

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 Qiu³
Hailong Qin³ Mu Yang³ Dongxu Wei² Peidong Liu^{2,*}

¹ Zhejiang University ² Westlake University ³ Afari Intelligent Drive

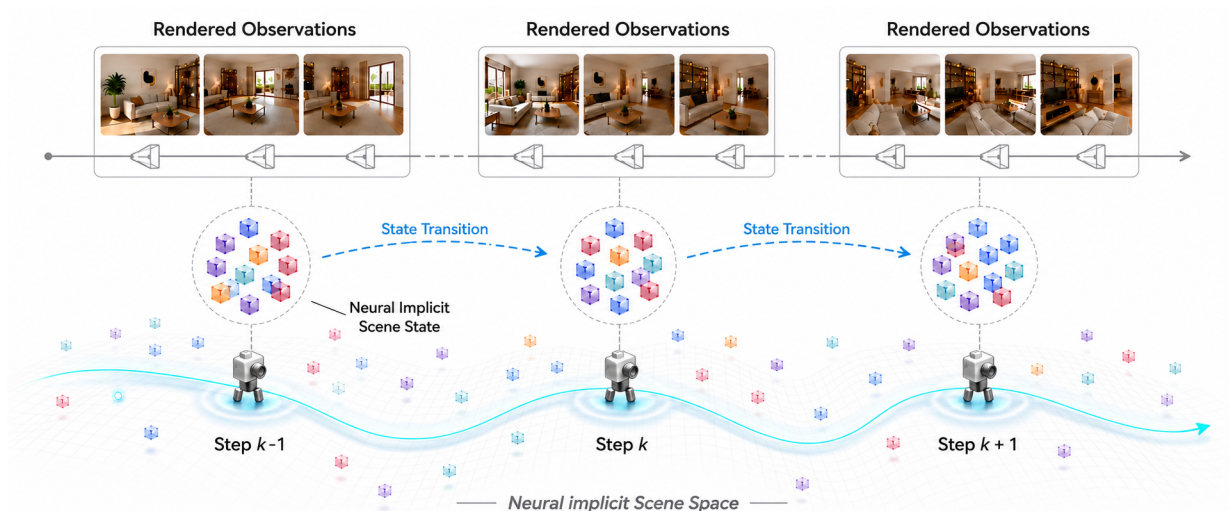


Figure 1 Walking in the Implicit. NeuWorld rolls out a fixed-length Neural Implicit Scene state and renders queried observations from it under camera control.

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 Page: <https://lizhiqi49.github.io/NeuWorld>

Date: June 30, 2026

[†] 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 [1–4]. 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 [5] and 3D Gaussian Splatting [6] provide strong geometric inductive biases, and reconstruction-based interactive systems repeatedly rebuild a 3D/4D representation to support revisitation [7–10]. 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 [11] and RayZer [12] 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.

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 [13] and DL3DV [14], 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 [5] and its variants advance neural volumetric rendering in quality [15–17], speed [18–20], and generalization to sparse inputs [21–23]. Explicit structures like voxels [24, 25], hash grids [26], point-based representations [27–29], and 3D Gaussian Splatting [6] trade off compactness for rendering speed. Data-driven methods further learn generalizable NVS from posed views [30–36]. Closest to our representation, SRT [37], LVSM [11], and RayZer [12] 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 [38–40]. Existing methods inject camera information into video generators through motion adapters [41, 42], 6DoF camera encoders [43, 44], cross-video synchronization [45], trajectory-aware generation [46, 47], or diffusion-transformer camera knowledge [48]. 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 [49–51]. Recent interactive video generation systems extend camera or action signals into repeated rollouts for world exploration [1–4, 52, 53]. 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 [54–56], using reconstruction or geometry modules [57–61] to organize historical evidence [7–10, 62, 63], or retrieving relevant past observations based on geometry or field-of-view overlap [64, 65]. 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.

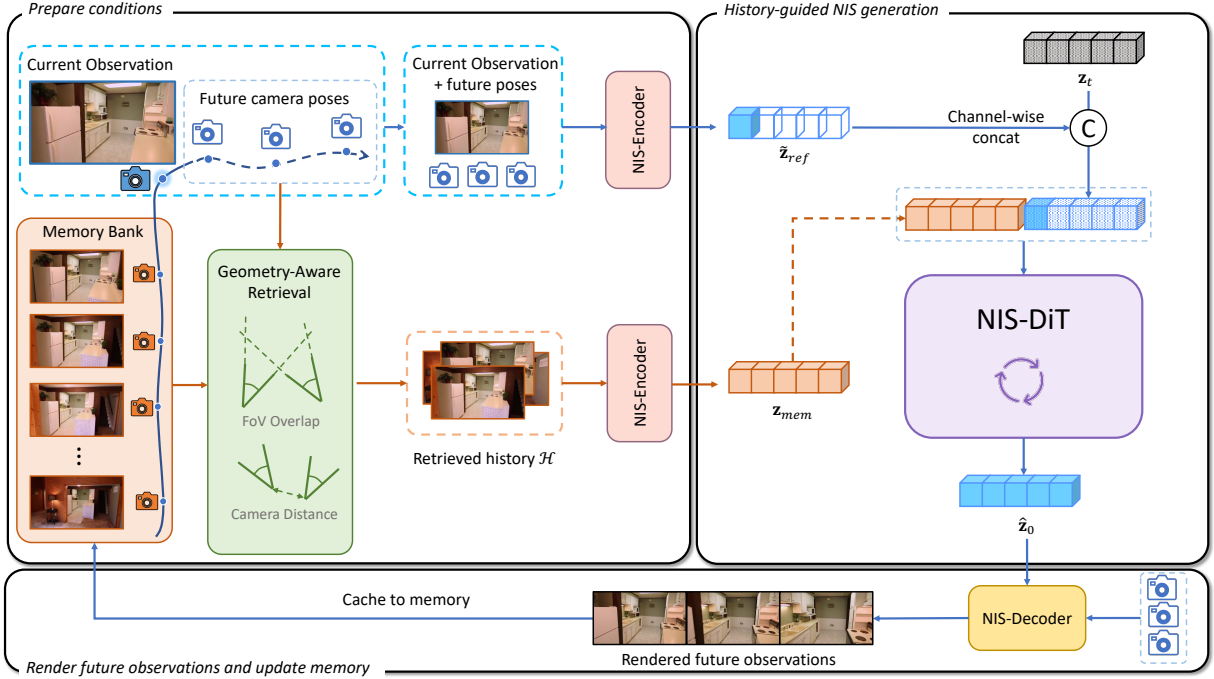


Figure 2 Method overview. At an interaction step, the frozen NIS-VAE encoder maps the current observation and a sparse future pose trajectory to a partial NIS condition $\tilde{\mathbf{z}}_{ref}$, which is injected by channel-wise concatenation with the noised latent \mathbf{z}_t . Geometry-aware retrieval selects a history set \mathcal{H} and encodes it as memory NIS tokens \mathbf{z}_{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.

3.1 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 $\mathcal{V}_{ctx} = \{(I_{ctx}^{(i)}, \mathbf{T}_{ctx}^{(i)})\}_{i=1}^M$ and target views $\mathcal{V}_{tgt} = \{(I_{tgt}^{(j)}, \mathbf{T}_{tgt}^{(j)})\}_{j=1}^N$. 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 $\mathcal{T}_{fut}^{(k)} = \{\mathbf{T}_j^{(k)}\}_{j=1}^S$. 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 $\mathcal{T}_{fut}^{(k)}$ 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 $\hat{\mathbf{z}}^{(k)} = \mathcal{G}_\theta(O_k, \mathcal{T}_{fut}^{(k)}, \mathcal{H}_k)$, where \mathcal{G}_θ is the diffusion sampling procedure parameterized by NIS-DiT. The frozen NIS decoder then renders future observations under queried poses: $\hat{I}_j^{(k)} = \mathcal{D}_\phi(\hat{\mathbf{z}}^{(k)}, \mathbf{T}_j^{(k)})$, $\mathbf{T}_j^{(k)} \in \mathcal{T}_{fut}^{(k)}$. 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.

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 [66] 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:

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

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 [67–70], we train NIS-VAE with reconstruction, perceptual, adversarial, and KL regularization losses:

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

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:

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

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

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

Retrieved history frames are encoded in the same way as memory NIS tokens, $\mathbf{z}_{mem} = \mathcal{E}(\mathcal{H}_k)$. 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) [71] 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 [72]. 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 [73, 74] with long skip connections, AdaLN-style timestep modulation [71], RMSNorm [75], and Q, K normalization [76].

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 $\tilde{\mathbf{z}}_{pose}$, $\tilde{\mathbf{z}}_{ref}$, and the denoising target share the same token shape, we inject partial NIS by channel-wise concatenation:

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

The concatenated tokens are projected to the DiT hidden width by a linear input projection. Memory NIS tokens \mathbf{z}_{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 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, construct $\mathbf{z}_t = t\mathbf{z}_0 + (1-t)\epsilon$, and train NIS-DiT to predict the velocity

$\mathbf{v}^* = \mathbf{z}_0 - \epsilon$. The DiT predicts $\hat{\mathbf{v}} = f_\theta(\mathbf{z}_t, t, \mathbf{c})$, and we optimize

$$\mathcal{L}_{DiT} = \mathbb{E} \left[\|\hat{\mathbf{v}} - \mathbf{v}^*\|_2^2 \right]. \quad (3)$$

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) [77].

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}_{cond} denotes the conditioning latents before noise injection, including partial NIS conditions and, when available, memory latents. γ is the augmentation strength sampled from a scaled Beta distribution. During training, we condition DiT on γ 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 γ with interaction step k to account for increasingly noisy generated history: $\gamma_k = \gamma_{\min} + \min\left(\frac{k}{K_{\text{ramp}}}, 1\right) (\gamma_{\max} - \gamma_{\min})$, where γ_{\min} and γ_{\max} are the lower and upper augmentation bounds, and K_{ramp} is the ramp length.

Hybrid geometry-aware memory retrieval. To provide relevant history for long-horizon roll-out, 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.

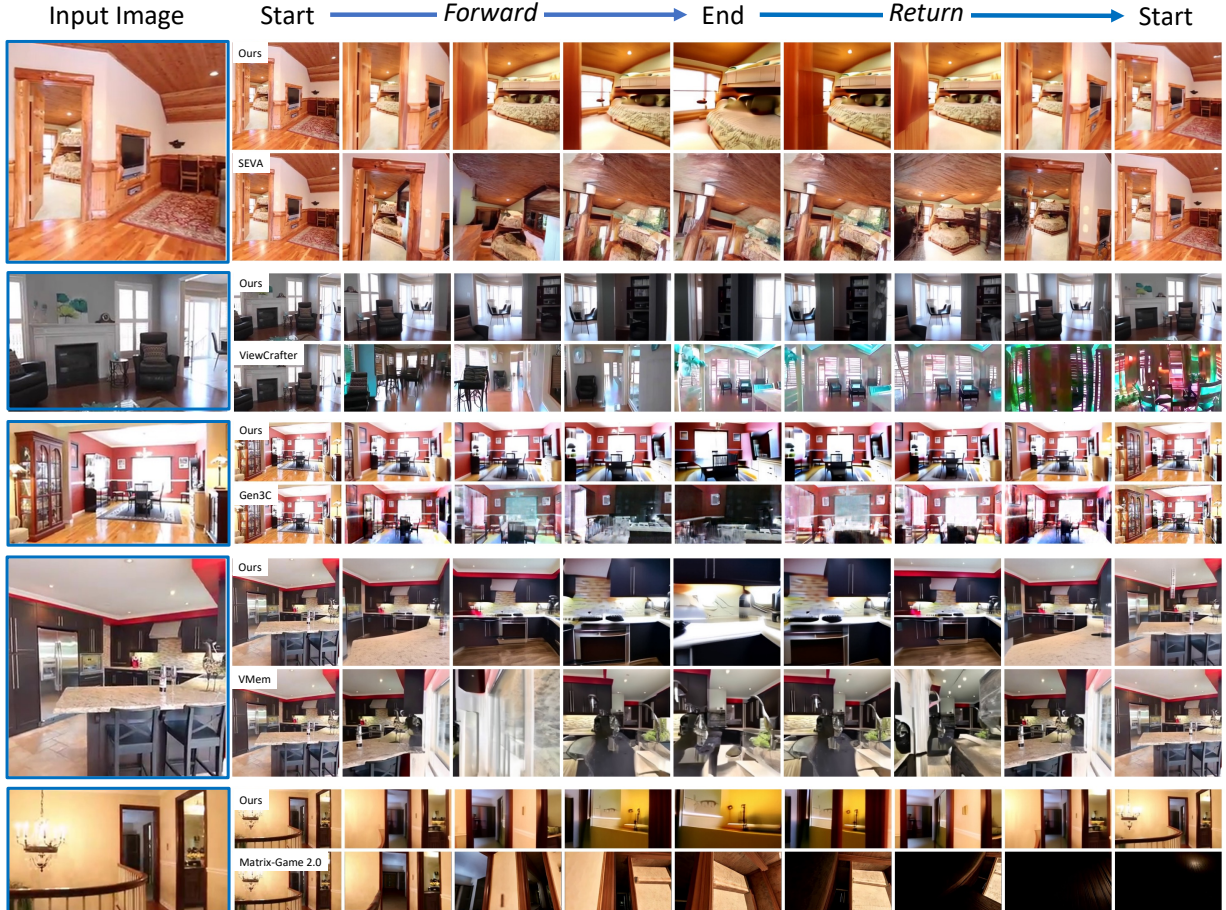


Figure 3 Qualitative comparison. We compare NeuWorld with representative baselines under the same camera-controlled trajectories. Compared with video-latent and reconstruction-based baselines, NeuWorld better preserves scene geometry and appearance consistency under both short-horizon and long-horizon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

4.1 Experimental Setup

4.1.1 Implementation details.

We train our main models (NIS-VAE and NIS-DiT) from scratch on posed multi-view data from Re10K [13] and DL3DV-10K [14], and evaluate interactive exploration on both test splits. All images are center-cropped and resized to 256×256 . 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.

4.1.2 Evaluation.

We evaluate NeuWorld under two camera-controlled protocols and report PSNR, SSIM, LPIPS, FID, and pose errors (R_{dist} and T_{dist}) computed from generated-frame pose estimates. R_{dist} denotes the rotation distance in degrees and T_{dist} denotes the normalized translation distance. For

Dataset: Re10K

	Short-term (50 th frame)						Long-term (200 th frame)					
	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓
SEVA [78]	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 [79]	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 [9]	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 [10]	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 [53]	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_{dist} ↓	T_{dist} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓
SEVA [78]	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 [79]	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 [9]	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 [10]	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 [53]	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 50th/200th frames on Re10K and the 20th/80th frames on DL3DV. We mark the **best**, **second-best**, and **third-best** results in each column.

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 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 50th/200th frames on Re10K and the 20th/80th 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 [10], SEVA [78], Gen3C [9], ViewCrafter [79], and Matrix-Game 2.0 [53].

4.2.1 Forward trajectory generation.

Table 1 shows that NeuWorld achieves strong camera-controlled novel view generation with low pose drift over long rollouts. On Re10K, NeuWorld is best across all metrics at the 50th frame and achieves the lowest pose errors at the 200th 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

Dataset: Re10K

	Return (vs. GT)					Revisit self-consistency					Runtime
	LPIPS ↓	PSNR ↑	SSIM ↑	R_{dist} ↓	T_{dist} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	R_{dist} ↓	T_{dist} ↓	ART (min) ↓
SEVA [78]	0.565	12.27	0.352	0.255	0.405	0.354	17.99	0.553	0.212	0.484	7.75
ViewCrafter [79]	0.668	9.91	0.175	0.324	0.488	0.540	11.69	0.276	0.254	0.616	5.65
Gen3C [9]	0.611	11.61	0.290	0.378	0.411	0.332	19.13	0.620	0.115	0.444	47.62
VMem [10]	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 [53]	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_{dist} ↓	T_{dist} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	R_{dist} ↓	T_{dist} ↓	ART (min) ↓
SEVA [78]	0.583	12.82	0.286	0.848	0.534	0.480	14.28	0.349	0.732	0.568	11.71
ViewCrafter [79]	0.655	11.66	0.234	0.841	0.580	0.514	14.76	0.365	0.464	0.586	1.62
Gen3C [9]	0.479	14.41	0.362	0.938	0.601	0.267	21.69	0.701	0.094	0.427	13.61
VMem [10]	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 [53]	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.

the best long-horizon translation consistency ($T_{\text{dist}}=0.274$). The visual comparison in Fig. 3 is consistent with these quantitative trends.

4.2.2 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 $14\times$ faster than VMem and Gen3C (47.62 minutes each) on the Re10K cycle 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.

4.3.1 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 [59] 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-

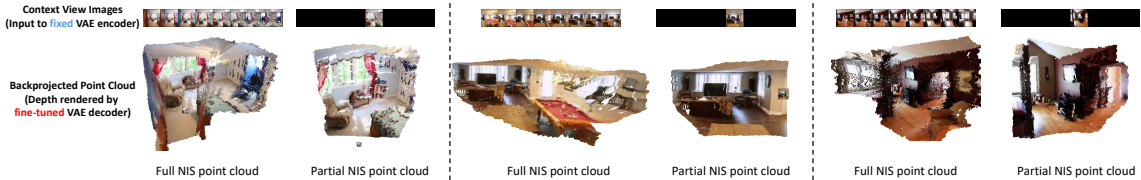


Figure 4 Ablation on geometry decodable from frozen NIS. We freeze the NIS-VAE encoder, train only a decoder depth head by distilling Depth-Anything-3 [59], and backproject rendered depths into point clouds. Both full NIS and pose+reference partial NIS retain a coherent geometric scaffold, supporting partial NIS as a geometry-preserving condition. Please zoom-in for details.

Method	FVD ↓	R_{dist} (deg) ↓	T_{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.

collapsing geometric condition for NIS-DiT.

4.3.2 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 [40], train the latent-frame DiT from scratch with PRoPE camera conditioning [80], and train both priors for 50k steps on $8 \times A100$ 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_{dist} (degrees) and T_{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_{dist} 3.26° vs. 4.20°), with a small regression in translation error (T_{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.

4.3.3 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/S-SIM/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 tokens L . In contrast, increasing the channel width D yields only marginal gains (PSNR $26.25 \rightarrow 26.82$ 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.

4.3.4 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

L	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	D	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
512	25.896	0.832	0.179	32	26.245	0.839	0.170
1024	26.590	0.844	0.165	64	26.590	0.844	0.165
2048	27.801	0.858	0.154	128	26.721	0.845	0.163
3072	28.352	0.862	0.148	256	26.822	0.846	0.162

(a) Token length L ($D=64$). (b) Channel width D ($L=1024$).

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

Settings	Short-term			Long-term		
	LPIPS \downarrow	R_{dist} \downarrow	T_{dist} \downarrow	LPIPS \downarrow	R_{dist} \downarrow	T_{dist} \downarrow
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.

Settings	Return vs. GT			Revisit self-consistency		
	LPIPS \downarrow	R_{dist} \downarrow	T_{dist} \downarrow	LPIPS \downarrow	R_{dist} \downarrow	T_{dist} \downarrow
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.

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_{dist} 0.030 \rightarrow 0.095 and long-term $R_{\text{dist}}/T_{\text{dist}}$ 0.109/0.153 \rightarrow 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_{dist} 0.153 \rightarrow 0.680), indicating its importance for imperfect generated history.

4.3.5 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 cycle trajectories. Table 6 reports return-path quality and revisit self-consistency. Recent-only retrieval fails badly on the return path (R_{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] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.
- [2] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *ICLR*, 2025.
- [3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, pages 15791–15801, 2025.
- [4] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *ICCV*, 2025.
- [5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [7] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024.
- [8] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.
- [9] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025.

- [10] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *ICCV*, 2025.
- [11] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2025.
- [12] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *ICCV*, 2025.
- [13] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.
- [14] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024.
- [15] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021.
- [16] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [17] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, pages 19697–19705, 2023.
- [18] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, pages 14335–14345, 2021.
- [19] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *ICCV*, pages 5875–5884, 2021.
- [20] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (ToG)*, 42(4):1–12, 2023.
- [21] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, pages 5480–5490, 2022.
- [22] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021.
- [23] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf-: Neural radiance fields without known camera parameters. 2021.
- [24] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5459–5469, 2022.
- [25] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022.
- [26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.
- [27] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, pages 5438–5448, 2022.

- [28] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia*, pages 1–12, 2022.
- [29] Wanquan Feng, Jin Li, Hongrui Cai, Xiaonan Luo, and Juyong Zhang. Neural points: Point cloud representation with neural fields for arbitrary upsampling. In *CVPR*, pages 18633–18642, 2022.
- [30] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021.
- [31] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, pages 4690–4699, 2021.
- [32] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021.
- [33] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pages 19457–19467, 2024.
- [34] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, pages 370–386. Springer, 2024.
- [35] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025.
- [36] Zhiqi Li, Chengrui Dong, Yiming Chen, Zhangchi Huang, and Peidong Liu. 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.
- [37] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, pages 6229–6238, 2022.
- [38] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [39] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024.
- [40] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2024.
- [41] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.
- [42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [43] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

- [44] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *ICLR*, 2025.
- [45] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024.
- [46] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *ICLR*, 2025.
- [47] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.
- [48] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, pages 22875–22889, 2025.
- [49] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. URL <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/>.
- [50] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pages 2451–2463. Curran Associates, Inc., 2018. URL <https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution>. <https://worldmodels.github.io>.
- [51] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.
- [52] Etched Decart. Oasis: A universe in a transformer. <https://oasis-model.github.io/>, 2024.
- [53] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.
- [54] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.
- [55] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.
- [56] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pages 6613–6623, 2024.
- [57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024.
- [58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025.
- [59] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

- [60] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025.
- [61] Jingnan Gao, Zhe Wang, Xianze Fang, Xingyu Ren, Zhuo Chen, Shengqi Liu, Yuhao Cheng, Jiangjing Lyu, Xiaokang Yang, and Yichao Yan. More: 3d visual geometry reconstruction meets mixture-of-experts. *arXiv preprint arXiv:2510.27234*, 2025.
- [62] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *SIGGRAPH Asia*, pages 1–12, 2025.
- [63] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, pages 100–111, 2025.
- [64] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *SIGGRAPH Asia*, 2025.
- [65] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 2025.
- [66] Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155):725–791, 1865.
- [67] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.
- [68] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.
- [69] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.
- [70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
- [71] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.
- [72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [74] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, pages 22669–22679, 2023.
- [75] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.
- [76] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023.
- [77] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

- [78] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishtha, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.
- [79] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [80] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025.

A Appendix

A.1 Additional Analysis of Partial NIS and NIS Space

A.1.1 Visual Evidence from Masked Reconstruction

Motivation. We provide visual evidence for the key empirical property used by our unified conditioning interface: the frozen NIS-VAE encoder can still produce a useful geometric scaffold when most image content is removed. This property allows camera trajectory, reference appearance, and retrieved history to be mapped into the same NIS latent modality before being injected into NIS-DiT.

Protocol. To visualize this property, we compare two NIS variants encoded from the same posed-view input:

- **Full NIS:** the default latent with all context-view information preserved, which serves as an upper bound on the encoder’s representational capacity.
- **Pose+reference partial NIS:** a masked latent where the camera poses are preserved, only one reference image is retained, and all other view pixels are dropped.

We reconstruct the context images by rendering both NIS variants under the context poses and compare the results qualitatively.

Observations. As shown in Figure 5, the full NIS reconstructs the input context views with high fidelity, as expected when the encoder has access to complete multi-view evidence. In contrast, the pose+reference partial NIS cannot recover unsupported regions with comparable detail, since most appearance observations have been removed. Importantly, however, its decoded views still preserve a camera-consistent coarse layout around the reference anchor: dominant scene structure, coarse depth ordering, and cross-view geometric alignment remain largely stable, while unsupported regions become uncertain or texture-poor rather than collapsing into geometrically implausible content. This is the behavior needed from a unified conditioning interface. When partial NIS is used as the condition for NIS-DiT, it provides a stable geometric scaffold without over-specifying erroneous details, allowing the denoiser to propagate supported reference-view appearance and infer newly exposed content only where extrapolation is required. Together with the geometry probing ablation in the main paper, this evidence supports the use of partial NIS as a non-collapsing, geometry-preserving condition for NIS-DiT.

A.1.2 Latent Interpolation as NIS-Space Smoothness Evidence

Motivation. Since NeuWorld represents the scene state as a fixed-length NIS token set, it is useful to examine whether nearby points in this latent space decode into visually coherent intermediate scene states. While interpolation alone does not prove that the latent manifold is globally well structured, it provides a qualitative probe of local smoothness.

Protocol. For all interpolation visualizations, we linearly interpolate between the posterior mode of two endpoint latents:

$$\mathbf{z}_\lambda = (1 - \lambda)\mathbf{z}_\alpha + \lambda\mathbf{z}_\beta, \quad \lambda \in [0, 1]. \quad (4)$$

We consider two complementary settings.

- **Same-sequence reference-shift interpolation.** We first sample a single sequence with N context views. We then rebuild the same sampled views under two different canonicalizations: once using the first context view as the reference frame, and once using the last context view



Figure 5 Visual comparison between full-context NIS and pose+reference partial NIS. Full NIS reconstructs the observed inputs almost faithfully, while pose+reference partial NIS retains a coherent coarse layout and view-consistent geometric scaffold around the reference anchor despite losing most appearance detail in unsupported regions. This behavior supports its role as a unified geometric conditioning interface for NIS-DiT.

as the reference frame. Encoding these two re-canonicalized inputs gives endpoint latents \mathbf{z}_α and \mathbf{z}_β . For each interpolation coefficient λ , we decode \mathbf{z}_λ and render only the canonical reference-view image. This setting probes whether the internal coordinate system of NIS changes smoothly when the same underlying scene evidence is re-anchored to different local reference frames.

- **Cross-sequence reference-view interpolation.** We sample two different sequences with N context views each, and canonicalize both using the first context view as the reference frame. Encoding these two inputs gives endpoint latents \mathbf{z}_α and \mathbf{z}_β . For each interpolation coefficient λ , we decode \mathbf{z}_λ and inspect only the rendered reference-view image, yielding a compact strip that isolates how canonical reference-view content changes along the interpolation path.

The former isolates how NIS behaves under a change of local coordinate anchor within the same scene evidence, while the latter probes whether meaningful decoded reference-view contents remain across different sequences.

Observations. In the same-sequence reference-shift setting, the decoded reference-view image changes smoothly as the canonical anchor moves from the first view to the last. The dominant

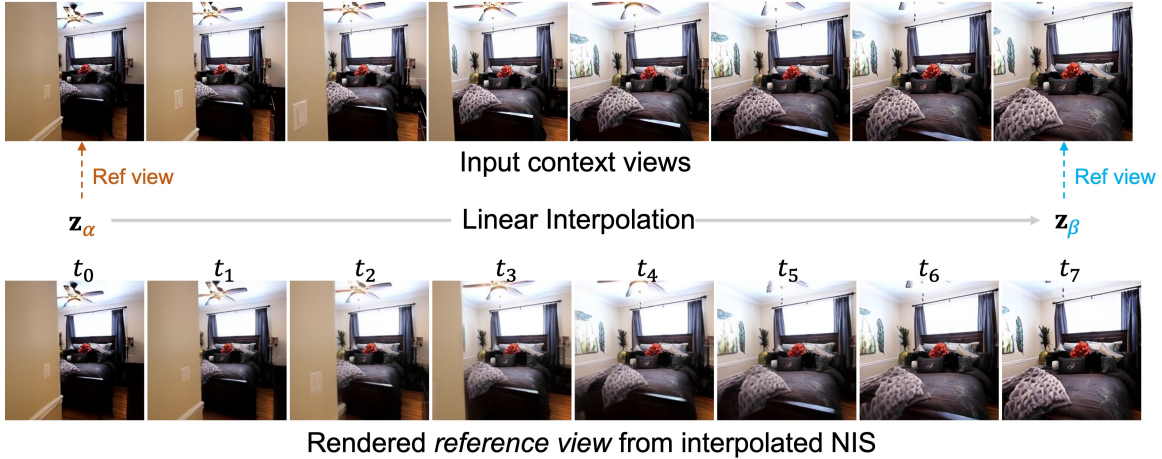


Figure 6 Same-sequence reference-shift interpolation in NIS space. We sample one sequence once, re-canonicalize the same context-view set using the first and last view as reference frames, and interpolate between the resulting latents. At each interpolation step, we render only the canonical reference-view image. The smoothness of the decoded transition is used as a qualitative probe of whether the internal coordinate system of NIS changes continuously under a shift of local reference frame.

scene layout and major structures remain coherent throughout the interpolation strip, rather than exhibiting abrupt jumps or inconsistent geometry. This suggests that the internal coordinate system of NIS adjusts continuously when the same underlying scene evidence is re-anchored to different local reference frames, which is consistent with the local-anchoring design of our representation.

In the cross-sequence setting, interpolation between two different sequences also produces a gradual evolution of rendered reference-view content instead of immediate collapse into meaningless images. We do not interpret these intermediate states as physically faithful trajectories between unrelated world states. Rather, this experiment serves as a diagnostic showing that the decoder responds continuously to latent perturbations even across distinct sequences, indicating a non-trivial degree of smoothness and robustness in the learned NIS space.

Taken together, the two interpolation results support the view that the various conditioning inputs used in NeuWorld can be mapped into a common NIS space with enough structure and continuity to support stable local anchoring and interactive generation.

A.2 Additional Implementation Details

This subsection collects implementation details that are omitted from the main Experimental Setup section for brevity. Unless otherwise stated, the settings below are shared across all main experiments.

A.2.1 Data Processing and Pose Normalization

Image preprocessing. All frames are center-cropped and resized to 256×256 before being fed into the models. Camera intrinsics are adjusted accordingly as the aspect ratio changes.

Extrinsics normalization and local-state definition. For each training sample, we choose one context view as the reference frame and transform all camera poses into this local coordinate system. We further normalize translation magnitude using scene-extent statistics, with the farthest camera distance as the default scale and an additional random rescaling factor during training. Under this normalization, each NIS represents a local scene neighborhood anchored at the chosen reference view.

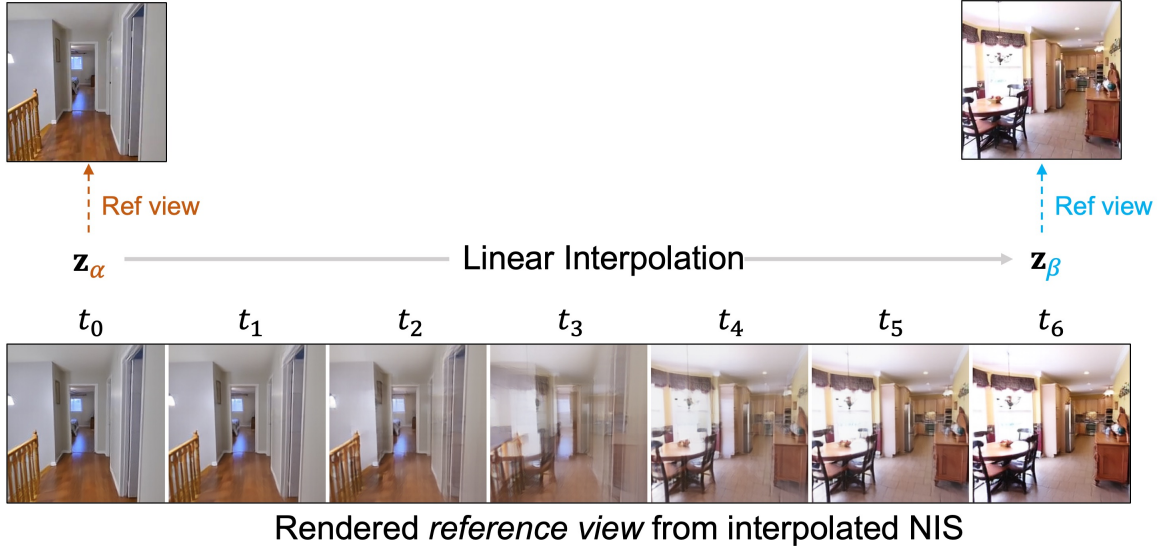


Figure 7 Cross-sequence reference-view interpolation in NIS space. We encode two different sequences, linearly interpolate between their endpoint latents, and show only the rendered reference-view image at each interpolation step. This visualization is intended as a diagnostic of whether the decoder responds continuously to latent perturbations across distinct sequences, rather than as evidence of a physically faithful interpolation between unrelated world states.

A.2.2 View Sampling Details

Re10K sequential sampling. For Re10K, context views are sampled from a contiguous pose-image sequence. We first sample a temporal gap g within predefined bounds, then choose a left endpoint and set the right endpoint to left+ g . Context indices are obtained by uniformly spacing samples between the two endpoints, while target indices are drawn from a window around the context range.

Stage-specific configurations. Unless otherwise stated:

- Stage 1 and Stage 2 use $M=8$ context views, with a temporal gap $g \sim \text{Unif}[80, 128]$ and a random reference view.
- Stage 3 uses $M=15$ context views with a larger gap $g \sim \text{Unif}[180, 300]$ and fixes the reference view to the middle index. The left half of the sampled sequence, including the reference, serves as history, whereas the right half defines the future trajectory segment.

A.2.3 Model and Optimization Details

NIS-VAE. We use a transformer encoder-decoder to learn a fixed-length NIS with $L=1024$ latent tokens and channel width $D=64$ (patch size $p=16$). Training combines pixel-space reconstruction, perceptual supervision, KL regularization, and patch-based adversarial supervision after a warmup phase. We start with deterministic latents for stability and gradually increase latent stochasticity. To improve robustness of encoder-based conditioning, each context image is optionally dropped with probability $p=0.2$.

NIS-DiT. We use a set-based DiT trained with a velocity-prediction objective under flow matching. Stage 1 uses pose-only trajectory conditioning, Stage 2 additionally conditions on the reference image, and Stage 3 further conditions on retrieved history; the three stages are trained with the weak-to-strong curriculum described in Section A.2.4. All reported conditioning branches are

represented through the NIS encoder interface, keeping the denoising input aligned with the NIS rollout state.

Training setup and optimization. The full training pipeline uses 16 A100 GPUs for roughly one week in total. All models are trained with bf16 mixed precision and gradient checkpointing. For DiT training, we apply gradient clipping with norm 1.0. Learning-rate schedules follow a constant-with-warmup policy with a 1k-step warmup. Our default learning rates are $1e-4$ for Stage 1 and $5e-5$ for Stage 2/3.

A.2.4 NIS-DiT Training Curriculum and Regularization

Motivation. The weak-to-strong curriculum is an optimization-stabilization device rather than an additional modeling assumption. It prevents early shortcut copying from strong appearance conditions. In local scenes, a single view often covers a large portion of visible content. If DiT is trained from scratch with strong reference-image conditions, it can align to visible appearance before learning a useful NIS prior. We therefore introduce conditions progressively from weak to strong.

Stage 1: NIS prior learning with pose-only conditioning. We pretrain DiT with weak conditioning using pose-only partial NIS $\tilde{\mathbf{z}}_{\text{pose}}$, which provides a geometric scaffold without any reference pixels. This stage encourages learning the NIS distribution and camera-aligned structure before introducing strong appearance supervision, and prevents shortcut learning by copying visible content.

Stage 2: Alignment to reference appearance. We switch to pose+reference partial NIS to align sampled states with the reference view. To preserve the Stage 1 prior, we apply Stage 2 updates with probability 30%; otherwise, we fall back to pose-only conditioning. Equivalently, the reference-image signal is dropped with probability 70% during this stage.

Stage 3: History-aware interactive generation. We append memory latent \mathbf{z}_{mem} for history-conditioned training. To avoid over-reliance on strong conditions, we keep two stochastic fallback paths: (i) jointly dropping the reference-image and history conditions with probability 50%, which falls back to weak conditioning, and (ii) dropping history only with another 25% probability to preserve cold-start capability when history is limited.

Classifier-free guidance training. We enable classifier-free guidance by dropping all conditioning branches jointly with probability 10%. In addition, each individual condition branch is independently dropped with probability 10%.

A.2.5 History Corruption and Latent Condition Augmentation

History-side corruption. To narrow the train-test gap in long rollouts, Stage 3 corrupts history images during training with four modes:

- Gaussian blur (30%),
- downsample-then-upsample degradation (30%),
- VAE reconstruction replacement (30%), and
- clean history (10%).

Latent condition noise. After encoding, we perturb conditioning latents with additive Gaussian noise. The noise level is sampled as $\gamma \sim \text{Beta}(2, 5)$ and scaled by $\gamma_{\max}=0.5$, while 10% of samples remain clean. For reported long-horizon rollouts, we use the same noise-level conditioning interface at inference and ramp the augmentation magnitude over rollout steps; ablations that remove anti-drift augmentation disable both history-side corruption and latent condition noise.

A.2.6 Inference Settings

Unless otherwise stated, we use 50 denoising steps with CFG scale $s=4.0$ with Euler sampler.

A.2.7 Hybrid Geometry-Aware Memory Retrieval

Default hybrid retrieval. Hybrid retrieval selects history that is both temporally recent and geometrically relevant to the upcoming trajectory. At interaction step k , we maintain a memory bank of posed frames

$$\mathcal{M}_k = \{(I_i, \mathbf{T}_i)\}_{i=0}^{N_k-1}$$

and retrieve $M_{\text{ret}}=8$ frames to form \mathcal{H}_k , matching the number of context views used by the Stage 3 pipeline. Retrieval combines *recent context* for temporal continuity and *global context* for geometric recall: $M_{\text{ret}} = M_r + M_g$. Here M_r and M_g denote the numbers of retrieved recent and global frames, respectively. The split between recent and global history is fixed across datasets and main experiments.

Trajectory-aware scoring. For each candidate history index i , we compute a trajectory-aware relevance score

$$S_i(\mathbf{T}_q) = w_p S_i^{\text{pose}}(\mathbf{T}_q) + w_o S_i^{\text{fov}}(\mathbf{T}_q) + w_r S_i^{\text{rec}},$$

where w_p , w_o , and w_r are scalar mixing weights for pose, FoV-overlap, and recency terms, and S_i^{rec} is a small recency prior to favor temporally closer frames when other signals are ambiguous. Pose similarity is defined as

$$S_i^{\text{pose}}(\mathbf{T}_q) = \exp(-d_{\text{geo}}(\mathbf{T}_i, \mathbf{T}_q)),$$

where d_{geo} is a normalized pose-space distance combining rotational and translational differences. FoV overlap is estimated by visibility overlap:

$$S_i^{\text{fov}}(\mathbf{T}_q) = \frac{|\mathcal{P}_{q \rightarrow i}|}{|\mathcal{P}_q|}.$$

Here \mathbf{T}_q is a sampled query pose, \mathcal{P}_q denotes Monte Carlo 3D samples drawn in the query frustum, and $\mathcal{P}_{q \rightarrow i} \subseteq \mathcal{P}_q$ denotes the subset of samples that also project inside the candidate view frustum of \mathbf{T}_i . This asymmetric estimate measures how much of the query view can be supported by the candidate history frame.

Trajectory aggregation and diversity. Instead of querying only an endpoint pose, we sample a sparse query pose set $\mathcal{T}_{\text{query}} \subset \mathcal{T}_{\text{fut}}^{(k)}$ and aggregate:

$$\bar{S}_i = (1 - \alpha) \text{Avg}_{\mathbf{T}_q \in \mathcal{T}_{\text{query}}} S_i(\mathbf{T}_q) + \alpha S_i(\mathbf{T}_{\text{start}}),$$

where α controls the balance between trajectory-level relevance and starting-pose relevance, and $\mathbf{T}_{\text{start}}$ is the starting pose of the step. The score mixing weights and start-trajectory aggregation weight are kept fixed across datasets and main experiments. Global candidates are selected by top- M_g with a pose-space diversity constraint, then merged with recent indices and encoded into

Method	Base model	Training regime
SEVA	Stable Diffusion 2.1	Large public NVS data, including Re10K and DL3DV
ViewCrafter	DynamiCrafter	Fine-tuned on Re10K and DL3DV
Gen3C	Stable Video Diffusion / Cosmos	Re10K and DL3DV, plus WOD and Kubric4D
VMem	SEVA	Fine-tuned on Re10K
Matrix-Game 2.0	SkyReels-V2-I2V	Hundreds to >1000h action-annotated video data
NeuWorld	–	Trained from scratch on Re10K and DL3DV

Table 7 Training regimes of all compared methods. Most baselines inherit strong large-scale pretrained priors and are trained or fine-tuned on data distributions closely related to the evaluation setting. In contrast, NeuWorld is trained from scratch on Re10K and DL3DV without pretrained video backbones or auxiliary 3D reconstructors.

$\mathbf{z}_{\text{mem}} = \mathcal{E}(\mathcal{H}_k)$ for history-conditioned denoising. This retrieval module only selects relevant evidence for the next local NIS state; it does not bypass NIS transition or directly determine rendered pixels.

Retrieval ablations. For retrieval ablations, we vary the retrieval rule among *temporal*, *camera distance (pose)*, *FoV overlap*, *hybrid* to isolate the contribution of each signal, while keeping the retrieval budget and all other rollout settings fixed.

A.2.8 Evaluation Protocol Knobs and Metrics

Protocols. We evaluate two protocols: (i) forward trajectory novel view generation and (ii) cycle-trajectory revisitation with a return path. By default, trajectories are subsampled with a 10-frame interval. For the forward protocol, evaluation is performed at the 50th/200th frames from the start on Re10K and at the 20th/80th frames on DL3DV.

Evaluation subset and shared settings. All quantitative results in the main paper are averaged over 100 randomly sampled long trajectories from the corresponding test split. For fair comparison, all methods are evaluated on the same trajectories, at their native image resolution, and with the same protocol-specific checkpoints.

Runtime reporting. For cycle revisitation, the reported ART (average runtime per forward-and-return trajectory) is measured with the same evaluation runner and hardware across all methods. The reported runtime corresponds to the full trajectory rollout under the shared protocol, rather than per-frame latency in isolation.

Metrics. We report image metrics (PSNR, SSIM, and LPIPS) whenever ground-truth frames are available. We also report pose errors, including rotation distance (R_{dist}) and translation distance (T_{dist}), computed from poses estimated on generated frames. Poses are expressed relative to the first frame, and translations are normalized by the furthest ground-truth frame. For cycle revisitation, we additionally report revisit self-consistency by comparing each return frame with its paired frame from the forward pass.

A.3 Baseline Training Regimes

Table 7 summarizes the base models and training regimes of all compared methods. The compared baselines generally inherit large-scale pretrained image or video priors, and several are further fine-tuned on datasets that are closely aligned with Re10K and DL3DV. In contrast, NeuWorld is trained from scratch on Re10K and DL3DV without any pretrained video backbone or auxiliary reconstruction module. This heterogeneous regime makes the comparison conservative for assessing the representation-level contribution of NeuWorld.