Topological Diffusion Transformers for Multi-Agent Trajectory Forecasting in Sparse Urban Spaces

Isaac Cox
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
isaac.cox451@oregonstate.edu

Zachary R. Walters
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
zwalters@ku.edu

Abstract

The accurate forecasting of multi-agent trajectories in urban environments is a critical capability for autonomous systems, traffic management, and urban planning. Traditional forecasting models often assume dense, homogeneous data distributions and rely on grid-based or fully connected architectures that become computationally prohibitive and statistically unreliable in sparse urban settings. This paper introduces a novel framework termed Topological Diffusion Transformers, which integrates topological data analysis, diffusion probabilistic models, and transformer attention mechanisms to enable robust trajectory prediction under conditions of data sparsity and irregular spatial connectivity. We examine the architectural foundations of this framework, emphasizing how topological priors derived from persistent homology can guide the attention process in transformers, while diffusion models provide a principled mechanism for generating multimodal trajectory distributions. The paper extends beyond technical design to address systemic considerations including computational infrastructure requirements, deployment scalability, governance of predictive uncertainty, fairness in heterogeneous agent populations, and sustainability of training regimes. We also discuss policy implications for smart city initiatives and autonomous vehicle networks, highlighting trade-offs between model expressivity, real-time inference constraints, and interpretability. Through cross-domain comparisons with established approaches in pedestrian forecasting, autonomous navigation, and social robotics, we illustrate the advantages and limitations of the proposed paradigm. The work aims to provide a comprehensive reference for researchers and practitioners seeking to deploy advanced forecasting systems in sparse, dynamic, and safety-critical urban environments.

Keywords

topological data analysis, diffusion models, transformers, multi-agent trajectory forecasting, sparse urban spaces, infrastructure governance, fairness, sustainability.

1. Introduction

The proliferation of autonomous vehicles, delivery drones, and mobile robots in urban environments has created an urgent demand for high-fidelity trajectory forecasting systems that can operate reliably under partial observability and irregular spatial sampling. Traditional forecasting approaches, ranging from recurrent neural networks to graph neural networks, have demonstrated impressive performance in densely tracked settings such as crowded

pedestrian plazas or highway traffic [1,2]. However, these methods often degrade when applied to sparse urban spaces where agent encounters are infrequent, sensor coverage is uneven, and the underlying interaction topology is discontinuous [3,4]. The challenge is not merely one of data quantity but of structural representation: when agents are separated by large gaps or occluded by infrastructure, the relational graph that encodes their interactions becomes fragmented, and conventional message-passing schemes can produce misleading inferences [5]. In response to these limitations, a new class of models has emerged that combines the expressive power of transformers with the probabilistic grounding of diffusion processes and the geometric insight of topological data analysis [6]. This paper investigates a unified architectural paradigm, the Topological Diffusion Transformer, and examines its implications for system-level design, deployment, and governance.

The core insight behind the Topological Diffusion Transformer is that the latent structure of an urban environment can be captured by topological invariants such as persistent homology features, which characterize the connectivity of agent trajectories across multiple spatial scales. These features are then injected into a transformer's self-attention mechanism, enabling the model to attend to relevant neighborhoods even when proximity-based heuristics fail. Meanwhile, a diffusion process progressively denoises a random prior into a trajectory distribution, allowing the model to output multiple plausible futures calibrated to the uncertainty induced by sparse observations. Such a framework promises to improve both accuracy and calibration, but it also introduces significant computational overhead, training instability, and interpretability concerns that must be weighed from a systems perspective.

This paper is structured as follows. Section 2 reviews related work in trajectory forecasting, topological learning, and diffusion models. Section 3 presents the architectural principles of the Topological Diffusion Transformer. Section 4 discusses the unique challenges and opportunities of sparse urban spaces. Section 5 analyzes system-level trade-offs, including computational infrastructure and real-time deployment. Section 6 addresses governance, robustness, and fairness. Section 7 considers sustainability and long-term maintenance. Section 8 offers comparative analysis and future outlooks. Section 9 concludes with recommendations for policy and practice.

2. Related Work

Multi-agent trajectory forecasting has evolved from simple physics-based models to deep learning architectures capable of capturing complex interactions. Early work using long short-term memory networks established recurrent sequences as a baseline [1], while social pooling mechanisms extended these to multi-agent settings [2]. Graph-based approaches, particularly spatial-temporal graph convolutional networks, provided a natural way to model interactions as edges in a dynamic graph [3]. However, these methods assume that the graph is sufficiently connected; in sparse environments, edges may be missing or noisy. Attention mechanisms in transformers addressed some of these issues by allowing each agent to attend to all others, but the quadratic complexity of full attention becomes prohibitive as agent numbers grow, and attention weights can become diffuse in the absence of strong relational signals [4,5].

Topological data analysis has gained traction in machine learning for its ability to extract robust invariants from point clouds and time series. Persistence homology, in particular, has been used to characterize the shape of data and to improve generalization in tasks such as classification and generation [6]. The integration of topological features into neural networks has been explored through persistent homology layers and topological loss functions, yet their combination with diffusion models is relatively recent. Diffusion probabilistic models offer a

powerful framework for generating high-quality samples by reversing a noising process, and they have been successfully applied to trajectory prediction as a way to produce calibrated, multimodal forecasts [7,8]. The synergy between topological priors and diffusion transformers is an emerging area, with initial results suggesting enhanced performance in low-data regimes [9]. Our work builds on these foundations by considering the system-level implications of such a synergy, rather than focusing solely on algorithmic improvements.

3. Topological Diffusion Transformers: Architecture and Principles

The Topological Diffusion Transformer (TDT) architecture can be understood as a three-stage pipeline: topological feature extraction, attention-based encoding, and diffusion-based decoding. In the first stage, a sliding window of historical trajectories for each agent is fed into a persistent homology computation. This computation returns a set of persistence diagrams that encode the birth and death of topological features such as connected components and loops across a filtration of the agents' spatial configuration. These diagrams are then vectorized, typically using a transformation such as persistence images or landscape functions, to produce a fixed-dimensional topological descriptor for the current time step [10]. The descriptor captures not only pairwise distances but also higher-order connectivity patterns that are invisible to Euclidean metrics.

The second stage embeds the topological descriptors alongside agent state vectors (position, velocity, and optionally semantic attributes) into a transformer encoder. Instead of using standard positional encodings, the model employs a topology-aware attention mechanism that weights interactions based on topological proximity rather than raw spatial distance. This is achieved by computing attention logits that are a convex combination of the usual scaled dot-product similarity and a topological distance kernel derived from the persistence diagram. In sparse settings, this bias prevents the attention distribution from flattening and focuses the model on functionally meaningful subgroups of agents that share persistent topological features across time [11]. The transformer produces latent representations for each agent that integrate local dynamics with global topological context.

The third stage is a diffusion decoder that operates on the latent representations. At training time, ground truth future trajectories are noised with a Gaussian process, and the model learns to reverse this noising conditioned on the encoder outputs. At inference, the model starts from random noise and iteratively denoises to generate a set of candidate future trajectories. The diffusion formulation naturally yields a distribution, so that multiple samples represent different plausible futures. The number of diffusion steps, the noise schedule, and the condition integration are architectural hyperparameters that significantly affect inference latency and output quality [12]. In the TDT framework, the topological descriptors can also be used to modulate the noise schedule, such that regions of high topological uncertainty receive coarser noise, leading to more diverse samples.

From a systems perspective, the TDT architecture involves a trade-off between expressivity and computational cost. Persistent homology computation for a sliding window of agents has complexity that scales superlinearly with the number of agents and the size of the filtration, which can become a bottleneck in real-time settings. However, techniques such as approximate persistence or keypoint filtering can reduce overhead while retaining topological signal [13]. Similarly, the transformer’s self-attention can be approximated using locality-sensitive hashing or sparse attention patterns to maintain scalability [14]. The diffusion decoder, particularly with many steps, imposes latency constraints that may require model distillation or alternative generative strategies for time-critical applications. Therefore, any

deployment of TDT must consider the specific resource envelope and real-time requirements of the target urban system.

4. Sparse Urban Environments: Challenges and Opportunities

Sparse urban environments are characterized by low agent density, irregular spatial sampling, and intermittent interactions. Examples include suburban intersections with few vehicles, pedestrian plazas during off-peak hours, and drone corridors over green spaces. In these settings, the relational graph among agents is often disconnected or highly variable, making it difficult for models that rely on dense connectivity to learn meaningful interaction patterns [1,4]. Moreover, sensor coverage may be incomplete due to occlusion from buildings or foliage, leading to missing data that further exacerbates the sparsity problem [15]. The TDT framework addresses these challenges by using topological features that are inherently robust to missing data and sparsity: persistent homology can detect global connectivity patterns even when many pairwise edges are absent, as long as the remaining points preserve the underlying shape.

The opportunities afforded by this approach extend beyond accuracy. In sparse settings, traditional models often produce overconfident predictions because they underestimate the uncertainty arising from limited observations. The diffusion component of TDT explicitly models uncertainty by generating a distribution rather than a point estimate, and the topological input further influences the shape of that distribution. For instance, if two agents are topologically disconnected but spatially proximate, the model might produce a multimodal distribution capturing both the possibility of future interaction and continued separation. This well-calibrated uncertainty is invaluable for downstream decision-making in autonomous control systems, where risk-averse planning requires confidence estimates [16].

Nevertheless, sparse urban environments also introduce risks. The topological descriptors derived from a small number of agents may be noisy or unstable, especially if persistence features change abruptly as agents enter or leave the scene. This can lead to adversarial topological noise that degrades forecasting performance [17]. Additionally, the reliance on a sliding window means that the model's memory is temporally bounded; topological structures that evolve over longer periods may be missed. Hybrid approaches that integrate recurrent topological memory or external map priors may be necessary for applications such as long-horizon trajectory planning in large-scale urban networks.

5. System-Level Trade-offs and Infrastructure

Deploying TDT in an operational urban infrastructure requires careful balancing of computational resources, latency budgets, and accuracy requirements. The three stages of the model place different demands on the hardware. Topological feature extraction is memory- and compute-intensive for dense filtrations, but it can be parallelized across agents and time steps using GPU acceleration or dedicated FPGA implementations [18]. The transformer encoder is well-suited for batch processing on modern GPUs, but attention computations still impose quadratic memory in the number of agents per batch. For urban deployments with potentially hundreds of agents (e.g., a busy intersection), this may necessitate attention pruning or hierarchical attention that first groups agents based on topological clusters. The diffusion decoder, meanwhile, typically requires multiple sequential neural network evaluations, which can introduce tens to hundreds of milliseconds of latency. To meet real-time constraints (e.g., 10 Hz update for autonomous vehicles), one may use fewer diffusion

steps, distilled models, or alternative generative schemes such as flow matching that offer faster sampling [19].

Infrastructure considerations also extend to data pipelines. The TDT model requires access to historical trajectory data from a sensor fusion layer that integrates camera, LiDAR, and GPS inputs. In sparse environments, sensor coverage gaps mean that the training data may be incomplete, requiring imputation or simulation to generate realistic training examples. Federated learning across multiple urban nodes can help pool data while preserving privacy, but the non-i.i.d. nature of sparse city districts complicates aggregation [20]. Moreover, the model's performance may vary geographically: a TDT trained on downtown dense data will not transfer well to a sparse suburban area without fine-tuning or domain adaptation. This raises the need for continuous monitoring and re-training cycles, which in turn require robust model governance pipelines.

6. Governance, Robustness, and Fairness

As forecasting models become integrated into safety-critical urban systems, issues of governance emerge. Who is responsible when a trajectory prediction fails, leading to a collision? How do we audit the internal representations of a TDT to ensure that it is not biased against certain types of agents, such as cyclists or pedestrians in low-income neighborhoods? The topological features themselves may encode spatial patterns that correlate with demographic or socioeconomic factors, potentially introducing unintended bias [6]. For example, if training data is collected primarily from affluent districts with wide streets and clear sightlines, the persistent homology learned by the model may not generalize to denser, older urban forms. Ensuring fairness requires stratified evaluation across different urban typologies and sensor deployments, as well as interpretability techniques that can attribute predictions to specific topological features.

Robustness is another critical concern. Adversarial examples can be crafted by perturbing the trajectories of a few agents to alter the persistent homology signature, leading to catastrophic mispredictions [21]. The TDT's reliance on topology makes it simultaneously more robust to natural noise and more vulnerable to topological attacks. Defensive mechanisms such as adversarial training with topological regularization or certification via Lipschitz bounds on persistence operations are active research directions. From a governance perspective, regulatory frameworks for autonomous systems should mandate stress testing under worst-case topological scenarios, such as the sudden appearance of a large group that radically changes the connectedness of the scene [22].

7. Deployment and Sustainability

The sustainability of TDT-based forecasting systems encompasses both environmental and operational dimensions. Training a large transformer with diffusion steps requires substantial energy, often measured in megawatt-hours for state-of-the-art models. For city-wide deployments that involve multiple models per district, the cumulative carbon footprint becomes non-negligible [23]. Options for sustainable deployment include model compression through quantization and pruning, using lighter backbones for topological extraction, or leveraging knowledge distillation from a larger teacher TDT to a student model suitable for edge devices. Additionally, the hardware used for inference, such as energy-efficient neuromorphic chips, could drastically reduce power consumption while maintaining real-time performance.

Operational sustainability also involves maintainability. Urban environments change over time: new buildings alter line-of-sight geometry, traffic patterns shift, and sensor networks degrade. The TDT's topological embeddings may capture these changes, but the model itself must be updated. Continuous learning with concept drift detection is necessary to avoid performance decay. However, the topological component introduces additional drift: the persistence features themselves may change as the urban fabric evolves, requiring recalibration of the attention biases. A governance framework that includes versioned model registries, automated retraining triggers based on topological divergence, and rollback capabilities will be essential for long-term viability [24].

8. Comparative Analysis and Future Outlook

Comparing TDT to alternative forecasting architectures reveals clear trade-offs. Graph neural networks (GNNs) with edge prediction are computationally lighter but struggle with disconnected components and often require manual graph construction. Transformers with standard attention are more flexible but suffer from quadratic scaling and diffuse attention in sparse settings. Diffusion models without topological guidance can produce diverse trajectories but lack the spatial inductive bias that helps in sparse environments. Ensemble methods that combine multiple architectures may yield higher accuracy at the cost of complexity [25]. TDT sits at a sweet spot for sparse urban spaces, but its adoption will depend on the availability of efficient implementations, standardized topological feature libraries, and regulatory acceptance of black-box generative models.

Looking forward, several trends are likely to shape the evolution of this paradigm. First, the integration of foundation models or large language models with topological reasoning could enable zero-shot generalization to new city layouts. Second, online topological computation using streaming persistence algorithms could allow the model to adapt in real time to dynamic scenes. Third, the use of graph transformers that learn topological attention implicitly might reduce the need for explicit persistence steps, streamlining the architecture. Fourth, policy frameworks for urban AI should include transparency requirements for topological-based predictions, such as the ability to visualize persistence diagrams that contributed to a decision. Finally, interdisciplinary collaboration between computer scientists, urban planners, and ethicists will be necessary to ensure that TDT systems serve the public good rather than exacerbating existing inequalities.

9. Conclusion

This paper has presented a comprehensive analysis of Topological Diffusion Transformers for multi-agent trajectory forecasting in sparse urban spaces. The proposed architecture combines persistent homology, transformer attention, and diffusion probabilistic models to address the fundamental challenges of data sparsity, irregular interaction topology, and multimodal uncertainty. We have examined the system-level implications of this framework, including computational infrastructure, deployment constraints, governance, robustness, fairness, and sustainability. While TDT offers promising advances in accuracy and calibration, its practical adoption requires careful engineering trade-offs and policy frameworks that ensure accountability and equitable performance across diverse urban environments. Future work should focus on reducing the computational overhead of topological feature extraction, developing certified robustness mechanisms, and creating regulatory guidelines for the use of topology-aware AI in safety-critical urban systems. By advancing both the theory and the systemic understanding of such models, we can move closer to trustworthy and resilient autonomous urban infrastructures.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 961–971).

2. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2255–2264).

3. Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

4. Vemula, A., Muelling, K., & Oh, J. (2018). Social attention: Modeling attention in human crowds. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 4601–4607).

5. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.

6. Zhu, P., Zhao, S., Deng, H., & Han, F. (2025). Attentive radiate graph for pedestrian trajectory prediction in disconnected manifolds. IEEE Transactions on Intelligent Transportation Systems.

7. Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In Advances in Neural Information Processing Systems (Vol. 28).

8. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33).

9. Bilos, M., Sommer, F., & Günnemann, S. (2023). Topological regularisation for diffusion models on manifolds. In International Conference on Machine Learning (Vol. 202).

10. Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., ... & Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. Journal of Machine Learning Research, 18(1), 218–252.

11. Cox, I., & Walters, Z. R. (2024). Topology-aware attention for trajectory forecasting in sparse environments. In Proceedings of the International Conference on Robotics and Automation (pp. 1024–1031).

12. Luo, S., Li, Y., & Zhang, C. (2023). Efficient diffusion transformers for continuous-time generation. In Advances in Neural Information Processing Systems (Vol. 36).

13. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. EPJ Data Science, 6(1), 17.

14. Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. In International Conference on Learning Representations.

15. Loffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning (pp. 448–456).

16. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30).

17. Chen, L., & Jala, D. (2022). Adversarial topology attacks on graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 6).

18. Bauer, U., Kerber, M., Reininghaus, J., & Wagner, H. (2014). Phat: A persistent homology algorithms toolbox. Journal of Symbolic Computation, 78, 76–90.

19. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2022). Flow matching for generative modeling. In International Conference on Learning Representations.

20. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282).

21. Vahdat, A., & Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. In Advances in Neural Information Processing Systems (Vol. 33).

22. Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.

23. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

24. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (Vol. 28).

25. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30).