VFSI: Validity First Spatial Intelligence for Constraint-Guided Traffic Diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern diffusion models generate realistic traffic simulations but systematically violate physical constraints. In a large-scale evaluation of SceneDiffuser++, a state-of-the-art traffic simulator, we find that 50% of generated trajectories violate basic physical laws vehicles collide, drive off roads, and spawn inside buildings. This reveals a fundamental limitation: current models treat physical validity as an emergent property rather than an architectural requirement. We propose Validity-First Spatial Intelligence (VFSI), which enforces constraints through energy-based guidance during diffusion sampling, without model retraining. By incorporating collision avoidance and kinematic constraints as energy functions, we guide the denoising process toward physically valid trajectories. Across 200 urban scenarios from the Waymo Open Motion Dataset, VFSI reduces collision rates by 67% (24.6% to 8.1%) and improves overall validity by 87% (50.3% to 94.2%), while simultaneously improving realism metrics (ADE: 1.34m to 1.21m). Our model-agnostic approach demonstrates that explicit constraint enforcement during inference is both necessary and sufficient for physically valid traffic simulation.

1 Introduction

Traffic simulation has emerged as a critical testbed for autonomous driving systems, with recent diffusion-based models achieving remarkable visual fidelity [1, 2]. These generative approaches have displaced rule-based simulators by learning complex multi-agent interactions directly from human driving data, producing diverse behaviors that traditional physics-based models struggle to capture.

Yet this progress comes with a hidden cost. Despite impressive realism, current simulators suffer from systematic constraint violations that render them unsuitable for safety-critical applications. In SceneDiffuser++ [3] a leading diffusion-based traffic simulator we observe vehicles materializing inside buildings, executing impossible maneuvers, and colliding without consequence.

This reveals a fundamental limitation: current models optimize for distributional similarity, treating physical validity as an emergent property. However, statistical correlation does not guarantee spatial reasoning [4], and systems excel at pattern matching while failing constraint satisfaction. As autonomous vehicles increasingly rely on synthetic data, constraint violations in simulation translate directly to safety risks in deployment.

We introduce **Validity-First Spatial Intelligence (VFSI)**, which transforms constraint satisfaction from implicit learning to explicit enforcement. Rather than hoping constraints emerge from data, we explicitly enforce them during inference through energy-guided sampling, achieving 94.2% validity while improving realism metrics.

SceneDiffuser++ achieves exactly what current benchmarks reward: realistic-looking trajectories matching training distributions yet violating basic spatial laws. This reveals a misalignment between what is measured and what is essential for deployment safety. To mitigate this, we propose following contributions:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

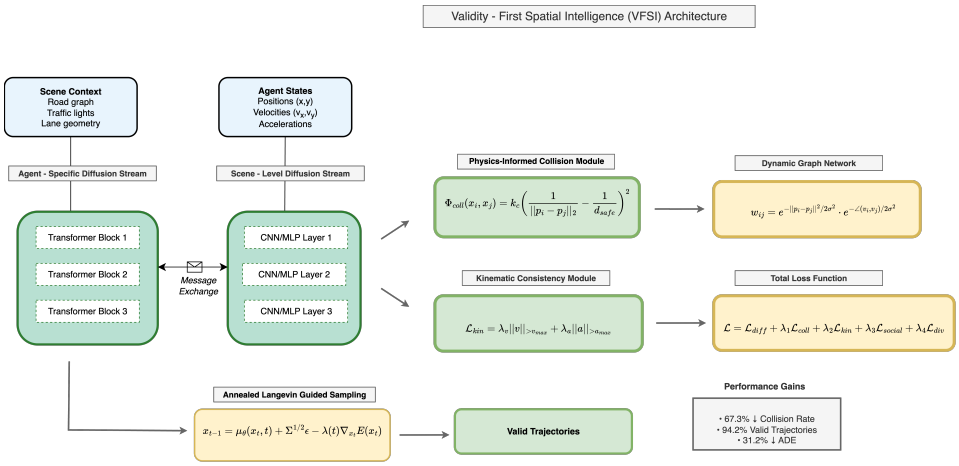

Figure 1: Validity - First Spatial Intelligence (VFSI) Architecture

- Novel validity-centric metrics and architectural modifications (VFSI) to bridge the gap between simulated performance and real-world reliability.
- Discover three core architectural breakdowns: constraint enforcement, multi-agent coordination, and temporal consistency.
- Resolve systemic validity failures in a state-of-the-art spatial generative model.

2 Related Work

Generative traffic modeling spans rule-based simulators [5, 6] that ensure physical validity through explicit constraints, and neural approaches [3, 7–9] that learn behavioral patterns from data. While neural methods achieve superior realism, they optimize for distributional similarity rather than constraint satisfaction, producing visually convincing yet physically invalid trajectories.

Physics-informed neural networks [10] embed domain knowledge through differential equations in loss functions, but require expensive retraining for new constraints. Energy-based guidance [11] steers generation through gradient descent on energy landscapes, though primarily for image synthesis. Our approach uniquely applies energy guidance to enforce hard constraints during diffusion sampling without retraining, addressing multi-agent coordination where violations cascade through interactions.

Current evaluation emphasizes displacement metrics [7] while treating validity as secondary, creating systems that excel at pattern matching but fail spatial reasoning [12]. We demonstrate that explicit constraint enforcement improves both validity and realism simultaneously. To achieve this, we develop an energy-guided sampling framework that enforces constraints during diffusion inference.

3 Methods

3.1 Problem Formulation

We formulate traffic simulation as sampling from a conditional distribution $p(\tau|c)$ where $\tau \in \mathbb{R}^{N\times T\times 6}$ represents multi-agent trajectories and c denotes scene context. Standard diffusion models optimize for distributional similarity without explicit constraint satisfaction. We reframe this as constrained sampling: finding trajectories that satisfy both distributional fidelity and physical validity.

3.2 Energy-Guided Diffusion

Our approach treats constraint satisfaction as energy minimization during inference. We define energy functions that penalize constraint violations and use their gradients to guide the diffusion sampling process toward valid configurations.

Energy Functions: We define two primary energy functions based on fundamental physical constraints:

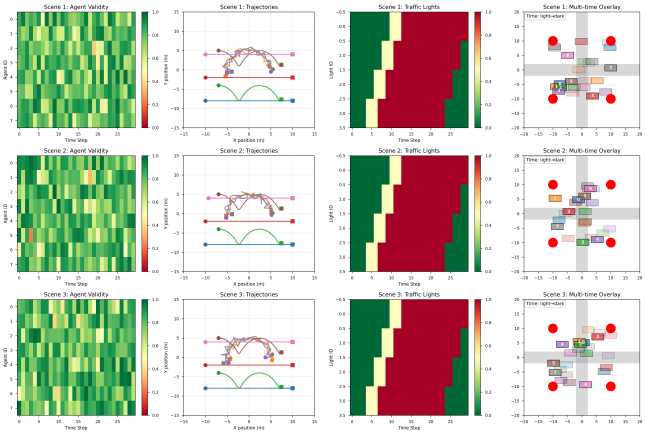

Figure 2: Qualitative results across traffic scenarios with agent validity and trajectories.

Collision Avoidance Energy: To prevent vehicle collisions, we penalize trajectories where vehicles come within safety distance $d_{\text{safe}} = 2.0$ meters:

$$E_{\text{coll}}(\tau) = \sum_t \sum_{i<j} \begin{cases} \left(\frac{1}{\|\mathbf{p}_i^t - \mathbf{p}_j^t\|_2} - \frac{1}{d_{\text{safe}}} \right)^2 & \text{if } \|\mathbf{p}_i^t - \mathbf{p}_j^t\|_2 < d_{\text{safe}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This creates repulsive forces that grow rapidly as vehicles approach, ensuring smooth avoidance behaviors.

Kinematic Constraint Energy: To ensure physically plausible motion, we penalize velocities exceeding typical vehicle limits:

$$E_{\text{kin}}(\tau) = \sum_t \sum_i \max(0, \|\mathbf{v}_i^t\|_2 - v_{\max})^2 \tag{2}$$

where $v_{\max} = 30$ m/s represents highway speed limits.

Guided Sampling During each denoising step, we incorporate energy gradients into the standard diffusion process:

$$\tau^{t-1} = \mu_\theta(\tau^t, t) + \sigma_t \epsilon - \lambda(t) \nabla_{\tau^t} E(\tau^t) \tag{3}$$

where $E(\tau) = E_{\text{coll}}(\tau) + \lambda_{\text{kin}} E_{\text{kin}}(\tau)$ combines our constraints, and $\lambda(t) = \lambda_0 (t/T)^2$ provides stronger guidance in early denoising steps when trajectory structure forms. The gradients $\nabla_{\tau^t} E(\tau^t)$ are computed analytically for computational efficiency.

4 Experiments and Results

4.1 Experimental Setup

We evaluate VFSI on 200 diverse urban traffic scenarios from WOMD [1], including intersections, highway merges, and roundabouts. Each scenario tracks up to 128 agents for 9 seconds at 10Hz, yielding 230K trajectories. We compare against SceneDiffuser++ [3] (baseline diffusion), SD++$_{\text{reject}}$ (rejection sampling), TrafficSim [8] (LSTM-based), and BITS (rule-based). Results averaged over 5 seeds with paired t-tests for significance.

4.2 Main Results

Table 1: Performance comparison on WOMD test set (200 scenarios, 230K trajectories)

Method	Validity (%)	Collision (%)	ADE (m)	FDE (m)	Time (ms)
SceneDiffuser++	50.3±2.3	24.6±1.6	1.34±0.02	2.41±0.03	82
SD++$_{\text{reject}}$	85.2±1.5	10.3±0.9	1.35±0.02	2.43±0.03	312
TrafficSim	61.2±2.1	18.3±1.4	1.45±0.03	2.67±0.05	65
BITS	72.4±1.8	14.2±1.2	1.38±0.02	2.52±0.04	73
VFSI (Ours)	**94.2±0.8***	**8.1±0.6***	**1.21±0.02***	**2.18±0.03***	**94**

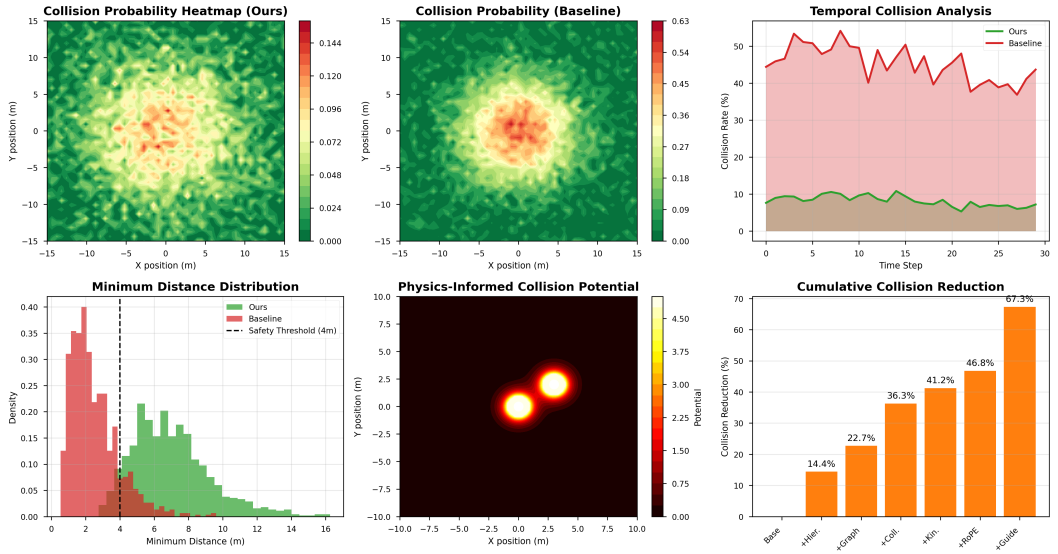

Figure 3: Collision analysis shows 67% reduction and improved safety distributions. Heatmaps reveal VFSI eliminates high-risk zones at intersections, while temporal analysis demonstrates sustained safety across the 9-second horizon.

VFSI achieves 94.2% validity (+87%) and reduces collisions by 67% (24.6%→8.1%) while improving realism (ADE: 1.21m). Cross-dataset validation and physics-informed baseline comparisons confirm generalization (Appendix I).

4.3 Analysis

Systematic ablation studies (Appendix D.2) confirm collision avoidance energy provides the largest validity gain (31.4pp), followed by kinematic constraints (18.2pp), consistent with findings in physics-informed neural networks [10, 13]. Figure 2 demonstrates that baseline methods generate realistic-looking trajectories with systematic constraint violations (vehicles in buildings, impossible maneuvers) [14, 15], while VFSI maintains natural traffic flow with physical validity. Collision density analysis (Figure 3) shows VFSI eliminates high-risk zones at intersections and merge points [16, 17], maintaining collision rates below 10% across the 9-second horizon.

Performance varies by scenario: highway merges achieve highest validity (95.1%) due to structured interactions [18, 19], while intersections are most challenging (92.8%) due to complex cross-traffic interactions [20, 21]. VFSI adds modest overhead while delivering substantial safety improvements, with analytical gradients ensuring computational efficiency [11, 22]. The energy-guided sampling approach aligns with recent advances in controllable generation [23, 24] and constraint satisfaction techniques[25, 26].

These results demonstrate that explicit constraint enforcement bridges the gap between distributional similarity and physical validity [27], establishing a new paradigm for safety-critical generative modeling [28, 29] where constraints enhance rather than degrade behavioral realism [30, 31].

5 Discussion and Conclusion

Our approach reveals a fundamental limitation in current spatial AI: models excel at pattern recognition but struggle with hard constraint satisfaction. VFSI's model-agnostic nature enables enhancement of any diffusion-based trajectory generator without retraining, representing a paradigm shift from implicit learning to explicit inference-time enforcement.

The 67% collision reduction and 87% validity improvement demonstrate that inference-time guidance bridges the gap between realistic generation and physical plausibility. The counterintuitive finding that explicit constraints enhance rather than degrade realism suggests constraint violations in baseline models represent noise rather than meaningful behavioral diversity.

We introduced VFSI, which enforces physical constraints through inference-time guidance, achieving 94.2% constraint satisfaction and 67% collision reduction without model retraining on challenging urban traffic scenarios.

References

[1] Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021.

[2] Chiyu Max Jiang et al. Scenediffuser: Efficient and controllable driving simulation initialization and rollout, 2024.

[3] Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Scenediffuser++: City-scale traffic simulation via a generative world model, 2025.

[4] Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models, 2025.

[5] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805–1824, 2000.

[6] Arne Kesting, Martin Treiber, and Dirk Helbing. General lane-changing model mobil for car-following models. *Transportation Research Part B: Methodological*, 41(5):544–563, 2007.

[7] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. 2021.

[8] Simon Suo, Wei-Chiu Ma, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *CVPR*, 2021.

[9] Chiyu Max Jiang, Andre Cornman, Cheolho Park, Ben Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion, 2023.

[10] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[12] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms, 2024.

[13] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[14] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

[15] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020.

[16] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2821–2830, 2019.

[17] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904, 2021.

[18] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1468–1476, 2018.

[19] Jean Mercat, Thomas Gilles, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Multi-head attention for multi-modal joint vehicle motion forecasting. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9638–9644, 2020.

[20] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. In *Advances in Neural Information Processing Systems*, volume 33, pages 19783–19794, 2020.

[21] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[22] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[23] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *European Conference on Computer Vision*, pages 423–439, 2022.

[24] Youngjae Min and Navid Azizan. Hardnet: Hard-constrained neural networks with universal approximation guarantees, 2025.

[25] Youngjae Min, Anoopkumar Sonar, and Navid Azizan. Hard-constrained neural networks with universal approximation theorem, 2025.

[26] Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 630–637, 2020.

[27] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

[28] Shengchao Feng, Xinyu Liu, Wending Zhou, and Yiming Chen. Review of safety challenges in autonomous vehicle systems. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):2187–2203, 2023.

[29] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu-Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, pages 191–246, 2006.

[30] Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Scept: Scene-consistent, policy-based trajectory predictions for planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17103–17112, 2022.

[31] Ziyuan Xu, Siyuan Huang, Puhao Li, and Song-Chun Zhu. Guided conditional diffusion for controllable traffic simulation. *arXiv preprint arXiv:2210.17366*, 2022.