# MC-WORLD: MEMORY-GUIDED VIDEO DIFFUSION WITH COLLISION-AWARE EXPLORATION FOR SCENE-CONSISTENT 3D WORLD GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion-based 3D scene generation has achieved remarkable progress. However, existing methods still face two critical challenges: (1) They struggle to remember the historical scenes and their spatial relationships in long-range video generation, making it difficult to maintain long-term 3D scene consistency; (2) They are prone to trajectory collisions with objects due to the lack of evasion capabilities for objects, leading to unreasonable geometric structures and inconsistent temporal sequences, thereby further limiting realistic and smooth scene exploration. To address these challenges, we propose **MC-World**, a novel memory-guided and collision-aware video diffusion framework for explorable 3D world generation. Specifically, MC-World consists of two core designs: Spatio-Temporal Memory Retrieval and Collision-Aware Exploration Control. The former retrieves temporally coherent and spatially relevant historical frames from a global memory cache, providing long-range contextual guidance for consistent multi-view generation. Collaboratively, the latter leverages collision detection strategy to adaptively truncate high-risk trajectory and switch generation directions, thus ensuring geometrically reasonable and temporally consistent scene content. Benefiting from the above designs, our method enables consistent 3D long-range scene generation and supports unrestricted realistic world exploration. Extensive experiments show that MC-World significantly superior to existing methods in various scenes.

## 1 INTRODUCTION

3D scene generation has been a crucial task in the field of 3D vision, holding considerable potential in various practical applications, such as virtual reality (Jiang et al., 2024), video gaming (Kim & Lee, 2024) and robotic simulation (Zhang et al., 2024). Although several data-driven studies (Liu et al., 2023; Zhao et al., 2025) have achieved significant progress in generating 3D objects (Shi et al., 2023; Xiang et al., 2025) or simple scenes (Huang et al., 2025b; Tang et al., 2025), they still struggle to scale complex scenes due to the scarcity of high-quality 3D scene data (Xiang et al., 2025; Xie et al., 2024). This issue highlights the urgent demand to build scalable generative frameworks that maintain 3D scene consistency while enabling users to freely explore 3D world.

Recently, numerous studies have explored the potential of 360° panorama (Zhou et al., 2024; Huang et al., 2025c) and iterative image generation (Höllein et al., 2023; Shriram et al., 2024) for novel view synthesis as alternative paradigms for world modeling. Specifically, panoramic-based methods leverage the image model to synthesize 360° panorama image and employ 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) to reconstruct 3D scene, which yields high-quality view synthesis near the scene center but constrains large-scale free exploration. Moreover, it is prone to geometric distortion and visual inconsistencies due to occlusion and missing multi-view information when large viewpoint changes. Furthermore, iterative image generation methods build upon existing viewpoints and leverage depth estimation to complete missing regions, thereby generating novel view images and progressively expanding the scene coverage. Although such methods generate scenes beyond panoramic scope, the accumulation of depth errors leads to geometric stretching and distortions, resulting in unrealistic rendering results.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To mitigate these limitations, trajectory-based video diffusion methods (Zhou et al., 2025; Yu et al., 2025b) have been explored to reduce artifacts by leveraging continuous motion priors, thereby generating locally plausible object geometric structures (Chen et al., 2025a; Huang et al., 2025a). For example, SEVA (Zhou et al., 2025) adopts diverse view sampling to achieve robust generalization and employs a two-pass inference strategy to enable flexible novel view synthesis along arbitrary camera trajectories. In addition, FlexWorld (Chen et al., 2025a) leverages a video model with geometry-aware integration and 3DGS optimization mechanism to generate visually consistent 3D scene with flexible viewpoints. However, the above methods still remain two challenges: (1) The y neglect the historical scenes and their spatial relationships during long video generation, thereby failing to maintain the long-term 3D scene consistency; (2) They struggle to effectively avoid trajectory collisions with objects due to the lack of awareness of object positions within the scene, thus disrupting local geometric plausibility and affecting temporal consistency, as shown in Fig. 1.

Based on the above observations, we propose a memory-guided and collision-aware video diffusion scheme for scene-consistent 3D world generation, termed **MC-World**, which maintains long-term 3D consistency while supporting flexible and unrestricted exploration. Specifically, we design a Spatio-Temporal Memory Retrieval (STMR) mechanism that exploits temporal continuity and spatial pose proximity between the current viewpoint and previously generated viewpoints to retrieve coherent historical scene context, thereby facilitating scene-consistent multi-view generation. Furthermore, to avoid geometrically unreasonable and temporally inconsistent results, we devise a Collision-Aware Exploration Control (CAEC) strategy that utilizes adaptive collision detection to regulate the generation process, thus suppressing potential collision conflicts. Attributing to these collaborative designs, MC-World not only achieves high-quality, well-structured geometry and long-term consistent 3D scene generation, but also enables unrestricted world exploration in complex scenes. Extensive experiments on diverse scenes demonstrate that our method yields state-of-the-art performance both qualitatively and quantitatively. The main contributions of this paper are as follows:

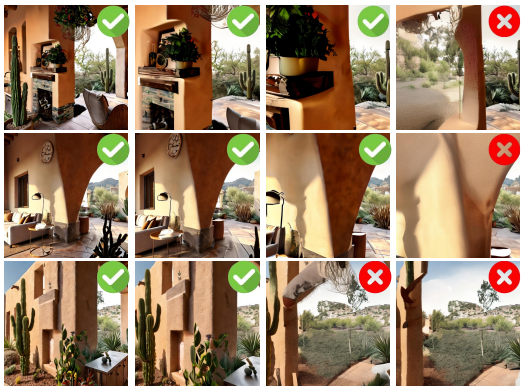


Figure 1: **Examples of trajectory-object collision in generated videos.** Non-collision Frame represents uncollided results that maintain better geometric coherence and temporal consistency, whereas Collision Frame indicates trajectory collisions with objects, resulting in unreasonable geometrical structures and noticeable temporal inconsistencies.

- We propose MC-World, a memory-guided and collision-aware video diffusion scheme for scene-consistent 3D world generation, which maintains long-term 3D consistency while supporting unrestricted exploration.
- We design an STMR mechanism that retrieves long-term consistent historical scene context based on temporal continuity and spatial pose proximity, thereby facilitating consistent multi-view generation.
- To reduce potential conflicts, we devise a CAEC strategy that performs adaptive collision detection and regulates high-risk trajectories, thereby ensuring geometric plausibility and temporal consistency.
- Extensive experiments demonstrate that our MC-World is superior to existing methods in both scene consistency and interactive exploration.

## 2 Related Works

**Image Diffusion-Based 3D Scene Generation.** Diffusion models (Ho et al., 2020; Lv et al., 2025b) have achieved significant breakthroughs in image generation (Liu et al., 2025; Song et al., 2020), enabling the synthesis of high-resolution and realistic visual content guided by given text prompts

(Lv et al., 2025a; Gong et al., 2025). Building upon these advances, recent works have extended image diffusion models to 3D scene generation by leveraging 360° panorama (Zhou et al., 2024; Chen et al., 2025b; Yang et al., 2025) and iterative image expansion strategy (Höllein et al., 2023; Shriram et al., 2024; Li et al., 2025) to construct spatially coherent scenes. Specifically, 360° panorama-based approaches such as DreamScene360 (Zhou et al., 2024) and LayerPano3D (Yang et al., 2025) first employ the text-to-image diffusion model (Black Forest Labs, 2023) to synthesize 360° panoramic image and then reconstruct 3D scene via 3DGS. They can achieve high-quality view synthesis near the scene center by inheriting strong generative priors from 2D diffusion models. However, their exploration capability is inherently constrained by the fixed panoramic coverage, and large viewpoint changes often introduce geometric distortions and visual inconsistencies due to occlusions and insufficient multi-view constraints. To mitigate these issues, iterative image expansion methods such as Text2Room (Höllein et al., 2023) and WonderWorld (Yu et al., 2025a) have been proposed, which progressively generate novel views based on existing images to achieve incremental scene expansion. By combining image diffusion with depth estimation models, these approaches incrementally complete unseen regions and expand scene coverage beyond the original field of view. Despite their flexibility in extending the scene scope, the iterative accumulation of monocular depth errors frequently results in geometric stretching and structural distortions, making it challenging to preserve global multi-view consistency. In contrast, we exploit the temporal and spatial continuity priors inherent and employ an efficient iterative generation mechanism to construct large-scale 3D scene, thereby addressing the aforementioned issues while effectively supporting unrestricted interactive exploration.

**Video Diffusion-Based 3D Scene Generation.** Recent video generation models (Yang et al., 2024; Blattmann et al., 2023; Liang et al., 2025) have exhibited remarkable potential to generate 3D scene (Yu et al., 2024; **?**; Xu et al., 2024; Yu et al., 2025b). For example, SEVA (Zhou et al., 2025) adopts diverse view sampling strategy to enhance cross-view consistency, while employing a two-pass inference mechanism to enable flexible novel view synthesis. Subsequently, GEN3C (Ren et al., 2025) reconstructs point clouds with pixel-level depth maps, utilizing them as the 3D cache for enable more precise camera manipulation during viewpoint extrapolation. FlexWorld (Chen et al., 2025a) leverages a depth-fine-tuned video model to generate novel views under large camera pose changes and progressively integrates new 3D content into the global scene via geometry-aware scene fusion. Moreover, Voyager (Huang et al., 2025a) further explicitly incorporates depth information as spatial priors and jointly generates RGB sequences to mitigate visual artifacts in video generation. However, the aforementioned methods still suffer from two dilemmas: (1) They fail to memorize the historical scenes and their spatial relationships in long-range video generation, resulting in long-term incoherent 3D scenes; (2) They are prone to trajectory collisions with objects due to the lack of evasion capabilities for objects, leading to unreasonable geometric structure and poor temporal consistency, thereby limiting interactive exploration. In this paper, we propose MC-World, which integrates spatio-temporal memory retrieval and collision-aware exploration control to maintain long-range 3D consistency while ensuring geometric plausibility and temporal coherence, thereby enabling realistic and unrestricted world exploration along arbitrary trajectories.

## 3 METHODOLOGY

The overview of the proposed MC-World is displayed in Fig. 2. Our method aims to address two core challenges: (1) maintaining long-term 3D scene consistency during large-range video generation, and (2) preventing trajectory–object collisions to ensure geometric plausibility and temporal coherence. Specifically, given a text prompt, we first utilize the text-to-image Flux model (Black Forest Labs, 2023) to generate four initial views covering different regions of the scene. Then, these views are expanded into a complete 360° panoramic scene overview using an inpainting model. Furthermore, the camera trajectories and panoramic images are jointly input into the video diffusion model to drive the subsequent 3D scene generation.

To maintain long-term 3D scene consistency, we design a Spatio-Temporal Memory Retrieval (STMR) mechanism that exploits temporal continuity and spatial pose proximity to retrieve coherent historical video frames, thereby providing consistent contextual guidance for multi-view generation. In addition, to prevent geometrically implausible and temporally inconsistent exploration, we develop a Collision-Aware Exploration Control (CAEC) strategy that estimates trajectory-specific collision risks and adaptively regulates the diffusion process to mitigate potential trajectory–object

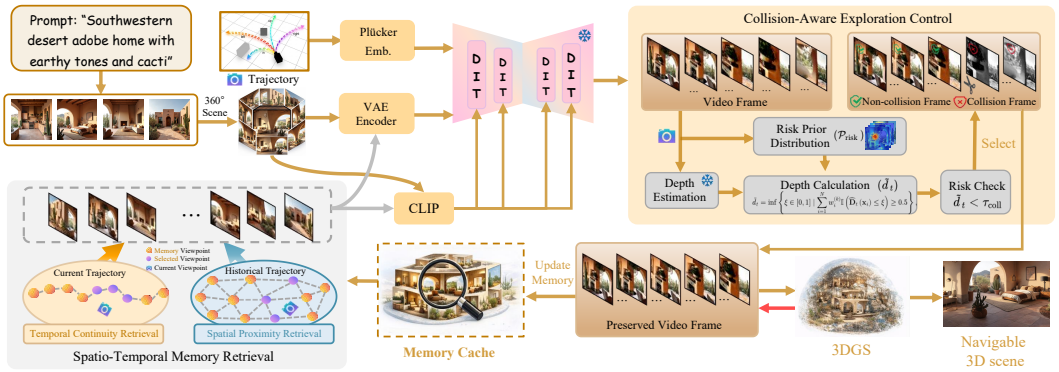


Figure 2: **Overview of MC-World.** MC-World integrates Spatio-Temporal Memory Retrieval and Collision-Aware Exploration Control to address long-range scene inconsistency and trajectory–object collisions. The former retrieves coherent historical context to preserve 3D consistency, while the latter adaptively regulates high-risk trajectories to improve geometric plausibility.

conflicts. Notably, STMR and CAEC are jointly designed to form a closed-loop generation process: STMR preserves global scene consistency by retrieving reliable historical context, while CAEC maintains geometric plausibility and temporal coherence by preventing collisions, consequently providing higher-quality frames for subsequent memory retrieval. In the following, we first introduce the preliminaries and then present the details of STMR and CAEC.

### 3.1 Preliminary

**Diffusion Model.** Diffusion models (Nichol & Dhariwal, 2021; Croitoru et al., 2023) are powerful generative models that synthesize data by learning a gradual denoising process. In the forward diffusion process, starting from a clean sample $\mathbf{x}_0$, Gaussian noise is progressively added over $T$ timesteps according to a defined noise schedule, such that the data distribution gradually approaches an isotropic Gaussian distribution. At an arbitrary timestep $t \in [0, T]$, the noisy sample can be expressed as:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\alpha_t$ and $\sigma_t$ are predefined scheduling coefficients controlling the signal-to-noise ratio.

In the reverse denoising process, the model learns to progressively remove noise and recover samples that follow the data distribution from pure Gaussian noise. Specifically, a parameterized noise prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is trained to estimate the injected noise at each timestep. The optimization objective is formulated as:

$$\min_\theta \ \mathbb{E}_{t\sim\mathcal{U}(0,T),\, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})} \left[ \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_2^2 \right]. \tag{2}$$

**3D Gaussian Splatting.** As one of the most widely adopted explicit 3D representations, 3DGS models a scene as a collection of 3D Gaussian primitives, which can be defined as:

$$\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^N. \tag{3}$$

Each Gaussian primitive $\mathcal{G}_i$ is parameterized by color $c_i \in \mathbb{R}^k$, center position $\mu_i \in \mathbb{R}^3$, opacity $\alpha_i \in \mathbb{R}$, scale $s_i \in \mathbb{R}^3$, and rotation quaternion $q_i \in \mathbb{R}^4$. During optimization, 3DGS learns the scene parameters by jointly minimizing the reconstruction loss $\mathcal{L}_1$ and structural similarity loss $\mathcal{L}_{\text{SSIM}}$. The overall objective is formulated as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{SSIM}}, \tag{4}$$

where $\lambda$ is a hyperparameter, and it is set to 0.2 by default.

### 3.2 Spatio-Temporal Memory Retrieval

To ensure long-term 3D scene consistency, we propose a STMR mechanism that retrieves scene-consistent historical content based on temporal continuity and spatial pose proximity, thereby providing coherent contextual guidance for large-scale 3D scene generation. Specifically, we organize

all previously generated video frames and their corresponding camera viewpoints into a unified global memory cache:

$$\mathcal{M} = \{(\mathbf{F}_i, \mathbf{V}_i)\}_{i=1}^{N}, \tag{5}$$

where $\mathbf{F}_i$ denotes the $i$-th historical video frame, and $\mathbf{V}_i \in SE(3)$ denotes its corresponding camera viewpoint, which consists of a rotation matrix $\mathbf{R}_i \in SO(3)$ and a translation vector $\mathbf{T}_i \in \mathbb{R}^3$. The initial 8 input images are also included in the global memory cache.

Given a target viewpoint $\mathbf{V}_j^*$, STMR aims to retrieve historical viewpoints that are consistent with the target viewpoint in terms of both temporal continuity and spatial pose proximity. The video frames corresponding to the retrieved viewpoints are then employed as conditioning inputs, providing continuous historical context for the subsequent diffusion process. STMR consists of two complementary components: Temporal Continuity Retrieval and Spatial Proximity Retrieval, which are described below.

**Temporal Continuity Retrieval.** To maintain temporal consistency during long-range video generation, we first select locally coherent historical viewpoints along the temporal dimension. Specifically, we compute the rotational distance between the target viewpoint $\mathbf{V}_j^*$ and each historical viewpoint $\mathbf{V}_i$ in the global memory cache:

$$d^{\mathrm{rot}}(i,j) = \arccos\left(\frac{\mathrm{tr}\left(\mathbf{R}_i^\top \mathbf{R}_j^*\right) - 1}{2}\right), \tag{6}$$

where $\mathbf{R}_i$ and $\mathbf{R}_j^*$ denote the rotation components of $\mathbf{V}_i$ and $\mathbf{V}_j^*$, and $\mathrm{tr}(\cdot)$ denotes the trace operator. Then, we identify the historical viewpoint closest to the target viewpoint in the rotation space and use it as the rotation anchor:

$$i_j^{\mathrm{anc}} = \arg\min_i d^{\mathrm{rot}}(i,j). \tag{7}$$

The resulting anchor corresponds to the historical viewpoint whose viewing direction is most similar to that of the target viewpoint and serves as a temporal reference for current generation step. Based on this anchor, we construct a local temporal neighborhood to select candidate historical viewpoints. Given a predefined temporal window size $K_{\mathrm{temp}}$, the temporal candidate set is defined as

$$\mathcal{C}_j^{\mathrm{temp}} = \left\{ i \;\middle|\; \left|i - i_j^{\mathrm{anc}}\right| \le K_{\mathrm{temp}} \right\}. \tag{8}$$

By jointly considering viewing-direction similarity and temporal proximity, Temporal continuity retrieval restricts candidate viewpoints to a locally coherent historical region, reducing temporal discontinuities during long-range generation.

**Spatial Proximity Retrieval.** Although the temporal neighborhood preserves local continuity, it cannot explicitly guarantee global 3D consistency under large camera motions or when the camera revisits previously explored regions. We therefore introduce spatial proximity retrieval to identify historical viewpoints that are geometrically close to the target viewpoint. In addition to the rotational distance, we compute the translation distance between the target viewpoint $\mathbf{V}_j^*$ and each historical viewpoint $\mathbf{V}_i$:

$$d^{\mathrm{trans}}(i,j) = \left\|\mathbf{T}_i - \mathbf{T}_j^*\right\|_2, \tag{9}$$

where $\mathbf{T}_i$ and $\mathbf{T}_j^*$ denote the translation components of $\mathbf{V}_i$ and $\mathbf{V}_j^*$, respectively. Then, we define a unified spatial pose distance by combining the normalized translation and rotation differences:

$$d^{\mathrm{spatial}}(i,j) = \frac{d^{\mathrm{trans}}(i,j)}{D_{\max}} + \frac{d^{\mathrm{rot}}(i,j)}{\pi}, \tag{10}$$

where $D_{\max}$ denotes the maximum translation distance in the viewpoint space. Based on this metric, we select the $K_{\mathrm{scene}}$ historical viewpoints with the smallest spatial pose distances:

$$\mathcal{C}_j^{\mathrm{spatial}} = \mathrm{TopK}_{\min}\left(\left\{d^{\mathrm{spatial}}(i,j)\right\}_{i=1}^{N}, K_{\mathrm{scene}}\right), \tag{11}$$

where $\mathrm{TopK}_{\min}(\cdot)$ returns the indices corresponding to the $K_{\mathrm{scene}}$ smallest values. By jointly considering rotational similarity and translation proximity, spatial proximity retrieval identifies globally relevant historical viewpoints and improves geometric consistency during large-scale exploration.

Finally, we obtain the spatio-temporally consistent candidate set by intersecting temporal and spatial candidate sets:

$$\mathcal{C}_j^{\mathrm{ST}} = \mathcal{C}_j^{\mathrm{temp}} \cap \mathcal{C}_j^{\mathrm{spatial}}. \tag{12}$$




Figure 3: **Qualitative comparison of our MC-World and other methods.** Existing methods suffer from blurry artifacts, 3D inconsistencies, and geometrically implausible structures under viewpoint changes. In contrast, MC-World generates high-quality, geometrically coherent results with strong 3D consistency and supports unrestricted exploration across viewpoints.

After determining the spatio-temporally consistent historical viewpoints, the corresponding video frames are encoded through complementary semantic and latent pathways. Specifically, the semantic memory features are obtained using the CLIP encoder:

$$\mathcal{F}_j^{\text{mem}} = \left\{ \text{CLIP}(\mathbf{F}_i) \,\middle|\, i \in \mathcal{C}_j^{\text{ST}} \right\}, \tag{13}$$

where $\mathcal{F}_j^{\text{mem}}$ denotes the semantic memory features used as cross-attention context for the video diffusion model. Meanwhile, the retrieved frames are encoded by the VAE encoder to obtain fine-grained latent memory representations:

$$\mathcal{Z}_j^{\text{mem}} = \left\{ E_{\text{VAE}}(\mathbf{F}_i) \,\middle|\, i \in \mathcal{C}_j^{\text{ST}} \right\}, \tag{14}$$

where $\mathcal{Z}_j^{\text{mem}}$ denotes the latent memory representations, which are inserted into corresponding conditional positions, providing explicit appearance and structure guidance for subsequent generation.

### 3.3 Collision-Aware Exploration Control

During trajectory-based video generation, strictly following predefined camera trajectories may lead to trajectory–object collisions. Such collisions not only produce geometrically implausible scene structures, but also disrupt temporal coherence and restrict realistic exploration, as shown in Fig. 1. To address this issue, we propose a CAEC strategy that dynamically estimates collision risks and adaptively regulates the exploration trajectory.

Specifically, given the current exploration trajectory $\pi_k$, we first construct a trajectory-aligned spatial risk distribution over the discrete image grid $\Omega = \{\mathbf{x}_i\}_{i=1}^{N}$. The unnormalized collision risk score at each pixel is defined as:

$$r_i^{(k)} = \exp\left( -\frac{1}{2} \left( \mathbf{x}_i - \boldsymbol{\mu}_{\pi_k} \right)^{\top} \boldsymbol{\Sigma}_{\pi_k}^{-1} \left( \mathbf{x}_i - \boldsymbol{\mu}_{\pi_k} \right) \right), \tag{15}$$

Table 1: **Quantitative comparison with existing methods.** Our method achieves the best performance across all automatic metrics. ↑: higher is better, and ↓: lower is better.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | NIQE ↓ | MUSIQ ↑ | NIMA ↑ | BRISQUE ↓ | HYPER ↑ | CS ↑ | IS ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| DreamScene360 | 16.27 | 0.5894 | 0.4931 | 6.54 | 51.90 | 4.70 | 44.41 | 0.4271 | 0.3864 | 2.20 |
| SEVA | 16.85 | 0.6029 | 0.4841 | 4.50 | 64.01 | 5.15 | 26.37 | 0.5209 | 0.4349 | 1.37 |
| LayerPano3D | 18.76 | 0.6426 | 0.3964 | 7.16 | 34.82 | 4.73 | 48.34 | 0.3386 | 0.2353 | 1.34 |
| WorldExplorer | 20.18 | 0.6534 | 0.4630 | 7.49 | 45.52 | 4.92 | 43.46 | 0.2560 | 0.2363 | 2.27 |
| FlexWorld | 18.33 | 0.6249 | 0.4148 | 5.60 | 54.02 | 4.95 | 34.97 | 0.4690 | 0.3151 | 1.56 |
| **MC-World (Ours)** | **22.52** | **0.6910** | **0.3713** | **3.39** | **71.60** | **5.73** | **21.41** | **0.5993** | **0.5744** | **2.31** |

where $\boldsymbol{\mu}_{\pi_k}$ and $\boldsymbol{\Sigma}_{\pi_k}$ denote the mean and covariance of spatial risk distribution aligned with the motion direction of trajectory $\pi_k$. The collision risk scores are normalized as:

$$w_i^{(k)} = \frac{r_i^{(k)}}{\sum_{n=1}^{N} r_n^{(k)}}, \qquad \sum_{i=1}^{N} w_i^{(k)} = 1. \tag{16}$$

Thus, pixels located along the forward motion direction are assigned larger weights, allowing CAEC to focus on image regions that are more likely to encounter scene objects.

Subsequently, we employ a video depth estimation network to obtain the normalized depth map $\widehat{\mathbf{D}}_t(\mathbf{x}_i) \in [0, 1]$ for the $t$-th generated frame, where smaller values indicate scene structures closer to the camera. Directly averaging the estimated depths may be sensitive to isolated depth errors and temporal flickering. Therefore, we adopt a risk-weighted median to robustly characterize the foreground depth along the camera motion direction:

$$\widetilde{d}_t = \inf \left\{ \xi \in [0,1] \;\middle|\; \sum_{i=1}^{N} w_i^{(k)} \mathbb{I}\left(\widehat{\mathbf{D}}_t(\mathbf{x}_i) \le \xi\right) \ge 0.5 \right\}, \tag{17}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Intuitively, $\widetilde{d}_t$ is the depth value at which the cumulative collision-risk weight of pixels with smaller or equal depth reaches $0.5$. In practice, it is computed by sorting the estimated depth values in ascending order and selecting the first depth whose cumulative risk weight is no smaller than $0.5$.

Based on the estimated risk-aware depth, we determine the collision status of the current frame as

$$\text{Status}_t = \begin{cases} \text{collision}, & \widetilde{d}_t < \tau_{\text{coll}}, \\ \text{non-collision}, & \text{otherwise}, \end{cases} \tag{18}$$

where $\tau_{\text{coll}}$ denotes the collision threshold. When $\widetilde{d}_t < \tau_{\text{coll}}$, CAEC identifies a potential collision along the current trajectory. It then terminates the remaining portion of the high-risk trajectory, stores the previously generated non-collision frames in the global memory cache, and switches to a new exploration direction. We set $\tau_{\text{coll}} = 0.4$ based on the sensitivity analysis presented in Table 3.

Through this adaptive trajectory regulation, CAEC prevents the camera trajectory from passing through scene objects, thereby improving geometric plausibility and temporal coherence while supporting unrestricted exploration. Moreover, the valid frames preserved by CAEC provide reliable historical content for subsequent STMR, establishing a closed-loop interaction between memory retrieval and collision-aware exploration.

After completing the exploration process, we optimize 3DGS using all valid generated frames according to Eq. (4). By jointly optimizing the geometry and appearance across different viewpoints, we obtain a consistent and renderable explicit 3D representation for unrestricted world exploration.

# 4 Experiments

## 4.1 Experimental Settings

**Implementation Details.** We employ the default trajectory sampling scheme to generate all experimental results, comprising 8 initial images and 4 predefined trajectories. All experiments are executed on a single NVIDIA RTX 4090 GPU, with scenes rendered at a resolution of 576 × 576 pixels. In the experiments, we evaluate our MC-World using a variety of text prompts and scene configurations (covering both indoor and outdoor scenes) for comprehensive assessment.




Figure 4: **Visual results of our MC-World.** MC-World generates high-fidelity, geometrically consistent 3D scenes across diverse environments and camera trajectories, demonstrating strong effectiveness and generalization.

**Evaluation Metrics.** We use PSNR, SSIM (Bakurov et al., 2022), and LPIPS (Zhang et al., 2018) to assess 3D consistency and geometric accuracy. Meanwhile, we report NIQE (Mittal et al., 2012b), MUSIQ (Ke et al., 2021), NIMA (Talebi & Milanfar, 2018), BRISQUE (Mittal et al., 2012a), and HYPER (Su et al., 2020) to assess perceptual visual quality. In addition, we include CLIP Score (CS) (Radford et al., 2021) and Inception Score (IS) (Salimans et al., 2016) as complementary metrics. All metrics are calculated on renderings from multiple novel viewpoints to comprehensively assess visual quality, cross-view consistency and generative diversity. More experimental details are provided in the *Supplementary Material*.

### 4.2 Experimental Results

**Qualitative comparison.** We compare our MC-World with existing approaches in terms of novel-view scene quality. As shown in Fig. 3, existing methods suffer from blurry artifacts, 3D scene inconsistencies, and geometrically implausible structures across viewpoints. Specifically, DreamScene360 and LayerPano3D struggle to generate plausible content for unobserved regions beyond the panoramic view, leading to visible distortions and occlusion artifacts. SEVA neglects the potential risk of trajectory–object collisions due to insufficient modeling of spatial relationships among scene objects, thereby leading to geometrically unreasonable structures in the generated results. In addition, FlexWorld ignores the effective memory content, resulting in noticeable 3D inconsistencies across viewpoints and consequently restricting free navigation within the scene. In contrast, our MC-World achieves superior visual quality and geometrically plausible results while maintaining robust 3D consistency and supporting unrestricted exploration along arbitrary trajectories. We also present our independent experimental results in Fig. 4, including scene overviews and novel views that deviate from the camera center. These results further demonstrate that our MC-World consistently produces high-fidelity, structurally stable, and geometrically coherent 3D scene across diverse scenarios.

**Quantitative Comparison.** We further validate the effectiveness of MC-World through quantitative experiments. As shown in Table 1, MC-World outperforms existing methods across all automatic metrics. PSNR, SSIM, and LPIPS directly reflect cross-view reconstruction fidelity and 3D consistency, demonstrating our ability to generate geometrically consistent novel views. The superior NIQE, MUSIQ, NIMA, BRISQUE, and HYPER scores highlight its perceptual quality, while the best CS and IS scores indicate strong prompt alignment and structural coherence beyond the panoramic field of view. In addition, we also report user study results in Table 4 of the *Supplementary Material*, where MC-World achieves the highest scores in visual quality, 3D consistency,

Table 2: **Quantitative results of the ablation study.** Both STMR and CAEC play essential roles in improving visual fidelity, structural stability, perceptual quality, and 3D consistency. ↑ indicates that higher is better, while ↓ indicates that lower is better.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | NIQE ↓ | MUSIQ ↑ | NIMA ↑ | BRISQUE ↓ | HYPER ↑ | CS ↑ | IS ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| w/o STMR | 19.98 | 0.6342 | 0.4025 | 6.38 | 59.93 | 5.02 | 35.78 | 0.3982 | 0.2353 | 2.06 |
| w/o CAEC | 20.31 | 0.6548 | 0.4216 | 5.82 | 55.98 | 4.86 | 33.45 | 0.2866 | 0.2801 | 2.19 |
| **MC-World (Ours)** | **22.52** | **0.6910** | **0.3713** | **3.39** | **71.60** | **5.73** | **21.41** | **0.5993** | **0.5744** | **2.31** |

geometric plausibility, and interactive exploration, further confirming its ability to generate high-fidelity, structurally coherent 3D scenes with unrestricted world exploration.

### 4.3 Ablation Study

**Effectiveness of Spatio-Temporal Memory Retrieval.** We verify the contribution of STMR in Table 2 and Fig. 5. It shows without STMR leads to noticeable degradation across all evaluation metrics. Specifically, NIQE and BRISQUE increase significantly, while MUSIQ, NIMA, HYPER, CS and IS decrease compared to the full model, which indicate that without STMR, the generated results perform poorly in perceptual quality, semantic consistency and 3D consistency across views. After incorporating STMR, all metrics show substantial improvements, demonstrating its effectiveness in enhancing visual fidelity, structural stability and 3D consistency.

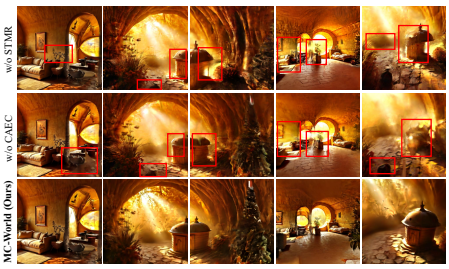


Figure 5: **Visual ablation analysis.** Full MC-World achieves superior visual quality and stronger 3D structural consistency.

**Effectiveness of Collision-Aware Exploration Control.** We perform validation without CAEC, where the generation process lacks a risk-aware geometric constraint. As illustrated in Table 2 and Fig. 5, removing CAEC results in unreasonable geometries and inconsistent sequences due to the absence of collision detection strategy, thereby hindering free exploration. Moreover, MUSIQ, NIMA, HYPER, CS and IS scores decrease, while NIQE and BRISQUE increase, indicating consistent degradation across all evaluation metrics. This performance drop stems from the lack of risk-aware regulation during generation, which fails to effectively prevent trajectory–object conflicts. These results further confirm the critical role of CAEC in enhancing geometric plausibility, temporal consistency and overall generation quality.

**Effectiveness of Different Collision Thresholds $\tau_{\text{coll}}$.** We conduct an ablation study on the collision threshold $\tau_{\text{coll}}$, as shown in Table 3. When $\tau_{\text{coll}} = 0.2$, all evaluation metrics exhibit noticeable degradation, indicating that insufficient collision constraint leads to reduced visual fidelity and geometric structure. While a larger threshold $\tau_{\text{coll}} = 0.6$ results in overly conservative collision constraint, limiting scene exploration ability and leading to overall performance degradation. In contrast, $\tau_{\text{coll}} = 0.4$ achieves the best results in automatic metrics and user study. Therefore, we set $\tau_{\text{coll}} = 0.4$ in all experiments.

Table 3: **Ablation on $\tau_{\text{coll}}$.** Our method ($\tau_{\text{coll}} = 0.4$) achieves the best performance. ↑: higher is better. ↓: lower is better.

| $\tau_{\text{coll}}$ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | NIQE ↓ | MUSIQ ↑ | IS ↑ |
|---|---|---|---|---|---|---|
| 0.2 | 21.38 | 0.6619 | 0.4156 | 5.11 | 63.68 | 2.03 |
| **0.4 (Ours)** | **22.52** | **0.6910** | **0.3713** | **3.39** | **71.60** | **2.31** |
| 0.6 | 21.84 | 0.6793 | 0.4012 | 4.83 | 56.21 | 1.97 |

## 5 Conclusion

In this paper, we propose MC-World, a novel memory-guided and collision-aware video diffusion framework for scene-consistent 3D world generation. Unlike existing methods that neglect long-term historical context and trajectory-object collision risks, our MC-World leverages effective historical memories and collision-aware detection strategy to achieve long-term consistency while enabling realistic and unrestricted 3D world exploration. Specifically, MC-World consists of two core components: Spatio-Temporal Memory Retrieval and Collision-Aware Exploration Control. The former retrieves coherent historical scene information from the global memory cache by exploiting temporal continuity and spatial pose proximity, thereby preserving long-term 3D scene consistency. The latter combines trajectory-aware collision detection with depth information to identify potential trajectory–object conflicts and adaptively truncate high-risk trajectories, thereby improving geometric plausibility and temporal coherence. Extensive experimental results show that MC-World significantly outperforms existing methods both qualitatively and quantitatively in diverse 3D scenes.

### AI USE STATEMENT

In this work, we used generative AI tools to polish the content. We did not use generative AI tools for data generation or concept design. We have reviewed all work completed with the assistance of AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

The code will be released after the paper is accepted.

## REFERENCES

Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi. Structural similarity index (ssim) revisited: A data-driven approach. *Expert Systems with Applications*, 189:116087, 2022.

Black Forest Labs. Flux, 2023. URL `https://github.com/black-forest-labs/flux`.

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.

Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025a.

Zheng Chen, Chenming Wu, Zhelun Shen, Chen Zhao, Weicai Ye, Haocheng Feng, Errui Ding, and Song-Hai Zhang. Splatter-360: Generalizable 360 gaussian splatting for wide-baseline panoramic images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21590–21599, 2025b.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhengyi Gong, Mingwen Shao, Chang Liu, Xiang Lv, and Huan Liu. Fcat-diff: Flexible and consistent appearance transfer based on training-free diffusion model. *Computers & Graphics*, pp. 104247, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7909–7920, 2023.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025a.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23646–23657, 2025b.

Zilong Huang, Jun He, Junyan Ye, Lihan Jiang, Weijia Li, Yiping Chen, and Ting Han. Scene4u: Hierarchical layered 3d scene reconstruction from single panoramic image for your immerse exploration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26723–26733, 2025c.

Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–1, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Hyunjeong Kim and In-Kwon Lee. Is 3dgs useful?: Comparing the effectiveness of recent reconstruction methods in vr. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 71–80. IEEE, 2024.

Bohan Li, Xingyi Li, Yangwen Liang, Shuangquan Wang, and Kee-Bong Song. Leveraging latent diffusion in 3d gaussian splatting for novel view synthesis. In *International Conference on Multimedia Modeling*, pp. 150–157. Springer, 2025.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 798–810, 2025.

Chang Liu, Mingwen Shao, Zhengyi Gong, Xiang Lv, and Lingzhuang Meng. Mirrordiff: Prompt redescription for zero-shot grounded text-to-image generation with attention modulation. *Engineering Applications of Artificial Intelligence*, 153:110741, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

Xiang Lv, Mingwen Shao, Lingzhuang Meng, Chang Liu, Yecong Wan, and Xinyuan Chen. Suv: Suppressing undesired video content via semantic modulation based on text embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18357–18366, 2025a.

Xiang Lv, Mingwen Shao, Yecong Wan, Yue Ma, and Yuanshuo Cheng. Bm-edit: Background retention and motion consistency for zero-shot video editing. *Knowledge-Based Systems*, pp. 113784, 2025b.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Xiang Tang, Ruotong Li, and Xiaopeng Fan. Zeroscene: A zero-shot framework for 3d scene generation from a single image and controllable texture editing. *arXiv preprint arXiv:2509.23607*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025.

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9666–9675, 2024.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pp. 1–10, 2025.

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. *arXiv preprint arXiv:2408.06072*, 2024.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5916–5926, 2025a.

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 *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025b.

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. *arXiv preprint arXiv:2409.02048*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Tianyi Zhang, Kaining Huang, Weiming Zhi, and Matthew Johnson-Roberson. Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12864–12871. IEEE, 2024.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12405–12414, 2025.

Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, pp. 324–342. Springer, 2024.

## OVERVIEW

In this supplementary material, we provide additional implementation details and more experimental results to comprehensively validate the effectiveness of our MC-World. Specifically, Sec. A elaborates on the detailed implementation of our framework. Subsequently, Sec. B presents extended qualitative comparisons with existing approaches, further demonstrating the advantages of our method in terms of high fidelity, structural stability, and geometric consistency. Finally, Sec. C showcases additional independent results across diverse scenes, highlighting the robustness and high-quality generation capability of our MC-World.

## A. MORE IMPLEMENTATION DETAILS

**Additional Setup Details.** In our method, we adopt SEVA (Zhou et al., 2025) as the pretrained video diffusion backbone for novel-view synthesis. All videos are generated sequentially in a clockwise order according to the initial images. For each video, we first generate the "in" trajectory, followed by the "left", "right", and "up" trajectories. All experiments are conducted on Ubuntu 22.04.1 using an Intel Core i5-13600KF CPU, one NVIDIA RTX 4090 GPU, and 32 GB of RAM. We will open access to the code after the paper is accepted.

**Comparison Methods.** To conduct a comprehensive comparison, we evaluate our method against several representative 3D scene generation approaches, including DreamScene360 (Zhou et al., 2024), SEVA (Zhou et al., 2025), LayerPano3D (Yang et al., 2025), WorldExplorer (Schneider et al., 2025), and FlexWorld (Chen et al., 2025a). All comparison methods are evaluated using their default settings as specified in the papers.

## B. MORE COMPARISON RESULTS

To more intuitively demonstrate the effectiveness of our MC-World, we conduct extensive comparisons with existing representative 3D scene generation approaches across diverse scenes. Specifically, we compare our approach with five state-of-the-art baselines: DreamScene360 (Zhou et al., 2024), SEVA (Zhou et al., 2025), LayerPano3D (Yang et al., 2025), WorldExplorer (Schneider et al., 2025), and FlexWorld (Chen et al., 2025a). All methods are evaluated using the default configurations provided in their original papers to ensure a fair comparison.

Through extensive evaluation, our MC-World demonstrates significant advantages across various scenarios. As shown in Fig. 6, existing methods are prone to issues such as blur artifacts, inconsistent 3D structures, and unreasonable geometric results under different viewpoints. For instance, DreamScene360 (Zhou et al., 2024) and LayerPano3D (Yang et al., 2025) struggle to generate realistic content in regions not directly observed in the synthetic panorama input, resulting in distortion and occlusion-related artifacts. Furthermore, SEVA (Zhou et al., 2025) fails to adequately model spatial relationships between scene objects, which may lead to trajectory collisions with objects and thus generate geometrically unreasonable scene structures. Additionally, FlexWorld (Chen et al., 2025a) struggles to effectively leverage memory representations, causing significant inconsistencies in scene reconstruction and thereby limiting the flexibility of free-viewpoint navigation. In contrast, our MC-World generates visually coherent and geometrically plausible scenes while maintaining strong 3D consistency, enabling unrestricted exploration along arbitrary trajectories.

Meanwhile, our method also supports depth estimation. During rendering, the corresponding depth maps are produced simultaneously with the generated results. As shown in Fig. 11, the predicted depths are well aligned with the scene geometry, further demonstrating the effectiveness of our method in modeling scene geometry and its strong potential for practical applications.

To further evaluate the perceptual quality and exploration experience of the generated 3D world, we conduct a user study with 100 participants. The assessment focuses on four aspects: (1) viewpoint-aware visual quality, referring to realism and detail preservation across different views; (2) 3D consistency, measuring structural coherence across viewpoints; (3) geometric plausibility, evaluating the reasonableness of the generated scene geometry; and (4) interactive exploration, reflecting scene completeness and visual appeal during navigation. All evaluations are conducted in a randomized and anonymized manner to minimize potential bias.




Figure 6: **Qualitative comparison of our MC-World and other methods.** Existing methods suffer from artifacts and 3D inconsistencies across viewpoints, whereas our MC-World achieves geometrically coherent results with strong 3D consistency.

The final results are obtained by averaging the scores from all participants. Table 4 presents the results of the user study. Our MC-World achieves the highest scores in visual quality, 3D consistency, geometric plausibility, and interactive exploration, further confirming our superiority in generating high-fidelity and structurally coherent 3D scene while enabling unrestricted world exploration.




Figure 7: **Visual results of our MC-World.** The results show that our MC-World achieves high-fidelity and geometrically consistent 3D scene generation across diverse scenes and camera trajectories.

Table 4: **User study comparison with existing methods.** Our MC-World achieves the best performance across all metrics.

| Method | Visual Quality ↑ | 3D Consistency ↑ | Geometric Plausibility ↑ | Interactive Exploration ↑ |
|---|---|---|---|---|
| DreamScene360 (Zhou et al., 2024) | 3.48 | 3.02 | 3.66 | 3.53 |
| SEVA (Zhou et al., 2025) | 4.03 | 3.25 | 3.03 | 3.76 |
| LayerPano3D (Yang et al., 2025) | 3.25 | 3.47 | 3.54 | 3.33 |
| WorldExplorer (Schneider et al., 2025) | 3.86 | 3.78 | 3.92 | 4.18 |
| FlexWorld (Chen et al., 2025a) | 4.15 | 3.72 | 3.61 | 3.94 |
| **MC-World (Ours)** | **4.53** | **4.24** | **4.32** | **4.49** |

## C. More Visual Results

To further demonstrate the advantages and generalization capability of MC-World, we present additional independent experimental results in Fig. 7, Fig. 8 and Fig. 9, including scene overviews and supplementary novel views that deviate from the camera center. These results further verify that MC-World consistently generates high-fidelity, structurally stable, and geometrically coherent 3D scenes across diverse scenes while enabling unrestricted exploration along arbitrary trajectories. Furthermore, we provide accompanying video materials to better visualize the generated results, allowing readers to observe the advantages of our method more intuitively.

Meanwhile, we further present out-of-distribution results in Fig. 10 to evaluate the generalization capability of MC-World in previously unseen environments. Despite substantial differences from the scenes used in our standard evaluation, MC-World consistently generates high-fidelity novel views while preserving stable scene structures under long-range camera movements. Moreover, the generated views maintain coherent geometry and strong 3D consistency across diverse viewpoints, with




Figure 8: **Visual results of our MC-World.** The results show that our MC-World achieves high-fidelity and geometrically consistent 3D scene generation across diverse scenes and camera trajectories.




Figure 9: **Visual results of our MC-World.** The results show that our MC-World achieves high-fidelity and geometrically consistent 3D scene generation across diverse scenes and camera trajectories.

fewer structural distortions and temporal inconsistencies. These results demonstrate the robustness of MC-World in unfamiliar environments.

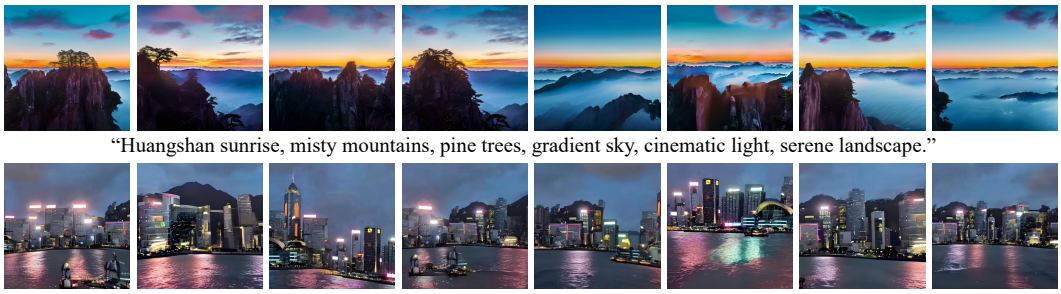


Figure 10: **Out-of-distribution (OOD) visual results of our method.** MC-World consistently generates high-fidelity novel views in previously unseen environments while maintaining strong long-range 3D consistency and coherent scene geometry, demonstrating its robust generalization capability.

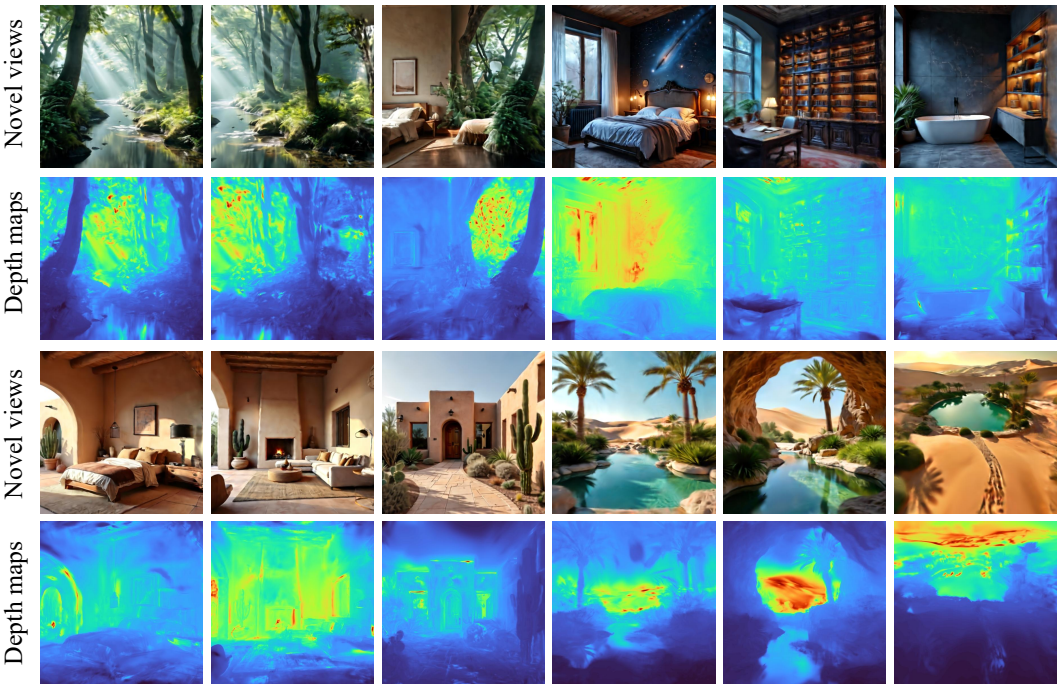


Figure 11: **Depth estimation results.** MC-World renders novel views with depth maps that faithfully capture scene geometry.