# Walking in the Implicit: Interactive World Exploration via Neural Scene Representation

Zhiqi Li[1,2], Chengrui Dong[1,2], Zhenhua Du[1,2], Hangning Zhou[3†], Cong Qiu[3], Hailong Qin[3], Mu Yang[3], Dongxu Wei[2], and Peidong Liu[2⋆]

[1]Zhejiang University [2]Westlake University [3]Afari Intelligent Drive

https://lizhiqi49.github.io/NeuWorld

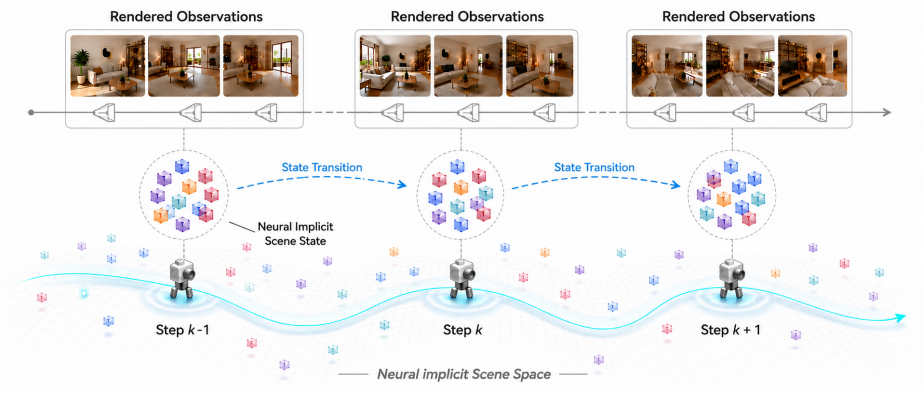


**Fig. 1: Walking in the Implicit.** NeuWorld rolls out a fixed-length Neural Implicit Scene state and renders queried observations from it under camera control.

**Abstract.** Interactive video generation systems for camera-controlled world exploration roll out growing sequences of latent video frames, entangling state transition with high-frequency observation synthesis. We propose *Walking in the Implicit*, a scene-centric paradigm that changes the rollout variable from frame latents to a fixed-length, renderable implicit state, termed *Neural Implicit Scene* (NIS). This factorizes interactive generation into stochastic transition of a compact scene state and deterministic pose-conditioned rendering given the sampled state. We instantiate this paradigm as NeuWorld: a transformer VAE learns locally anchored NIS from sparse posed frames, and a diffusion transformer evolves NIS conditioned on future camera trajectories and geometry-aware retrieved history. By reusing the VAE encoder as a unified conditioner, NeuWorld maps camera, reference-image, and history cues into the same NIS modality, avoiding external heterogeneous encoders. Trained from scratch on public posed-view data without pretrained video backbones or auxiliary 3D reconstructors, NeuWorld achieves strong long-horizon consistency with favorable inference efficiency.

**Keywords:** Interactive World Exploration · Camera-Controlled Generation · Implicit Representation

† Project Lead. ⋆ Corresponding Author.

## 1 Introduction

A useful camera-controlled world model should let an agent move through a scene, synthesize future observations along queried camera motions, and retain consistency upon revisitation. Recent interactive systems build on powerful pretrained video diffusion backbones and steer them with actions or camera trajectories [1,5,55,71]. These systems generate visually plausible videos, but directly rolling out observations entangles state transition with high-frequency appearance synthesis in a single process. This entanglement makes long-horizon consistency increasingly difficult to maintain.

A natural route toward consistency is to introduce explicit 3D structure. Handcrafted scene representations such as NeRF [42] and 3D Gaussian Splatting [33] provide strong geometric inductive biases, and reconstruction-based interactive systems repeatedly rebuild a 3D/4D representation to support revisitation [20,36,40,50]. This route is powerful, but metric reconstruction is heavier than necessary for local camera-controlled exploration. A more direct abstraction is a compact, renderable scene state that lies between frame latents and explicit reconstruction: it should preserve sufficient geometry for consistent view synthesis while remaining suitable for generative latent transition.

Recent NVS models such as LVSM [31] and RayZer [30] offer such a candidate by encoding sparse posed views into a fixed-length latent token set which can support local novel view synthesis through a decoder. We refer to this renderable token set as a *Neural Implicit Scene* (NIS). In this work, NIS is not used merely as an NVS condition, but as the state variable of interaction. This leads to *Walking in the Implicit*: instead of rolling out future frame latents, the model rolls out a locally anchored NIS state that can be queried by camera poses. Each interaction step is thereby factorized into generative transition in NIS space and pose-conditioned rendering from the sampled state, as summarized in Figure 1.

We instantiate this formulation as NeuWorld, a camera-controlled exploration system built around NIS. A transformer VAE (NIS-VAE) learns to encode posed views into NIS states and decode queried target views, while a diffusion transformer (NIS-DiT) samples the next local NIS state under camera control. To align conditioning with the rollout variable, NeuWorld reuses the NIS encoder as a unified conditioner: camera-only cues, camera-and-reference-image cues, and retrieved history are all mapped into *partial NIS* or memory NIS tokens, rather than handled by separate image, camera, or reconstruction encoders. The sampled NIS is then rendered by the frozen decoder to produce future observations.

We validate NeuWorld on static-scene camera-controlled exploration, which isolates the representation question of whether a compact scene state can support local re-anchoring, memory-conditioned rollout, and revisitation consistency. Both NIS-VAE and NIS-DiT are trained *from scratch* on public posed-view datasets, Re10K [77] and DL3DV [39], without pretrained video foundation models or auxiliary 3D reconstructors. Experiments on forward trajectory generation and cycle revisitation show strong long-horizon pose and revisit consistency with favorable inference efficiency, and ablations confirm the roles of NIS rollout, partial-NIS conditioning, anti-drift augmentation, and geometry-aware retrieval.

*Contributions.* Our contributions are threefold. First, we propose *Walking in the Implicit*, a scene-centric rollout formulation that replaces growing video-latent trajectories with a fixed-length, renderable NIS state and decouples latent state transition from pose-conditioned rendering. Second, we instantiate this formulation as NeuWorld, where NIS-VAE provides a renderable latent scene interface and NIS-DiT samples local NIS states under unified NIS-modality conditioning from camera, reference-image, and history cues. Third, we validate this formulation by training NeuWorld from scratch on public posed-view datasets, showing favorable long-horizon consistency and inference efficiency.

## 2 Related Work

**Latent Scene Representations for NVS.** Novel view synthesis (NVS) has long relied on scene representations that can render images under queried viewpoints. NeRF [42] and its variants advance neural volumetric rendering in quality [6, 7, 57], speed [25, 48, 49], and generalization to sparse inputs [41, 44, 61]. Explicit structures like voxels [19, 54], hash grids [43], point-based representations [18, 63, 75], and 3D Gaussian Splatting [33] trade off compactness for rendering speed. Data-driven methods further learn generalizable NVS from posed views [10,11,14,37,59,67,69]. Closest to our representation, SRT [53], LVSM [31], and RayZer [30] encode sparse posed observations into a fixed-length latent token set and decode target views from these tokens, demonstrating that a compact tokenized scene representation can be both renderable and Transformer-compatible. Our work builds on this representational insight but changes its role: rather than using latent scene tokens only for NVS, NeuWorld uses them as a local rollout state that is transitioned under camera motion, conditioned on generated history, and queried again under revisitation.

**Camera-Controlled Video Generation.** Camera control has become an important steering interface for video generation [8,13,66]. Existing methods inject camera information into video generators through motion adapters [21,27], 6DoF camera encoders [23,60], cross-video synchronization [34], trajectory-aware generation [3, 65], or diffusion-transformer camera knowledge [2]. These methods improve controllability by enabling a video model to synthesize a coherent clip from a reference image or prompt under a specified camera trajectory. These works provide the camera-conditioned generation foundation for visual world simulation, where camera motion becomes the primary control signal.

**Interactive and Consistent World Models.** World models predict future observations or states from past observations and actions [22, 45, 64]. Recent interactive video generation systems extend camera or action signals into repeated rollouts for world exploration [1, 5, 15, 24, 55, 71]. This setting requires efficient iterative inference and consistency when the camera revisits previously observed regions. Existing methods improve long-horizon behavior by distilling video generation backbones for faster rollout [12,28,68], using reconstruction or geometry

modules to organize historical evidence [9, 20, 36, 40, 50, 72], or retrieving relevant past observations based on geometry or field-of-view overlap [62, 70]. These approaches provide important mechanisms for repeated visual generation, but largely keep the rollout variable in the form of latent video frames. Our work instead explores a different rollout variable: a compact, renderable implicit state that can be transitioned under camera motion and queried again for observation synthesis.

## 3 Method

We study camera-controlled interactive exploration in static scenes, where an agent observes the current posed view, specifies a short future camera trajectory, and synthesizes future observations that remain consistent over long rollouts and revisitation. NeuWorld instantiates the *Walking in the Implicit* formulation by representing the interaction state as a compact, fixed-length *Neural Implicit Scene* (NIS): a locally anchored scene state that covers the upcoming trajectory segment and is re-anchored as the agent moves. Each rollout step is factorized into **scene-state transition** in NIS space and **pose-conditioned rendering** from the sampled NIS. Concretely, we first train a transformer VAE (NIS-VAE) to encode sparse posed context views into NIS tokens and render target views (Section 3.2); we then freeze NIS-VAE and train a set-based diffusion transformer (NIS-DiT) to sample future NIS states under camera and history conditions (Section 3.3). At inference, geometry-aware retrieval is utilized to build highly relevant history and reduce long-horizon drift (Section 3.4). The overall pipeline is illustrated in Figure 2.

### 3.1 Problem Formulation

Each posed view is a pair $(I, \mathbf{T})$, where $I \in \mathbb{R}^{H\times W\times 3}$ is an RGB image and $\mathbf{T} \in SE(3)$ is the camera pose. For NIS-VAE training, we use context views $\mathcal{V}_{ctx} = \{(I^{(i)}_{ctx}, \mathbf{T}^{(i)}_{ctx})\}_{i=1}^{M}$ and target views $\mathcal{V}_{tgt} = \{(I^{(j)}_{tgt}, \mathbf{T}^{(j)}_{tgt})\}_{j=1}^{N}$. During interactive inference at step $k$, the agent observes the current posed view $O_k = (I_k, \mathbf{T}_k)$ and specifies a future camera segment $\mathcal{T}^{(k)}_{\text{fut}} = \{\mathbf{T}^{(k)}_j\}_{j=1}^{S}$. We take $\mathbf{T}_k$ as the local coordinate origin, retrieve a small history set $\mathcal{H}_k$ from a memory bank $\mathcal{M}_k$ of past posed views, and use sparse poses from $\mathcal{T}^{(k)}_{\text{fut}}$ when constructing encoder-based conditions.

Each rollout step is factorized into stochastic NIS state sampling and deterministic pose-conditioned rendering. We first sample the next local scene state $\hat{\mathbf{z}}^{(k)} = \mathcal{G}_\theta(O_k, \mathcal{T}^{(k)}_{\text{fut}}, \mathcal{H}_k)$, where $\mathcal{G}_\theta$ is the diffusion sampling procedure parameterized by NIS-DiT. The frozen NIS decoder then renders future observations under queried poses: $\hat{I}^{(k)}_j = \mathcal{D}_\phi\left(\hat{\mathbf{z}}^{(k)}, \mathbf{T}^{(k)}_j\right), \quad \mathbf{T}^{(k)}_j \in \mathcal{T}^{(k)}_{\text{fut}}$. Here $\hat{\mathbf{z}}^{(k)}$ is a locally anchored, fixed-capacity NIS state for the upcoming trajectory segment. This formulation keeps the rollout variable compact while exposing consistency through rendering multiple queried views from the same sampled state.

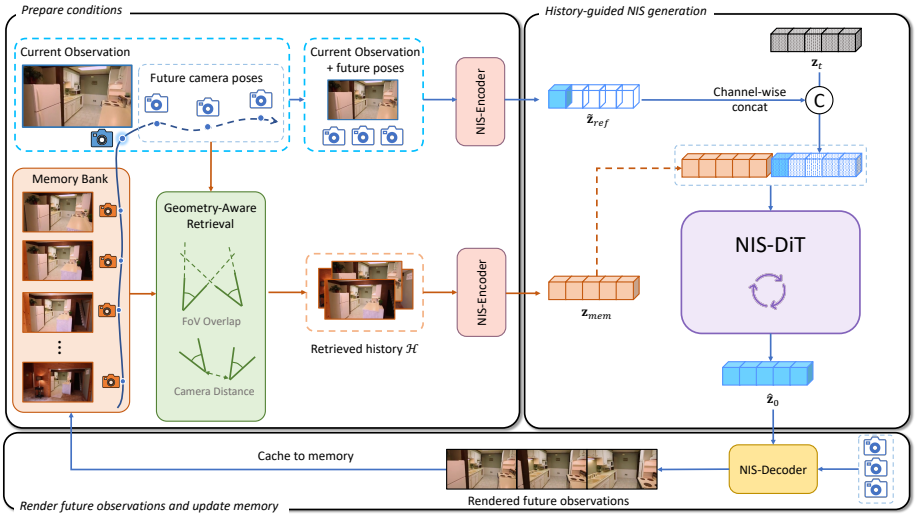


**Fig. 2: Method overview.** At an interaction step, the frozen NIS-VAE encoder maps the current observation and a sparse future pose trajectory to a partial NIS condition $\tilde{\mathbf{z}}_{\text{ref}}$, which is injected by channel-wise concatenation with the noised latent $\mathbf{z}_t$. Geometry-aware retrieval selects a history set $\mathcal{H}$ and encodes it as memory NIS tokens $\mathbf{z}_{\text{mem}}$, which are appended by token-wise concatenation. NIS-DiT samples the next local NIS state $\hat{\mathbf{z}}_0$, and the frozen decoder renders future views under the corresponding future poses.

### 3.2 NIS-VAE: Learning Renderable Scene States

**Architecture.** NIS-VAE follows the encoder-decoder paradigm of LVSM-style latent scene models, with both encoder and decoder implemented as transformer stacks. Given context views $\mathcal{V}_{ctx}$, we convert each camera pose into a pixel-wise Plücker ray embedding [47] and concatenate it with RGB channels. The resulting image-ray fields are patchified into tokens and processed by a transformer encoder together with $L$ learnable query tokens. The output query tokens parameterize a diagonal Gaussian posterior, from which we sample the NIS latent by reparameterization:

$$\mathbf{z} = \mu + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{z} \in \mathbb{R}^{L \times D}.$$

The latent $\mathbf{z}$ is a fixed-length token set, independent of the number of target frames to be rendered. We denote the encoder mapping as $\mathbf{z} = \mathcal{E}(\mathcal{V}_{ctx})$.

Given NIS $\mathbf{z}$ and a target pose $\mathbf{T}$, the decoder predicts $\hat{I} = \mathcal{D}_\phi(\mathbf{z}, \mathbf{T})$ by attending target ray tokens to the NIS tokens and decoding RGB patches. Following common image-level autoencoder training [17, 29, 32, 51], we train NIS-VAE with reconstruction, perceptual, adversarial, and KL regularization losses:

$$\mathcal{L}_{VAE} = \mathcal{L}_{MSE} + \lambda_1 \mathcal{L}_{Percep} + \lambda_2 \mathcal{L}_{GAN} + \lambda_3 \mathcal{L}_{KL}. \tag{1}$$

**NIS encoder as a unified conditioner.** Beyond encoding full context views into renderable scene states, the frozen NIS encoder also provides a common interface for constructing conditions in the same modality as the rollout state. For pose-only conditioning, we sample sparse poses $\mathcal{T}_{sparse} = \{\mathbf{T}^{(i)}\}_{i=1}^{M}$ along the future trajectory and zero all RGB images:

$$\tilde{\mathbf{z}}_{\text{pose}} = \mathcal{E}\Big(\{(\mathbf{0}, \mathbf{T}^{(i)})\}_{i=1}^{M}\Big).$$

For pose+reference conditioning, we keep one reference image $I^*$ at index $r$ and zero the remaining images:

$$\tilde{\mathbf{z}}_{\text{ref}} = \mathcal{E}\Big(\{(\tilde{I}^{(i)}, \mathbf{T}^{(i)})\}_{i=1}^{M}\Big), \qquad \tilde{I}^{(i)} = \begin{cases} I^*, & i = r, \\ \mathbf{0}, & i \neq r. \end{cases}$$

Retrieved history frames are encoded in the same way as memory NIS tokens, $\mathbf{z}_{\text{mem}} = \mathcal{E}(\mathcal{H}_k)$. Thus, camera-only cues, camera-and-reference-image cues, and history cues are all represented as NIS tokens before being consumed by NIS-DiT. Fig. 4 and additional visual results in Appendix A.1 provide evidence that such partial NIS inputs remain non-collapsing and retain useful camera-aligned structure.

### 3.3 NIS-DiT: Conditional Latent Dynamics

**Set-based diffusion transformer.** We model latent state evolution with a diffusion transformer (DiT) [46] operating directly on NIS token sets. Since NIS tokens are not tied to a raster grid, temporal order, or explicit 3D grid, we omit spatial and temporal positional encodings in the denoiser and let self-attention operate on the token set [56]. The learned query slots of NIS-VAE provide a shared canonical interface across clean, noised, and partial NIS latents, making slot-wise conditioning well defined. We use a U-shaped transformer backbone [4, 52] with long skip connections, AdaLN-style timestep modulation [46], RMSNorm [74], and $Q, K$ normalization [16].

**Denoising with NIS conditions.** NIS-DiT samples the next local NIS state conditioned on partial NIS and optional memory NIS tokens produced by the frozen encoder. Since $\tilde{\mathbf{z}}_{\text{pose}}$, $\tilde{\mathbf{z}}_{\text{ref}}$, and the denoising target share the same token shape, we inject partial NIS by channel-wise concatenation:

$$\mathbf{z}_t^{in} = \text{Concat}(\mathbf{z}_t, \tilde{\mathbf{z}}_{\text{partial}}), \qquad \tilde{\mathbf{z}}_{\text{partial}} \in \{\tilde{\mathbf{z}}_{\text{pose}}, \tilde{\mathbf{z}}_{\text{ref}}\}. \tag{2}$$

The concatenated tokens are projected to the DiT hidden width by a linear input projection. Memory NIS tokens $\mathbf{z}_{\text{mem}}$ are appended by token-wise concatenation, allowing the denoiser to attend to retrieved history as latent evidence for state sampling. In this way, all conditioning signals are consumed in the NIS modality rather than through separate image, camera, or reconstruction encoders.

**Flow-matching objective.** We obtain clean target latents $\mathbf{z}_0$ by encoding the corresponding ground-truth posed views with the frozen NIS-VAE encoder. Given $\mathbf{z}_0$, we sample $t \sim p(t)$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, construct $\mathbf{z}_t = t\mathbf{z}_0 + (1-t)\boldsymbol{\epsilon}$, and train NIS-DiT to predict the velocity $\mathbf{v}^\star = \mathbf{z}_0 - \boldsymbol{\epsilon}$. The DiT predicts $\hat{\mathbf{v}} = f_\theta(\mathbf{z}_t, t, \mathbf{c})$, and we optimize

$$\mathcal{L}_{DiT} = \mathbb{E}\Big[\|\hat{\mathbf{v}} - \mathbf{v}^\star\|_2^2\Big]. \tag{3}$$

Here $\mathbf{c}$ denotes the NIS-condition bundle, including partial NIS and optional memory NIS tokens. We drop all conditions with 10% probability to enable classifier-free guidance (CFG) [26].

**Training curriculum.** We train NIS-DiT with a weak-to-strong curriculum to stabilize from-scratch learning and avoid early shortcut copying from strong appearance conditions. The model first learns an NIS prior with pose-only partial NIS, then incorporates pose+reference partial NIS to align sampled states with the current observation, and finally adds memory NIS for history-aware interactive generation. During the later stages, we randomly fall back to weaker conditions to preserve the pose-conditioned NIS prior and maintain cold-start robustness when retrieved history is limited. The detailed motivation, stage schedule, and fallback probabilities are provided in our appendix.

### 3.4 Geometry-Consistent Long-Horizon Interaction

**Anti-drift condition augmentation.** The key train-test gap in long-horizon interaction is history-quality mismatch: training history comes from ground truth, while inference history is generated and may contain blur, aliasing, or local drift. During Stage-3 training, we stochastically degrade history images using Gaussian blur, downsample-then-upsample, or VAE-reconstruction replacement. After encoding conditions, we further inject latent condition noise: $\tilde{\mathbf{z}}_{\text{cond}} = \mathbf{z}_{\text{cond}} + \gamma\boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z}_{\text{cond}}$ denotes the conditioning latents before noise injection, including partial NIS conditions and, when available, memory latents. $\gamma$ is the augmentation strength sampled from a scaled Beta distribution. During training, we condition DiT on $\gamma$ via AdaLN-like modulation so that it adapts to imperfect conditions across noise levels. At inference, we use the same noise-level conditioning interface and ramp $\gamma$ with interaction step $k$ to account for increasingly noisy generated history: $\gamma_k = \gamma_{min} + \min\Big(\frac{k}{K_{\text{ramp}}}, 1\Big)(\gamma_{max} - \gamma_{min})$, where $\gamma_{min}$ and $\gamma_{max}$ are the lower and upper augmentation bounds, and $K_{\text{ramp}}$ is the ramp length.

**Hybrid geometry-aware memory retrieval.** To provide relevant history for long-horizon rollout, we maintain a memory bank of past generated frames and their camera poses, $\mathcal{M}_k = \{(I_i, \mathbf{T}_i)\}_{i=0}^{N_k-1}$, at interaction step $k$ and retrieve a small history set $\mathcal{H}_k$ for memory conditioning. Following common practice in

retrieval-based interactive generation, we combine two types of evidence: recent frames for local temporal continuity and globally retrieved frames for geometric recall. The global retrieval score considers pose distance, estimated field-of-view overlap with the upcoming trajectory, and a weak recency prior. Instead of querying only the endpoint pose, we score candidate history frames against a sparse set of poses sampled from the future trajectory segment. The selected global frames are filtered with a pose-space diversity constraint and then merged with recent frames to form $\mathcal{H}_k$. The retrieved history is encoded by the frozen NIS encoder as memory NIS tokens, $\mathbf{z}_{\text{mem}} = \mathcal{E}(\mathcal{H}_k)$, which are used as latent evidence for history-conditioned denoising. For detailed scoring functions and retrieval hyperparameters please refer to the appendix.

## 4 Experiments

We evaluate NeuWorld on static-scene interactive exploration under camera control, focusing on long-horizon geometric consistency. Experiments are conducted on Re10K and DL3DV-10K. This section summarizes datasets, training details, inference-time settings, and evaluation protocols used throughout the paper.

### 4.1 Experimental Setup

**Implementation details.** We train our main models (NIS-VAE and NIS-DiT) from scratch on posed multi-view data from Re10K [77] and DL3DV-10K [39], and evaluate interactive exploration on both test splits. All images are center-cropped and resized to $256 \times 256$. We use a fixed-length NIS with $L$=1024 tokens and $D$=64 channels; Stage 1/2 use 8 context views and Stage 3 uses 15 to expose both history and a future trajectory segment. Training uses 16 A100 GPUs for roughly one week in total, without relying on pretrained video generation backbones or auxiliary 3D reconstruction modules. Unless otherwise stated, we sample NIS-DiT with 50 steps and CFG scale $s$=4.0. Most compared baselines inherit pretrained image/video generation priors or auxiliary reconstruction modules, whereas NeuWorld is trained from scratch on the same public posed-view datasets. We therefore report both accuracy and runtime to contextualize the accuracy–efficiency trade-off under this different training regime.

**Evaluation.** We evaluate NeuWorld under two camera-controlled protocols and report PSNR, SSIM, LPIPS, FID, and pose errors ($R_{\text{dist}}$ and $T_{\text{dist}}$) computed from generated-frame pose estimates. $R_{\text{dist}}$ denotes the rotation distance in degrees and $T_{\text{dist}}$ denotes the normalized translation distance. For forward generation and return-path quality, they are computed between estimated and ground-truth camera trajectories expressed relative to the first frame. For revisit self-consistency, they are computed between the estimated return trajectory and the paired forward trajectory. Since these errors are obtained from an external pose estimator, they should be interpreted as pose-consistency proxies rather than direct camera-control errors; we therefore report them alongside image

**Dataset: Re10K**

| | Short-term ($50^{th}$ frame) | | | | | | Long-term ($200^{th}$ frame) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ |
| SEVA [76] | 0.566 | 11.80 | 0.343 | 52.25 | 0.110 | 0.262 | 0.666 | 10.34 | 0.293 | 79.56 | 0.210 | 0.295 |
| ViewCrafter [73] | 0.664 | 9.91 | 0.195 | 54.60 | 0.173 | 0.605 | 0.697 | 9.47 | 0.160 | 96.13 | 0.226 | 0.609 |
| Gen3C [50] | 0.667 | 10.71 | 0.273 | 76.88 | 0.203 | 0.429 | 0.672 | 10.61 | 0.278 | 131.88 | 0.226 | 0.443 |
| VMem [36] | 0.547 | 12.57 | 0.374 | 55.49 | 0.102 | 0.326 | 0.669 | **10.64** | **0.329** | 64.06 | 0.172 | 0.364 |
| Matrix-Game 2.0 [24] | 0.636 | 9.56 | 0.202 | 86.69 | 0.334 | 0.358 | 0.717 | 7.76 | 0.148 | 114.24 | 0.567 | 0.432 |
| NeuWorld | **0.431** | **15.11** | **0.476** | **34.55** | **0.026** | **0.098** | **0.665** | 10.12 | 0.308 | **54.08** | **0.083** | **0.141** |

**Dataset: DL3DV**

| | Short-term ($20^{th}$ frame) | | | | | | Long-term ($80^{th}$ frame) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ |
| SEVA [76] | 0.607 | 11.86 | 0.254 | 74.68 | 0.401 | 0.420 | 0.645 | 11.34 | 0.245 | **91.41** | 0.647 | 0.488 |
| ViewCrafter [73] | 0.657 | 11.69 | 0.235 | 69.93 | 0.408 | 0.645 | 0.701 | 10.94 | 0.230 | 110.60 | 0.536 | 0.669 |
| Gen3C [50] | **0.548** | 13.00 | 0.267 | 88.99 | **0.150** | 0.627 | **0.598** | **11.91** | 0.262 | 187.72 | **0.182** | 0.640 |
| VMem [36] | 0.674 | 11.63 | 0.287 | 78.09 | 0.489 | 0.646 | 0.697 | 11.45 | **0.291** | 110.81 | 0.717 | 0.697 |
| Matrix-Game 2.0 [24] | 0.672 | 9.50 | 0.160 | 98.09 | 0.349 | 0.378 | 0.725 | 8.22 | 0.127 | 100.76 | 0.742 | 0.432 |
| NeuWorld | 0.580 | **13.48** | **0.353** | **69.31** | 0.178 | **0.217** | 0.656 | 11.66 | 0.281 | 96.19 | 0.376 | **0.274** |

**Table 1: Short-/long-term forward trajectory generation on Re10K and DL3DV.** Each subtable reports mean metrics on 100 test trajectories, evaluated at the $50^{th}/200^{th}$ frames on Re10K and the $20^{th}/80^{th}$ frames on DL3DV. We mark the best, second-best, and third-best results in each column.

metrics instead of using them in isolation. **Forward trajectory generation (short/long horizon):** starting from the first frame, we autoregressively synthesize novel views along the ground-truth (GT) camera trajectory. We subsample trajectories with a 10-frame interval and evaluate at the $50^{th}/200^{th}$ frames on Re10K and the $20^{th}/80^{th}$ frames on DL3DV by comparing generated frames against the corresponding GT frames. **Cycle revisitation:** the camera moves from the start pose to the end pose and then returns along the same path. We report return-path quality (vs. GT when available) and revisit self-consistency by comparing each return frame to its paired frame from the forward pass (start → end), along with ART, the average runtime per forward-and-return trajectory.

### 4.2 Main Results

We report quantitative comparisons under short-/long-term forward generation and cycle-trajectory revisitation, averaged over 100 randomly sampled long trajectories from each dataset's test split. We emphasize long-horizon pose drift, revisitation self-consistency, and runtime because they are particularly diagnostic for camera-controlled interactive exploration. For the cycle setting, runtime is measured by the same evaluation runner and hardware for all methods. We compare against representative interactive world models and camera-controlled baselines, including VMem [36], SEVA [76], Gen3C [50], ViewCrafter [73], and Matrix-Game 2.0 [24].

**Forward trajectory generation.** Table 1 shows that NeuWorld achieves strong camera-controlled novel view generation with low pose drift over long

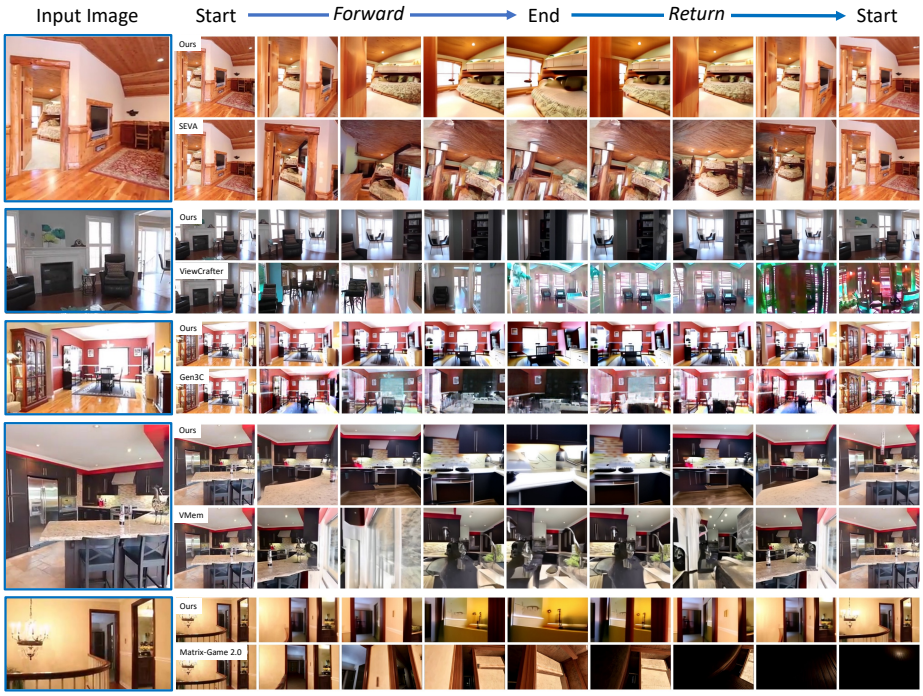


**Fig. 3: Qualitative comparison.** We compare NeuWorld with representative baselines under the same camera-controlled trajectories. Compared with video-latent and reconstruction-based baselines, NeuWorld better preserves scene geometry and appearance consistency under both short-horizon and long-horizon rollout checkpoints.

rollouts. On Re10K, NeuWorld is best across all metrics at the $50^{th}$ frame and achieves the lowest pose errors at the $200^{th}$ frame ($R_{\text{dist}}$=0.083, $T_{\text{dist}}$=0.141) together with the best LPIPS/FID, while remaining competitive on PSNR/SSIM. On DL3DV, NeuWorld ranks within the top two on five of six short-term metrics. Under the more challenging 80th-frame setting, it remains competitive, ranking within the top group on most metrics and achieving the best long-horizon translation consistency ($T_{\text{dist}}$=0.274). The visual comparison in Fig. 3 is consistent with these quantitative trends.

**Cycle revisitation.** Cycle revisitation is particularly diagnostic for interactive exploration, because it tests whether the model can return to previously visited regions without relying on a growing frame buffer or explicit reconstruction. On Re10K, NeuWorld achieves the best revisit self-consistency in LPIPS/SSIM (0.208/0.692) and the lowest return-path translation error ($T_{\text{dist}}$=0.382), while remaining competitive on return-path image quality. Crucially, NeuWorld is also efficient at inference time: 3.24 minutes per forward-and-return trajectory, about 14× faster than VMem and Gen3C (47.62 minutes each) on the Re10K cy-

**Dataset: Re10K**

| | Return (vs. GT) | | | | | Revisit self-consistency | | | | | Runtime |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | LPIPS ↓ | PSNR ↑ | SSIM ↑ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | ART (min) ↓ |
| SEVA [76] | **0.565** | **12.27** | 0.352 | 0.255 | 0.405 | 0.354 | 17.99 | 0.553 | 0.212 | 0.484 | 7.75 |
| ViewCrafter [73] | 0.668 | 9.91 | 0.175 | 0.324 | 0.488 | 0.540 | 11.69 | 0.276 | 0.254 | 0.616 | 5.65 |
| Gen3C [50] | 0.611 | 11.61 | 0.290 | 0.378 | 0.411 | 0.332 | 19.13 | 0.620 | **0.115** | **0.444** | 47.62 |
| VMem [36] | 0.599 | 12.12 | 0.360 | **0.155** | 0.411 | 0.242 | **24.02** | 0.674 | 0.185 | 0.529 | 47.62 |
| Matrix-Game 2.0 [24] | 0.730 | 6.72 | 0.092 | 0.594 | 0.424 | 0.575 | 12.41 | 0.298 | 0.570 | 0.668 | **1.33** |
| NeuWorld | **0.565** | 12.21 | **0.373** | 0.165 | **0.382** | **0.208** | 19.30 | **0.692** | 0.128 | 0.466 | 3.24 |

**Dataset: DL3DV**

| | Return (vs. GT) | | | | | Revisit self-consistency | | | | | Runtime |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | LPIPS ↓ | PSNR ↑ | SSIM ↑ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | ART (min) ↓ |
| SEVA [76] | 0.583 | 12.82 | 0.286 | 0.848 | 0.534 | 0.480 | 14.28 | 0.349 | 0.732 | 0.568 | 11.71 |
| ViewCrafter [73] | 0.655 | 11.66 | 0.234 | 0.841 | 0.580 | 0.514 | 14.76 | 0.365 | 0.464 | 0.586 | 1.62 |
| Gen3C [50] | **0.479** | **14.41** | **0.362** | 0.938 | 0.601 | **0.267** | 21.69 | **0.701** | **0.094** | 0.427 | 13.61 |
| VMem [36] | 0.645 | 12.65 | 0.312 | 0.582 | 0.555 | 0.304 | **24.43** | 0.648 | 0.613 | 0.431 | 14.17 |
| Matrix-Game 2.0 [24] | 0.730 | 7.58 | 0.096 | 0.910 | 0.562 | 0.609 | 11.84 | 0.255 | 0.894 | 0.629 | **0.73** |
| NeuWorld | 0.638 | 12.12 | 0.313 | **0.410** | **0.507** | 0.335 | 21.71 | 0.620 | 0.241 | **0.315** | 1.14 |

**Table 2: Cycle-trajectory revisitation on Re10K and DL3DV.** Models generate frames from the start pose to the end pose and then return along the same path. Each subtable reports return-path quality, revisit self-consistency, and ART (*average runtime per forward-and-return trajectory*, minutes). We mark the best, second-best, and third-best results in each column.

cle protocol under the same evaluation runner. The only faster baseline in our benchmark is Matrix-Game 2.0, which uses a distilled few-step diffusion model. On DL3DV, NeuWorld achieves the best return-path pose errors ($R_{\text{dist}}$=0.410, $T_{\text{dist}}$=0.507) and the best revisit translation consistency ($T_{\text{dist}}$=0.315), while being the second fastest overall (1.14 minutes per forward-and-return trajectory), yielding a favorable accuracy–latency trade-off for interactive exploration.

### 4.3 Ablation Studies

We present five targeted ablations to isolate the roles of (i) decodable geometry in frozen NIS, (ii) the latent representation, (iii) the NIS capacity, (iv) the unified conditioning interface together with anti-drift augmentation, and (v) the memory retrieval strategy.

**Ablation 1: Geometry probing in frozen NIS.** We first test whether NIS tokens encode geometry beyond appearance reconstruction by freezing the NIS-VAE encoder, adding a decoder depth head, and fine-tuning only this head with Depth-Anything-3 [38] distillation. We backproject the rendered depths into point clouds to visualize geometry decodable from full and pose+reference partial NIS. As shown in Fig. 4, full NIS decodes into a meaningful 3D layout, and pose+reference partial NIS still preserves a coherent geometric scaffold when most appearance observations are removed. This supports partial NIS as a non-collapsing geometric condition for NIS-DiT.

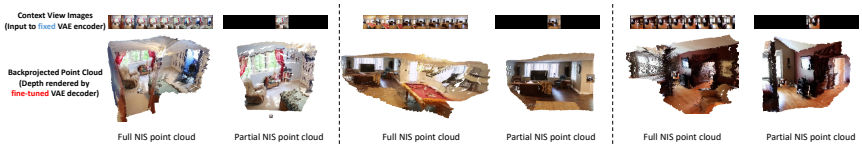


**Fig. 4: Ablation on geometry decodable from frozen NIS.** We freeze the NIS-VAE encoder, train only a decoder depth head by distilling Depth-Anything-3 [38], and backproject rendered depths into point clouds. Both full NIS and pose+reference partial NIS retain a coherent geometric scaffold, supporting partial NIS as a geometry-preserving condition. Please zoom-in for details.

| **Method** | **FVD** ↓ | $R_{\text{dist}}$ (deg) ↓ | $T_{\text{dist}}$ ↓ | **Prior Time (h)** ↓ |
|---|---|---|---|---|
| Video baseline (latent frames) | 88.03 | 4.20 | **0.141** | 78.0 |
| NeuWorld (NIS latent) | **86.20** | **3.26** | 0.157 | **17.2** |

**Table 3: Stage-1 prior-only ablation on latent representation.** This controlled proxy compares latent-prior learning efficiency rather than full system-level performance. We report FVD, VGGT-based pose proxies ($R_{\text{dist}}$/$T_{\text{dist}}$) [58], and 50k-step prior training time.

**Ablation 2: Latent representation (NIS vs. latent video frames).** We ablate the latent representation in Stage 1 camera-controlled generation by comparing our NIS latents against conventional latent video frames while keeping the prior architecture and explicit camera-trajectory conditioning fixed. We use the pretrained video VAE from [66], train the latent-frame DiT from scratch with PRoPE camera conditioning [35], and train both priors for 50k steps on 8×A100 with total batch size 128. We evaluate *pure Stage-1 prior sampling* on Re10K, using 8 context views for NeuWorld and 29 video-baseline context views compressed by the video VAE into 8 latent frames. We report FVD and VGGT-based camera controllability proxies [58], $R_{\text{dist}}$ (degrees) and $T_{\text{dist}}$. The reported training time measures only the Stage 1 latent-prior training to 50k steps under this controlled setup, excluding representation autoencoder pretraining. Table 3 summarizes the comparison. Under aligned conditions, NIS improves video quality (FVD 86.20 vs. 88.03) and rotation trajectory error ($R_{\text{dist}}$ 3.26° vs. 4.20°), with a small regression in translation error ($T_{\text{dist}}$ 0.157 vs. 0.141). The latent-frame prior is also substantially slower to train: reaching 50k steps takes ∼78.0 hours versus 17.2 hours for NIS, suggesting that the set-based NIS prior is more compute-efficient for camera-controlled generation.

**Ablation 3: NIS capacity (tokens $L$ and channels $D$).** We study how NIS capacity affects view synthesis by varying token length $L$ with $D$=64 and channel width $D$ with $L$=1024, while keeping the rest of the recipe fixed. We report PSNR/SSIM/LPIPS on Re10K. Results are reported in Table 4. As shown in Table 4, NVS quality improves steadily as we increase the number of NIS

| $L$ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| 512 | 25.896 | 0.832 | 0.179 |
| 1024 | 26.590 | 0.844 | 0.165 |
| 2048 | 27.801 | 0.858 | 0.154 |
| 3072 | 28.352 | 0.862 | 0.148 |

**(a)** Token length $L$ ($D$=64).

| $D$ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| 32 | 26.245 | 0.839 | 0.170 |
| 64 | 26.590 | 0.844 | 0.165 |
| 128 | 26.721 | 0.845 | 0.163 |
| 256 | 26.822 | 0.846 | 0.162 |

**(b)** Channel width $D$ ($L$=1024).

**Table 4: Ablation on NIS capacity.** We vary token length $L$ and channel width $D$.

| | Short-term | | | Long-term | | |
|---|---|---|---|---|---|---|
| **Settings** | **LPIPS** ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | **LPIPS** ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ |
| Alternative ref.+traj. cond. | 0.433 | 0.095 | 0.150 | **0.663** | 0.144 | 0.203 |
| w/o anti-drift aug. | 0.455 | 0.116 | 0.205 | 0.720 | 0.495 | 0.680 |
| Full | **0.422** | **0.030** | **0.114** | 0.671 | **0.109** | **0.153** |

**Table 5: Ablation on unified conditioning and anti-drift augmentation** (forward trajectory, Re10K). The alternative conditioning injects DINOv2 reference-image tokens and lightweight camera-encoder tokens through cross-attention instead of partial NIS. Unified partial-NIS conditioning mainly improves pose consistency, while removing anti-drift augmentation substantially hurts long-horizon robustness.

tokens $L$. In contrast, increasing the channel width $D$ yields only marginal gains (PSNR 26.25→26.82 from $D$=32 to 256 at fixed $L$=1024). We therefore use $L$=1024 and $D$=64 as a compute–quality trade-off that preserves interactive efficiency and stable from-scratch diffusion-prior training.

**Ablation 4: Unified conditioning and anti-drift augmentation.** This ablation isolates the *conditioning interface* and *rollout robustness* while keeping the NIS representation fixed. We evaluate two Re10K forward-trajectory variants (Table 5): a heterogeneous conditioning design and a model without anti-drift condition augmentation (Section 3.4). Replacing unified conditioning with cross-attention degrades pose consistency, e.g., short-term $R_{\text{dist}}$ 0.030→0.095 and long-term $R_{\text{dist}}/T_{\text{dist}}$ 0.109/0.153→0.144/0.203, showing that mapping camera and reference cues into NIS provides a stronger geometric scaffold. Removing anti-drift augmentation has little short-horizon effect but sharply degrades long-horizon rollout ($T_{\text{dist}}$ 0.153→0.680), indicating its importance for imperfect generated history.

**Ablation 5: Memory retrieval strategy.** Finally, we ablate the hybrid geometry-aware retrieval (Section 3.4) used to form the revisitation history set, keeping memory size and rollout settings fixed. We compare *recent-only*, *camera-distance-only*, *FoV-overlap-only*, and the *full* hybrid score on Re10K cy-

| | Return vs. GT | | | Revisit self-consistency | | |
|---|---|---|---|---|---|---|
| **Settings** | **LPIPS** ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ | **LPIPS** ↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ |
| Recent only | 0.755 | 0.940 | 0.619 | 0.709 | 0.886 | 0.630 |
| Cam. dist. only | 0.572 | 0.148 | 0.382 | 0.327 | 0.260 | 0.412 |
| FoV only | 0.562 | 0.151 | 0.379 | 0.335 | 0.251 | 0.370 |
| Full | **0.553** | **0.145** | **0.379** | **0.302** | **0.212** | **0.361** |

**Table 6: Ablation on memory retrieval strategy** (cycle revisitation, Re10K), which mainly influences the return path. Performance on the return path drops significantly when geometry-aware retrieval is replaced by recent-only history. The hybrid geometry-aware retrieval, considering both FoV overlap and camera distance, achieves the best performance under the controlled ablation setting.

cle trajectories. Table 6 reports return-path quality and revisit self-consistency. Recent-only retrieval fails badly on the return path ($R_{\text{dist}}$ 0.940, return LPIPS 0.755), since temporal neighbors carry little information about a revisited region. Either geometric signal alone (camera distance or FoV overlap) recovers most of the performance, and the full hybrid score is best across all reported metrics, confirming that pose and FoV cues are complementary for loop-closure recall.

## 5 Conclusion

We presented NeuWorld, an instantiation of the *Walking in the Implicit* formulation for camera-controlled interactive exploration. NeuWorld represents each rollout state as a fixed-length, locally anchored Neural Implicit Scene (NIS) and renders queried observations from this state with a frozen decoder, thereby separating latent scene-state sampling from high-frequency observation synthesis. This scene-centric design also provides a unified NIS modality for camera, reference-image, and history conditions. Experiments on Re10K and DL3DV show strong long-horizon pose and revisit consistency, a favorable accuracy–latency trade-off under cycle revisitation, and robustness without relying on pretrained video foundation models or auxiliary 3D reconstructors. Ablations further confirm the complementary roles of NIS rollout, partial-NIS conditioning, anti-drift augmentation, and geometry-aware retrieval. Together, these results suggest that compact renderable scene states offer an effective alternative to frame-latent rollout for camera-controlled exploration.

*Future directions.* Our evaluation is deliberately scoped to static scenes under camera control, isolating the representation question from the complexities of object dynamics. The current NIS state is local and bounded: it is anchored at the current reference frame and re-anchored as the agent moves, covering a trajectory segment rather than maintaining a persistent global map. Extending this local-state formulation to dynamic environments, richer action spaces, and larger-scale scene-state composition is a promising direction for future work.

## References


1. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024)
2. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. pp. 22875–22889 (2025)
3. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: Vd3d: Taming large video diffusion transformers for 3d camera control (2025)
4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. pp. 22669–22679 (2023)
5. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. pp. 15791–15801 (2025)
6. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. pp. 5855–5864 (2021)
7. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. pp. 19697–19705 (2023)
8. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
9. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: SIGGRAPH Asia. pp. 1–12 (2025)
10. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. pp. 19457–19467 (2024)
11. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. pp. 14124–14133 (2021)
12. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)
13. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. pp. 7310–7320 (2024)
14. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. pp. 370–386. Springer (2024)
15. Decart, E.: Oasis: A universe in a transformer. https://oasis-model.github.io/ (2024)
16. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International conference on machine learning. pp. 7480–7512. PMLR (2023)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. pp. 12873–12883 (2021)

18. Feng, W., Li, J., Cai, H., Luo, X., Zhang, J.: Neural points: Point cloud representation with neural fields for arbitrary upsampling. pp. 18633–18642 (2022)
19. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. pp. 5501–5510 (2022)
20. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
21. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (2024)
22. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems 31, pp. 2451–2463. Curran Associates, Inc. (2018), https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution, https://worldmodels.github.io
23. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation (2025)
24. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
25. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. pp. 5875–5884 (2021)
26. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
28. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. pp. 1125–1134 (2017)
30. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., et al.: Rayzer: A self-supervised large view synthesis model (2025)
31. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias (2025)
32. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. pp. 694–711. Springer (2016)
33. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
34. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. Advances in Neural Information Processing Systems **37**, 16240–16271 (2024)
35. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding. arXiv preprint arXiv:2507.10496 (2025)
36. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory (2025)
37. Li, Z., Dong, C., Chen, Y., Huang, Z., Liu, P.: Vicasplat: A single run is all you need for 3d gaussian splatting and camera estimation from unposed video frames. arXiv preprint arXiv:2503.10286 (2025)

38. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
39. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. pp. 22160–22169 (2024)
40. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: Reconx: Reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024)
41. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. pp. 7210–7219 (2021)
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
43. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
44. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. pp. 5480–5490 (2022)
45. Parker-Holder, J., Ball, P., Bruce, J., Dasagi, V., Holsheimer, K., Kaplanis, C., Moufarek, A., Scully, G., Shar, J., Shi, J., Spencer, S., Yung, J., Dennis, M., Kenjeyev, S., Long, S., Mnih, V., Chan, H., Gazeau, M., Li, B., Pardo, F., Wang, L., Zhang, L., Besse, F., Harley, T., Mitenkova, A., Wang, J., Clune, J., Hassabis, D., Hadsell, R., Bolton, A., Singh, S., Rocktäschel, T.: Genie 2: A large-scale foundation world model (2024), https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. pp. 4195–4205 (2023)
47. Plucker, J.: Xvii. on a new geometry of space. Philosophical Transactions of the Royal Society of London (155), 725–791 (1865)
48. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. pp. 14335–14345 (2021)
49. Reiser, C., Szeliski, R., Verbin, D., Srinivasan, P., Mildenhall, B., Geiger, A., Barron, J., Hedman, P.: Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. ACM Transactions on Graphics (ToG) **42**(4), 1–12 (2023)
50. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. pp. 6121–6132 (2025)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. pp. 10684–10695 (2022)
52. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
53. Sajjadi, M.S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., et al.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. pp. 6229–6238 (2022)

54. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. pp. 5459–5469 (2022)
55. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines (2025)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
57. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
58. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. pp. 5294–5306 (2025)
59. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. pp. 4690–4699 (2021)
60. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
61. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters (2021)
62. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: Worldmem: Long-term consistent world simulation with memory. Advances in Neural Information Processing Systems (2025)
63. Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U.: Point-nerf: Point-based neural radiance fields. pp. 5438–5448 (2022)
64. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114 **1**(2), 6 (2023)
65. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
66. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer (2024)
67. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images (2025)
68. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. pp. 6613–6623 (2024)
69. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. pp. 4578–4587 (2021)
70. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: SIGGRAPH Asia (2025)
71. Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., Liu, X.: Gamefactory: Creating new games with generative interactive videos (2025)
72. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. pp. 100–111 (2025)
73. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

74. Zhang, B., Sennrich, R.: Root mean square layer normalization. Advances in neural information processing systems **32** (2019)
75. Zhang, Q., Baek, S.H., Rusinkiewicz, S., Heide, F.: Differentiable point-based radiance fields for efficient view synthesis. In: SIGGRAPH Asia. pp. 1–12 (2022)
76. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025)
77. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics (TOG) **37**(4), 1–12 (2018)