ISSN: 2581-4311
DOI: doi.org/10.30967/IJCRSET/Alberto-Hojel/152

URF PUBLISHERS
connect with research world

International Journal of Current Research in Science, Engineering & Technology

https://urfpublishers.com/journal/ijcrset

Vol: 7 & Iss: 4

Review

Lucid v1: Real-Time Latent World Models

Rami Seid, Alberto Hojel*

Lucid Simulations, Corp

Citation: Hojel A, Seid R. Lucid v1: Real-Time Latent World Models. *Int J Cur Res Sci Eng Tech* 2024; 7(4), 116-119. DOI: doi.org/10.30967/IJCRSET/Alberto-Hojel /152

Received: 21 December, 2024; **Accepted:** 24 December, 2024; **Published:** 26 December, 2024

***Corresponding author:** Alberto Hojel, Lucid Simulations, Corp. North, E-mail: alberto@lucidsim.com

ABSTRACT

Recent advances in neural world models have demonstrated impressive capabilities in simulating complex virtual environments. However, achieving real-time performance on consumer hardware remains a significant challenge, limiting their practical applications for studying causal relationships in interactive settings. We present Lucid v1, a world model for Minecraft that achieves real-time performance through extreme latent space compression. Our approach combines a VAE that reduces each frame to just 15 tokens with a causally-trained diffusion model that captures temporal dynamics. While previous approaches require hundreds of tokens per frame, our highly compressed representation enables inference speeds of over 20 FPS on consumer GPUs and 60 FPS on H100 accelerators. We demonstrate that despite this aggressive compression, the model maintains visual fidelity and causal consistency, successfully capturing complex game mechanics like physics interactions and environmental persistence. Our results show that extreme token compression can enable truly real-time neural world models while preserving the causal relationships necessary for interactive environments. You can interact with Lucid v1 at: https://lucidv1-demo.vercel.app/

1. Introduction

Recent advances in generative AI have shown remarkable progress in creating interactive virtual environments, with models like Genie[1] and GameNGen[2] demonstrating the potential of neural networks to simulate complex game worlds. However, achieving real-time interaction remains a significant challenge, with current approaches typically operating at less than 5 frames per second on consumer hardware. This limitation severely restricts their practical applications and ability to serve as testbeds for studying causal relationships in virtual environments.

We present Lucid v1, a world model for Minecraft that achieves real-time performance through extreme latent space compression. Our approach combines a compact VAE that reduces each frame to just 15 tokens with a causally trained diffusion model that captures the temporal dynamics of the environment. While previous approaches like Oasis require hundreds of tokens per frame, our highly compressed representation enables inference speeds of over 20 FPS on consumer GPUs and 60 FPS on H100 accelerators, making real-time interaction feasible.

The key to our approach lies in the diffusion model's causal structure, which conditions each frame generation on both past observations and actions. This allows the model to learn and respect the causal relationships inherent in the environment - from basic physics to complex game mechanics. Unlike previous work that sacrifices temporal consistency for compression, our model maintains coherent long-term behavior while operating at a fraction of the computational cost. Our main contributions is a real-time world model architecture achieving unprecedented inference speeds through extreme token compression.

These advances represent a step toward practical neural world models that can serve as platforms for studying causal relationships in complex virtual environments. The real-time performance of Lucid v1 enables immediate feedback loops critical for understanding cause and effect in interactive settings.

2. Related Work

Interactive Neural Environments. Recent work has demonstrated significant progress in creating neural networks that can simulate interactive environments. World Models[3] introduced the concept of learning environment dynamics from raw pixels, while GameGAN[4] showed how GANs could generate game frames conditioned on actions. More recently, Genie[1] and GameNGen[2] have shown how diffusion models can generate high-quality game environments from prompts. DIAMOND[5] further demonstrated how diffusion models can serve as world models for reinforcement learning. However, these approaches typically require significant computational resources and operate far below realtime speeds on consumer hardware.

Latent Compression. A key challenge in world modeling is efficiently representing high-dimensional observations. Recent video generation models like Phenaki[6] and Sora use discrete autoencoders with hundreds of tokens per frame, while IRIS[7] and DreamerV3[8] leverage smaller discrete latent spaces for reinforcement learning. Our approach pushes the boundaries of compression by reducing each frame to just 15 tokens while maintaining visual fidelity and temporal consistency. This extreme compression is crucial for achieving real-time performance.

Causal World Models. The field of causal modeling in virtual environments has gained increasing attention, with works like TWM[9] and STORM[10] demonstrating how transformer architectures can capture causal relationships in game dynamics. These models typically require substantial context lengths to maintain temporal consistency. In contrast, our highly compressed latent space allows the diffusion model to efficiently process longer temporal sequences while preserving causal dependencies between states and actions.

Real-time Neural Rendering. Previous attempts at real time neural game engines have faced significant challenges in balancing visual quality with inference speed. Video GPT[11] and Godiva[12] demonstrated early success in generating game-like videos but were limited to offline generation. More recent work like Oasis has shown progress toward real-time interaction but remains constrained to specialized hardware. Our work bridges this gap by achieving real-time performance on consumer GPUs through aggressive latent compression.

3. Method

Our approach consists of two primary components: (1) a highly compressed VAE[13] that reduces Minecraft frames to a minimal latent representation and (2) a diffusion transformer[14] that models the temporal dynamics in this compressed space. This design enables real-time interaction by dramatically reducing the computational complexity of frame generation **(Figure 1)**.

We train a VAE to encode each frame $x_t \in R^{H\times W\times 3}$ into a compact latent representation $z_t \in R^{h\times w\times d}$ where h,w are the spatial dimensions of the latent space and d is the embedding dimension. Unlike previous approaches that use hundreds of tokens, we aggressively compress each frame to just 15 latent tokens, reducing the quadratic attention cost in the subsequent diffusion model. The VAE is trained using a combination of reconstruction loss and GAN-based perceptual loss:

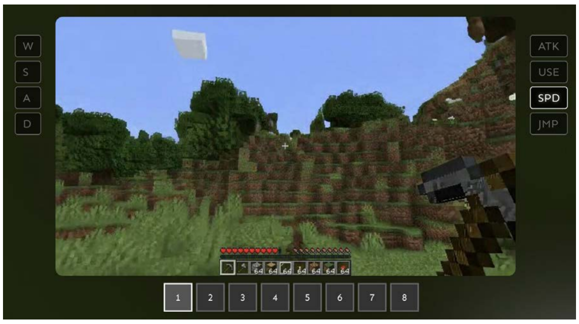

Figure 1: Lucid v1: A real-time playable world model that emulates minecraft.

$$L_{VAE} = E_{x\sim p_{\text{data}}}[\|x-\hat{x}\|_2^2 + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}] \qquad (1)$$

where xˆ is the reconstructed frame. The GAN objective helps preserve fine details despite the extreme compression ratio.

Our world model operates in the compressed latent space using a diffusion transformer (DiT) conditioned on both past observations and actions. Given the current latent state z_t and action a_t, the model predicts the next latent state z_{t+1}. We use Rectified Flow (Liu, Gong and Liu 2022) for training, which offers better stability than traditional diffusion approaches:

$$\mathcal{L}_{\text{DiT}} = E_{z_t,a_t,z_{t+1}}[\|\epsilon_\theta(z_t^\tau,\tau,c)-\epsilon\|_2^2] \qquad (2)$$

where $c = [z_{<t}, a_{<t}]$ is the causal context, τ is the diffusion time and ϵ is Gaussian noise. The model architecture uses alternating spatial and temporal attention layers to efficiently process the sequence of compressed tokens.

The resulting pipeline achieves frame generation rates of 20+ FPS on NVIDIA RTX 4090 and 60+ FPS on NVIDIA H100. **(Figure 2 and 3)**

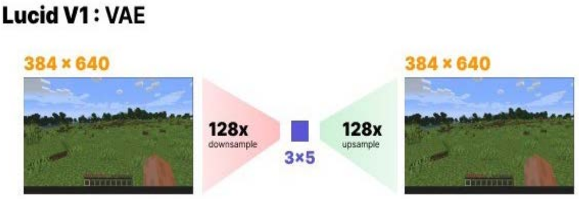

Figure 2: Visualization of our VAE.

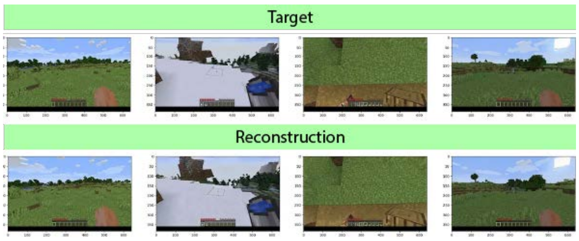

Figure 3: Final VAE reconstructions, after 150k training steps.

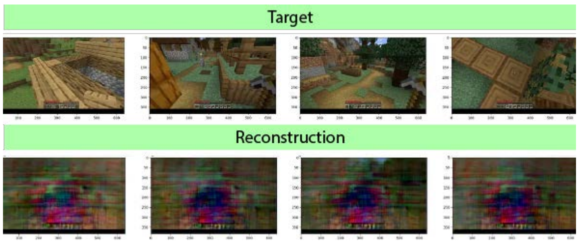

Figure 4: Intermediate VAE reconstructions, after 30k training steps.

The VAE training initially focuses on individual frames before introducing the GAN discriminator to improve reconstruction quality. Our final model achieves somewhat stable long-term generation while maintaining frame rates suitable for real-time interaction. **(Table 1)**

Table 1: Frame generation speed comparison (FPS).

Model	RTX 4090	H100	Token Count	
Oasis	4	20	256	
DIAMOND		*	15	*
Lucid v1 (Ours)		20	60	15

4. Results

We evaluate Lucid v1 on both quantitative metrics of performance and qualitative assessments of visual quality and causal consistency. Our primary baseline comparisons are against recent neural game engines Oasis and GameNGen, as they represent the current state-of-the-art in real-time interactive environments. **(Figure 5)**

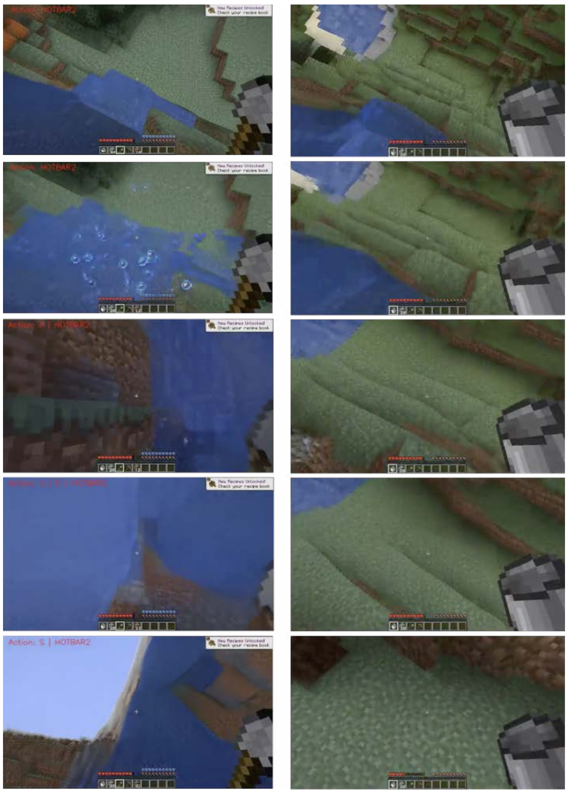

Figure 5: Frame sequences from two different trajectories, on the left the player falls onto water and receives no damage and on the right the player falls onto the ground and does receive damage.

4.1. Performance Analysis

Inference Speed. Table 1 shows frame generation speeds across different hardware configurations. Our model achieves significantly higher frame rates than existing approaches, making it the first neural game engine capable of true real-time interaction on consumer hardware.

4.2. Visual Quality

VAE Reconstruction. Figure 3 shows the reconstruction of our final VAE after training. Despite the aggressive compression to 15 tokens, our model successfully preserves key visual details necessary for gameplay:

- Structural elements (blocks, terrain)
- Interactive objects (items, entities)
- UI elements (inventory, health bar)

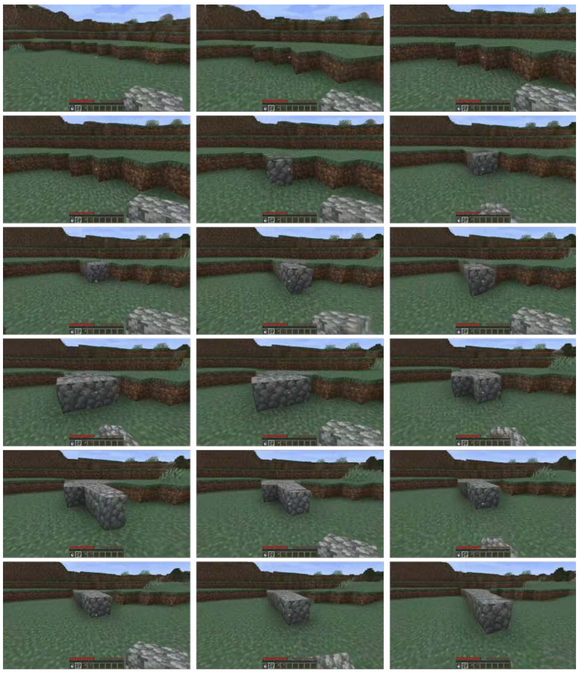

Figure 6: Action rollout showcasing building ability.

In Figure 4 we show the reconstructions of the VAE during training, at step 30k out of a total 150k, showing bad reconstructions.

Temporal Consistency. Figure 5 presents two different sequences of generated frames, where the player falls onto water and onto ground showcasing the ability to adhere to the ground truth logic of the Minecraft game. When landing on water, the player takes no damage. On the other hand, when landing on the ground, the player does take damage accordingly. Figure 6 shows a sequence of generated frames demonstrating the model's ability to maintain causal consistency over time and the ability to place down blocks to build structures from Cobblestone. The causally-trained diffusion model successfully captures:

- Physics-based interactions (gravity, collisions)
- Player actions and their consequences
- Environmental persistence across frames

5. Conclusion

Our results demonstrate that extreme token compression combined with causally-trained diffusion models can enable truly real-time neural world models. While previous approaches have focused on maximizing visual fidelity through larger latent spaces, we show that aggressive compression to just 15 tokens per frame can maintain sufficient quality for interactive environments while dramatically improving inference speed.

Despite the advances presented, several limitations remain. First, our model occasionally exhibits temporal inconsistencies during rapid scene changes, suggesting that the extreme compression may occasionally discard relevant causal information. Second, while we achieve real-time performance on consumer hardware, the model still requires significant computational resources for training. Third, the current approach is specific to Minecraft's visual domain and generalization to more complex 3D environments remain an open challenge.

Lucid v1 demonstrates that extreme token compression can enable real-time neural world models while maintaining visual quality and causal consistency. By achieving over 20 FPS on consumer hardware, our approach makes interactive neural environments accessible to a broader research community. We hope this work encourages further exploration of efficient architectures for real-time world modeling and advances our understanding of causal relationships in virtual environments.

6. References

1. Bruce J, Dennis M, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald, R, Apps C, Aytar Y, Bechtle S, Behbahani F, Chan S.
2. Valevski D, Leviathan Y, Arar M and Fruchter S. 2024. Diffusion Models Are Real-Time Game Engines.
3. Ha D and Schmidhuber J. World Models, 2018.
4. Kim SW, Zhou Y, Philion J, Torralba A and Fidler S. Learning to Simulate Dynamic Environments with GameGAN, 2020.
5. Alonso E, Jelley A, Micheli V, Kanervisto A, Storkey A, Pearce T and Fleuret F. 2024. Diffusion for World Modeling: Visual Details Matter in Atari, 2024.
6. Villegas R, Babaeizadeh M, Kindermans PJ, Moraldo H, Zhang H, Saffar MT, Castro S, Kunze J and Erhan D. Phenaki: Variable Length Video Generation from Open Domain Textual Description, 2022.
7. Micheli V, Alonso E and Fleuret F. Transformers are Sample-Efficient World Models, 2023.
8. Hafner D, Pasukonis J, Ba J and Lillicrap T. Mastering Diverse Domains through World Models, 2024.
9. Robine J, Hoftmann M, Uelwer T and Harmeling S. Transformer-based World Models Are Happy With 100k Interactions, 2023.
10. Zhang J, Zolna K, Clune J, de Freitas N, Singh S and Rocktaschel T. Genie: Generative Interactive Envi-¨ ronments, 2024.
11. Yan W, Zhang Y, Abbeel P and Srinivas A. VideoGPT: Video Generation using VQ-VAE and Transformers, 2021.
12. Wu C, Huang L, Zhang Q, Li B, Ji L, Yang F, Sapiro G and Duan N. GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions, 2021.
13. Kingma DP and Welling M. Auto-Encoding Variational Bayes, 2022.
14. Peebles W and Xie S. Scalable Diffusion Models with Transformers, 2023.