Cross-Dimensional Phantasm

An architectural formation approach based on Generative AI and 3D reconstruction algorithms

Yingxu Wang[1], Dasong Wang[2], Peng Tang[3]
[1,2,3] Southeast University
[1,2,3]{213203243|dasong.wang| tangpeng} @seu.edu.cn

This paper proposes a generative design workflow that fuses generative artificial intelligence with three-dimensional reconstruction techniques to enable novel architectural formation processes. Drawing on recent advances in spatial intelligence and their successful applications in fields beyond architecture, the study seeks to bridge multi-dimensional AI-generated content—such as text-to-image synthesis and video generation—with architectural spatial logic. It further explores a computational pathway that transitions design input from static 2D imagery to constructible, high-fidelity 3D architectural scenarios. The workflow consists of four key stages: (1) multi-style text-to-image generation guided by natural language prompts; (2) synthesis of perspective-consistent 3D video sequences from selected images using a diffusion transformer (DiT)-based model with iterative artifact correction; (3) photogrammetric reconstruction of the video into a detailed 3D mesh; and (4) topology optimization and fabrication-informed material testing to evaluate constructability through robotic fabrication processes. By integrating AI-generated architectural scenes with 3D reconstruction and fabrication-aware geometry optimization, the framework introduces a potential methodology for aligning generated architectural forms with real-world material and structural constraints. This research contributes to the expansion of the generative design paradigm and opens new directions for the practical application of AI-assisted architectural design.

Keywords: *Generative design,Diffusion Transformer,3D reconstruction,AI-aided design.*

INTRODUCTION

With the rapid development of multimodal machine learning models, generative artificial intelligence is demonstrating extraordinary creativity and enhanced perception of the real world. Beginning with first-generation models like Cycle-GAN (Zhu et al.,2017) that demonstrated intentional image synthesis, the field has progressed through transformative developments including the Diffusion Model (Rombach,R. et al.,2022) with its enhanced resolution and qualitative control, to the recent Diffusion Transformer (Peebles,W. and Xie,2023) capable of generating perspective-continuous videos and 3D virtual environments. This dimensional advancement presents significant opportunities for architectural applications that are fundamentally rooted in 3D spatial logic.

Complementing these developments, spatial intelligence technologies - particularly advanced photogrammetric reconstruction techniques as implemented in RealityCapture - have achieved remarkable precision in converting multi-view images into high-fidelity 3D models through point cloud generation and manifold mesh optimization.

Figure 1
Comparison between an AI-generated image (left) and its corresponding topologically optimized 3D model (right)

This study proposes a novel hypothesis at the intersection of these technologies: that combining Diffusion Transformer's generative capabilities with photogrammetric reconstruction techniques could enable the transformation of AI-generated virtual scenes into physically constructible architectural forms, thereby addressing current limitations in handling complex 3D architectural spatial logic.

The paper, titled "Cross-Dimensional Phantasm", draws conceptual inspiration from 18th-century optical illusion techniques to frame its investigation of AI's capacity to generate inconceivable spatial experiences through computational approaches. The term "Phantasm" metaphorically represents the AI's generation of visionary spaces, while "Cross-Dimensional" emphasizes emerging computational technology including 3d-reconstruction, geometric topological optimization approach, etc. - that translate these virtual creations into physically realizable architectural tectonics. Figure 1 presents the AI-generated image and the resulting topologically optimized 3D model.

We present an integrated generative workflow comprising three key phases: (1) two-dimensional AI generation with style fusion, (2) spatial aware three-dimensional AI generation (3) topological optimization and materialized experiments with AI generated forms. This approach seeks to expand the boundaries of generative design in architecture while fully leveraging the creative potential of contemporary AI systems.

The remainder of this paper is organized as follows : Section two provides a technical background and literature review, Section three details our methodological framework and implementation, and Section four discusses implications, limitations, and future research directions in this emerging field.

BACKGROUND AND REVIEW

This study reviews recent advancements in generative AI technologies. As a emerging research direction in applied fields, numerous research teams have conducted studies on AI-based architectural generative design. Our work specifically examines the technical frameworks of two state-of-the-art models - Diffusion Model and Diffusion Transformer - and investigates their latest applications in architectural generative design. By integrating techniques from video generation (Brooks et al.,2024; Chen et al.,2024) , 3D model generation (Poole, B. et al., 2023) , and advanced 3D reconstruction techniques, we ultimately develop the generative design workflow presented in this research.

Generative AI and 3D reconstruction technical background

Diffusion Models (DMs) and Diffusion Transformers (DiTs) represent cutting-edge approaches in generative AI, marking a significant paradigm shift in generative modeling. Rooted in non-equilibrium thermodynamics and statistical physics, diffusion models were first adapted for machine learning applications by Sohl-Dickstein et al. (2015). These models establish a Markov chain of bidirectional mappings between data and noise spaces,

demonstrating superior performance to traditional GANs in terms of image fidelity and mode coverage through iterative denoising.

The field advanced significantly with Ho et al.'s (2020) DDPM framework, which enabled practical implementation via parameterized Markov chains. Open-source projects like Stable Diffusion later extended these models to text-to-image generation and inpainting, establishing them as core technologies for multimodal content creation.

DiTs introduced architectural innovation by replacing U-Nets with Transformer-based designs. Their global attention mechanisms and positional encoding strategies enhance long-range semantic modeling. Current DiT applications span video generation and 3D modeling, with modular designs enabling multi-task training possibilities.

3D reconstruction remains a fundamental challenge in computer vision and graphics, focusing on recovering 3D geometry and appearance from 2D observations (e.g., images or point clouds). Recent advances span three dominant paradigms: Neural Radiance Fields (NeRF) leverage implicit neural representations for view synthesis but face computational intensity; Gaussian Splatting (Kerbl et al., 2023) employs explicit anisotropic Gaussians to balance quality and efficiency; and Structure-from-Motion (SfM) with Multi-View Stereo (MVS), as implemented in tools like RealityCapture, combines geometric triangulation with dense matching for scalable photogrammetric reconstruction. Current research highlights their respective strengths in dynamic scene modeling (e.g., 4D Gaussian Splatting), neural rendering (NeRF), and industrial-scale applications (SfM/MVS).

Generative AI applications in architectural formation

With the rapid advancement of generative AI technologies, their applications in generative design are continually expanding the boundaries of design innovation. Generative adversarial networks (GANs) have been increasingly applied in architectural design. Del Campo (2022) uses StyleGAN2 to generate defamiliarized architectural forms from modernist datasets. Bolojan et al. (2022) created a multimodal GAN system combining text prompts with 3D modeling for controlled architectural form generation. Ennemoser and Mayrhofer-Hufnagl (2023) develop a 3DGAN workflow that hybridizes voxelized spatial data and 2D GAN training, enabling the generation of multiscale architectural geometries with hybrid spatial configurations beyond conventional typologies.

Due to its demonstrated advantages in image generation quality and diversity, diffusion models are finding increasing applications in generative architectural design. Bank et al. (2023) implemented text-to-image diffusion models (Midjourney) in architectural education, showing their effectiveness in generating targeted visual vocabularies for early design ideation.

DiT enhances diffusion models via Transformers' global modeling, improving generation quality and scalability. Extensions like Sora (OpenAI, 2024) enable high-quality long videos. In architecture, CityDreamer (Xie et al., 2024) introduces a compositional diffusion-based framework capable of generating large-scale, unbounded 3D urban scenes with high view consistency.

Research questions and objectives

This study addresses the following core limitations in generative design: its reliance on encoding intentions as algorithms and constrained morphological exploration within predefined parameter spaces of formation systems. Leveraging the multimodal generation capabilities of DiTs and advanced 3D reconstruction techniques, we aim to push beyond conventional parametric design boundaries.

METHODOLOGY

This section presents our generative workflow (Figure 2), demonstrating how AI-driven concepts evolve into three-dimensional architectural manifestations. The process comprises four key stages:

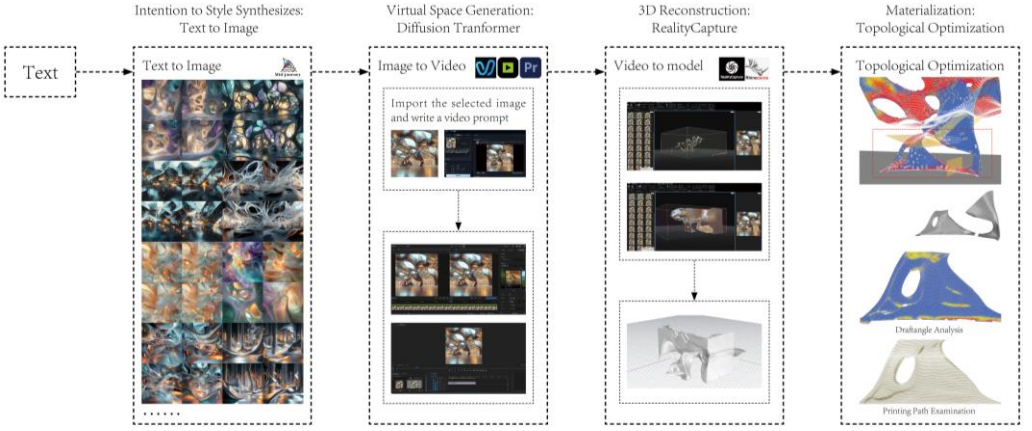


Figure 2
Integrated generative design workflow

- Intention to Style Synthesizes: Text-to-image workflow generates diverse design options through tailored prompts, iteratively refining them via architectural style blending and abstract element integration to produce optimized solutions.
- Virtual Space Generation: Image-to-video workflow converts 2D images to video sequences using Vidu's diffusion transformer, iteratively optimizes outputs to resolve artifacts, enhances quality with Topaz, and recursively generates video segments for complete 3D visualization.
- 3D Reconstruction: Video-based 3D reconstruction workflow processes three video sequences (>60% frame overlap) in RealityCapture through: feature detection → sparse/dense point cloud generation → mesh optimization → texture mapping, yielding a precise 3D model.
- Materialization: This step tested digital-to-physical transformation through mesh topology optimization and robotic fabrication verification.

Detailed implementation specifics will be presented in subsequent sections.

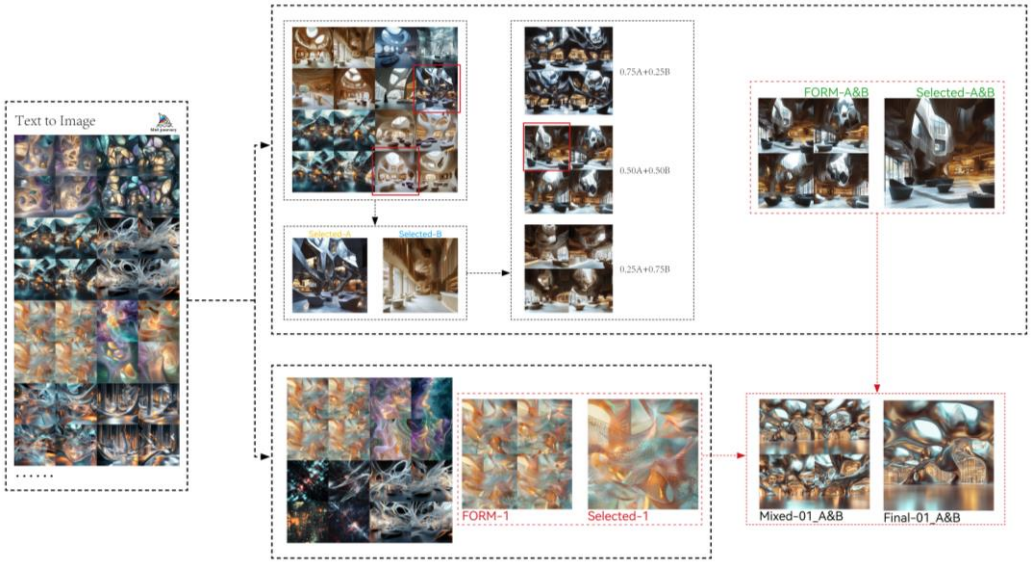
Intention to style synthesizes

The text-to-image generation process is illustrated in Figure 3. We first employ Midjourney to produce diverse architectural spatial concepts by inputting multiple sets of textual prompts describing different architectural spatial imagery. From these generated results, we select two distinct stylistic groups labeled as "Style A" and "Style B", and subsequently identify the most design-aligned images from each group, naming them "Selected-A" and "Selected-B" respectively.

To achieve stylistic fusion, we blend these two selected images with varying weight ratios, generating a series of hybrid images. The most balanced fusion result with equal 50%/50% weighting is designated as "FORM-A&B". To explore non-architectural abstract forms, we generate additional 2D imagery using prompts describing concepts such as "biological tissue", "particle collisions", and "lunar regolith material". From these abstract outputs, we select the most promising set labeled as "FORM-1". This abstract form is then blended with the architectural "FORM-A&B" image through weighted fusion, producing a new series of hybrid spatial concepts. The optimal fusion result from this stage is named "Mixed-01_A&B", from

which we ultimately select the final conceptual image "Final-01_A&B" for further development.

Figure 3
Text-to-image generation and multi-image fusion workflow



Virtual Space Generation

Figure 4 illustrates the complete workflow for transforming 2D images into 3D videos. The process begins by importing the selected 2D image "Final-01_A&B" into Vidu, a diffusion transformer-based video generation platform. Through carefully crafted prompts controlling camera movement, scene composition, and lighting parameters, we generate preliminary videos that simulate spatial environments and architectural elements.

To address common generation artifacts (e.g., physical inconsistencies, design deviations, camera/lighting flaws), we employed iterative optimization through prompt refinement and output filtering. The resulting video demonstrates smooth left-to-right rotational pans that progressively reveal spatial details while preserving physical coherence with the original design. The post-processing phase included:

- Resolution enhancement (4× upscaling with detail preservation) using Topaz Video AI
- Frame-by-frame decomposition via temporal sampling in Premiere Pro
- Recursive generation of two additional 4-second segments using terminal frames as subsequent inputs

The output comprises three seamlessly connected 4-second videos, which together form a fluid 12-second animation. These segments collectively visualize the architectural space, dynamically revealing depth, materials, and lighting from the original 2D concept while maintaining spatial coherence throughout the camera movement.

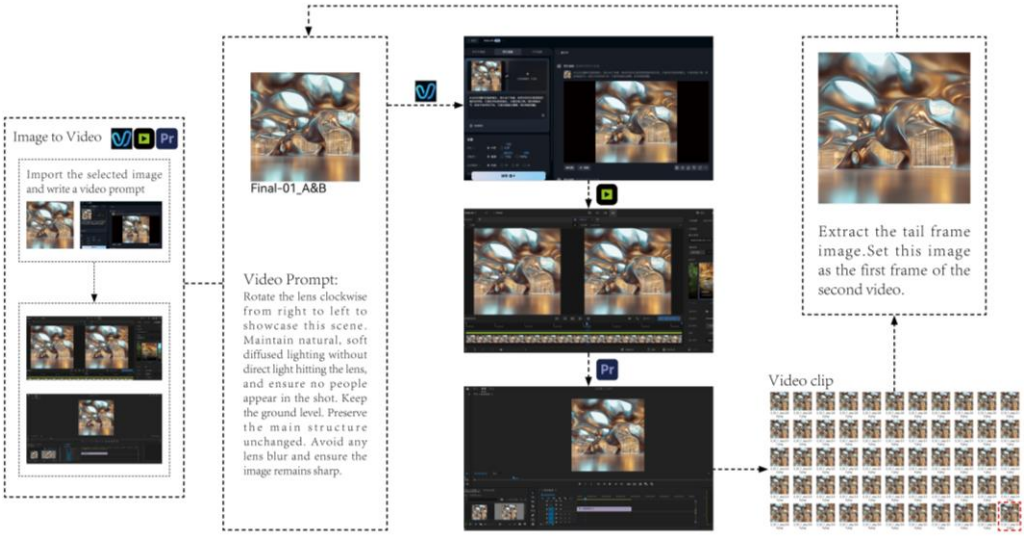


Figure 4
Image-to-3D video generation workflow

3D reconstruction

The 3D reconstruction workflow from multiple video segments is systematically illustrated in Figure 5. The processing pipeline began with extracting frame sequences at optimal intervals from the three video segments, which were then imported into Reality Capture for photogrammetric analysis. The continuous camera motion with controlled translation and rotation ensured sufficient frame overlap (>60%), providing the necessary conditions for reliable feature matching throughout the sequence. The software's advanced feature detection module automatically identified and tracked key image features (including edges, corners, and texture patterns) across consecutive frames, establishing robust correspondences to generate an initial sparse point cloud representation of the scene geometry.

Subsequent processing stages involved iterative refinement of camera pose estimation using bundle adjustment techniques, followed by dense point cloud generation through multi-view stereo matching algorithms. The reconstructed point data was then converted into an optimized triangular mesh structure, with surface regularization applied to maintain geometric coherence. For the final output, high-fidelity texturing was achieved by intelligently projecting and blending the original image colors onto the reconstructed geometry, preserving the visual authenticity of the architectural space while resolving potential artifacts at occlusion boundaries. This comprehensive reconstruction pipeline effectively transformed the 2D video inputs into a fully-realized 3D model with accurate spatial and visual properties.

Figure 5
Video-based 3D reconstruction workflow

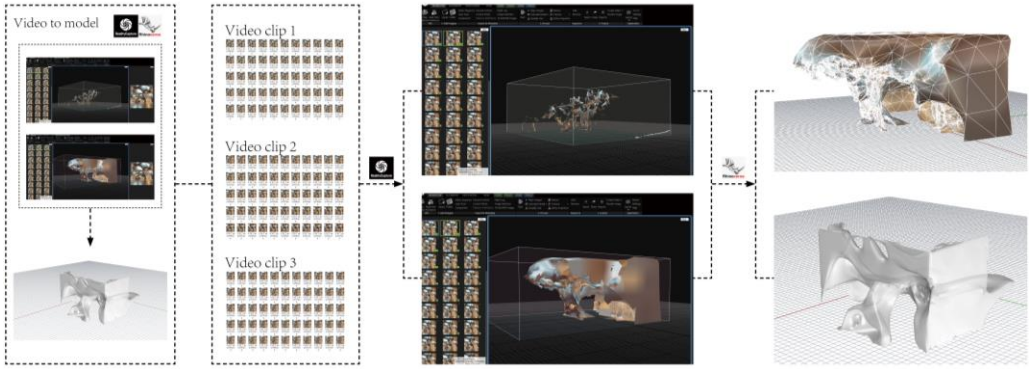


Figure 6
Topology optimization and segmentation position determination

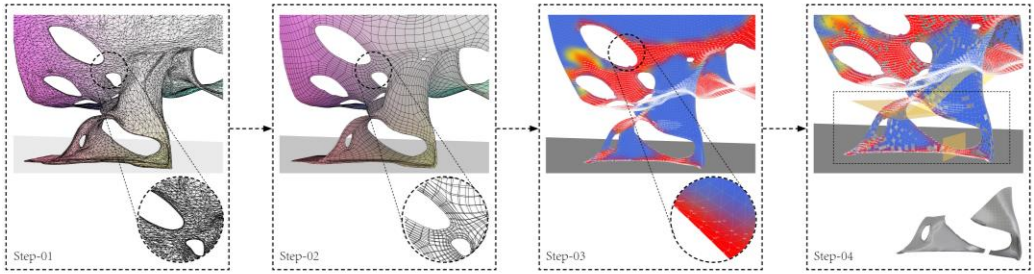
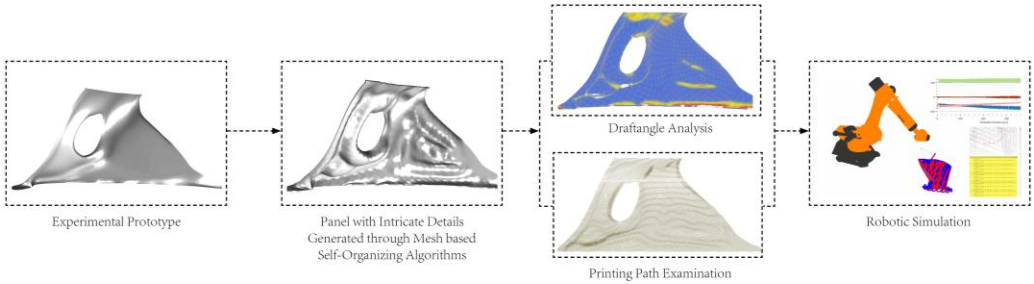


Figure 7
Secondary detail texture generation and fabrication feasibility verification



Optimization & Materialization

After obtaining the mesh model generated by RealityCapture, we conducted topology optimization with fabrication-aware constraints and secondary form-finding, ultimately materializing the geometry through robotic additive manufacturing to validate its constructability. As shown in Figure 6, the process began with topological reconstruction of the mesh using Rhino's QuadRemesh command with fixed vertex count parameters. Subsequently, we calculated the global draft angle conditions through a custom plugin, which explicitly addressed 3D printing limitations (e.g., avoiding overhangs exceeding material self-supporting capacity). These conditions dynamically determined the positioning and angles of splitting planes to ensure manufacturability of all components. A representative component was then subjected to detailed materialization and optimization.

Figure 7 reveals that the remeshing and 3D reconstruction process caused significant loss of formal details, prompting the application of alternative algorithms for secondary detail generation. Mesh differential expansion was employed to generate intricate corrugations while preserving overall geometry integrity. All algorithmically generated textures underwent iterative draft angle validation and non-planar toolpath simulation to guarantee their printability. These complex textures not only enhanced design resolution but also improved structural robustness through surface articulation. The prototype's manufacturing feasibility was further verified through robotic simulation. Future research will further explore integrated topology generation and materialization strategies, with practical applications to test and refine this workflow.

DISCUSSIONS

In this section, we summarize the challenges encountered and insights gained during our research, aiming to provide valuable references for researchers in related fields. We will discuss both the innovative aspects and limitations of our study. Furthermore, based on the current development of generative AI, we propose potential future research directions from the perspective of architectural design.

Reflections on the formation procedure

The key difference between this generative design workflow and traditional approaches lies in its natural language-based control of forms and styles, moving beyond conventional direct manipulation of images or 3D models. This shift enables the creation of innovative hybrid forms that surpass typical human imagination. However, effectively optimizing text prompts to achieve desired design outcomes presents a significant new challenge.

When converting 2D images to 3D spaces through AI-generated videos, the results often lack proper physical representations and logical consistency. This leads to noticeable discrepancies between the original image and generated structures when adjusting camera angles, requiring repeated prompt refinements and multiple generation attempts that introduce inherent randomness.

Comparisons to 3D Generative AI approach

With the advancement of AI technologies in multimodal generation, our research has investigated approaches for direct 3D model generation from images or text prompts using diffusion transformers. Notable developments in this domain include Meta 3D Gen (2024) for text-to-3D conversion and the Edify 3D system (2024), both primarily focused on generating individual models. Currently, Tripo AI has developed a nodal workflow for three-view to 3D model generation, providing additional possibilities for direct spatial generation from images. However, as these text/image-to-model techniques are currently limited to single-object generation, they still exhibit insufficient spatial predictability. Consequently, we have maintained our established workflow that first generates videos from images before performing 3D reconstruction and topological optimization.

Compared to these approaches, our method utilizes multi-view video as an intermediate medium, which better captures spatial continuity and structural rationality, providing a more reliable 3D foundation for subsequent architectural applications.

While existing 3D generative AI approaches excel at single-object modeling, they still face challenges when dealing with complex architectural spatial systems. Our workflow effectively addresses geometric inconsistency caused by single-view input through a multi-stage process of video generation and 3D reconstruction. This approach is particularly suitable for architectural scenarios requiring precise spatial relationships and structural rationality, bridging the gap in current 3D generative AI applications for architecture. Furthermore, our topological optimization step additionally ensures the constructibility of generated results, representing a critical factor that most general 3D generation systems have not yet fully considered.

Future works

Current generative design approaches using predictive models predominantly focus on aesthetic considerations. Building upon this foundation, our research will further explore methods for integrating practical construction principles into design workflows. By strengthening the connections between AI-generated outcomes and real-world engineering constraints, including structural and material limitations, we aim to better align these approaches with actual architectural requirements. Future investigations will prioritize developing more intelligent physical simulation integration methods, enabling generative systems to automatically evaluate structural rationality and construction feasibility.

At the fabrication level, we will continue investigating and expanding the capabilities of 3D printing technologies to improve the buildability of complex architectural components. Simultaneously, we will continue tracking technological advancements in 3D asset generation, particularly breakthroughs in predictive models' understanding of physical principles and spatial generation capabilities, which will further enhance the integration between generative design and practical construction.

CONCLUSIONS

This study presents a comprehensive AI-generated workflow that integrates 2D-to-3D dimensional transformation with awareness of physical construction constraints. By Diffusion Transformers' creative generation and 3D reconstruction algorithms, we proposed novel approach to directly realize design intentions across dimensions. This approach transcend the limitations of current generative AI in large scale and complex 3D spatial articulation. While challenges persist in cross-dimensional physical coherence, the integration of natural language control with multimodal generation capabilities establishes a scalable and applicable new approach for architectural generative design. The core contribution of the paper lies in the deep convergence of AI with generative design methodology through the "Cross-Dimensional Phantasm" framework, which not only expands the application scope of generative design but also provides a technical foundation for innovative architectural formation process in the age of spatial intelligence.

REFERENCES

Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A. (2017) 'Unpaired image-to-image translation using cycle-consistent adversarial networks', 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22-29 October. Piscataway: IEEE, pp. 2242-2251. doi: 10.1109/ICCV.2017.244.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) 'High-resolution image synthesis with latent diffusion models', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18-24 June.

Piscataway: IEEE, pp. 10674-10685. doi: 10.1109/CVPR52688.2022.01042.
Peebles, W. and Xie, S. (2023) 'Scalable diffusion models with transformers', Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2-6 October. Piscataway: IEEE, pp. 4172-4182. doi: 10.1109/ICCV51070.2023.00387.
Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R. and Ramesh, A. (2024) Video generation models as world simulators. Available at: https://openai.com/research/video-generation-models-as-world-simulators(Accessed: 14 June 2025)
Poole, B., Jain, A., Barron, J.T. et al. (2023) 'DreamFusion: text-to-3D using 2D diffusion', ICLR 2023. arXiv: 2209.14988. Available at: https://arxiv.org/abs/2209.14988
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. (2015) 'Deep unsupervised learning using nonequilibrium thermodynamics', Proceedings of the 32nd International Conference on Machine Learning, 6-9 July, Lille, France. PMLR 37, pp. 2256-2265. Available at: https://proceedings.mlr.press/v37/sohl-dickstein15.html(Accessed: 14 June 2025)
Ho, J., Jain, A. and Abbeel, P. (2020) 'Denoising diffusion probabilistic models', Advances in Neural Information Processing Systems, 33, pp. 68406851.Available at: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html(Accessed: 14 June 2025)
Kerbl, B., Kopanas, G., Leimkühler, T. and Drettakis, G. (2023) '3D Gaussian splatting for real-time radiance field rendering', ACM Transactions on Graphics, 42(4), 139. doi:10.1145/3592433.
del Campo, M., 2022. Deep House – datasets, estrangement, and the problem of the new. Architectural Intelligence, 1(12). Available at:https://link.springer.com/article/10.1007/s44223-022-00013-w(Accessed: 14 June 2025)
Bolojan, D., Vermisso, E. and Yousif, S., 2022. Is Language All We Need? A Query Into Architectural Semantics Using a Multimodal Generative Workflow. In: CAADRIA 2022: Post-Carbon, Sydney, pp.353-362. Available at:https://doi.org/10.52842/conf.caadria.2022.1.353(Accessed: 14 June 2025)
Ennemoser, B. and Mayrhofer-Hufnagl, I., 2023. Design across multi-scale datasets by developing a novel approach to 3DGANs. International Journal of Architectural Computing, 21(2), pp.358 – 373. Available at:https://journals.sagepub.com/doi/10.1177/14780771231168231(Accessed: 14 June 2025)
Bank, M., Moisi, A., Rasoulzadeh, S., Schinegger, K. and Rutzinger, S., 2023. AI Diffusion as Design Vocabulary – Investigating the use of AI image generation in early architectural design and education. In: eCAADe 41 – Digital Design Reconsidered, pp.587-596. Available at:https://www.researchgate.net/publication/37640340(Accessed: 14 June 2025)
Xie, H. et al. (2024) 'CityDreamer: Compositional Generative Model of Unbounded 3D Cities', 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Computer Vision and Pattern Recognition (CVPR), 2024 IEEE/CVF Conference on, CVPR, pp. 9666 – 9675. doi:10.1109/CVPR52733.2024.00923.
Bala, M., Cui, Y., Ding, Y., et al. (2024) Edify 3D: scalable high-quality 3D asset generation. arXiv: 2411.07135. Available at:https://arxiv.org/abs/2411.07135(Accessed: 14 June 2025)
Bensadoun, R., Monnier, T., Kleiman, Y. et al. (2024) Meta 3D Gen. arXiv: 2407.02599. Available at: https://arxiv.org/abs/2407.02599(Accessed: 14 June 2025)