This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# WorldStereo: Bridging Camera-Guided Video Generation and Scene Reconstruction via 3D Geometric Memories

Yisu Zhang[1,2*], Chenjie Cao[2*], Tengfei Wang[2†],
Xuhui Zuo[2], Junta Wu[2], Jianke Zhu[1‡], Chunchao Guo[2]
[1]Zhejiang University [2] Tencent Hunyuan

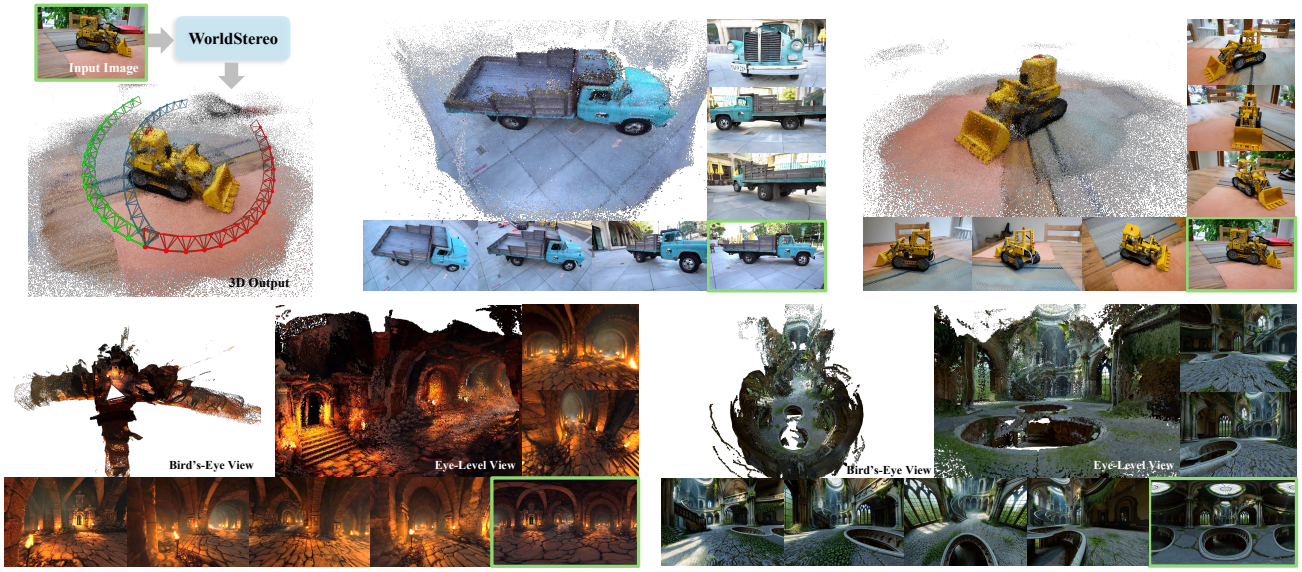



Figure 1. **WorldStereo enables high-quality 3D scene generation based on single-view or panoramic inputs.** The input reference views are framed in green. We present point clouds reconstructed from videos generated by WorldStereo: the top two perspective scenes use WorldMirror [57], while the bottom two panoramic scenes are aligned via monocular depth maps [83].

## Abstract

*Recent advances in foundational Video Diffusion Models (VDMs) have yielded significant progress. Yet, despite the remarkable visual quality of generated videos, reconstructing consistent 3D scenes from these outputs remains challenging, due to limited camera controllability and inconsistent generated content when viewed from distinct camera trajectories. In this paper, we propose **WorldStereo**, a novel framework that bridges camera-guided video generation and 3D reconstruction via two dedicated geometric memory modules. Formally, the global-geometric memory enables precise camera control while injecting coarse structural priors through incrementally updated point clouds. Moreover, the spatial-stereo memory constrains the model's attention receptive fields with 3D correspondence to focus on fine-grained details from the memory bank. These components enable WorldStereo to generate multi-view-consistent videos under precise camera control, facilitating high-quality 3D reconstruction. Furthermore, the flexible control branch-based WorldStereo shows impressive efficiency, benefiting from the distribution matching distilled VDM backbone without joint training. Extensive experiments across both camera-guided video generation and 3D reconstruction benchmarks demonstrate the effectiveness of our approach. Notably, we show that WorldStereo acts as a powerful world model, tackling diverse scene generation tasks (whether starting from perspective or panoramic images) with high-fidelity 3D results. Models will be released on https://github.com/FuchengSu/WorldStereo.*

*Equal Contribution. †Project Lead. ‡Corresponding author

Table 1. **Different video generation schemes for 3D reconstruction. Long-Bi** VDMs produce long trajectories in a single pass to cover diverse viewpoints. **AR** models sequentially generate long videos in an autoregressive manner. **Multi-Bi-Mem** (ours) achieves multiple consistent generations based on a powerful open-released VDM [77] with complementary viewpoints and memory mechanisms for integrated reconstruction.

| **Paradigms** | **Long-Bi** | **AR** | **Multi-Bi-Mem** |
|---|---|---|---|
| Receptive Field | Bidirectional | Autoregressive | Bidirectional |
| Trajectory Length/Num | Long/Single | Long/Single | Medium/Multiple |
| Consistency | 😐 | 😐 | 🙂 |
| Precise Camera Control | 🙂 | 😐 | 🙂 |
| Video Quality | 😐 | ☹ | 🙂 |
| Efficiency | ☹ | 🙂 | 🙂 |
| Community Progress | ☹ | ☹ | 🙂 |

## 1. Introduction

Recently, remarkably developed Video Diffusion Models (VDMs) [7, 8, 46, 47, 66, 77, 98] have made impressive advances, achieving remarkable performance in photorealistic video synthesis. These models have showcased significant potential across a broad range of applications, including virtual reality, digital content creation, and embodied AI. Meanwhile, camera-controllable VDMs have also enabled substantial progress, incorporating various camera and action controls [3, 31, 32, 53, 115], as well as leveraging explicit guidance like point clouds [13, 22, 50, 59, 63, 65, 105], optical flow [9, 43], and tracking points [28, 78].

Despite these breakthroughs, current camera-guided VDMs remain limited in recovering consistent and reliable 3D scene reconstructions, even when paired with advanced feed-forward 3D reconstruction approaches [44, 81, 82, 95], failing to step toward generalizable world models. A fundamental challenge lies in enabling VDMs to *generate videos that capture sufficiently diverse and comprehensive viewpoints of the target scene, while maintaining precise camera control and high-fidelity visual quality*. Existing camera-guided VDMs struggle to preserve consistency across varied camera trajectories, leading to ambiguous and blurry reconstructions. Intuitively, extending them into longer sequences can capture richer viewpoint coverage with natural consistency from the global attention of diffusion transformers (DiTs) [62], but this often comes at inferior video quality and prohibitive computation for both fine-tuning and inference. On the other hand, autoregressive (AR) VDMs [14, 40] improved the efficiency for long video generation via sequential synthesis, yet they suffer from limited camera precision [71] and error accumulation [40]. Moreover, both long video generation and AR models lack mature open-source communities compared to regular VDMs like HunyuanVideo [46], CogVideoX [98], and Wan [77].

To overcome these challenges, we present **WorldStereo**, a novel framework that bridges the gap between camera-guided VDMs and 3D scene reconstruction by enabling *consistent multi-trajectory video generations with geometry-aware memories*, as summarized in Table 1. Specifically, two complementary mechanisms, *global-geometric* and *spatial-stereo* memory, are incorporated into off-the-shelf VDMs to memorize coarse structures and fine-grained details, respectively. These two components enable precise and coherent video synthesis across diverse trajectories. WorldStereo subtly sidesteps the long sequence generation and largely preserves the generalization and usability of pre-trained VDMs, resulting in impressive 3D reconstruction as shown in Figure 1.

Formally, WorldStereo is built upon the camera-guided VDM, Uni3C [13], with the extended capability to memorize consistent generation along different trajectories. We first augment Uni3C's point cloud guidance as an incrementally updated Global-Geometric Memory (GGM) through iterative feed-forward 3D reconstruction. However, we find that GGM fails to preserve fine-grained details. Inspired by stereo matching [60], we introduce the Spatial-Stereo Memory (SSM): learning spatial coherence between the generated novel views and retrieved views from a deliberate memory bank by establishing explicit 3D correspondences. Furthermore, we constrain the attention receptive fields of each novel view to focus on the specific retrieved one, thereby enhancing detailed consistency. Additionally, WorldStereo inherits the flexibility of Uni3C: all pixel-wise aligned conditions are injected via the ControlNet branch, which is highly compatible with the distribution matching distillation (DMD) [100, 101] and eliminates the need for joint training. This technique significantly accelerates WorldStereo under 4-step inference without compromising generalization and consistency.

Extensive experiments confirm the effectiveness of WorldStereo, including both in-domain and out-of-distribution benchmarks. Our approach achieves superior camera-motion accuracy and higher-quality video generation. To further demonstrate its contributions to 3D reconstruction, we present a new 3D reconstruction benchmark to evaluate the output quality of camera-guided VDMs. We meticulously process and crop the ground-truth point clouds for Tanks-and-Temples [45] and MipNeRF360 [4], enabling comprehensive assessment of 3D consistency, visual quality, and camera trajectory fidelity. Our main contributions are summarized as follows:

- **Geometry-aware memory enhanced VDM.** WorldStereo contains two complementary memory mechanisms to generate multi-trajectory consistent videos tailored for 3D reconstruction.
- **Flexible framework with efficient inference.** Leveraging pixel-wise aligned ControlNet injection, WorldStereo can be accelerated via DMD without requiring joint training or sacrificing generalization.

- **Customized evaluation.** We present a new 3D reconstruction benchmark to evaluate the outcome quality of camera-guided VDMs.

## 2. Related Work

**Camera-Guided Video Generation.** Taming high fidelity video generations [46, 55, 77, 94, 98, 116] under controllable viewpoints and camera trajectories serves as a pivotal pathway for world simulation. Some works tried to implicitly control the VDM under specific camera motion via tailored LoRA tuning [6, 30, 73, 112]. Meanwhile, implicitly discrete controls are also widely adopted in action-based VDMs [48, 113] and AR models [20, 29, 33]. MotionCtrl [87] first explicitly injected camera poses into the pre-trained VDM. Subsequent works further extend the camera presentation into Plücker ray [3, 31, 32, 50, 96]. To enhance the camera control precision under metric scales, point clouds [13, 22, 50, 59, 63, 65, 105], mesh [37, 97], optical flow [9, 43], and tracking points [28, 78] have been employed for camera-guided VDMs, spanning both training-based and training-free manners [35, 36, 106]. Despite the precise camera controllability achieved by these pioneering efforts, we note that they struggle to produce convincing 3D reconstruction due to constrained video lengths and inadequate consistency caused by memoryless visual conflicts.

**Memory-based Video Generation.** Existing video generation fails to maintain geometrically accurate structures during long-trajectory synthesis. A straightforward solution involves training VDMs with extended context lengths [16, 72, 76] to memorize more visual clues, yet computationally costly. To reduce the of long contexts computation, concurrent works either compress prior frames into a condensed context window [27, 109] or inject historical frames via attention mechanisms [67, 93, 117], which inevitably cause information loss and compromise 3D consistency. Another research direction iteratively reconstructs 3D representations from historical frames, utilizing these as either conditional guidance [28, 59, 65, 90, 105] or retrieved references [49, 104] for future contexts synthesis. But they also suffer from error accumulation and degraded fine-grained details. Our geometry-aware memory preserves both 3D consistency and high-fidelity details by unifying 3D correspondence modeling and a tailored attention mechanism.

**Feed-Forward 3D Reconstruction.** In contrast to traditional Structure from Motion (SfM) [24, 61, 68], Multi-View Stereo (MVS) [11, 26, 99, 107], and learning-based SLAM [38, 52], recent feed-forward 3D reconstruction approaches efficiently predict both camera poses, point clouds, and depth maps within a single forward pass. Dust3R [84] pioneered learning two-view correspondence through powerful transformer models pre-trained by cross-view inpainting [88]. Many follow-ups further investigated this realm, including multi-view-based [95, 95], sequential-based [19, 82], dynamic-based [58, 108] and conditional priors-based [42, 44, 57] reconstructions. For impressive visual reconstruction, some works have focused on feed-forward 3D Gaussian Splatting (3DGS) generation [18], enabling fast, optimization-free 3DGS representations. However, 3D reconstruction methods follow "what you see is what you get", *i.e.*, they require sufficient images to capture 3D structures, suffering from inferior results explored beyond observed viewpoints.

**3D Scene Generation.** One popular scene generation workflow involves iterative *warp-and-inpaint*, combining depth estimation, alignment, and image inpainting [23, 34, 54, 70, 79, 103]. While iterative scene generations perform well in novel view synthesis (NVS), they suffer from prohibitive per-scene optimization and inconsistent geometry across views. Another prominent paradigm can be summarized as *generate first, then reconstruct*, yielding various 3D outcomes from multi-view images [12, 25, 117], panoramas [41] and camera-guided videos [53, 73, 105, 114]. However, both of them face distinct limitations: multi-view images cover adequate viewpoints but lack view consistency, while generated videos are insufficiently long to capture disparate viewpoints. Recent works have unified generation and reconstruction in end-to-end frameworks, simultaneously modeling depth [17, 39, 119], 3DGS [2, 51, 96], and point map [74, 111] rather than relying solely on RGB frames. Though these methods largely alleviate 3D consistency issues, they are highly data-hungry and require substantial training, potentially hindering the generalization of foundation models. Our method preserves the original output formulation of VDMs with retained generalization, while capturing more consistent viewpoints via a novel memory mechanism for 3D reconstruction.

## 3. Method

**Overview.** Given a single image, WorldStereo aims to reconstruct a complete 3D scene following the *"generate first, then reconstruct"* paradigm. We summarize the generation pipeline in Figure 2. We first introduce the camera-guided VDM with point cloud guidance and our basic memory storage in Section 3.1. Next, the global-geometric (Section 3.2) and spatial-stereo memories (Section 3.3) are proposed to generate multiple video sequences with consistent 3D geometry and textures for subsequent 3D reconstruction. To improve inference efficiency, WorldStereo adopts a modified distribution matching distillation (Section 3.4).

### 3.1. Preliminaries

**Uni3C** [13] is used as the base camera-guided VDM in WorldStereo. Formally, Uni3C is built upon the pre-trained Wan-I2V [77], integrated with a lightweight ControlNet branch [110] trained from scratch. Except for the camera Plücker rays, Uni3C incorporates point clouds extracted

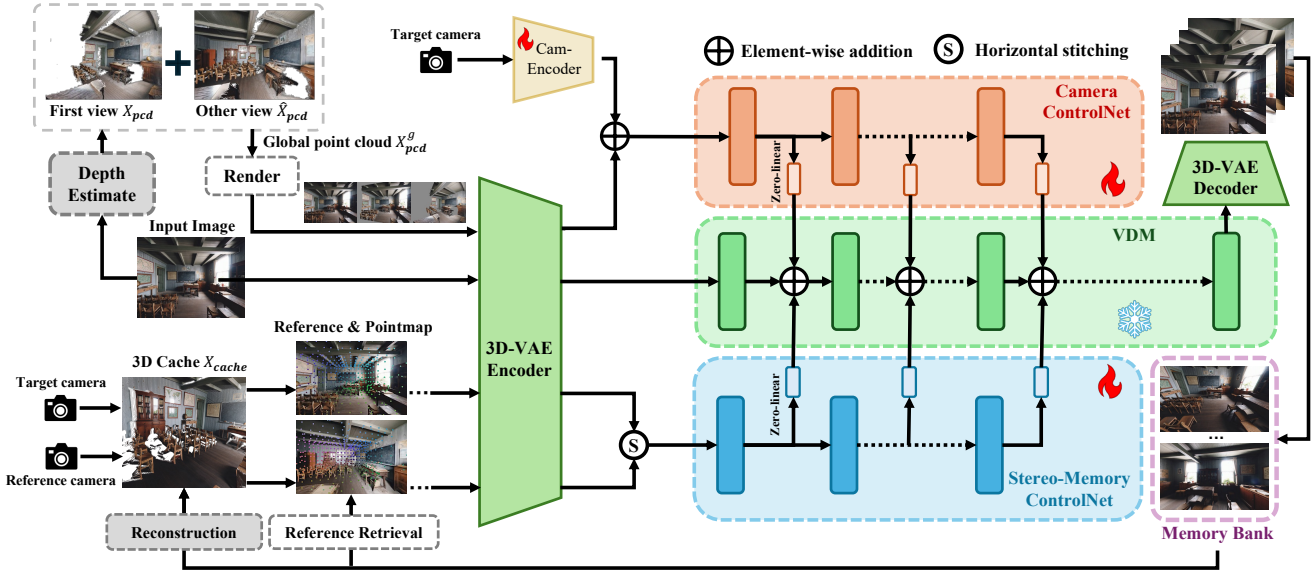



Figure 2. **Overview of WorldStereo.** WorldStereo comprises two ControlNet branches. The camera branch ensures precise camera control and Global-Geometric Memory (GGM), depending on global point clouds; the Spatial-Stereo Memory (SSM) branch leverages retrieved reference frames and pointmap (3D correspondence) guidance obtained from the 3D cache to further preserve fine-grained consistency. We omit the diffusion noise part for simplicity.

from the reference frame as 3D geometric priors, derived via back-projected monocular depth. Point clouds $X_{pcd}$ of the reference image can be presented as:

$$X_{pcd}(x) \simeq R_{c\to w} D(x) K^{-1} \hat{x}, \tag{1}$$

where $R_{c\to w}$ denotes the camera-to-world pose matrix; $D(\cdot)$ is the depth at pixel $x$ estimated by MoGe [83]; $K^{-1}$ is the inverse of the camera intrinsic matrix; $\hat{x}$ is the homogeneous pixel coordinate. Point clouds serve as strong geometric guidance to facilitate fast convergence and precise camera control, avoiding compromising the generalization of frozen foundational VDMs.

**Memory Bank & 3D Cache.** WorldStereo incorporates two memory components: a 2D memory bank and a 3D cache, as in Figure 2. Formally, generated video frames are first temporally downsampled and stored in the memory bank as $\{I_{mem}\}_{m=0}^{M}$, serving to retrieve spatially similar reference views for the subsequent spatial-stereo memory. Note that the initial condition image and perspective views split from the 360° panorama are also included in the memory bank. The 3D cache saves global point cloud set $X_{cache}$, reconstructed based on memory bank images using the feed-forward 3D reconstruction method: WorldMirror [57]. Specifically, we incrementally reconstruct point clouds for each generated video. To tackle long sequences, disparate 3D caches are merged by aligning point clouds from overlapping views via Umeyama transformation [75].

### 3.2. Global-Geometric Memory

We propose Global-Geometric Memory (GGM) that iteratively updates point cloud conditions, serving as global 3D priors for generating multiple consistent videos, iterative video continuation, and even supporting panorama-based 3D generation, as verified in Section 4.4. As discussed in [13], point clouds just provide camera guidance rather than forcing the VDM to overfit to these 3D presentations. While this phenomenon is reasonable for camera-guided VDMs to preserve overall generalization from the degradation caused by inferior monocular depth, it causes our model to ignore most geometric structures brought by point clouds, even when the point clouds themselves are perfectly reconstructed.

To overcome this, we fine-tune the original control branch of WorldStereo using extended global point clouds $X^g_{pcd}$ beyond the reference points ($X_{pcd}$ from Eq. 1) as:

$$X^g_{pcd} = [X_{pcd}, \hat{X}_{pcd}], \tag{2}$$

where $\hat{X}_{pcd}$ denotes point clouds from other views. To avoid overfitting novel views' point clouds during training, we introduce a point cloud masking strategy: randomly dropping a subset of points from target views to confirm robustness to partially geometric missing. We detail the sampling strategy in supplementary. For inference, we use the incrementally updated 3D cache as $\hat{X}_{pcd}$, which is aligned to the same coordinate as $X_{pcd}$ via the Umeyama transformation [75] among the overlapping point clouds from monocular depth estimation and WorldMirror. Notably, GGM is also compatible with panorama-based 3D generation. Since panoramas capture 360° views, VDMs have to maintain visually consistent content during viewpoint transitions. To this end, we follow the MoGe panorama depth

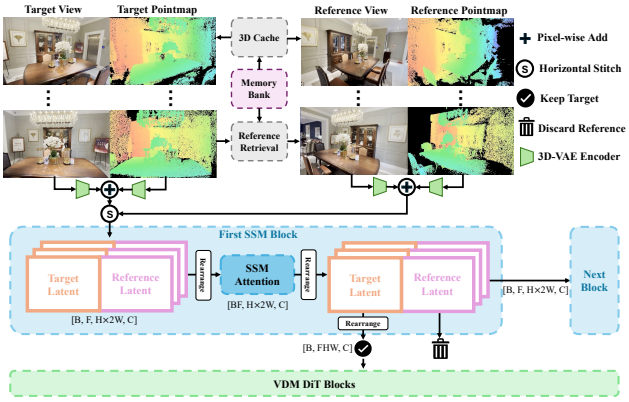



Figure 3. **Spatial-Stereo Memory (SSM).** Reference views are retrieved from the memory bank, while pointmaps for both target and reference views are constructed based on the 3D cache. In SSM attention, we horizontally stitch each target-reference pair and rearrange the tensor shape to make each target frame's features focus on the specifically retrieved reference. B, F, H, W, C indicate dimensions of batch, frame, height, width, and channels.

estimation [83] to construct panoramic point clouds as the initialized 3D cache for our panoramic 3D generation.

### 3.3. Spatial-Stereo Memory

Although GGM maintains coarse structures using point clouds from 3D cache, it struggles to preserve fine-grained details, as shown in Figure 5. Many previous studies [49, 67, 104, 117] retrieve historical reference frames and jointly model all frames via full-attention. However, this formulation requires substantial post-training to enable VDMs to adapt to long sequences. Moreover, we cannot guarantee the continuity of retrieved frames (*e.g.*, panoramic scenarios). These disparate, unordered reference views further hinder the VDM learning process. Thus, we get inspirations from the traditional stereo matching [60] and the reference-based inpainting [10] and propose the Spatial-Stereo Memory (SSM) as illustrated in Figure 3. Formally, we discretely retrieve reference views and spatially stitch each with its corresponding target view. We then constrain the attention receptive field to each reference-target pair, facilitating enhanced fine-grained detail recovery.

In practice, given $N$ target poses, we first sample $F = N/4$ poses uniformly and retrieve their nearest-neighbor frames from the memory bank as references. We extend the retrieval strategy of [104] from 2D planes to 3D spaces, selecting views where the volumetric overlapping fields-of-view (FoV) between the target and reference camera frustums are maximized. To ensure that fine-grained details are preserved after 3D-VAE encoding, we separately encode each retrieved reference view as an independent image, yielding latent features $\{z_{ref}\}_{i=1}^{F}$. Then we horizontally stitch target and reference latent features as $z_{stitch} = [z_{tar}; z_{ref}] \in \mathbb{R}^{F\times 2HW\times C}$, where $H, W, C$ denote the height, width, and channel dimension of latent features. To improve the geometry-aware perception, we incorporate a *pointmap* to each stitched target-reference pair $z_{stitch}$, which indicates 3D corresponding information derived from our 3D cache. Specifically, the pointmap records point cloud positions of the target-reference pair in the 3D world coordinate. We normalize and colorize this into an RGB format, which is then encoded into latent features by 3D-VAE—denoted as $\hat{z}_{tar}$ and $\hat{z}_{ref}$ for target and reference pointmaps, respectively. The stitched pointmap latents are thus formulated as $\hat{z}_{pm} = [\hat{z}_{tar}; \hat{z}_{ref}] \in \mathbb{R}^{F\times 2HW\times C}$. Subsequently, we achieve the final inputs for the SSM branch as $z_{ssm} = z_{stitch} + \hat{z}_{pm}$. Our ablation studies in Figure 5(c)(d) confirm that this 3D correspondence information is critical to SSM's performance. The model architecture of the SSM branch is similar to the camera branch, comprising 20-layer DiT blocks trained from scratch. A key distinction lies in the SSM attention constraint: instead of full attention, we limit the attention receptive field such that each target-reference pair attends exclusively to its own features—facilitating more precise fine-grained learning. As shown in Figure 3, we rearrange the feature to $[BF, H*2W, C]$ and restrict SSM attention to operate solely along the $H*2W$ dimension. Then, only the target features will be added to the main VDM block.

**Data Curation.** Intuitively, training SSM needs multi-view videos to construct reference-target pairs, which is difficult to achieve in the real world. To overcome this, we generate training pairs by temporally misaligned sampling existing multi-view data [1, 56, 85, 91], ensuring that the reference video and target video have a temporal overlap of 30% to 90%. Then we randomly shuffle and apply masks to reference views to confirm the robustness, as well as simulate the disorder and discreteness of real-world retrieval scenarios. More details are discussed in the supplementary.

### 3.4. Acceleration via DMD

We apply the modified Distribution Matching Distillation (DMD) [100] to accelerate the inference of WorldStereo. DMD extends the idea of Variational Score Distillation (VSD) [86], distilling a few-step diffusion student $G_\theta$ through the approximate Kullback-Liebler (KL) divergence built from the difference between the frozen real score function $s_{real}$ and the trainable fake score function $s_{fake}$. The update gradient of DMD can be written as:

$$\nabla \mathcal{L}_{\text{DMD}} = -\mathbb{E}_t \left( \int \left( s_{\text{real}}(x_t, t) - s_{\text{fake}}(x_t, t) \right) \frac{dx_t}{d\theta} dz \right), \quad (3)$$

where $x = G_\theta(z)$ denotes the student generation given random Gaussian noise $z \sim \mathcal{N}(0; \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$, while $x_t \sim q_t(x_t|x, t)$ indicates the forward diffusion process.

The generator $G_\theta$ of WorldStereo is distilled into a 4-step DiT. $G_\theta$, $s_{real}$, $s_{fake}$ are all initialized from the camera-guided VDM (Uni3C): $s_{real}$ is frozen, while $G_\theta$ and $s_{fake}$ are trainable. Following [100], we train $s_{fake}$ 5 times per generator update. The stochastic gradient truncation [40] is employed to stabilize the training phase. We omit the GAN loss, as we found its impact to be insignificant while substantially slowing down training. Moreover, the DMD training is based on pure camera-guided video generation without any memory training. To decouple the control and the basic few-step generation capabilities, we freeze the camera-guided control branch of the generator $G_\theta$, leaving only its main backbone trainable. Notably, both the camera-guided and memory-based control branches can be generalized to the distilled generator $G_\theta$ without any joint fine-tuning. This largely simplifies the DMD training process while preserving the generalization (annotated memory data requires well-aligned depth maps, which is less than the camera control data). Importantly, all parameters of the fake score function remain trainable, facilitating $s_{fake}$ to track the generator's distribution. Additionally, we empirically find that retaining high-quality and relatively easy trajectories is critical for the stable training of DMD. The student model prefers to attend to some artifacts (*e.g.*, oversaturation and hallucination) from the teacher model, especially under challenging trajectories or low-quality reference images. As verified in Table 2, this data filter strategy does not degrade the camera controllability.

# 4. Experiments

## 4.1. Implementation Details

The frozen VDM of WorldStereo is built upon Wan2.1-14B-I2V [77], with the camera ControlNet retrained under the same setting as [13] for 8,000 steps under a batch size of 32. For the GGM training in the second phase, we fine-tune the camera ControlNet with the global point cloud augmentation for 4,000 steps. For the third phase of SSM training, we train a new branch from scratch for 6,000 steps based on tailored memory-retrieval data as detailed in Section 3.3. Training both memory mechanisms costs 60 hours on 64 NVIDIA H20 GPUs. For DMD training, we set the classifier-free guidance (CFG) scale to 5.0 for the real score function, leading to a CFG-free generator with 2x inference efficiency. Furthermore, we reduce the inference denoising steps from 40 to 4, achieving an overall speedup of 20×. We train DMD for 1,000 steps within 13 hours under the same resources. All training data are resized to 480p with flexible aspect ratios. We also verify that our method can be generalized to 720p inference in supplementary.

**Datasets.** The training data of our camera control includes DL3DV [56], Real10k [118], Tartanair [85], Map-Free-Reloc [1], WildRGBD [91], and UE5 rendering data. For memory training, we discard Real10k due to its low-quality images and overly simple camera trajectories. We further eliminate the Tartanair and narrow the frame interval of DL3DV for stable DMD training. More details are discussed in the supplementary.

## 4.2. Results of Camera Control and Visual Quality

**Settings and Metrics.** To fairly evaluate the camera controllability of camera-guided models trained on different datasets, we introduce a new out-of-distribution (OOD) benchmark as shown in Table 2. We select 100 images from the static subset of WorldScore [21] as the initial frames of this benchmark, including high-quality samples across real-world, stylized, indoor, and outdoor scenarios. Next, we randomly combine translation, rotation, and panning to construct complex camera trajectories. To verify the camera precision, we employ WorldMirror [57] on the generated videos to extract predicted cameras and compare Rotation Error (RotErr), Translation Error (TransErr), and Absolute Trajectory Error (ATE) against ground-truth camera guidance. We further evaluate various quality assessments (Q-Align-Image&Video [89], CLIP-Image&Text [64], CLIP-IQA+ [80], Laion-Aes [69]) via IQA-PyTorch [15] to verify the overall image and video quality.

**Analysis.** From Table 2, the basic version WorldStereo* outperforms other competitors with superior camera control and visual quality. Moreover, we provide ablation studies on our memory components (GGM, SSM), where the memory bank and 3D cache only store the first frame's information. While the memory mechanism provides no benefit in single-image conditioned camera control, these results confirm that our memory-based training preserves the model's generalization and visual quality. In particular, GGM even improves the overall quality of the generated videos, while SSM slightly degrades performance, but enables strong fine-grained details recovery, as verified in Figure 5(d). We also include the result of DMD, replacing the VDM backbone of our full model with the distilled DiT, showcasing impressive camera control and consistent quality, and a rapidly reduced inference cost.

## 4.3. Single-View Reconstruction Benchmark

**Settings and Metrics.** To comprehensively evaluate the quality of video generation and multi-trajectory consistency for reconstruction, we propose a novel 3D reconstruction benchmark based on single-view generation. Our benchmark comprises the training split of Tanks-and-Temples [45] and MipNeRF360 [4]. The training split of Tanks-and-Temples has ground-truth point clouds. For

*All methods are evaluated under their pre-defined resolutions and frame numbers, while the settings of Uni3C and WorldStereo series are the same for fairness (512p and 81-frame).

Table 2. **Quantitative results of OOD benchmark with WorldScore [21] images**. * indicates the baseline version of our method without any memory mechanism, while the 'full' version denotes adding both GGM and SSM. The 'DMD' version is based on 'WorldStereo-full'. Inference times are all tested with 8 H20 GPUs. *

| | RotErr↓ | TransErr↓ | ATE↓ | Q-Align-I↑ | Q-Align-V↑ | CLIP-I↑ | CLIP-T↑ | CLIP-IQA+↑ | Laion-Aes↑ | Time (sec) |
|---|---|---|---|---|---|---|---|---|---|---|
| Voyager [39] | 0.678 | 0.630 | 1.343 | 3.279 | 0.664 | 82.63 | **11.399** | 0.414 | 4.994 | 343 |
| SEVA [117] | 0.171 | 0.540 | 1.023 | 3.907 | 0.782 | 91.74 | 11.194 | 0.514 | 5.331 | 90 |
| Gen3C [65] | 0.220 | 0.275 | 1.071 | 4.094 | 0.820 | 91.78 | 11.270 | 0.518 | 5.332 | 158 |
| Uni3C [13] | 0.155 | 0.192 | 0.572 | 4.202 | 0.846 | 91.63 | 11.115 | 0.549 | 5.350 | 162 |
| WorldStereo* | 0.132 | 0.178 | 0.542 | 4.273 | 0.860 | 91.94 | 11.145 | 0.559 | 5.352 | 162 |
| WorldStereo-GGM | **0.129** | **0.162** | 0.706 | **4.339** | **0.875** | **92.24** | 11.189 | 0.572 | **5.408** | 162 |
| WorldStereo-Full | 0.145 | 0.253 | 0.667 | 4.287 | 0.866 | 91.80 | 11.169 | 0.561 | 5.369 | 173 |
| WorldStereo-DMD | 0.146 | 0.203 | **0.504** | 4.338 | 0.874 | 91.15 | 11.183 | **0.573** | 5.393 | 9 |

Table 3. **Quantitative results of 3D reconstruction** based on Tanks-and-Temples [45] and MipNeRF360 [4]. * indicates the baseline version of our method without any memory mechanism.

| | Method | F1-Score ↑ | AUC ↑ | RotErr ↓ | TransErr ↓ | ATE ↓ |
|---|---|---|---|---|---|---|
| Tanks&Temples | Uni3C [13] | 0.424 | 0.378 | 0.362 | 0.1017 | 0.1572 |
| | Gen3C [65] | 0.416 | 0.380 | 0.342 | 0.0949 | 0.1704 |
| | SEVA [117] | 0.286 | 0.293 | 0.379 | 0.0949 | 0.1815 |
| | Lyra [2] | 0.227 | 0.193 | – | – | – |
| | VMem [49] | 0.386 | 0.375 | 0.533 | 0.1510 | 0.1922 |
| | WorldStereo* | 0.447 | 0.389 | 0.377 | 0.0990 | 0.1545 |
| | WorldStereo-GGM | 0.485 | 0.411 | **0.224** | **0.0885** | **0.1350** |
| | WorldStereo-Full | **0.578** | **0.437** | 0.247 | 0.0927 | 0.1501 |
| | WorldStereo-DMD | 0.534 | 0.410 | 0.291 | 0.1001 | 0.1547 |
| MipNeRF360 | Uni3C [13] | 0.352 | 0.347 | 0.112 | 0.0086 | 0.0104 |
| | Gen3C [65] | 0.356 | 0.340 | 0.349 | 0.0220 | 0.0318 |
| | SEVA [117] | 0.332 | 0.311 | 0.282 | 0.0138 | 0.0295 |
| | Lyra [2] | 0.203 | 0.263 | – | – | – |
| | VMem [49] | 0.256 | 0.245 | 0.403 | 0.0392 | 0.0752 |
| | WorldStereo* | 0.350 | 0.342 | **0.097** | **0.0076** | **0.0099** |
| | WorldStereo-GGM | 0.342 | 0.346 | 0.107 | 0.0079 | 0.0206 |
| | WorldStereo-Full | **0.406** | **0.402** | 0.114 | 0.0080 | 0.0132 |
| | WorldStereo-DMD | 0.390 | 0.387 | 0.159 | 0.0106 | 0.0267 |

MipNeRF360, we first reconstruct the global point clouds scene by MVS [11] and then crop the centric foreground areas as pseudo ground-truth point clouds. Each scene of both datasets only provides a single image as the initial frame. More details about the data curation of this benchmark are discussed in the supplementary. The evaluation workflow includes: 1) generating videos along 4 pre-defined trajectories of up, left, right rotations, and orbit as shown in Figure 4(a); 2) reconstructing point clouds through WorldMirror [57]; 3) aligning the reconstructed point clouds to the ground-truth ones. For MipNeRF360, we can leverage the initial frame's MVS depth as an anchor to align point-to-point matched WorldMirror points via Umeyama translation [75]. For Tanks-and-Temples, we first align the rotation and translation via the first frame's camera. Then, we apply ICP [5] to optimize a refined scale between the reconstructed and ground-truth point clouds. Except for camera metrics, we include two widely used point cloud metrics. Specifically, we get the precision and recall of point cloud within scene-wise distance thresholds manually adjusted as [45], which evaluate the accuracy and completeness of reconstructed point clouds. Besides, we incorporate the point cloud AUC, calculated as the area under the ROC curve of precision and recall with varying thresholds.

**Analysis.** From Table 3, WorldStereo* without any memory mechanisms already outperforms other methods, while our full model, enhanced with GGM and SSM, achieves substantial improvements on both reconstruction and camera precision. Moreover, WorldStereo-DMD still performs well in 3D reconstruction, showing impressive consistency with significant acceleration.

### 4.4. Extensive Qualitative Results

**Qualitative Comparison.** We show qualitative results in Figure 4. As shown in Figure 4(a), our 3D reconstruction benchmark features high-quality ground-truth point clouds that focus on foreground objects. The compared methods must retain consistent outcomes with symmetrical and logically coherent structures to achieve good point cloud scores. Thus, this benchmark serves as a reliable evaluation for 3D scene generation. SEVA [117] suffers from distorted structures and blurry backgrounds, while Gen3C [65] faces challenges in producing consistent videos under different trajectories—both methods yield ambiguous, incomplete reconstructions. In contrast, point clouds reconstructed from WorldStereo's outputs enjoy remarkable completeness and precision, while its novel views remain consistent across different trajectories and exhibit superior quality.

**Memory Mechanisms.** We show the qualitative ablation studies in Figure 5 to verify the individual effectiveness of each memory component. The baseline without any memory mechanism randomly hallucinates new objects in novel views, while GGM largely retains the structural consistency and improves camera control for disparate viewpoint changes (second row), benefiting from the incrementally updated 3D cache. The ability of SSM to preserve fine-grained consistency with the memory bank reference stems from its attention mechanism, which leverages images with 3D guidance instead of drawing information

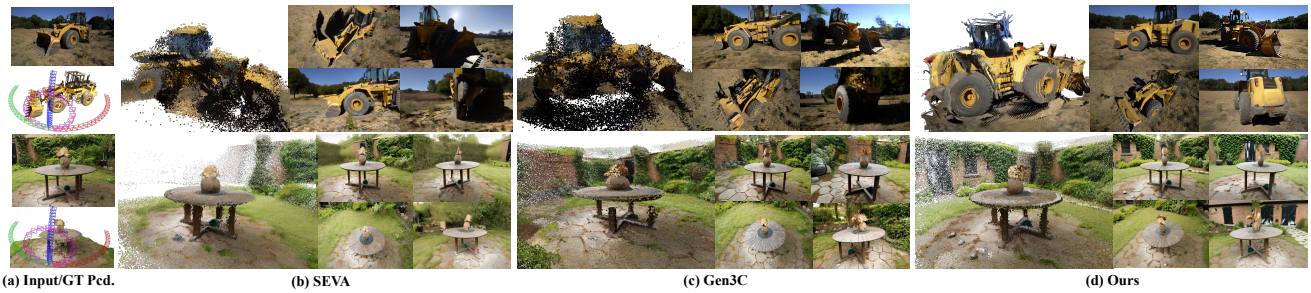



Figure 4. **Results of 3D reconstruction benchmark.** The column (a) shows input views and ground-truth point clouds with pre-defined four trajectories (up, left, right rotations, and orbit). We compare the qualitative results of reconstructed point clouds (left) and generated novel views (right) for each method.

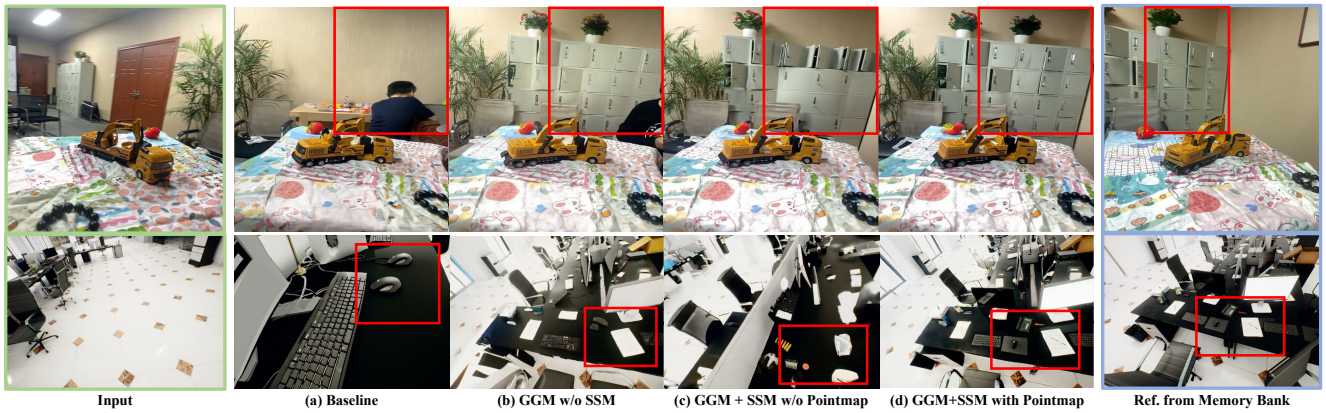



Figure 5. **Ablation studies of memory components.** Please see the red-framed regions to check the consistency compared to retrieved references. Baseline results are generated without any memory. GGM can capture coarse structures, but loses fine-grained details. Moreover, the incorporation of pointmap significantly enhances the consistency gained via the reference frames retrieved from the memory bank.

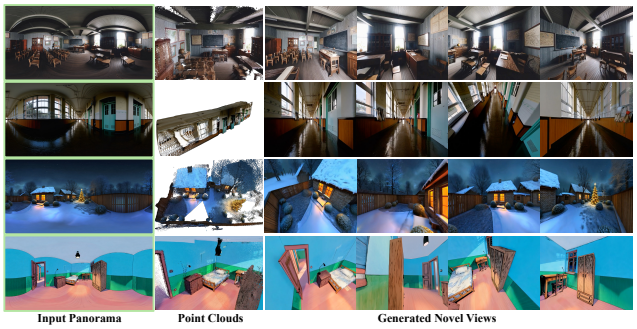



Figure 6. **Results of 3D panorama generation.** Please zoom-in for more details of reconstructed point clouds and novel views.

solely from coarse-grained point clouds. We clarify that the 3D correspondence-based pointmap guidance is critical for the SSM to focus on correct matching regions.

**3D Panorama Generation.** Thanks to memory capabilities, WorldStereo can be easily extended to 3D panorama generation as shown in Figure 6. Different from panoramic video generation [92, 97, 102], our method largely preserves the generalization of the foundational VDM while producing high-resolution perspective views (all panorama inferences are performed at 576p). Both aforementioned points are critical for 3D reconstruction. Formally, we split the panorama into 27 frames under FoV $90 \times 120$ as the initial memory bank, and leverage the panoramic depth estimation of MoGe [83] to build the 3D cache. We heuristically design wonder trajectories as used in our reconstruction benchmark (Section 4.3) and generate videos for the middle three non-overlap images via WorldStereo.

## 5. Conclusion

In this paper, we present WorldStereo, a novel camera-guided video generation framework tailored for 3D reconstruction. To address the challenge of capturing long-sequence viewpoints while maintaining 3D consistency, WorldStereo generates multiple consistent videos via two memory mechanisms: Global-Geometry Memory (GGM) and Spatial-Stereo Memory (SSM). GGM incrementally updates a point cloud-based 3D cache to enhance the coherent 3D structures, while SSM learns coherence between generated and retrieved views through 3D correspondence to preserve fine-grained details. Moreover, we modify the distribution matching distillation (DMD) strategy to accelerate WorldStereo for fast inference with negligible performance drop. Additionally, we develop a new 3D reconstruction benchmark for evaluating video generation performance. WorldStereo demonstrates strong performance and generalizes effectively to diverse tasks, such as object-centric, face-forward, and 3D panorama generation.

## Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2023YFF0905104) and also by National Natural Science Foundation of China under Grants (62376244).

## References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 5, 6

[2] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025. 3, 7

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2, 6, 7

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 7

[6] 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. 3

[7] 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. 2

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[9] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025. 2, 3

[10] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7705–7715, 2024. 5

[11] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations*, 2024. 3, 7

[12] Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6045–6056, 2025. 3

[13] 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. 2025. 2, 3, 4, 6, 7

[14] 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. 2

[15] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 6

[16] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3

[17] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025. 3

[18] 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 *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 3

[19] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2025. 3

[20] Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2024. 3

[21] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025. 6, 7

[22] Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. In *International Conference on Learning Representations*, 2025. 2, 3

[23] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36:39897–39914, 2023. 3

[24] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 3

[25] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3d: Create anything in 3d with multi-view diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[26] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 3

[27] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 3

[28] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2, 3

[29] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025. 3

[30] 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. *International Conference on Learning Representations*, 2024. 3

[31] 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 *International Conference on Learning Representations*, 2025. 2, 3

[32] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025. 2, 3

[33] 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. 3

[34] 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*, pages 7909–7920, 2023. 3

[35] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. In *International Conference on Learning Representations*, 2025. 3

[36] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 3

[37] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025. 3

[38] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 3

[39] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 3, 7

[40] 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. 2, 6

[41] Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025. 3

[42] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1071–1081, 2025. 3

[43] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025. 2, 3

[44] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 3

[45] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2, 6, 7

[46] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[47] Kuaishou. Kling. *https://klingai.kuaishou.com*, 2024. 2

[48] Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin

Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025. 3

[49] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025. 3, 5, 7

[50] Teng Li, Guangcong Zheng, Rui Jiang, Tao Wu, Yehao Lu, Yining Lin, Xi Li, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 2, 3

[51] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025. 3

[52] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 3

[53] 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. *arXiv preprint arXiv:2412.12091*, 2024. 2, 3

[54] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6517–6526, 2024. 3

[55] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 3

[56] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5, 6

[57] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025. 1, 3, 4, 6, 7

[58] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22820–22830, 2025. 3

[59] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2016–2029, 2025. 2, 3

[60] David Marr and Tomaso Poggio. Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. *Science*, 194(4262):283–287, 1976. 2, 5

[61] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[62] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[63] Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. *arXiv preprint arXiv:2501.06006*, 2025. 2, 3

[64] 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*, pages 8748–8763. PmLR, 2021. 6

[65] 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*, 2025. 2, 3, 7

[66] RunwayML. Gen-3 alpha. *https://runwayml.com/research/introducing-gen-3-alpha*, 2024. 2

[67] Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025. 3, 5

[68] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, 2016. 3

[69] Christoph Schuhmann. improved-aesthetic-predictor: Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 6

[70] 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. 3

[71] Chonghyuk Song, Michal Stary, Boyuan Chen, George Kopanas, and Vincent Sitzmann. Generative view stitching. *arXiv preprint arXiv:2510.24718*, 2025. 2

[72] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025. 3

[73] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[74] Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24846–24857, 2025. 3

[75] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 4, 7

[76] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 3

[77] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 6

[78] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025. 2, 3

[79] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multiview consistent images for single-view scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26772–26782, 2025. 3

[80] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 6

[81] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 2

[82] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 2, 3

[83] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 1, 4, 5, 8

[84] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[85] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 5, 6

[86] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023. 5

[87] 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. 3

[88] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3

[89] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6

[90] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025. 3

[91] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 5, 6

[92] Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. Panowan: Lifting diffusion video generation models to 360 {\deg} with latitude/longitude-aware mechanisms. *Advances in Neural Information Processing Systems*, 2025. 8

[93] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025. 3

[94] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 3

[95] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2, 3

[96] 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. 3

[97] Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang

Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025. 3, 8

[98] 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 *International Conference on Learning Representations*, 2025. 2, 3

[99] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 3

[100] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 2, 5, 6

[101] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 2

[102] Yuyang Yin, HaoXiang Guo, Fangfu Liu, Mengyu Wang, Hanwen Liang, Eric Li, Yikai Wang, Xiaojie Jin, Yao Zhao, and Yunchao Wei. Panoworld-x: Generating explorable panoramic worlds via sphere-aware video diffusion. *arXiv preprint arXiv:2509.24997*, 2025. 8

[103] 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*, pages 5916–5926, 2025. 3

[104] 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. *arXiv preprint arXiv:2506.03141*, 2025. 3, 5

[105] 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. 2, 3

[106] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2050–2062, 2025. 3

[107] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214, 2023. 3

[108] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3

[109] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 3

[110] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[111] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21685–21695, 2025. 3

[112] Yisu Zhang, Chenjie Cao, Chaohui Yu, and Jianke Zhu. Lion-lora: Rethinking lora fusion to unify controllable spatial and temporal generation for video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14569–14579, 2025. 3

[113] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025. 3

[114] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024. 3

[115] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. In *International Conference on Learning Representations*, 2025. 2

[116] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3

[117] Jensen Jinghao 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. *arXiv e-prints*, pages arXiv–2503, 2025. 3, 5, 7

[118] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), 2018. 6

[119] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8535–8546, 2025. 3