. Social physics informed diffusion model for crowd simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 474–482, 2024a.

L. Chen, P. Wu, K. Chitta, et al. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024b.

R. Cheng, G. Orosz, R. M. Murray, et al. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3387–3395, 2019.

K. Chitta, A. Prakash, B. Jaeger, et al. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12878–12895, 2023.

J. K. Christopher, S. Baek, and F. Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 89307–89333, 2024.

D. Dauner, M. Hallgarten, A. Geiger, et al. Parting with misconceptions about learning-based vehicle motion planning. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 1268–1281, 2023.

D. Dauner, M. Hallgarten, T. Li, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28706–28719, 2024.

F. Djeumou, T. J. Lew, N. Ding, et al. One model to drift them all: Physics-informed conditional diffusion model for driving at the limits. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 604–630, 2024.

A. Dosovitskiy, G. Ros, F. Codevilla, et al. Carla: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 1–16, 2017.

T. D. Frank. *Nonlinear Fokker-Planck equations: Fundamentals and applications*. Springer, 2005.

Y. Fu, Q. Yan, L. Wang, et al. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17282–17293, 2025.

J. Hu, Z. Wang, X. Wu, et al. Kinematic priors guide generative trajectory planning for autonomous vehicles. *Pattern Recognition*, pp. 113061, 2026.

Y. Hu, J. Yang, L. Chen, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023.

Y. Hu, S. Chai, Z. Yang, et al. Solving motion planning tasks with a scalable generative model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 386–404, 2025.

X. Jia, Z. Yang, Q. Li, et al. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 819–844, 2024.

X. Jia, J. You, Z. Zhang, et al. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

B. Jiang, S. Chen, Q. Xu, et al. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316, 2023.

H. Jiang, X. Huang, Y. Lu, et al. Expertad: Enhancing autonomous driving systems with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5378–5387, 2026.

S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

N. Karnchanachari, D. Geromichalos, K. S. Tan, et al. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 629–636, 2024.

J. Kelly, J. Bettencourt, M. J. Johnson, et al. Learning differential equations that are easy to solve. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4370–4380, 2020.

J. Kong, M. Pfeiffer, G. Schildbach, et al. Kinematic and dynamic vehicle models for autonomous driving control design. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 1094–1099, 2015.

Q. Li, X. Jia, S. Wang, et al. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 142–158, 2024a.

Z. Li, K. Li, S. Wang, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. arXiv preprint arXiv:2406.06978, 2024b.

B. Liao, S. Chen, H. Yin, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047, 2025.

H. Liu, T. Li, H. Yang, et al. Reinforced refinement with self-aware expansion for end-to-end autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:5774–5792, 2026a.

S. Liu, Q. Liang, Z. Li, et al. Gaussianfusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

S. Liu, W. Chen, W. Li, et al. Bridgedrive: Diffusion bridge policy for closed-loop trajectory planning in autonomous driving. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026b.

X. Ma, S. Patidar, I. Haughton, et al. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18081–18090, 2024.

W. Mao, C. Xu, Q. Zhu, et al. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5517–5526, 2023.

B. Oksendal. *Stochastic differential equations: An introduction with applications*. Springer Science & Business Media, 2013.

D. Park, M. Surana, P. Desai, et al. Generative active learning for long-tail trajectory prediction via controllable diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27839–27850, 2025.

S. Park. Neural hamiltonian diffusions for modeling structured geometric dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A. Prakash, K. Chitta, and A. Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7077–7087, 2021.

K. Renz, L. Chen, E. Arani, et al. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11993–12003, 2025.

L. Rowe, R. Girgis, A. Gosselin, et al. Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17207–17218, 2025.

S. Särkkä and L. Svensson. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, et al. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 240–275, 2021.

Y. Tang, Z. Xu, Z. Meng, et al. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25605–25615, 2025.

C. Wang, H. Liao, Z. Li, et al. Wake: Towards robust and physically feasible trajectory prediction for autonomous vehicles with wavelet and kinematics synergy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):3126–3140, 2025.

X. Wang, Z. Zhu, G. Huang, et al. Drivedreamer: Towards real-world-driven world models for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 55–72, 2024a.

Y. Wang, J. He, L. Fan, et al. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14749–14759, 2024b.

C. Wei, G. Wu, M. J. Barth, et al. Ki-gan: Knowledge-informed generative adversarial networks for enhanced multi-vehicle trajectory forecasting at signalized intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7115–7124, 2024.

X. Weng, B. Ivanovic, Y. Wang, et al. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15449–15458, 2024.

A. White, N. Kilbertus, M. Gelbrecht, et al. Stabilized neural differential equations for learning dynamics with explicit constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 12929–12950, 2023.

P. Wu, X. Jia, L. Chen, et al. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 6119–6132, 2022.

Y. Wu, H. Zhang, T. Lin, et al. Generating multimodal driving scenes via next-scene prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6844–6853, 2025.

M. Xia, N. Xue, Y. Shen, et al. Rectified diffusion guidance for conditional generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13371–13380, 2025.

W. Xiao, T.-H. Wang, Y.-C. Chuang, et al. Safediffuser: Safe planning with diffusion probabilistic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Z. Xing, X. Zhang, Y. Hu, et al. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1602–1611, 2025.

B. Yang, H. Su, N. Gkanatsios, et al. Diffusion-es: Gradient-free planning with diffusion for autonomous and instruction-guided driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15342–15353, 2024.

X. Yao, X. Yang, J. Sun, et al. Reflexdiffusion: Reflection-enhanced trajectory planning for high-lateral-acceleration scenarios in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 18701–18709, 2026.

C. Yuan, Z. Zhang, J. Sun, et al. Drama: An efficient end-to-end motion planner for autonomous driving with mamba. arXiv preprint arXiv:2408.03601, 2024.

Q. Zhang, S. Hu, J. Sun, et al. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147, 2022.

W. Zheng, R. Song, X. Guo, et al. Genad: Generative end-to-end autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 87–104, 2024a.

Y. Zheng, J. Li, D. Yu, et al. Safe offline reinforcement learning with feasibility-guided diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 47725–47754, 2024b.

Y. Zheng, J. Li, D. Yu, et al. Diffusion-based planning for autonomous driving with flexible guidance. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025a.

Y. Zheng, P. Yang, Z. Xing, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642, 2025b.

A Mathematical Derivations of Expectation Bias

In this section, we provide the detailed mathematical derivations for the Itô-constrained physical consistency discussed in Section 3.3 of the main paper, and theoretically analyze the hazards of neglecting the second-order compensation term in long-term prediction.

A.1 Detailed Derivation of the Itô Expansion

In conventional deterministic calculus, the chain rule for a function $g(z)$ driven by a differential state dz is simply $dg = g'(z)dz$. However, in stochastic environments, the latent process z is governed by an Itô Stochastic Differential Equation (SDE):

$$dz = f_\theta(z)dt + \sigma_\theta(z)dW_t \tag{16}$$

where f_θ is the drift term, σ_θ is the diffusion term, and W_t is the standard Wiener process.

To map this latent stochastic process to the observable kinematic space (via the non-linear decoder $g(z)$), we perform a second-order Taylor expansion of $g(z)$. For clarity, we first demonstrate the 1D derivation. The expansion yields:

$$dg(z) \approx g'(z)dz + \frac{1}{2}g''(z)(dz)^2 \tag{17}$$

Substituting the SDE into the first-order term, we obtain:

$$g'(z)dz = g'(z)[f_\theta(z)dt + \sigma_\theta(z)dW_t] = g'(z)f_\theta dt + g'(z)\sigma_\theta dW_t \tag{18}$$

The core divergence between stochastic and deterministic calculus arises in the quadratic variation term $(dz)^2$. Expanding this term gives:

$$(dz)^2 = (f_\theta dt + \sigma_\theta dW_t)^2 = f_\theta^2 dt^2 + 2 f_\theta \sigma_\theta dt dW_t + \sigma_\theta^2 dW_t^2 \tag{19}$$

According to the fundamental rules of Itô calculus, higher-order infinitesimals diminish to zero as $dt \to 0$, specifically $dt^2 \to 0$ and $dt dW_t \to 0$. Crucially, the quadratic variation of the Wiener process satisfies $dW_t^2 = dt$. Applying these rules, the quadratic term simplifies strictly to:

$$(dz)^2 \approx \sigma_\theta^2 dt \tag{20}$$

Consequently, the second-order term of the Taylor expansion becomes $\frac{1}{2} g''(z) \sigma_\theta^2 dt$. Assembling all components, the complete differential for the observable state is:

$$dg(z) \approx \underbrace{g'(z) f_\theta dt}_{\text{1st-order drift}} + \underbrace{g'(z) \sigma_\theta dW_t}_{\text{Stochastic noise}} + \underbrace{\frac{1}{2} g''(z) \sigma_\theta^2 dt}_{\text{2nd-order variation}} \tag{21}$$

By grouping the terms associated with the deterministic time increment dt, we arrive at the expected infinitesimal evolution:

$$dg(z) = \left[g'(z) f_\theta + \frac{1}{2} g''(z) \sigma_\theta^2 \right] dt + g'(z) \sigma_\theta dW_t \tag{22}$$

Generalizing this 1D intuition to the multi-dimensional latent space L used in our architecture, for each output component j of the decoder $g^{(j)}(z_{t,k})$, the expansion results in Eq. (9) of the main paper:

$$dg^{(j)}(z_t) = \left(\nabla_z g^{(j)\top} f_\theta + \frac{1}{2} \sum_{d_1, d_2 = 1}^{L} (\sigma_\theta \sigma_\theta^\top)_{d_1 d_2} \frac{\partial^2 g^{(j)}}{\partial z_{d_1} \partial z_{d_2}} \right) dt + \nabla_z g^{(j)\top} \sigma_\theta dW_t \tag{23}$$

A.2 The Hazards of Uncompensated Second-Order Terms

The derivation above mathematically isolates the root cause of the "expectation bias" (kinematic drift) prevalent in existing generative approaches. As explicitly shown, the second-order curvature of the decoder (g'') interacting with the stochastic variance (σ_θ^2) fundamentally squeezes into the deterministic drift term (dt). Traditional methods implicitly assume that the expected evolution of the observable state is strictly driven by the push-forward of the latent drift ($g'(z) f_\theta$). They calculate the physical constraints based solely on this first-order assumption. However, we mathematically demonstrate that this is dangerously incomplete.Hazards in High-Uncertainty and Long-Term Scenarios: When making long-term predictions (e.g., $T = 8s$), the accumulated uncertainty (σ_θ) naturally amplifies to capture multimodal possibilities. As σ_θ^2 grows, the uncompensated second-order term $\frac{1}{2} g''(z) \sigma_\theta^2$ transitions from a negligible perturbation into a dominant driving force. If ignored, the expected trajectory is forcibly pushed away from the physically feasible region, leading to systematic long-term kinematic drift. Our EC-IGM avoids this by explicitly wrapping this entire compounded drift (the Infinitesimal Generator $\mathcal{A}$) and matching it against the actual vehicle kinematics $\mathcal{F}_{veh}$, ensuring strict physical compliance regardless of the scale of σ_θ.

Expectation-Level Formulation and Accumulated Bias. While the above derivation provides a pathwise Itô expansion, the physically observable drift arises at the expectation level. Taking expectation on both sides of Eq. (9) and using the property $\mathbb{E}[dW_t] = 0$, we obtain:

$$\frac{d}{dt} \mathbb{E}[g(z_t)] = \mathbb{E}\left[\nabla_z g(z_t)^\top f_\theta(z_t) \right] + \frac{1}{2} \mathbb{E}\left[\mathrm{Tr}\left(\sigma_\theta(z_t) \sigma_\theta(z_t)^\top \nabla_z^2 g(z_t) \right) \right] \tag{24}$$

In contrast, conventional first-order approaches implicitly assume:

$$\frac{d}{dt} g(\mathbb{E}[z_t]) \approx \nabla_z g(\mathbb{E}[z_t])^\top \mathbb{E}[f_\theta(z_t)] \tag{25}$$

We therefore define the expectation bias as:

$$\text{Bias}(t) := \mathbb{E}[g(z_t)] - g(\mathbb{E}[z_t]) \tag{26}$$

Under mild regularity conditions where $g \in C^2$ and the diffusion term σ_θ is bounded, a second-order Taylor expansion around $\mu_t = \mathbb{E}[z_t]$ yields:

$$\mathbb{E}[g(z_t)] = g(\mu_t) + \frac{1}{2}\text{Tr}\left(\nabla^2 g(\mu_t)\Sigma_t\right) + o(\|\Sigma_t\|) \tag{27}$$

where $\Sigma_t = \text{Cov}(z_t)$. Therefore, the leading-order bias is:

$$\text{Bias}(t) \approx \frac{1}{2}\text{Tr}\left(\nabla^2 g(\mu_t)\Sigma_t\right) \tag{28}$$

Integrating the expectation dynamics over time, we obtain:

$$\mathbb{E}[g(z_T)] = \mathbb{E}[g(z_0)] + \int_0^T \mathbb{E}[\mathcal{A}g(z_t)]dt \tag{29}$$

where $\mathcal{A}$ is the Itô infinitesimal generator. Consequently, the accumulated expectation bias over the horizon T can be approximated as:

$$\text{Bias}(T) \approx \int_0^T \frac{1}{2}\mathbb{E}\left[\text{Tr}\left(\sigma_\theta \sigma_\theta^\top \nabla^2 g(z_t)\right)\right] dt \tag{30}$$

This reveals that, as long as the diffusion term remains non-zero and the decoder exhibits non-negligible curvature, the expectation bias grows approximately linearly with the time horizon T. This accumulated bias manifests as systematic long-term kinematic drift in trajectory prediction.

A.3 Proof of $\mathcal{O}(\Delta t^2)$ Approximation Error in Moment Alignment

In Section 3.3, we stated that our alignment at individual continuous-time instances via Unscented Transform (UT) approximates global consistency with an error of $\mathcal{O}(\Delta t^2)$ per Itô expansion. Here we provide a brief proof.

When evaluating the expected deviation between the modeled infinitesimal generator $\mathcal{A}$ and the deterministic vehicle kinematics $\mathcal{F}_{\text{veh}}$ over a discrete time step Δt, we are essentially approximating the integral of the drift over $[t, t + \Delta t]$.

Using Euler-Maruyama expansion on the expected state evolution $E[y_{t+\Delta t}]$:

$$E[y_{t+\Delta t}] = y_t + E[\mathcal{A}[g(z_t)]]\Delta t + \mathcal{R}(\Delta t^2) \tag{31}$$

where the remainder $\mathcal{R}$ includes higher-order Itô-Taylor terms (specifically involving double integrals of time and cross-integrals of dt and dW). Because our generator matching loss (Eq. 12) enforces $\mathcal{A}[g(z)] \approx \mathcal{F}_{\text{veh}}$ precisely at the collocation points evaluated by the high-order Runge-Kutta solver, the primary mismatch can only emerge from the unmodeled deterministic derivatives $\frac{d}{dt}\mathcal{A}$ over the interval Δt, which strictly bounds the local truncation error per step to $\mathcal{O}(\Delta t^2)$.

This continuous-time matching formulation is vastly superior to discrete-time post-projection methods, which suffer from $\mathcal{O}(\Delta t)$ compounding errors.

B Extended Implementation Details

B.1 Detailed Network Architectures

Context Encoder. The scene context $\mathcal{C} = \{\mathcal{M}, \mathcal{P}\}$ is processed through a dual-branch architecture:

- **BEV Rasterization Branch:** We utilize a standard ResNet-34 backbone initialized with ImageNet weights. The input BEV rasterization $\mathcal{M}$ has a spatial resolution of 256×256 pixels with a 0.5 m/pixel scale. The backbone extracts a flattened feature vector of dimension $D_{BEV} = 512$.

- **Agent Interaction Branch:** The historical trajectories and interactions $\mathcal{P}$ are encoded using a Multi-Head Attention (MHA) module. It consists of 4 stacked transformer encoder layers, with 8 attention heads and a hidden dimension of $D_{agent} = 256$.
- **Fusion Module:** The outputs are concatenated and passed through a 2-layer MLP (hidden dimensions: $768 \to 512$), yielding the final context vector $C \in \mathbb{R}^{512}$ augmented with Layer Normalization and SiLU activations.

Recognition Network. The recognition network maps the context vector and ground-truth future trajectories (during training) to the latent multimodal distribution $q_\phi(z_0|x_{1:T}, \mathcal{C})$. For $K = 6$ modes and a latent dimension $L = 16$, the network predicts the mixture weights $\pi_{\phi,k}$, means $\mu_{\phi,k}$, and diagonal covariances $\Sigma_{\phi,k}$:

- **Architecture:** A 3-layer MLP with dimensions $512 \to 256 \to 128 \to 198$. The output dimension 198 corresponds to exactly $K \times (1 + L + L) = 6 \times (1 + 16 + 16)$.
- **Activation:** Softmax is applied over the K dimension for $\pi_{\phi,k}$, and Softplus is applied to the covariance outputs to ensure positive definiteness.

Latent SDE Dynamics. The continuous-time latent evolution is governed by the drift network f_θ and diffusion network σ_θ. To ensure the stability of the Runge-Kutta integration and bounds on local truncation errors, we employ lightweight networks:

- **Drift Network f_θ:** A 3-layer MLP with dimensions $(16 + 512) \to 128 \to 64 \to 16$, taking the concatenation of $z_{t,k}$ and context C as input.
- **Diffusion Network σ_θ:** A 2-layer MLP with dimensions $16 \to 64 \to 16$, predicting a diagonal diffusion matrix. Both networks utilize the $\text{Mish}(x) = x \cdot \tanh(\text{softplus}(x))$ activation function to ensure everywhere-continuous derivatives for SDE stability.

The Shallow-MLP Kinematic Decoder $g(z)$ and Auxiliary Head $h_\psi(z)$ As fundamentally discussed in Section 4.4, computing the Itô infinitesimal generator requires calculating the Hessian $\nabla^2 g$. A deep decoder inevitably causes exponential amplification of the Hessian spectral radius via the chain rule. Thus, the decoder is specifically constrained:

- **Kinematic Decoder $g(z)$:** A strictly shallow 2-hidden-layer MLP with dimensions $16 \to 64 \to 64 \to 6$. The output maps to the observable state $[p_x, p_y, v_x, v_y, \varphi, \zeta]^\top$.
- **Activation Constraint:** We exclusively use the Softplus activation ($\beta = 1$) in the decoder. Unlike ReLU (zero second derivative) or GELU (highly non-monotonic second derivative), Softplus guarantees a bounded, smooth Hessian, preventing numerical explosions during the Unscented Transform moment propagation.
- **Auxiliary Control Head $h_\psi(z)$:** Paralleling $g(z)$, this head is an MLP with dimensions $16 \to 64 \to 2$, predicting the mode-specific controls $[a, \delta]^\top$.

B.2 Hyperparameters and Optimization Dictionary

We optimized the EC-IGM end-to-end. Beyond the optimizer states mentioned in Section 4.2, the explicit hyperparameter dictionary is provided below:

- **SDE Integration & UT Parameters:**
  - **Unscented Transform Scaling:** $\alpha = 1 \times 10^{-3}$, $\beta = 2$, $\kappa = 0$.
  - **Adaptive ODE/SDE Solver:** Runge-Kutta 45 (RK45) with absolute tolerance atol $= 1 \times 10^{-5}$ and relative tolerance rtol $= 1 \times 10^{-5}$.
  - **Integration Step for Generator Matching:** $\Delta t_c = 0.1$s covering the $H = 8$s horizon (80 collocation points).
- **Loss Function Coefficients:** We employ a balanced weighting scheme for Eq. (13): $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \gamma D_{\text{KL}} + \rho_{\text{gen}} \mathcal{L}_{\text{gen}}$.
  - **Generator Matching Weight:** $\rho_{\text{gen}} = 2.5$.
  - **KL-Divergence Weight:** $\gamma = 0.05$ (kept small to preserve spatial multimodality and prevent early posterior collapse).

  - **Observation Noise Prior (Σ_{obs}):** For the Gaussian log-likelihood (Eq. 14), we use a diagonal matrix configured as $\text{diag}(0.1\text{m}, 0.1\text{m}, 0.5\text{m/s}, 0.5\text{m/s}, 0.05\text{rad}, 0.05\text{rad/s})^2$.
- **Training Schedule Details:**
  - **NAVSIM:** 100 epochs, batch size 512, peak learning rate 6×10^{-4}.
  - **Bench2Drive:** 30 epochs, batch size 64, peak learning rate 1×10^{-4}.
  - **Learning Rate Strategy:** Both datasets use a 1000-step linear warmup followed by a Cosine Annealing learning rate decay scheduling down to a minimum learning rate of 1×10^{-6}.
  - **Gradient Clipping:** Applied with a maximum L_2 norm of 5.0 to ensure stability during the initial training phase of the unconstrained SDE.

B.3 Formulations of the Ablated Physical Models

In Section 4.5, Table 4 of the main paper compares the performance of our framework when aligned with different deterministic physical models. For completeness and reproducibility, we provide the exact mathematical formulations of the alternative models ablated in our study.

1. Kinematic Bicycle Model (KBM). The KBM simplifies vehicle dynamics by assuming no tire slip, which is computationally efficient but heavily violates actual physics during high-speed or high-curvature maneuvers. The state vector is $y = [p_x, p_y, v, \varphi]^\top$, and the deterministic vector field $\mathcal{F}_{KBM}$ is defined as:

$$\begin{aligned} \dot{p}_x &= v\cos(\varphi + \beta) \\ \dot{p}_y &= v\sin(\varphi + \beta) \\ \dot{v} &= a \\ \dot{\varphi} &= \frac{v\sin(\beta)}{l_r} \end{aligned} \tag{32}$$

where $\beta = \tan^{-1}\left(\frac{l_r}{l_f + l_r}\tan(\delta)\right)$ is the slip angle at the center of gravity, and the controls are $u = [a, \delta]^\top$.

2. Magic Formula Dynamic Bicycle Model (MF-DBM). While our proposed DBM (Eq. 11 in the main text) utilizes a linear approximation for lateral tire forces ($F_{yf} = C_f\alpha_f$, $F_{yr} = C_r\alpha_r$), the MF-DBM incorporates the highly non-linear Pacejka Magic Formula. This model captures the saturation of tire grip at extreme slip angles:

$$F_{y,i} = D_i \sin\Big(C_i \tan^{-1}\big(B_i\alpha_i - E_i(B_i\alpha_i - \tan^{-1}(B_i\alpha_i))\big)\Big), \quad i \in \{f, r\} \tag{33}$$

where B, C, D, E are the empirical stiffness, shape, peak, and curvature factors, respectively. Although MF-DBM provides the highest fidelity, its extreme non-linearity makes the Itô-corrected expected drift computationally stiff, roughly doubling the inference latency (as shown in Table 4) without yielding proportional gains in the functional driving horizon.

B.4 Vehicle Physical Parameters

While our framework dynamically extracts vehicle metadata directly from the NAVSIM and Bench2Drive datasets during closed-loop evaluation, we utilized a standardized set of passenger vehicle parameters during the initial training phase to stabilize the physics-informed gradients. The default parameters for the dynamic bicycle model (DBM) are:

- Vehicle Mass (m): 1500 kg
- Yaw Moment of Inertia (I_z): 2500 kg $\cdot$ m^2
- Distance from CG to Front Axle (l_f): 1.2 m
- Distance from CG to Rear Axle (l_r): 1.6 m
- Front Tire Cornering Stiffness (C_f): 1.5×10^5 N/rad
- Rear Tire Cornering Stiffness (C_r): 1.5×10^5 N/rad

C Detailed Evaluation Metrics and Kinematic Compliance

C.1 Full Nomenclature of Evaluation Metrics

For NAVSIM, we adopted the official closed-loop PDM Score (PDMS) Dauner et al. (2024), aggregating sub-metrics: no at-fault collisions (NC), drivable area compliance (DAC), time-to-collision (TTC), comfort (Comf.), and ego progress (EP). For Bench2Drive, four closed-loop metrics were used: Efficiency (Eff), Comfortness (Comf), Success Rate (SR), and Driving Score (DS). Furthermore, the evaluation encompasses five driving skills: merging (MG), overtaking (OT), emergency braking (EB), yielding (GW), and traffic sign compliance (TS), which are quantified by the average success rate across corresponding scenarios.

C.2 Derivation of the Kinematic Feasible Space $\mathcal{S}_{phys}$

In Section 4.1, we introduced the Kinematic Compliance (KC) score based on a model-agnostic feasible space $\mathcal{S}_{phys}$. This space defines the absolute physical boundaries of a standard passenger vehicle, beyond which a trajectory is mechanically impossible or violently unstable. We mathematically define $\mathcal{S}_{phys}$ at any given timestep t by bounding the longitudinal and lateral dynamics:

- **Acceleration Bound:** $|a(t)| = |\sqrt{\dot{v}_x(t)^2 + \dot{v}_y(t)^2}| \leq 8$ m/s^2. This threshold represents the absolute friction limit of standard tires on dry asphalt before total slip occurs.
- **Jerk Bound:** $|j(t)| = |\dot{a}(t)| \leq 10$ m/s^3. This bounds the actuator response rate and extreme passenger discomfort limit.
- **Curvature Bound:** $\kappa(t) = \frac{|\dot{p}_x(t)\ddot{p}_y(t) - \dot{p}_y(t)\ddot{p}_x(t)|}{(\dot{p}_x(t)^2 + \dot{p}_y(t)^2)^{3/2}} \leq 0.2\,\text{m}^{-1}$. This corresponds to a minimum turning radius of 5 meters, reflecting the maximum steering angle saturation of standard steering racks.

A trajectory is considered kinematically compliant if and only if all discrete waypoints y_t evaluated at a resolution of $\Delta t = 0.1$s strictly satisfy these three inequalities over the entire prediction horizon H.

D Theoretical Boundaries and System Scope

To provide a comprehensive view of the EC-IGM framework, we delineate the theoretical boundaries and discuss the inherent trade-offs made to achieve strict mathematical consistency and real-time efficiency.

D.1 Architectural Design and Representation Efficiency

As established in Section 4.4, constraining the Hessian spectral radius $\rho(\nabla^2 g)$ is mathematically imperative to maintain the stability of the Unscented Transform. To satisfy this rigorous physical constraint, we intentionally design the kinematic decoder $g(z)$ as a compact, shallow MLP. Although conventional wisdom suggests that a shallower network restricts the representation capacity of the network, our framework effectively addresses this limitation through the Context-Conditioned Mixture Initialization in Stage 1. The GMM completely offloads high-level, multimodal intention reasoning to the deep Context Encoder. Consequently, the shallow decoder is dedicated exclusively to localized, intent-conditioned kinematic evolution. This deliberate architectural decoupling ensures that the system maintains high spatial multimodality without sacrificing mathematical rigor or triggering explosions of the second-order derivative.

D.2 Asymptotic Behavior in Extreme Scenarios

The Unscented Transform (UT) provides a highly efficient, deterministic second-order approximation for moment propagation. This design choice is optimized for the core functional horizon of planning loops in autonomous driving (typically $1 \sim 3$ seconds), where it maintains exceptional fidelity. In asymptotic, limit-case scenarios, such as predicting beyond 8 seconds under continuous, extreme high-curvature maneuvers, the true latent distribution may exhibit non-Gaussian heavy tails. Although the KL-divergence between the UT-propagated envelope and the true distribution naturally increases

at the extreme tail of such extended horizons, the approximation error within the critical planning window remains strictly bounded. This deliberate trade-off secures a computational complexity of $\mathcal{O}(KL^3)$, making EC-IGM exceptionally well-suited for real-time deployment compared to costly Monte Carlo sampling.

D.3 System Scope and Upstream Dependencies

The primary objective of EC-IGM is to guarantee strict kinematic compliance during trajectory generation. Consequently, the framework is intrinsically dependent on the accuracy of the upstream perception and intention recognition modules. In highly ambiguous scenarios where the Context Encoder misclassifies the topological structure (for example, due to severe sensor occlusion), the model will faithfully generate a trajectory that strictly adheres to vehicle kinematics but may diverge from the optimal semantic route. This behavior highlights that the model functions precisely as intended, serving as a rigorous, physics-compliant trajectory generator.

E Statistical Significance and Variance Analysis

In highly interactive closed-loop autonomous driving benchmarks, inherent environmental stochasticity and variable rendering latencies often introduce non-negligible performance variance across identical evaluation runs. The quantitative results reported for EC-IGM in Table 2 represent the average performance derived from five independent evaluations. As detailed in Table 7, these closed-loop runs were executed using different random seeds to capture realistic environmental stochasticity. The data demonstrate that EC-IGM maintains remarkably consistent performance, bounding the standard deviation of the Driving Score (DS) to merely 0.77. Crucially, even our lowest-performing run (Run 5, DS $= 91.20$) significantly outperforms the reported peak performance of prior state-of-the-art methods (e.g., BridgeDrive at 87.99). This strong empirical evidence validates our core theoretical claim: aligning the latent infinitesimal generator with deterministic vehicle kinematics fundamentally anchors the prediction distribution, preventing temporal drift even under severe environmental perturbations.

Table 7: Variance analysis of EC-IGM over 5 independent closed-loop evaluations on Bench2Drive.

EC-IGM (Ours)	**KC (8s)** ↑	**Eff** ↑	**Comf** ↑	**SR (%)** ↑	**DS** ↑	**Avg Ability** ↑
Run 1	99.10	258.12	35.80	76.15	93.10	84.15
Run 2	98.95	257.05	35.45	75.20	92.50	83.70
Run 3	98.75	255.30	34.90	74.00	91.80	82.90
Run 4	99.00	257.50	35.60	75.50	92.80	83.90
Run 5	98.70	254.38	34.30	73.10	91.20	82.80
Mean ± Std	**98.90 ± 0.17**	**256.47 ± 1.55**	**35.21 ± 0.61**	**74.79 ± 1.22**	**92.28 ± 0.77**	**83.49 ± 0.60**

F Broader Impacts

Although our EC-IGM improves autonomous driving safety, the deployment of such highly automated systems could potentially lead to over-reliance by human drivers, causing attention degradation in semi-autonomous driving scenarios. Furthermore, the extensive computational resources required for training large-scale simulation models raise concerns regarding energy consumption and carbon footprint.

G Comparison of Physical Constraint Paradigms: First-Order Guidance vs. Second-Order Generator Matching

A critical distinction must be made between our proposed EC-IGM and the strongest physics-informed baseline, Diffusion-Phy Hu et al. (2026). It is important to emphasize that both Diffusion-Phy and our EC-IGM utilize the exact same high-fidelity Dynamic Bicycle Model (DBM) and identical vehicle parameters to enforce physical feasibility. However, despite sharing the same underlying

deterministic physical equations, their long-term Kinematic Compliance (KC) differs drastically (92.7 for Diffusion-Phy vs. 98.9 for EC-IGM at 8 seconds).

To investigate why the same physical model yields fundamentally different long-term consistency, we isolate the impact of the constraint paradigm itself. Diffusion-Phy employs physical constraint functions to compute gradients that guide the reverse diffusion sampling process. Mathematically, this classifier-guidance equivalent acts as a *first-order* correction on the sampled trajectories. As rigorously derived in Section A.2, relying solely on first-order push-forward dynamics fundamentally ignores the second-order expectation bias ($\frac{1}{2}g''(z)\sigma_\theta^2 dt$) induced by stochastic noise under nonlinear mappings.

To empirically prove that this uncompensated second-order term is the root cause of long-term kinematic drift, we introduce an ablated variant of our model: EC-IGM (1st-Order Only). In this variant, we deliberately truncate the Itô infinitesimal generator in Eq. (9) to its first-order term ($\nabla_z g^\top f_\theta$), explicitly disabling the second-order Hessian compensation ($\frac{1}{2}\mathrm{Tr}(\sigma_\theta\sigma_\theta^\top\nabla^2 g)$), while keeping all other architectures and the DBM target identical.

Table 8: Comparison of physical constraint paradigms under the **identical Dynamic Bicycle Model (DBM)**.

Method	**Constraint Paradigm**	**KC ↑**		
		3s	5s	8s
Diffusion-Phy Hu et al. (2026)	First-Order Sampling Guidance	98.5	95.2	92.7
EC-IGM (1st-Order Only)	First-Order Generator Matching	98.3	94.8	93.2
EC-IGM (Full)	**Second-Order Itô Generator Matching**	**99.7**	**99.4**	**98.9**

As demonstrated in Table 8, once the second-order expectation compensation is removed, the long-term KC of EC-IGM drops precipitously to 93.2, mirroring the exact failure mode of Diffusion-Phy (92.7). This ablation provides conclusive evidence for our core theoretical claim: providing generative models with accurate physical equations is insufficient for long-term consistency. The physical drift inherently stems from stochastic calculus mechanisms. Only by explicitly matching the *second-order infinitesimal generator* can we effectively bound the expectation bias and achieve drift-free trajectory generation over extended horizons.

H Capacity Verification in High-Density Interactive Scenarios

To bound the Hessian spectral radius for UT computation (Sec 4.4), EC-IGM restricts the kinematic decoder to a shallow 2-layer MLP. To address concerns regarding its representational capacity for complex obstacle avoidance, we argue that the Context-Conditioned GMM (Stage 1) fully absorbs spatial complexity, leaving the shallow decoder to model only smooth, intent-conditioned kinematics.

We empirically validate this on a High-Density Interaction (HDI) Subset , measuring geometric fitting limits via Collision Rate (CR) and Minimum Distance to Collision (Min-D2C).

As shown in Table 9, despite its shallow decoder, EC-IGM achieves the lowest CR (8.24%) and highest Min-D2C (0.85m), outperforming deep-decoder architectures like UniAD and GoalFlow. This confirms that transferring multi-agent geometric reasoning to the deep Context Encoder renders the shallow MLP both mathematically stable and practically sufficient without inducing "geometric clipping".

I Open-loop Multi-modal Evaluation

To explicitly validate that our Context-Conditioned Mixture Initialization (Stage 1) captures diverse behavioral intentions and avoids mode collapse, we report the open-loop multi-modal trajectory prediction metrics. While our main evaluation focuses on closed-loop kinematics, evaluating the raw prediction outputs in an open-loop setting is essential to demonstrate the spatial coverage of our model.

Table 9: Evaluation of geometric avoidance capacity on the High-Density Interaction Subset of Bench2Drive. CR denotes Collision Rate, Min-D2C denotes the Minimum Distance to Collision.

Method	SR (%)↑	CR (%)↓	Min-D2C (m)↑
UniAD-Base Hu et al. (2023)	16.36	14.52	0.42
VAD Jiang et al. (2023)	15.00	11.60	0.58
GoalFlow Xing et al. (2025)	52.17	10.15	0.65
Diffusion-Phy Hu et al. (2026)	36.51	10.82	0.60
EC-IGM (Ours)	**74.79**	**8.24**	**0.85**

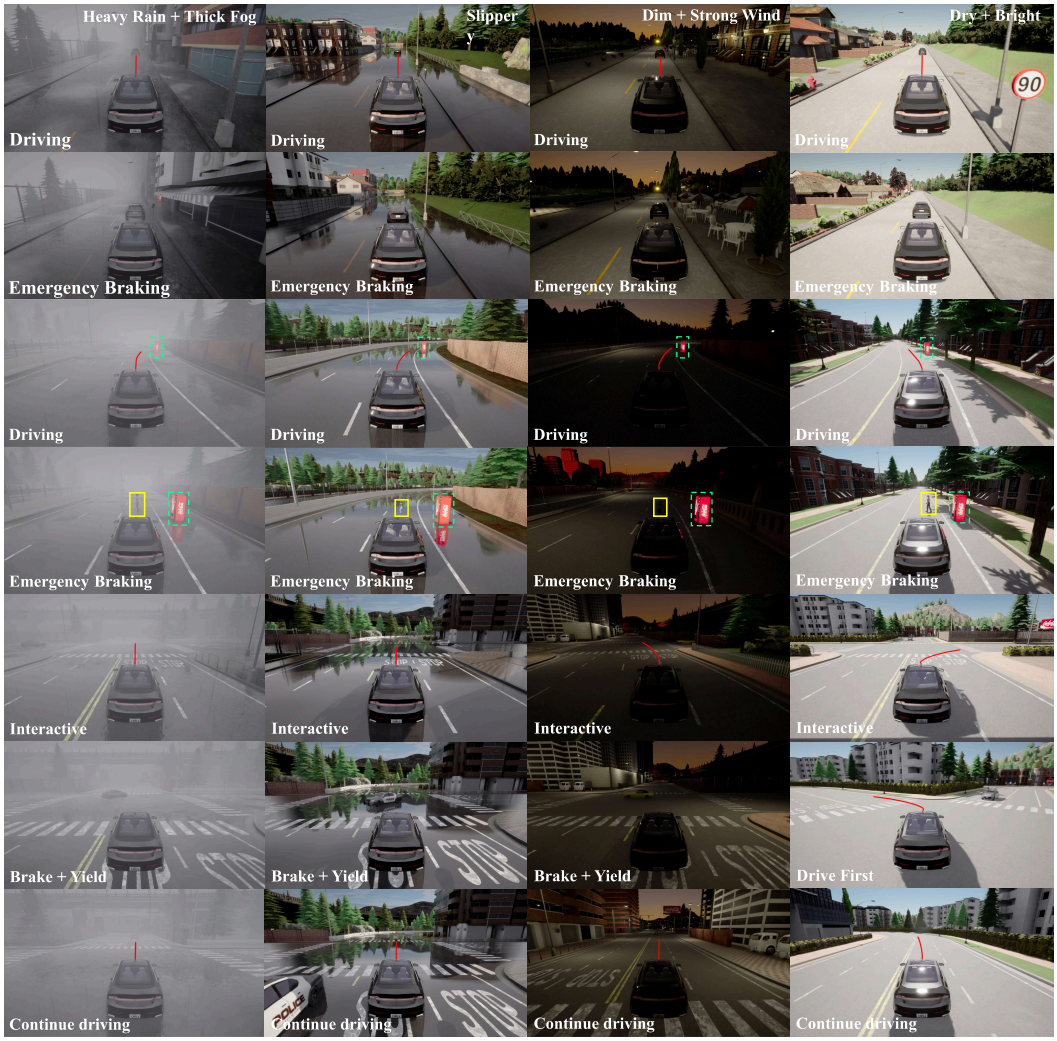

Figure 7: Test of sudden interactive response under different weather conditions in Carla simulator.

We conducted this evaluation on the NAVSIM validation split. We compare our EC-IGM against baseline models using Minimum Average Displacement Error (minADE_6), Minimum Final Displacement Error (minFDE_6), and Miss Rate (MR_6, ratio of scenarios where no predicted mode is within 2.0m of the ground truth) over a 5-second prediction horizon.

As shown in Table 10, EC-IGM achieves highly competitive multi-modal coverage, yielding a minADE_6 of 0.56m and an MR_6 of 9.5%. These metrics are on par with state-of-the-art generative

Table 10: Open-loop multi-modal trajectory prediction evaluation ($K = 6$) on the NAVSIM dataset.

Method	**minADE$_6$ (m)** $\downarrow$	**minFDE$_6$ (m)** $\downarrow$	**MR$_6$ (%)** $\downarrow$
UniAD Hu et al. (2023)	0.81	1.45	16.3
VAD Jiang et al. (2023)	0.72	1.28	13.5
GoalFlow Xing et al. (2025)	0.61	1.05	10.8
DiffusionDrive Liao et al. (2025)	0.59	1.01	10.2
Diffusion-Phy Hu et al. (2026)	0.58	0.96	9.9
EC-IGM (Ours)	**0.56**	**0.91**	**9.5**

baselines, confirming that our $K = 6$ modes effectively span distinct operational intents rather than collapsing into a single mean trajectory.

This evidence validates our architectural design: by offloading complex spatial reasoning entirely to the Context Encoder and GMM, the shallow kinematic decoder focuses solely on execution. Consequently, the infinitesimal generator alignment ensures that individual sampled modes remain strictly bound to physically viable kinematics, preserving both diversity and physical compliance.

J Further Qualitative Comparison

Robustness against Dynamic Physical Parameter Mismatch. To validate the robustness of our method against dynamic environmental perturbations, we conducted further closed-loop evaluations in CARLA under extreme conditions, including slippery roads and heavy rain. Unlike purely visual noise, slippery surfaces drastically alter the actual vehicle kinematics by severely reducing tire friction, creating a sudden mismatch with the pre-defined deterministic vector field $\mathcal{F}_{veh}$, as shown in Fig. 7. While conventional methods often suffer from kinematic collapse or drift off-road under such unmodeled dynamic friction drops, our approach consistently maintains safe and physically plausible trajectory generation. This demonstrates that aligning the infinitesimal generator inherently preserves robust kinematic bounds, even when real-world physical parameters dynamically diverge from the nominal model.

NAVSIM Dataset. We further visualize additional generated trajectories on the NAVSIM dataset to provide a more comprehensive qualitative assessment of the EC-IGM.



Figure 8: Trajectory visualizations of right-turn scenarios generated by the EC-IGM on the NavSim dataset.



Figure 9: Trajectory visualizations of multi-vehicle driving scenarios generated by the EC-IGM on the NavSim dataset.



Figure 10: Trajectory visualizations of multi-vehicle driving scenarios generated by the EC-IGM on the NavSim dataset.



Figure 11: Trajectory visualizations of left-turn scenarios generated by the EC-IGM on the NavSim dataset.