GeoWorldSim: Cross-Modal Geographic World Modeling for Embodied Urban Navigation via Diffusion-Based Spatial Scene Generation

Karan Dutta
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
contactkaran@uc.edu

Leif R. Mills
School of Computing, Clemson University, Clemson, SC, USA.
leifm@clemson.edu

Pascal Terry
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
pascalwork@uab.edu

Abstract

Embodied urban navigation requires agents to perceive, reason, and act within complex, dynamic city environments that are inherently multimodal and geographically structured. Existing approaches often rely on static map representations or single-modality sensor inputs, which fail to capture the richness of geographic context and the variability of real-world scenes. This paper introduces GeoWorldSim, a cross-modal geographic world modeling framework that integrates satellite imagery, street-level panoramas, LiDAR point clouds, and textual semantic labels through a diffusion-based spatial scene generation mechanism. The system constructs a unified, updatable world model that enables embodied agents to simulate plausible future views from unvisited viewpoints, thereby improving navigation robustness under occlusion, sensor noise, and environmental change. We present the architectural principles of GeoWorldSim, focusing on the trade-offs between geometric accuracy and perceptual realism, the governance of geospatial data fidelity across scales, and the policy implications of deploying such models in public urban infrastructures. The diffusion backbone is conditioned on geographic coordinates and partial observations, allowing the generation of spatially consistent scenes without requiring dense supervision. We discuss the sustainability of training such large-scale models, the fairness implications of geographic coverage biases, and the robustness of generated scenes under distribution shift. Through analytical case studies and cross-domain comparisons with prior world modeling systems, we demonstrate that GeoWorldSim offers a scalable and principled foundation for next-generation embodied navigation in urban environments. The paper concludes with forward-looking perspectives on the integration of geospatial world models with real-time edge computing and civic data governance frameworks.

Keywords

geographic world modeling, embodied navigation, diffusion models, cross-modal fusion, urban scene generation, spatial AI, geospatial data governance, sustainable AI infrastructure.

1. Introduction

Embodied navigation in urban environments poses a fundamental challenge for artificial intelligence systems: agents must operate in spaces that are large, heterogeneous, and subject to continuous change. Traditional approaches to navigation rely on precomputed maps, geometric Simultaneous Localization and Mapping (SLAM), or reinforcement learning policies trained in simulation. However, these methods exhibit critical limitations when deployed at geographic scale. Static maps cannot capture transient events such as road construction, pedestrian gatherings, or seasonal vegetation changes. SLAM-based methods suffer from drift over long trajectories and degrade in visually repetitive urban canyons. Reinforcement learning policies often fail to generalize to unseen city layouts because they lack a structured representation of the world that can be queried and updated [1], [2]. The emergence of world modeling as a paradigm offers a way forward: instead of treating perception and planning as separate modules, a world model learns a generative representation of the environment that can be used to predict future observations, reason about occluded regions, and simulate alternative trajectories [3].

GeoWorldSim is proposed as a cross-modal geographic world modeling system that explicitly addresses the urban navigation setting. The core insight is that urban environments are richly structured by geographic coordinates, satellite overhead imagery, and ground-level panoramic observations. By fusing these modalities into a unified latent space, and by leveraging diffusion-based generative processes to synthesize novel views from arbitrary viewpoints, GeoWorldSim enables an embodied agent to maintain a dynamic mental model of its surroundings. This paper describes the system architecture, the design choices that balance fidelity and efficiency, and the broader implications for governance, sustainability, and fairness. We argue that geographic world models represent a new class of socio-technical infrastructure, where the data used for training and the outputs generated have direct consequences for public safety, privacy, and equitable access to navigation services [4], [5].

2. Related Work

The research landscape relevant to GeoWorldSim spans several domains: geographic scene understanding, cross-modal representation learning, diffusion-based generative models, and embodied navigation. Early work on geographic scene understanding focused on classifying satellite imagery and extracting road networks, but these systems lacked the ability to synthesize ground-level views [6]. More recently, cross-modal learning methods have aligned satellite and street-level images through contrastive objectives, enabling tasks such as cross-view image retrieval and localization [7]. However, these approaches typically operate at the level of image pairs and do not construct a persistent world model that can be queried for arbitrary viewpoints.

Diffusion models have emerged as a powerful class of generative models capable of producing high-fidelity images conditioned on various inputs [8]. In the geospatial domain, diffusion has been applied to pan-sharpening, cloud removal, and super-resolution of satellite imagery [9]. However, its application to synthesizing street-level views conditioned on geographic context and partial observations is relatively new. The work leading to GeoWorldSim builds on conditional diffusion frameworks that incorporate spatial coordinates as conditioning signals [10]. In parallel, embodied navigation has advanced through the use of neural SLAM and learned policy architectures that aggregate observations into topological or metric maps [11]. Yet these map representations are often discrete and lack the generative capacity to imagine unseen perspectives, which is critical for planning under uncertainty.

GeoWorldSim distinguishes itself by unifying these threads into a single framework that not only generates scenes but also maintains a persistent, updatable world model. The model is designed to be trained on globally distributed urban data, raising important questions about geographic coverage and dataset bias that have been highlighted in prior fairness analyses of geospatial AI [12], [13]. Furthermore, the generative nature of the model introduces new challenges for verification and validation, as synthesized scenes must be consistent with the underlying geometry of the environment while remaining perceptually plausible [14].

3. System Architecture

The architecture of GeoWorldSim is organized around three principal components: a geographic encoder, a cross-modal fusion network, and a diffusion-based scene generator. The geographic encoder takes as input a query location specified by latitude, longitude, and optionally altitude, along with a set of metadata such as time of day, weather, and season. This encoder learns a compact latent representation of the geographic context using a positional encoding scheme that is invariant to global rotation and translation [15]. The cross-modal fusion network integrates multiple observation modalities that may be available at the query location: overhead satellite imagery, one or more street-level panoramas, LiDAR point cloud projections, and textual semantic descriptions (e.g., "busy intersection with traffic lights"). Each modality is processed through a dedicated encoder, and the resulting embeddings are aggregated via a cross-attention mechanism that learns to weight modalities according to their informativeness [16]. The fused latent code serves as the conditioning input to the diffusion generator.

The diffusion generator is based on a denoising diffusion probabilistic model (DDPM) architecture, extended to condition on the fused geographic latent and on a sparse set of observed pixel values from the target view, if available. The generator is trained to reverse a fixed noise process applied to panoramic images at a resolution of 1024x512 pixels. During inference, the agent can query the world model at any unvisited viewpoint within the trained geographic extent. The generator synthesizes a plausible panorama that respects the known geometric layout inferred from satellite and LiDAR data while also being perceptually coherent with the street-level observations used during conditioning [17]. This capability is particularly valuable for navigation tasks where the agent must anticipate what lies around a corner or behind a building before actually moving there.

A critical architectural trade-off exists between the quality of synthesized scenes and the computational cost of training and inference. High-resolution diffusion models demand substantial GPU memory and inference latency, which may be incompatible with real-time deployment on resource-constrained edge devices. GeoWorldSim addresses this through a hierarchical generation strategy: a coarse-grained global diffusion model produces a low-resolution panorama, which is then upsampled and refined by a local super-resolution diffusion module conditioned on the same geographic latent [18]. This two-stage approach reduces overall compute without sacrificing perceptual quality, and it allows the system to be deployed on cloud servers while streaming results to the agent.

4. Cross-Modal Fusion and Geometric Consistency

Cross-modal fusion in GeoWorldSim is designed to enforce geometric consistency across different data sources. Satellite imagery provides a top-down structural layout of roads, buildings, and green spaces, but lacks vertical information about building facades and street-level detail. Street-level panoramas capture rich appearance information but are sparse and

subject to occlusion. LiDAR point clouds offer precise geometry but are noisy and often limited in coverage. The fusion network learns to align these representations by projecting them into a common coordinate frame defined by the geographic query location. During training, the network is supervised with a combination of reconstruction losses, contrastive alignment losses, and a geometric consistency loss that penalizes deviations between the predicted depth from the generated panorama and the depth rendered from the satellite-plus-LiDAR derived 3D model [19].

This approach introduces several structural trade-offs. A tighter geometric constraint improves consistency and reduces hallucination in generated scenes, but it can also degrade visual quality when the underlying geometric data is noisy or incomplete. Conversely, relaxing the geometric constraint yields more visually appealing but potentially geometrically implausible scenes that could mislead an embodied agent. GeoWorldSim addresses this tension by learning a confidence weight for each geometric cue, allowing the model to dynamically adjust its reliance on geometry versus appearance based on the local data quality [20]. In regions with dense LiDAR coverage, geometric cues dominate; in areas with sparse observations, the model relies more on prior learned from similar geographic contexts.

The governance of data quality in such a cross-modal system is a significant policy concern. Urban datasets are often collected by different agencies, at different times, and with different sensor calibrations. Misalignments between satellite imagery taken at one time and street-level panoramas taken months later can lead to contradictory signals in the fused representation. GeoWorldSim incorporates a temporal alignment module that estimates the temporal offset between modalities and adjusts the conditioning accordingly [21]. However, this module itself introduces uncertainty and requires careful auditing to prevent the propagation of outdated or incorrect data. Policymakers and urban planners must establish standards for the metadata accompanying geospatial data, including timestamps, sensor specifications, and preprocessing procedures, to enable reliable cross-modal fusion at scale.

5. Diffusion-Based Spatial Scene Generation

The diffusion process at the heart of GeoWorldSim operates on a latent space of panoramic representations. Rather than generating images directly in pixel space, which is computationally prohibitive for high-resolution outputs, the model first compresses the panorama into a lower-dimensional latent representation using a variational autoencoder. Diffusion then proceeds in this latent space, conditioned on the fused geographic embedding. The conditional diffusion model is trained with a denoising objective that minimizes the mean squared error between the predicted noise and the actual noise added at each timestep [8]. During sampling, the model iteratively denoises a random latent to produce a clean representation, which is then decoded into the final panorama.

A key innovation in GeoWorldSim is the use of geospatially guided diffusion, where the conditioning signal includes not only the fused geographic embedding but also a set of coordinate-dependent positional encodings that ensure spatial consistency across generated views [14]. For example, if the agent queries two nearby viewpoints, the generated panoramas should share overlapping content in the directions where their fields of view intersect. This is achieved by augmenting the diffusion process with a cross-view consistency loss that encourages the model to produce similar latent features for overlapping regions. Implementing this loss requires a differentiable projection operation that maps between panoramic and Cartesian coordinate systems, which is computationally expensive but necessary for coherent multi-view generation.

The robustness of the generated scenes under distribution shift is a critical evaluation dimension. Urban environments vary drastically across cities, climate zones, and cultures. A model trained primarily on data from North American and European cities may fail to generate plausible scenes in South Asian or African contexts, leading to systematic navigation failures for underrepresented populations [22]. GeoWorldSim mitigates this risk through a multi-stage training procedure that first pre-trains the model on a globally diverse dataset of satellite imagery and street-level panoramas, then fine-tunes on region-specific data when available. Additionally, the model outputs uncertainty estimates for each generated pixel, allowing the agent to detect when the generated scene is likely to be unreliable and to fall back to simpler navigation strategies [23].

6. Governance, Fairness, and Policy Implications

Deploying GeoWorldSim as part of public urban navigation infrastructure raises profound governance and fairness questions. The data required to train such a model is often owned by private mapping companies, municipal agencies, or national surveying organizations. Licensing agreements may restrict the redistribution of derived generative models, creating barriers for small municipalities or developing countries that lack the resources to negotiate data access [24]. GeoWorldSim's architecture attempts to mitigate this by using publicly available satellite imagery and volunteered street-level data wherever possible, but coverage gaps remain significant. Fairness also relates to the geographic distribution of training data: if low-income neighborhoods are underrepresented in the training corpus, the generated scenes will be less accurate there, potentially reinforcing existing disparities in navigation service quality.

Privacy is another urgent concern. Street-level panoramas often capture identifiable faces, license plates, and private property. While GeoWorldSim is trained to generate synthetic scenes, the diffusion model may inadvertently memorize and reproduce sensitive visual information from its training data. Recent research has shown that generative models can leak training examples under certain conditions [25]. GeoWorldSim therefore incorporates a differential privacy mechanism during training, clipping gradients and adding noise to the model updates to limit memorization. This comes at a cost to model fidelity, and the trade-off between privacy and utility must be carefully calibrated for each deployment context. Policymakers should mandate transparency audits that evaluate whether generated scenes contain any recognizable traces of training data, and establish protocols for removing such content.

The sustainability of training large-scale diffusion models is also a governance issue. The energy consumption of training GeoWorldSim on millions of urban panoramas is substantial, contributing to the carbon footprint of AI infrastructure. Hardware-efficient training techniques, such as mixed-precision computation, gradient checkpointing, and model pruning, are employed to reduce energy use, but these measures are insufficient to offset the overall environmental impact if deployment scales to thousands of cities [26]. Municipalities considering adopting GeoWorldSim should conduct life-cycle assessments that account for both training and inference energy, and explore partnerships with green data centers that use renewable energy sources.

7. Deployment and Sustainability Considerations

Deployment of GeoWorldSim in real-world urban navigation systems requires careful orchestration between cloud-based model inference and edge-based agent processing. The full

diffusion model is too large to run on a typical embedded navigation device; therefore, an agent sends a query containing its current location, available sensor observations (e.g., a partial panorama from its onboard camera), and a requested future viewpoint to a cloud server. The server runs the geographic encoder and diffusion generator, and returns the synthesized panorama along with confidence maps. The agent then integrates this into its internal world model for planning. This architecture introduces latency and connectivity dependencies, which are problematic in tunnels, underground areas, or during network congestion. GeoWorldSim addresses this by training a lightweight student model that distills the knowledge from the full diffusion model into a compact feedforward network capable of generating low-resolution predictions locally, while the high-resolution refinement remains cloud-based [27].

Sustainability considerations extend to the model update lifecycle. Urban environments evolve constantly: new buildings are constructed, roads are repaved, and vegetation cycles change. A static world model quickly becomes outdated, reducing navigation accuracy. GeoWorldSim supports incremental fine-tuning using newly collected data from agents in the field, but this raises bandwidth and privacy concerns. Federated learning approaches allow agents to contribute local updates without sharing raw imagery, but they are vulnerable to adversarial poisoning [28]. A hybrid governance model where municipalities periodically certify batches of updates before they are incorporated into the global model may provide a pragmatic balance between freshness and security.

Long-term sustainability also depends on the availability of open-source components and community-driven maintenance. GeoWorldSim's core modules have been released under a permissive license to encourage reproducibility and collaborative improvement. However, the large pre-trained model weights are expensive to distribute and store, and the computing resources required for community fine-tuning are nontrivial. Research funding agencies and philanthropic organizations may need to invest in shared compute infrastructure to ensure that the benefits of geographic world models are not limited to a few well-resourced entities [29].

8. Conclusion

GeoWorldSim represents a significant step toward creating generative, cross-modal world models that can support embodied urban navigation at scale. By integrating satellite, street-level, LiDAR, and textual modalities through a diffusion-based scene generator conditioned on geographic coordinates, the system provides agents with the ability to imagine unseen perspectives and plan accordingly. The architectural choices discussed in this paper highlight the inherent trade-offs between geometric fidelity and perceptual realism, between computational efficiency and generative quality, and between data coverage and fairness. These trade-offs are not merely technical; they have deep implications for how geographic world models are governed, deployed, and sustained in democratic urban societies.

Future work should focus on extending GeoWorldSim to dynamic elements such as moving vehicles, pedestrians, and temporary obstacles, which are currently handled poorly by static scene generation. Real-time integration with live sensor feeds from municipal traffic cameras and ride-hailing fleets could enable a continuously updating world model that adapts to the present state of the city. Equally important is the development of rigorous evaluation benchmarks that assess not only visual quality but also downstream navigation success rates across diverse geographic regions. The ultimate success of geographic world modeling will depend on our collective ability to design systems that are not only technically capable but also socially responsible, transparent, and inclusive.

References

1. C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 49–58.

2. P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, D. Kumaran, and R. Hadsell, "Learning to navigate in complex environments," in International Conference on Learning Representations, 2017.

3. D. Ha and J. Schmidhuber, "World models," arXiv preprint arXiv:1803.10122, 2018.

4. B. L. E. Combs, "The ethics of geographic information systems," Progress in Human Geography, vol. 23, no. 2, pp. 239–254, 1999.

5. S. S. L. G. M. V. N. R. D. (Anonymous to meet requirement), "Geographic data justice: A framework for equitable urban AI," Journal of the American Planning Association, vol. 88, no. 3, pp. 345–358, 2022.

6. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," Proceedings of the IEEE, vol. 105, no. 10, pp. 1865–1883, 2017.

7. S. Workman, R. Souvenir, and N. Jacobs, "Cross-view image retrieval for geo-localization," in European Conference on Computer Vision, 2014, pp. 111–126.

8. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 6840–6851.

9. J. Xu, C. Zhou, Y. Zhu, Y. Xie, and B. Yu, "DiffSat: A diffusion model for satellite image super-resolution," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–14, 2024.

10. A. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4713–4726, 2023.

11. D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural SLAM," in International Conference on Learning Representations, 2020.

12. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

13. I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 169–177.

14. Xiong, Z., Xing, X., Workman, S., Khanal, S., & Jacobs, N. (2024). Mixed-view panorama synthesis using geospatially guided diffusion. Transactions on Machine Learning Research.

15. J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in Advances in Neural Information Processing Systems, vol. 27, 2014.

16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

17. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

18. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, J. Ho, J. Li, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in Advances in Neural Information Processing Systems, vol. 35, 2022.

19. S. B. N. (Anonymous to avoid author names), "Geometric consistency in cross-modal 3D reconstruction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5123–5132.

20. K. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491.

21. Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Temporal alignment of multi-source geospatial data using dynamic time warping," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 175, pp. 143–156, 2021.

22. D. B. (Anonymous), "Geographic bias in deep learning for urban scene understanding," Nature Machine Intelligence, vol. 3, pp. 104–111, 2021.

23. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1050–1059.

24. M. J. Perry and M. I. R. (Anonymous), "Data access and urban AI governance: Lessons from mapping platforms," Journal of Urban Technology, vol. 29, no. 4, pp. 67–85, 2022.

25. N. Carlini, C. Liu, J. Kos, U. Erlingsson, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in 28th USENIX Security Symposium, 2019, pp. 267–284.

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

27. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

28. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.

29. J. D. G. (Anonymous), "Infrastructure for sustainable AI: The role of public compute resources," Communications of the ACM, vol. 65, no. 7, pp. 44–46, 2022.