

# 3D4D: An Interactive, Editable, 4D World Model via 3D Video Generation

Yunhong He<sup>1,\*</sup>, Zhengqing Yuan<sup>2,\*</sup>, Zhengzhong Tu<sup>3</sup>, Yanfang Ye<sup>2</sup>, Lichao Sun<sup>4</sup>

<sup>1</sup>Independent Research  
<sup>2</sup>University of Notre Dame  
<sup>3</sup>Texas A&M University  
<sup>4</sup>Lehigh University

## Abstract

We introduce 3D4D, an interactive 4D visualization framework that integrates WebGL with Supersplat rendering. It transforms static images and text into coherent 4D scenes through four core modules and employs a foveated rendering strategy for efficient, real-time multi-modal interaction. This framework enables adaptive, user-driven exploration of complex 4D environments.

**Code** — <https://yunhonghe1021.github.io/NOVA/>

## Introduction

Advances in Generative Model and Multi-modal Learning have enabled more intuitive text-driven and multi-modal interactive generation, enhancing the usability and performance of 4D content generation (Fei et al. 2024; Liu et al. 2023; Rombach et al. 2022; Sun et al. 2024; Yuan et al. 2024a; Wei et al. 2025; Li, Zhang, and Sun 2023; Yan et al. 2024; Yuan et al. 2024b; Singer et al. 2023). Yet, conventional WebGL-based frameworks remain limited in their ability to handle real-time 4D rendering and fine-grained temporal navigation, often struggling with high computational costs, latency, and scalability issues (Xu et al. 2024; Li et al. 2024; Abbas, Hassan, and Sadiq 2025; Jang, Kim, and Seo 2022). More critically, existing systems fall short of providing genuinely interactive 4D environments, as technologies that seamlessly combine high-performance rendering with user interactivity are still largely absent (Miao et al. 2025; Zhou et al. 2025).

To address these challenges, we introduce 3D4D, an interactive 4D visualization framework that integrates WebGL with Supersplat (PlayCanvas and Contributors 2025) rendering to support real-time interaction and editing. As illustrated in Figure 1, the backend comprises four core modules—3D Scene Reconstruction, Image-to-Video Synthesis, Video-to-Frame Decomposition, and 4D Scene Generation—that transform static images and text into temporally coherent 4D scenes for adaptive, user-driven exploration.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

\*Equal contribution

†Yunhong He is an independent undergraduate student, remotely working with Lichao Sun.

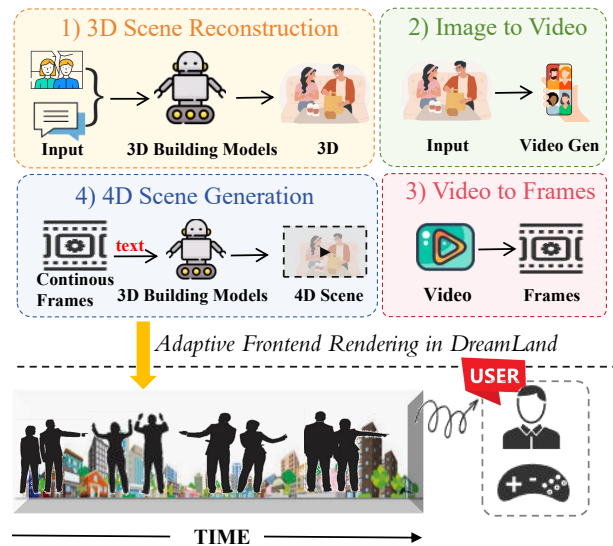


Figure 1: Overview of 3D4D pipeline. The system integrates multi-modal inputs with real-time 4D rendering to support interactive exploration.

This paper primarily focuses on the frontend component, which enables real-time interaction, editing, and visualization within the generated 4D environment.

3D4D further introduces a foveated rendering strategy inspired by human peripheral vision, progressively refining semantically important regions identified by a vision-language model while approximating low-saliency areas. This design aims to maintain semantic alignment and visual consistency while reducing GPU memory usage and latency, thereby enabling real-time 4D interaction even under constrained computational resources. For a more detailed illustration, please refer to our demonstration.

## System Framework

Our front-end is developed based on WebGL and Supersplat rendering technologies to achieve high-performance visualization of 4D environments. Since the standard WebGL framework does not support fine-grained temporal interac-

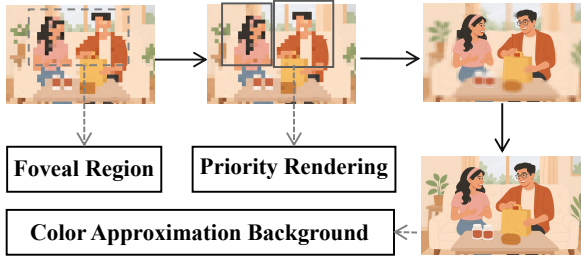


Figure 2: 3D4D’s adaptive frontend rendering: semantically important regions are rendered in high resolution, while peripheral areas are approximated to reduce cost.

tion, we have developed several customized functions on this basis, enabling users to manipulate, edit, and analyze the generated 4D scenes in real time.

### Design of Frontend

The frontend, developed with WebGL and Supersplat rendering, provides high-performance visualization and real-time editing of 4D environments. When a user inputs an image and a text prompt, the system triggers the backend to reconstruct the corresponding 3D scene and synthesize temporally coherent frames. These outputs are converted into multiple PLY point clouds and streamed to the frontend, where they are rendered sequentially or looped to form a continuous 4D video (Deng et al. 2024; Casas et al. 2014; Wu et al. 2024).

The interface allows users to dynamically explore the generated scene by adjusting camera pose, playback speed, and frame rate through an interactive timeline. Fine-grained temporal navigation enables users to balance visual quality and efficiency during analysis. Furthermore, the frontend supports direct scene editing through tools such as rectangle, brush, polygon, lasso, and sphere selection, enabling precise manipulation of objects and regions. All user interactions are synchronized with the backend via an API, ensuring seamless integration of rendering, computation, and user control.

### Rendering Video

When the user clicks “Rendering Video”, the system initiates a VLM-guided rendering pipeline that transforms uploaded PLY scenes into high-quality videos. After loading one or multiple Gaussian Splat point clouds and interpolating camera trajectories from user-defined keyframes, a Vision-Language Model (VLM) (Yuan et al. 2023) like Qwen2.5vl (Wang et al. 2024) analyzes each frame to generate an importance map identifying semantically critical regions such as humans or moving objects. As illustrated in Figure 2 During rendering, the WebGL shader adaptively allocates resources, rendering foveal regions at full precision while applying blurred, low-cost shading to background areas, thereby reducing GPU load without compromising perceptual quality. Each frame is captured from the framebuffer,

Model	CC	CS
WonderJourney (Yu et al. 2024)	27.34	0.9544
LucidDreamer	26.72	0.8972
Text2Room (Höllein et al. 2023)	24.50	0.9035
WonderWorld	29.47	0.9948
SV4D (Xie et al. 2024)	30.29	0.8856
4D-fy (Bahmani et al. 2024)	11.23	0.6147
<b>3D4D (Ours)</b>	<b>30.40</b>	<b>0.9951</b>

Table 1: Evaluation results for CLIP Consistency (CC) and CLIP Score (CS) across different models.

temporally smoothed, and encoded in real time through the browser’s MediaRecorder into .webm or .mp4 format. This fully client-side pipeline achieves semantic-aware, real-time 4D video generation that balances visual fidelity and computational efficiency.

### Evaluation

**Metrics.** We evaluate our system using both performance and efficiency metrics. Performance metrics include CLIP Score (CS), calculated as the CLIP (Radford et al. 2021) similarity between the textual scene prompt and the rendered image, indicating the semantic alignment quality; and CLIP Consistency (CC), measured by the cosine similarity between the CLIP embeddings of each novel view image and the central reference view, evaluating visual consistency across different viewpoints. Efficiency metrics include frames per second (fps), measuring the rendering speed, and a binary indicator of real-time interaction system.

Model	FPS	Real-time Interaction
SVD-4D (Xie et al. 2024)	40	✗
4D-fy (Bahmani et al. 2024)	16	✗
<b>3D4D (Ours)</b>	<b>60</b>	✓

Table 2: Performance comparison in terms of frame rate (FPS) and real-time interaction capability.

**Results.** Our proposed system, 3D4D, outperforms existing methods across both performance and efficiency metrics. Specifically, as demonstrated in Table 1, 3D4D achieves the highest CC of 30.40 and the best CS of 0.9951, indicating superior semantic alignment and visual consistency compared to other models. Moreover, Table 2 highlights 3D4D’s substantial improvement in rendering speed, delivering 60 frames per second (fps), and uniquely supporting real-time interaction, a critical difference for other methods.

### References

Abbas, M. A.; Hassan, E. B.; and Sadiq, M. T. 2025. Real-time visualization of high-dimensional data streams using efficient GPU techniques. *arXiv preprint arXiv:2503.14501*.

Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2024. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7996–8006.

- Casas, D.; Volino, M.; Collomosse, J.; and Hilton, A. 2014. 4d video textures for interactive character appearance. In *Computer Graphics Forum*, volume 33, 371–380. Wiley Online Library.
- Deng, Z.; Li, X.; Li, X.; Tong, Y.; Zhao, S.; and Liu, M. 2024. Vg4d: Vision-language model goes 4d video recognition. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5014–5020. IEEE.
- Fei, H.; Zhang, M.; Zhang, M.; and Chua, T.-S. 2024. XNLP: An Interactive Demonstration System for Universal Structured NLP. In Cao, Y.; Feng, Y.; and Xiong, D., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 19–30. Bangkok, Thailand: Association for Computational Linguistics.
- Höllein, L.; Cao, A.; Owens, A.; Johnson, J.; and Nießner, M. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7909–7920.
- Jang, H.; Kim, E.; and Seo, J. 2022. Scene manipulation and multimodal querying techniques for complex data exploration. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, 1535–1543.
- Li, X.; Zheng, H.; Chen, J.; Zong, Y.; and Yu, L. 2024. User interaction interface design and innovation based on artificial intelligence technology. *Journal of Theory and Practice of Engineering Science*, 4(03): 1–8.
- Li, Y.; Zhang, Y.; and Sun, L. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Miao, Q.; Li, K.; Quan, J.; Min, Z.; Ma, S.; Xu, Y.; Yang, Y.; Liu, P.; and Luo, Y. 2025. Advances in 4d generation: A survey. *arXiv preprint arXiv:2503.14501*.
- PlayCanvas; and Contributors. 2025. SuperSplat Online Editor. Accessed: November 2025. Browser-based tool for editing and visualizing 3D Gaussian Splats.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.
- Singer, U.; Sheynin, S.; Polyak, A.; Ashual, O.; Makarov, I.; Kokkinos, F.; Goyal, N.; Vedaldi, A.; Parikh, D.; Johnson, J.; et al. 2023. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*.
- Sun, W.; You, X.; Zheng, R.; Yuan, Z.; Li, X.; He, L.; Li, Q.; and Sun, L. 2024. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.
- Wei, L.; Li, Y.; Wang, C.; Wang, Y.; Kong, L.; Huang, W.; and Sun, L. 2025. Unsupervised Post-Training for Multi-Modal LLM Reasoning via GRPO. *arXiv preprint arXiv:2505.22453*.
- Wu, Z.; Yu, C.; Jiang, Y.; Cao, C.; Wang, F.; and Bai, X. 2024. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*, 361–379. Springer.
- Xie, Y.; Yao, C.-H.; Voleti, V.; Jiang, H.; and Jampani, V. 2024. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*.
- Xu, Z.; Peng, S.; Lin, H.; He, G.; Sun, J.; Shen, Y.; Bao, H.; and Zhou, X. 2024. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20029–20040.
- Yan, Z.; Sun, W.; Zhou, R.; Yuan, Z.; Zhang, K.; Li, Y.; Liu, T.; Li, Q.; Li, X.; He, L.; and Sun, L. 2024. Biomedical SAM 2: Segment Anything in Biomedical Images and Videos. *arXiv:2408.03286*.
- Yu, H.-X.; Duan, H.; Hur, J.; Sargent, K.; Rubinstein, M.; Freeman, W. T.; Cole, F.; Sun, D.; Snively, N.; Wu, J.; et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6658–6667.
- Yuan, Z.; Li, Z.; Huang, W.; Ye, Y.; and Sun, L. 2023. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*.
- Yuan, Z.; Liu, Y.; Cao, Y.; Sun, W.; Jia, H.; Chen, R.; Li, Z.; Lin, B.; Yuan, L.; He, L.; et al. 2024a. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*.
- Yuan, Z.; Zhou, R.; Wang, H.; He, L.; Ye, Y.; and Sun, L. 2024b. Vit-1.58 b: Mobile vision transformers in the 1-bit era. *arXiv preprint arXiv:2406.18051*.
- Zhou, X.; Xu, Y.; Tie, G.; Chen, Y.; Zhang, G.; Chu, D.; Zhou, P.; and Sun, L. 2025. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827*.