Personalized 3D Scene Generation with Spatially Grounded Diffusion Models for Immersive VR Content Creation

Finn Norris
Department of Computer Science, University of North Texas, Denton, TX, USA.
finn.work@unt.edu

Manoj Menon
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
manojmenon223@colostate.edu

Sagar M. Saini
School of Computing, Clemson University, Clemson, SC, USA.
sagar.m.saini@clemson.edu

Suraj Jain
Department of Computer Science, George Mason University, Fairfax, VA, USA.
jain2005@gmu.edu

Abstract

The emergence of diffusion models has revolutionized generative visual content creation, yet their application to personalized three-dimensional scene generation for immersive virtual reality environments remains fraught with systemic challenges. This paper examines the architecture and deployment of spatially grounded diffusion models designed to produce customized 3D scenes that respect geometric constraints and user-specific semantic preferences. We argue that achieving spatial grounding necessitates a tight coupling between text-to-image diffusion priors and volumetric scene representations, a coupling that introduces trade-offs in model expressiveness, computational efficiency, and controllability. The discussion extends beyond algorithmic design to consider the socio-technical infrastructure required for scalable VR content generation, including data governance, model robustness against distributional shifts, fairness in user-adaptive outputs, and the sustainability of large-scale training pipelines. By analyzing recent advances in grounding mechanisms and scene generation pipelines, we highlight the structural tensions between personalization fidelity and system generalization. The paper further explores policy implications surrounding intellectual property, algorithmic bias, and the environmental cost of high-resolution 3D generation. We conclude by outlining a research agenda that prioritizes transparent evaluation frameworks, equitable access to generative tools, and interdisciplinary governance models for immersive content ecosystems.

Keywords

Personalized 3D Scene Generation, Diffusion Models, Spatial Grounding, Virtual Reality, Immersive Content Creation, Socio-technical Systems.

1. Introduction

The rapid maturation of generative artificial intelligence has profoundly altered the landscape of digital content production, with diffusion models emerging as the dominant paradigm for synthesizing high-quality images and, more recently, three-dimensional assets [1][2][3]. At the same time, virtual reality systems are evolving from passive consumption platforms into interactive environments where users expect bespoke, dynamic scenes that adapt to individual preferences and contextual constraints. The convergence of these trends gives rise to a critical research challenge: how can diffusion models be repurposed to generate personalized 3D scenes that are not only visually compelling but also spatially consistent with the geometry and semantics of a target VR environment? Spatially grounded generation addresses this question by conditioning the generative process on explicit spatial information such as bounding boxes, depth maps, or semantic layouts. Unlike unconstrained image synthesis, grounded 3D scene generation demands that the model respect physical plausibility, occlusion relationships, and user-defined placement constraints, all while maintaining the high degree of variation and aesthetic quality that users expect from personalized content. This paper adopts a systems-level perspective, examining the architectural decisions, infrastructural dependencies, and socio-technical implications that arise when deploying such models in real-world VR content creation pipelines. We explore the trade-offs inherent in combining text-to-image diffusion priors with volumetric rendering techniques, the governance of training data that encodes spatial and semantic annotations, and the robustness of grounded generation against adversarial inputs or distributional shifts. By situating this technological development within broader discussions of fairness, sustainability, and policy, we aim to provide a comprehensive framework for researchers and practitioners working at the intersection of generative AI and immersive media.

2. Background and Related Work

The foundations of modern diffusion-based generative modeling were established by the introduction of denoising diffusion probabilistic models, which iteratively reverse a Markov chain of Gaussian noise additions to produce samples from a learned data distribution [1]. Subsequent work scaled these models to high-resolution image synthesis using latent diffusion architectures, enabling efficient training and inference on consumer-grade hardware [2]. The extension to 3D content generation has followed multiple trajectories, including score-based methods applied to neural radiance fields [4] and two-stage approaches that first generate a 2D image and then lift it to 3D using priors from large-scale datasets [3][19]. However, these methods typically operate in an unconstrained generative space, producing scenes that may violate geometric or semantic constraints imposed by a target VR environment. Spatial grounding techniques address this limitation by injecting explicit conditioning signals into the generative process. Early work in conditional image generation used semantic maps or bounding boxes as spatial inputs, but these approaches were largely limited to 2D domains [16]. The recent introduction of grounding modules that combine textual descriptions with spatial layouts within a diffusion framework represents a significant advance [18]. These modules learn to attend to both linguistic tokens and spatial coordinates, generating regions that align with user-provided placements while preserving the stylistic coherence of the pre-trained diffusion model. In parallel, the field of neural rendering has produced powerful scene representation methods such as NeRF and its variants, which allow for photorealistic novel view synthesis from sparse observations [4]. The integration of these representations with diffusion-based generation has enabled the creation of 3D scenes that can be navigated in real time, a critical requirement for VR applications. Nonetheless, the personalization of such scenes for individual users—accounting for preferences in style,

object arrangement, and interactivity—remains an under-explored area. This paper builds upon these prior contributions by analyzing the systemic challenges that arise when grounding diffusion models for personalized VR content, extending the discussion beyond algorithmic novelty to encompass infrastructure, governance, and societal implications.

3. System Architecture and Design Trade-offs

Designing a system for personalized 3D scene generation with spatial grounding involves a series of architectural trade-offs that directly impact model fidelity, computational cost, and user control. At the highest level, the system must integrate three core components: a text-to-image diffusion backbone, a spatial grounding module that interprets layout and depth constraints, and a 3D scene representation that can be rendered interactively. The diffusion backbone is typically a large pre-trained model, such as a latent diffusion model [2] conditioned on a text encoder like CLIP [13]. The spatial grounding module can be implemented either as an auxiliary conditioning network that modifies the internal activations of the diffusion model [16] or as a separate encoder that produces spatial embeddings fused with the text embedding [18]. The choice between these two approaches entails a trade-off between modularity and end-to-end optimization. A modular design allows reuse of existing diffusion weights and facilitates interchangeable grounding components, but may lead to suboptimal alignment between spatial cues and generative outputs. An integrated design, in which the grounding module is jointly trained with the diffusion backbone, can achieve tighter spatial correspondence but sacrifices the ability to upgrade the grounding module independently and increases training instability. Another critical architectural decision concerns the representation of the 3D scene. Volumetric representations such as signed distance fields or neural radiance fields [4] offer high geometric fidelity and support free-viewpoint rendering, but their computation is expensive and often requires per-scene optimization. In contrast, voxel-based or mesh-based representations are more amenable to real-time rendering but may lack the fine detail required for immersive VR. A common compromise is to use a hybrid approach: the diffusion model generates a set of multi-view images or a latent code that is subsequently decoded into a 3D representation via a fast feedforward network [19][20]. However, this introduces a decoupling between the generative and 3D reconstruction stages, potentially amplifying errors in spatial grounding. From a systems perspective, the latency of the generation pipeline is a primary concern for interactive VR content creation, where users expect feedback within seconds. This necessitates careful engineering of the denoising steps, possibly using fewer sampling steps or knowledge distillation to accelerate inference. Moreover, the computational demands of training such models require distributed infrastructure, often relying on large clusters of GPUs, which raises questions about equitable access and environmental sustainability. The trade-off between personalization depth and system scalability must be managed through adaptive sampling strategies that allocate more computational resources to regions of high user interest or semantic importance.

4. Spatial Grounding Mechanisms in Diffusion Models

Spatial grounding in diffusion models refers to the incorporation of explicit location, shape, and relational constraints into the generation process. These constraints are typically derived from user inputs, such as specifying that a chair should be placed in the left corner of a room or that a tree should stand three meters behind a bench. Early approaches to grounding in 2D image generation used cross-attention modulation to guide which regions of the latent space correspond to which textual tokens [16]. Extending this to 3D requires the modeling of depth,

scale, and occlusion. One strategy is to condition the diffusion process on a sparse set of 3D keypoints or bounding boxes, often aligned with a canonical coordinate frame [18]. The grounding module learns to map these spatial coordinates into a representation that is compatible with the attention layers of the diffusion backbone. A key challenge is handling variations in camera viewpoint and scene scale, which requires the grounding module to be either view-invariant or to be conditioned on camera parameters. Another mechanism involves the use of multi-view consistency during training: the diffusion model is trained to generate images from multiple camera angles that obey the same underlying spatial layout. This enforces a form of 3D reasoning without explicitly modeling geometry. However, such consistency constraints increase the complexity of the training data pipeline, as they require synchronized multi-view images with annotated spatial relationships. Furthermore, the grounding module must be robust to ambiguous or incomplete user specifications, a difficulty that arises frequently in personalized VR creation where users may provide rough sketches or verbal descriptions rather than precise positional coordinates. To address this, recent work has explored probabilistic grounding that outputs a distribution over possible spatial configurations, leaving the final selection to an interactive refinement step. The governance of such systems involves decisions about how user-defined spatial constraints are validated and potentially corrected to prevent physically impossible scenes. For example, a user may request an object to be partially occluded by another object that is not present in the scene, leading to contradictions that the model must resolve gracefully. Robustness to such edge cases requires the grounding mechanism to include fallback heuristics or to re-project the user's specification onto a feasible manifold. These considerations highlight the need for transparent model behavior and user feedback loops, which are essential for trust in personalized generative systems.

5. Deployment and Infrastructure Considerations

The deployment of spatially grounded diffusion models for VR content creation imposes stringent requirements on computing infrastructure, data management, and real-time serving pipelines. Unlike offline image generation, VR applications demand low-latency inference to maintain user presence and avoid motion sickness, with acceptable delays typically measured in tens of milliseconds. This is particularly challenging for 3D scene generation, where the model must not only produce pixel-level outputs but also reconstruct geometry and material properties. To meet latency constraints, many production architectures adopt a client-server split: a lightweight client runs on the VR headset or personal computer to perform real-time rendering, while a cloud-based server handles the computationally expensive diffusion and grounding computations. This split introduces dependencies on network bandwidth, latency, and reliability, which can vary significantly across deployment contexts. Edge computing paradigms propose moving parts of the inference pipeline to local hardware, but the memory footprint of large diffusion models often exceeds the capacity of consumer-grade VR devices. Model compression techniques, such as quantization, pruning, and knowledge distillation, are therefore critical for enabling on-device generation. However, compression may degrade the quality of spatial grounding, especially in fine-grained layout control. Another infrastructure challenge is the management of training data. Personalized 3D scene generation requires datasets that pair textual descriptions with 3D scene layouts, multi-view images, and explicit spatial annotations. Such datasets are scarce and expensive to produce, often relying on synthetic environments or manual annotation. The governance of these datasets involves questions of representativeness: if the training data is dominated by a narrow range of scene types, object categories, or cultural styles, the resulting model will exhibit bias in its

personalized outputs. Data augmentation strategies, including domain randomization and procedural scene generation, can mitigate some of these issues but may introduce artifacts. Furthermore, the storage and transmission of high-resolution 3D assets generated by the model pose their own engineering challenges, particularly when users expect to modify and reuse scenes over time. Versioning and persistence mechanisms must be designed to track changes in personalized content, and interoperability standards are needed to ensure that generated scenes can be imported into different VR platforms. These deployment considerations underscore the importance of a holistic system design that balances performance, cost, and quality.

6. Robustness, Fairness, and Governance

As spatially grounded diffusion models become integrated into VR content creation tools, their robustness and fairness must be evaluated from both technical and societal perspectives. Robustness concerns the model's ability to produce consistent, high-quality outputs under varying input conditions, including adversarial perturbations, distributional shifts, and ambiguous user queries. Spatial grounding can introduce unique failure modes, such as collisions between generated objects, physically implausible placements, or violations of scene semantics. For instance, a model might place a sofa floating in mid-air if the grounding constraint is loosely specified, or it might generate an object that obscures a critical navigation path in a VR environment. Ensuring robustness requires the development of verification checks that run after generation, potentially using physics simulators or collision detection routines to flag implausible configurations. These checks can be embedded as a post-processing step or used to guide the generative process itself via rejection sampling, though the latter increases computational cost. Another dimension of robustness is the model's behavior under out-of-distribution prompts. Users may request scenes that combine elements not seen together during training, such as a steampunk spaceship in a medieval castle, and the model should generate a coherent result without catastrophic failure. This highlights the tension between memorization and generalization: large diffusion models often rely on training data coverage, and underrepresented combinations may lead to low-quality outputs or amplification of stereotypes. Fairness in personalized 3D scene generation relates to the equitable treatment of users across different demographic, cultural, and geographic backgrounds. If the training data overrepresents Western interior design styles or particular architectural typologies, users from other contexts may receive outputs that feel alien or inappropriate. Moreover, personalization algorithms that adapt to user feedback can inadvertently reinforce existing biases, creating a feedback loop that narrows the diversity of generated content. Governance frameworks must therefore include mechanisms for auditing model outputs across diverse user groups and for enabling users to customize not only the content but the underlying aesthetic and functional priors. This could involve participatory design methods where community input shapes the training data and model objectives. Additionally, the intellectual property implications of generatorially produced 3D scenes are complex: if a model is fine-tuned on copyrighted 3D assets, the generated scenes may infringe on existing rights. Legal clarity is needed regarding ownership of personalized and grounded outputs, especially when users modify and share scenes. Transparency in model provenance, such as through datasheets and model cards, is a critical governance tool to document training data sources, intended use cases, and known limitations [9][10]. These considerations point toward a regulatory environment that encourages innovation while protecting users and creators.

7. Sustainability and Policy Implications

The environmental footprint of training and deploying large-scale diffusion models has become a pressing policy concern, and the extension to 3D generation exacerbates these challenges due to the increased dimensionality and computational demands of volumetric representations [11]. Training a grounded 3D diffusion model from scratch can require thousands of GPU-hours, resulting in substantial carbon emissions that vary with the energy mix of the computing infrastructure. Inference, while less demanding than training, still imposes a non-trivial energy cost when scaled to millions of VR users. Policy interventions could incentivize the adoption of energy-efficient hardware, such as specialized accelerators for diffusion operations, and the use of dynamic power management that reduces consumption during periods of low demand. Another sustainability dimension is the lifecycle of generated content: users may create, discard, and recreate 3D scenes frequently, leading to a proliferation of digital waste in the form of stored unused assets. While digital waste does not have the same physical impact as material waste, the associated storage and transmission energy should not be overlooked. From a policy perspective, there is an opportunity to develop standards for lightweight scene representation that minimize storage footprint without compromising immersion. Additionally, the democratization of VR content creation through personalized generative models raises questions of digital inclusion. High-quality generation currently requires access to expensive computing resources and large datasets, which may reinforce existing inequalities in technology access. Public investment in open-source models, shared infrastructure, and educational initiatives can help level the playing field. However, open-source release of powerful generative models also carries risks, such as the potential for misuse in creating deceptive or harmful VR experiences. Striking a balance between openness and safety is a delicate policy challenge that requires collaboration among researchers, platform companies, and regulators. The governance of personalized 3D content also intersects with data privacy, particularly when user preferences and spatial interactions are collected for model fine-tuning. Privacy-preserving techniques such as federated learning and differential privacy can mitigate some risks, but they introduce additional computational overhead and may reduce model quality. Policy frameworks must therefore be adaptive, recognizing that the appropriate level of regulation depends on the context of use—whether for entertainment, education, therapy, or professional design. Forward-looking approaches should include regular impact assessments, stakeholder engagement, and sunset clauses that allow regulations to evolve with the technology.

8. Conclusion

This paper has presented a comprehensive systems-level analysis of personalized 3D scene generation using spatially grounded diffusion models for immersive VR content creation. We have examined the architectural trade-offs inherent in integrating grounding mechanisms with pre-trained diffusion backbones and volumetric scene representations, emphasizing the need for careful balancing between expressiveness, controllability, and computational efficiency. The discussion extended to deployment infrastructure, where latency, data governance, and edge-cloud partitioning pose significant engineering challenges. Robustness and fairness considerations were explored, revealing that spatial grounding introduces particular vulnerabilities that must be addressed through verification tools and inclusive data practices. Finally, we considered the sustainability and policy implications of scaling these systems, highlighting the environmental costs, digital inclusion gaps, and regulatory uncertainties that accompany the democratization of generative VR tools. As the field advances, future research

should prioritize the development of evaluation benchmarks that capture both perceptual quality and spatial consistency, the design of interactive interfaces that allow users to iteratively refine grounding constraints, and the creation of governance frameworks that balance innovation with accountability. The convergence of diffusion models and VR represents a transformative opportunity, but realizing its full potential requires a sustained interdisciplinary effort that attends to the technical, social, and ethical dimensions of the systems we build.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

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

3. Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2023). DreamFusion: Text-to-3D using 2D diffusion. International Conference on Learning Representations.

4. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. European Conference on Computer Vision, 405-421.

5. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248-255.

6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

7. Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators and Virtual Environments, 6(6), 603-616.

8. Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems, 77(12), 1321-1329.

9. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

10. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

11. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

12. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. European Conference on Computer Vision, 694-711.

13. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning

transferable visual models from natural language supervision. International Conference on Machine Learning, 8748-8763.

14. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. International Conference on Machine Learning, 16784-16807.

15. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo-Lopes, R., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35, 36479-36494.

16. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 3836-3847.

17. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations.

18. Xiong, Z., Xiong, W., Shi, J., Zhang, H., Song, Y., & Jacobs, N. (2024). Groundingbooth: Grounding text-to-image customization. arXiv preprint arXiv:2409.08520.

19. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., & Vondrick, C. (2023). Zero-1-to-3: Zero-shot one image to 3D object. Advances in Neural Information Processing Systems, 36.

20. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., & Zhu, J. (2023). ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. arXiv preprint arXiv:2305.16213.