This CVPR Findings 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.

# Stepper: Stepwise Immersive Scene Generation with Multiview Panoramas

Felix Wimbauer[1,3,4,†] Fabian Manhardt[1] Michael Oechsle[1] Nikolai Kalischek[1]
Christian Rupprecht[2] Daniel Cremers[3,4] Federico Tombari[1,4]
[1]Google [2]University of Oxford [3]MCML [4]Technical University of Munich
felix.wimbauer@tum.de fabianmanhardt@google.com

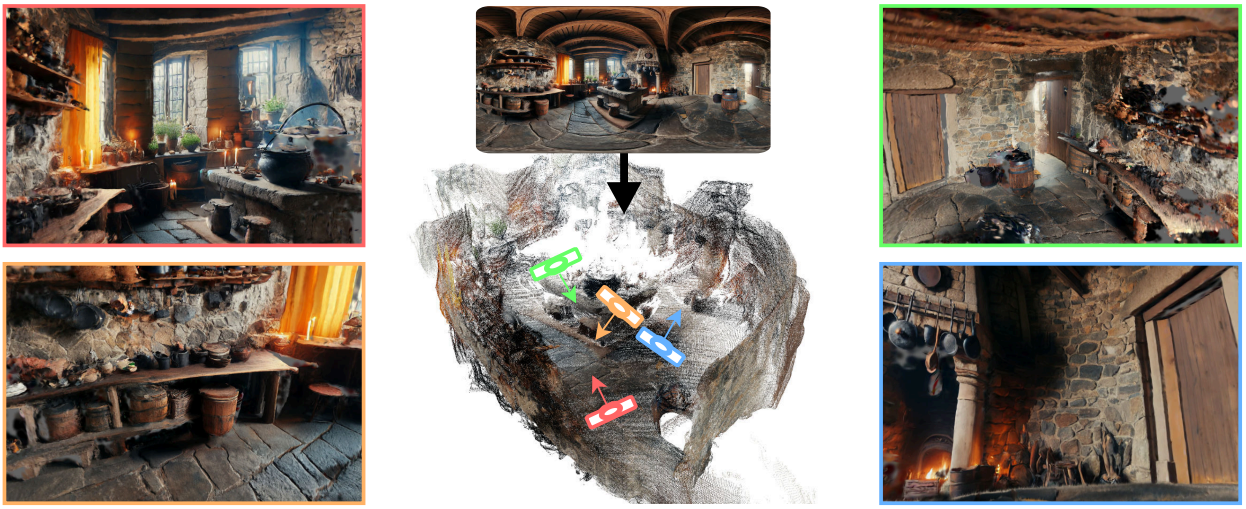

Figure 1. **Stepper** sets a new state-of-the-art quality level of generated explorable 3D scenes. Its core innovation is a novel cubemap-based multi-view panorama diffusion model that ensures high-resolution scene synthesis while facilitating step-wise, coherent scene expansion and high-quality scene reconstruction. Please check out our project page at: fwmb.github.io/stepper

## Abstract

*The synthesis of immersive 3D scenes from text is rapidly maturing, driven by novel video generative models and feed-forward 3D reconstruction, with vast potential in AR/VR and world modeling. While panoramic images have proven effective for scene initialization, existing approaches suffer from a trade-off between visual fidelity and explorability: autoregressive expansion suffers from context drift, while panoramic video generation is limited to low resolution. We present* Stepper*, a unified framework for text-driven immersive 3D scene synthesis that circumvents these limitations via stepwise panoramic scene expansion. Stepper leverages a novel multi-view 360° diffusion model that enables consistent, high-resolution expansion, coupled with a geometry reconstruction pipeline that enforces geometric coherence. Trained on a new large-scale, multi-view panorama dataset, Stepper achieves state-of-the-art fidelity and structural consistency, outperforming prior approaches, thereby setting a new standard for immersive scene generation.*

## 1. Introduction

The synthesis of immersive 3D scenes from text or images has rapidly evolved from a novel challenge to a central task in computer vision, mirroring the unprecedented success of generative image and video models. [46, 64, 65] This task serves as an essential stepping stone towards general world models, but has immediate applications in spatial computing, particularly for Mixed Reality and next-generation mapping applications. Crucially, in these settings, synthesized environments must satisfy strict perceptual criteria: high-fidelity rendering, visual consistency and unrestricted navigation within the synthesized environment.

To address this challenge, recent work has largely focused on two distinct paradigms. The first adopts an iterative strategy that leverages generative image or video models to autoregressively hallucinate and fuse novel views into an expanding scene representation [4, 5, 10, 16, 47]. While theoretically enabling large-scale exploration, this approach is susceptible to subtle inconsistencies and context drift, often resulting in accumulating geometric errors

[†]Work done during Felix' internship at Google.

and degraded visual fidelity. Alternatively, a second line of work [46, 48, 65, 76] targets lifting 360° panoramas directly into 3D space. Although these methods deliver superior visual quality near the projection center, they fundamentally struggle with occluded regions; rendering viewpoints far from the origin inevitably introduces artifacts such as blurring and stretched primitives.

To bridge this gap, we propose a method that enables the generation of high-quality 3D scenes, as seen in Fig. 1 with support for large-baseline navigation, achieved through three primary contributions. First, we introduce a multi-view panorama diffusion model. Leveraging an initial canonical view from [20], this model enables us to essentially "step" into the scene. Crucially, by processing full panoramic contexts rather than limited field-of-view perspective images, our approach minimizes the accumulation of geometric and semantic inconsistencies, *i.e*. context drift. Simultaneously, it circumvents the resolution bottlenecks of panoramic video generation, delivering the high-definition imagery that ensures superior immersion.

Second, we introduce a reconstruction framework that enforces geometric consistency across multiple panoramic views. To avoid distortions and undesired artifacts inherent to conventional monocular depth estimators on spherical data, we decompose our generated multi-view panoramas into perspective views and process them with a robust feed-forward SfM model [22] to recover a dense point cloud. Subsequently, we optimize a 3D Gaussian Splatting [23] representation for real-time exploration.

Third, we release a large-scale synthetic dataset to overcome the severe scarcity of multi-view panoramic data. Existing public collections suffer from limited scale, low resolution, and a lack of multi-view observations required to learn scene exploration. We extend the procedural generation framework Infinigen [41, 42] to render high-quality, multi-view panoramic trajectories across a diverse set of indoor and outdoor environments. Comprising approximately 230 000 samples at $4096 \times 2048$ resolution across 5,000 scenes, this dataset provides the geometric priors necessary for strong generalization. Additionally, we also curate a small test set of 3D scenes from Infinigen and the web, allowing us to benchmark our model against existing baselines. To summarize, our contributions are:

1. *A multi-view, high-resolution panorama diffusion model* for iterative scene expansion,
2. *A robust reconstruction framework* that synthesizes generated multi-view panoramas into a consistent, explorable representation.
3. *A large-scale, multi-view panorama dataset* with an accompanying benchmark set for improved training and rigorous evaluation of existing world models.

## 2. Related Work

In this section we introduce all relevant related work. To this end, we first start by discussing 2D generative models, before diving into the literature for 3D reconstruction and synthesis from images.

### 2.1. 2D Generative Models

**Image Generation.** With the introduction of diffusion models [15, 50, 51], the field of image generation has recently taken a huge leap forward. Facilitated by advances in latent space modelling [45] and classifier-free guidance [14], modern models [8, 27, 40] are capable of synthesizing photo-realistic images at high resolution from merely text prompts with reasonable hardware requirements. Methods like LoRA finetuning [18], ControlNets [73] and Adapters [33] allow users to introduce explicit constraints, enabling tasks ranging from inpainting [6, 30] to layout-guided synthesis [75].

**Controllable Video Generation.** Building upon image foundations, diffusion models have recently demonstrated remarkable success in video synthesis [17, 25, 35, 36, 52, 53, 59, 66]. To leverage these models for scene exploration, research has focused on disentangling camera movement from content generation. By injecting explicit trajectory and intrinsic control into the denoising process, recent approaches [13, 58, 71, 72, 74] can be used for controllable exploration of unobserved parts of a virtual scene.

**Panorama Generation.** Parallel to video generation, a distinct line of work focuses on 360° panorama synthesis [9, 11, 31, 54, 60], which offers a compact yet comprehensive representation of scene context and can serve as a strong 3D scene initialization. However, most existing methods rely on generating equirectangular panoramas, which introduces significant polar distortions and restricts resolution. To circumvent these limitations, CubeDiff [20] proposes repurposing multi-view diffusion models to jointly synthesize the six faces of a cubemap. We identify this cubemap-based paradigm as a robust foundation for 3D scene exploration and adopt it as the backbone for our proposed multi-view panorama generation model.

### 2.2. 3D Representations from 2D images

Obtaining a 3D scene representation from a set of images and videos is a long-standing problem in computer vision.

**Monocular Depth Estimation.** Historically, 3D reconstruction relied on estimating explicit geometry from single views [7, 28]. This field has recently matured into the era of foundation models; pioneered by MiDaS [43], current approaches demonstrate that training on massive, diverse datasets unlocks strong zero-shot generalization [62]. Modern depth predictors [3, 21, 63] are able to infer highly-

detailed geometry from unconstrained images, including metrically accurate depth maps [19, 37–39, 56, 68].

**Joint Pose and Pointmap Estimation.** While depth maps provide local geometry, reconstructing a coherent scene requires aligning multiple views in 3D space. Recent advances have precipitated a paradigm shift from traditional Structure-from-Motion pipelines to end-to-end foundation models. Methods like DUSt3R [57] and its successors propose to simultaneously regress camera poses and dense 3D pointmaps directly from image sets. Current state-of-the-art methods [22, 55] achieve robust reconstruction across large numbers of input images, leveraging the rich priors distilled from extensive, densely annotated training datasets.

**Novel View Synthesis.** To translate these geometric representations into immersive exploration, the field has largely adopted neural rendering techniques. Neural Radiance Fields (NeRFs) [1, 2, 32] established the standard for photorealistic view synthesis, while 3D Gaussian Splatting (3DGS) [23, 34, 67] has recently emerged as a dominant alternative, offering real-time rendering speeds with high visual fidelity. In this work, we bridge these paradigms by employing MapAnything [22] to lift our generated multi-view panoramas into a consistent geometric scaffold, which is subsequently optimized into a 3DGS representation for real-time exploration.

### 2.3. 3D Scene Synthesis

**Perspective-based Synthesis.** Synthesizing coherent 3D environments from sparse inputs is a fundamentally ill-posed problem. Nevertheless, recent methods have made progress by leveraging advances in generative models and 3D reconstruction. Approaches like DiffDreamer [4], Text2Room [16], and others [5, 10, 47] adopt an iterative paradigm, using depth estimation to lift images in 3D and image diffusion to autoregressively fill missing regions, with extensions to interactive scene generation [61, 69, 70]. These methods inherently rely on partial observations, leading to severe semantic drift and geometric inconsistencies when moving far from the origin. Alternative video-based methods [29] attempt to generate exploration paths directly but often struggle to maintain multi-view consistency without explicit geometric constraints.

**Panoramic Scene Synthesis.** To overcome these limitations, recent research has shifted toward panoramic generation. Methods like HoloDreamer [76] and RfG3D [48] lift a single generated panorama into 3D but degrade rapidly during translation due to the lack of disoccluded geometry. WorldExplorer [46] and Matrix-3D [65] address this by driving exploration with panoramic video models, with the latter achieving improved stability via a feed-forward 3DGS reconstruction model. However, the computational cost of video synthesis imposes severe resolution bottlenecks, limiting output fidelity. Inspired by the multi-view success of CAT3D [12] and CubeDiff [20], we propose to overcome these trade-offs by treating scene expansion as a multi-view cubemap generation problem. By iteratively synthesizing high-definition cubemaps rather than low-resolution panorama videos, our method minimizes drift while maintaining the photorealistic quality required for immersive exploration.

## 3. Method

In the following, we first introduce the required preliminaries before describing our multi-view panorama generation model for step-by-step 3D scene exploration. Finally, we present a specialized pipeline, visualized in Fig. 2, to obtain an immersive 3D scene from multiple exploration steps.

### 3.1. Preliminaries

Let $\mathbf{P} \in [0,1]^{H\times W\times 3}$ be a panoramic image in the equirectangular format, which covers the full sphere with $360^\circ \times 180^\circ$. A 3D point $\mathbf{x} \in \mathbb{R}^3$ can be mapped to the image plane using the panoramic projection function

$$\pi_{\text{pano}}(\mathbf{x}) = \begin{pmatrix} \frac{W}{360} \cdot \arctan\left(\mathbf{x}_z, \mathbf{x}_x\right) \\ \frac{H}{180} \cdot \arctan\left(\mathbf{x}_y, \sqrt{\mathbf{x}_x^2 + \mathbf{x}_z^2}\right) \end{pmatrix}. \quad (1)$$

A perspective image $\mathbf{I} \in [0,1]^{h\times w\times 3}$ with camera rotation $\mathbf{R} \in \mathrm{SO}(3)$ and focal length $f$ can be obtained from $\mathbf{P}$ through resampling (and vice versa)

$$\mathbf{I}_p(\mathbf{P}, \mathbf{R}, K) = \mathbf{P}\left[\pi_{\text{pano}}\left(\mathbf{R}K^{-1}\begin{pmatrix} p_x \\ p_y \\ 1 \end{pmatrix}\right)\right], \quad (2)$$

with $K$ denoting the camera intrinsics given the focal length $f$, and $\mathbf{I}_p$ being the resulting perspective image at pixel $p$.

### 3.2. Multi-view Panorama Generation

Most 3D scene synthesis methods start from an input image or panorama and then rely on off-the-shelf 2D image or video priors to fill in unobserved areas of the scene. However, as these models are usually conditioned on regular perspective images with a small field of view, they often struggle to grasp the entirety of the scene. This results in context drift and 3D inconsistencies. We argue that operating directly on panoramic images presents a promising alternative. First, panoramas, when projected to the right representation, are still 2D and fairly similar to perspective images, and can thus strongly benefit from pretrained 2D image and video models. Second, panoramic images always capture a significant portion of the scene context, ensuring more coherence and reducing drift.

To this end, we propose a multi-view panorama generation model $\Phi_d : \mathbf{P}_{\text{in}} \mapsto \mathbf{P}_{\text{nv}}$, which synthesizes a novel

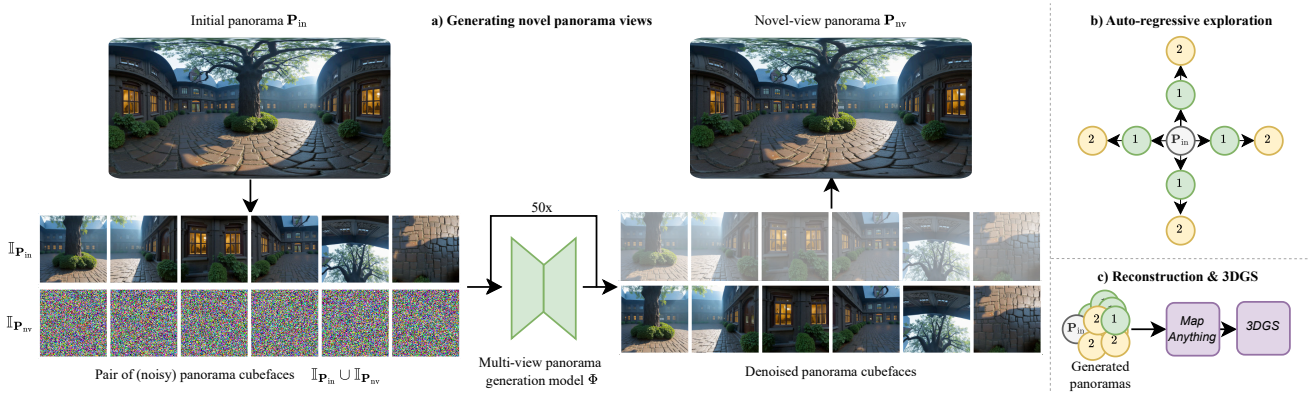


Figure 2. **Method overview.** a) Our model generates a new panoramic image from a previously unobserved viewpoint based on a given input panorama. To ensure high quality, we utilize a pre-trained diffusion model with expanded multi-view attention that is instrumental for jointly denoising the high-resolution cubefaces of the newly generated novel-view panorama. b) Our ability to generate novel-view panoramas enables auto-regressive scene generation in all directions of the scene yielding a set of high quality, consistent panoramas that effectively complete the representation of the 3D scene. c) The generated panoramas are processed with the feed-forward reconstruction model, *i.e.* MapAnything. The output pointcloud serves as the initialization of a custom 3D Gaussian Splatting reconstruction enabling high quality novel view synthesis of the generated 3D scene.

panoramic view by virtually moving a distance $d$ forward into the scene from the input panorama $\mathbf{P}_{\text{in}}$. Such a model can thus be naturally utilized for full 3D scene exploration. By rotating the input panorama $\mathbf{P}_{\text{in}}$ by an angle $\alpha$, which can be achieved by horizontally rolling the equirectangular image by $\alpha \cdot \frac{W}{360}$ pixels, we can adjust the movement direction. Through auto-regressive invocation, we can then take multiple steps and hence walk along longer paths.

**Multi-cubemap representation.** As the availability of multi-view panoramic image pairs is limited, it is not feasible to train such a model from scratch and still achieve good generalization capabilities. Therefore, we adopt a pretrained image diffusion model for conditional panorama generation. Drawing inspiration from [20], we represent panoramic images as cubemaps by sampling the six faces of a cube $\mathbb{I}_{\mathbf{P}} = \{\mathbf{I}(\mathbf{P}, \mathbf{R}_{\text{k}}, K) \mid k \in \{\text{Front}, \text{Left}, \ldots, \text{Bottom}\}\}$. Note that we set $f$ in $K$ such that each face has a FOV of 90 degrees. Hence, a given equirectangular panorama $\mathbf{P}$ can be fully described by these perspective images and vice-versa. A pair of panoramas is then the set $S \in [0,1]^{12 \times H \times W \times 3}$ of all twelve cubefaces with $S = \mathbb{I}_{\mathbf{P}_{\text{in}}} \cup \mathbb{I}_{\mathbf{P}_{\text{nv}}}$. Thanks to this representation, we can now feed our panoramas to regular image diffusion models by simply setting the batch size $t = 12$, without suffering from a domain gap due to distortions.

**Model architecture.** Similarly to [12], we base our model on an LDM with a latent space of $128 \times 128 \times 8$, which is pretrained on a large-scale image dataset. In particular, the LDM follows an architecture similar to Stable Diffusion [45] with multiple convolutional and self-attention layers. Inspired by [20], we modify the architecture to simultaneously generate multiple images, *i.e.* the different cubefaces $\mathbb{I}_{\mathbf{P}}$. To ensure cross-view and cross-panorama consistency, we inflate [49] the deeper self-attention layers of the LDM in order to enable the tokens of every cubeface to attend to the tokens of all other faces of its own as well as the other cubemap. In practice, the self-attention token sequence length is simply extended for those layers from $(bt) \times (hw) \times l$ to $b \times (thw) \times l$, where $b$ is the number of panoramas in the batch, $hw$ denotes the spatial dimensions, and $t = 12$, as noted above, denotes all cube faces in $S$. Note that we also concatenate a positional encoding $p$ and mask $m$ to each pixel in order to encode the pixel location and whether it needs to be generated. We compute UV coordinates on the unit cube and mark each pixel by its panorama of origin with

$$p = \pi_{\text{pano}}(x) \quad \text{and} \quad m_p = \begin{cases} -1 & \text{if } p \in \mathbf{P}_{\text{in}} \\ 1 & \text{otherwise} \end{cases}, \tag{3}$$

where $x$ denotes the 3D coordinates of the point on the cube face. Note that we do not need to condition the LDM on $d$ as we empirically found that a fixed stepping length works best for scene expansion. Finally, we finetune the model using a standard diffusion loss on ground truth panorama pairs converted to cubefaces.

### 3.3. 3D Gaussian Splats from Panoramas.

While the described model already enables us to freely generate 3D-coherent panoramas at different viewpoints, it does not yet allow real-time exploration of the scene. To this end, we additionally distill our multi-view panoramas into a 3D Gaussian Splatting (3DGS) [23] representation for real-time novel view synthesis.

**Pointcloud from feed-forward model.** Recently, several works have shown that a strong initialization prior, in the form of a 3D pointcloud, can be helpful for high-quality results and training stability [26, 34]. However, as aligning individual depth maps from a given monocular depth prior can be prone to errors, we instead apply a state-of-the-art feed-forward reconstruction model MapAnything [22] on perspective views that we extract from all of our generated panoramas. Nevertheless, the application of MapAnything to cubemap faces is not trivial. First, the lack of overlap with other views can lead to unsatisfying results, which is particularly prominent in the up- and downward facing views. To better mimic MapAnything's training data distribution, we design a different viewing pattern for reconstruction: we rotate 45 degrees up and down from the horizontal cubefaces to always ensure sufficient overlap among views. Note that we show resulting 3D pointclouds for different input patterns to MapAnything in the supplementary material. Second, the resulting point maps can easily become very large with a huge number of redundant points, leaving a large memory footprint and slowing down rendering. Hence, in an effort to remove these redundant points, we build the final pointcloud in an iterative fashion. Using the pointcloud renderer from PyTorch3D [44], we check if newly introduced points by a panorama are already visible in any of the previous panoramas. We then only add the previously unobserved points to the final pointcloud.

**3DGS optimization.** For photometric reconstruction, we use projected views of generated panoramas, consisting of the six cubefaces and eight additional perspective views. Further, we initialize the 3DGS representations with the accurate pointcloud from the feed-forward model and apply a simplified optimization strategy of MCMC-GS [24]. Due to the under-constrained nature of our setup and the dense initialization, we reduce the complexity of the optimization problem by assuming fixed Gaussian positions and only a color value per Gaussian as the appearance representation. For details of the 3DGS optimization, we refer to the supplementary materials.

### 3.4. Step-wise Scene Exploration

Our framework relies on a starting panorama $\mathbf{P}_{\text{init}}$ to initialize the scene for 3D exploration. If no panorama is available, we rely on the state-of-the-art method CubeDiff [20] to generate a high-resolution panorama image $\mathbf{P}_{\text{init}}$ given a text prompt and/or a reference image. From the initial panorama, we take $n$ auto-regressive steps using our model $\Phi$ into four directions to obtain $1 + 4n$ panoramic views, covering significant portions of the scene, which were initially unobserved in $\mathbf{P}_{\text{init}}$. Finally, by lifting these panoramas to 3D Gaussians, as described above, we enable full real-time 3D exploration of the generated scene.

Figure 3. **Dataset Samples.** The dataset generated with Infinigen consists of a diverse set of high quality synthetic panoramas of indoor and outdoor scenes. For every panorama, we rendered a pair from a novel viewpoint enabling the training of the multi-view panorama generation model. All panoramas are aligned to the horizontal line.

## 4. Experiments

In the following, we first describe the design of our training and test datasets, before qualitatively and quantitatively evaluating our approach against state-of-the-art baselines.

### 4.1. Data & Setup

Panoramic images are a powerful representation because they capture a significant part of a scene's context. Unfortunately, existing panorama datasets are generally quite small and contain only a single image per scene. To overcome this limitation, we thus develop a pipeline built on top of Infinigen [41, 42] to generate a custom, synthetic dataset of multi-view panoramas. Infinigen procedurally generates and populates indoor and outdoor scenes within the 3D rendering software Blender. We adopt Infinigen to generate 3D scenes and render high-resolution multi-view 360 panoramas, as can be seen in Fig. 3. In total, we collect around 230 000 pairs of panoramas at a resolution of $4096 \times 2048$ across 5 000 scenes to train our multi-view generation model. Furthermore, we find that existing works in 3D scene generation do not have a unified system for quantitative evaluation and often rely on hand-crafted solutions. Therefore, we also curate a small test set consisting of six photorealistic scenes from Blender and ten scenes generated with Infinigen. For every scene, nine panoramas and their corresponding depth maps are rendered in an area of $[-1m, 1m]$ around the scene origin and serve as references for evaluating visual quality. Both the training and testing dataset will be made publicly available to facilitate future research.

As introduced in Sec. 3.2, our multi-view panorama generation model is built on a customized version of the popu-

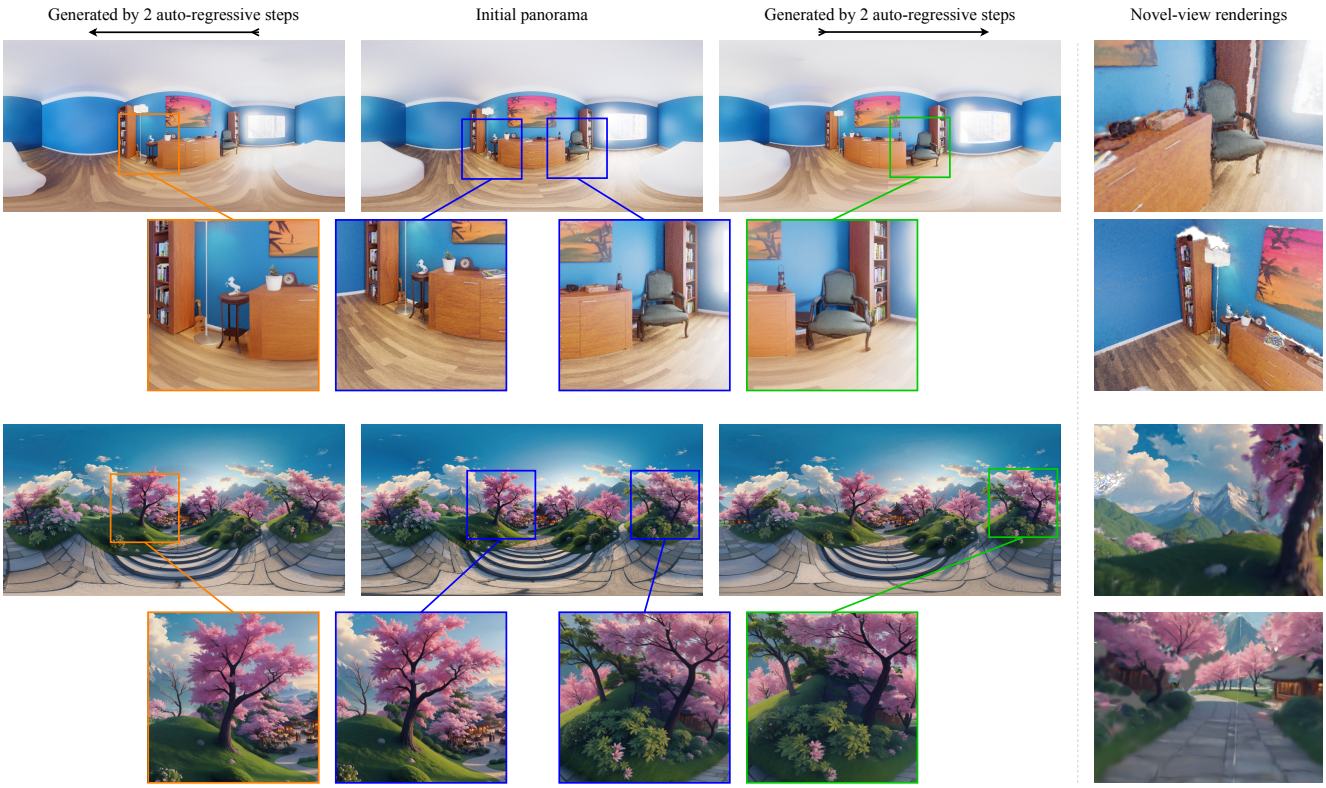


Figure 4. **3D Scene Generation.** We provide visual example of generated novel-view panoramas on the left side. The details of the initial panorama are well preserved and previously unseen regions are filled in. On the right side we show novel-view renderings of the reconstructed scenes indicating the 3D consistency of the generated panoramas.

lar LDM [45] model. It processes twelve individual cubemap faces at a resolution of $1024 \times 1024$, derived from two $4096 \times 2048$ input equirectangular panoramas. We initialize our model with pretrained weights and finetune for 90,000 steps ($\sim$ 2.5 days) at a batch size of 1 (= 12 cubefaces) sharded across 4 ViperFish TPUs with 64 TPUs in total, making for an effective batch size of 16. We empirically find that a step size of $d = 0.25m$ provides a good trade-off between scene exploration and robustness.

### 4.2. 3D Scene Generation

**Qualitative results.** To provide an overview of the capabilities of our model, we show a range of scenes in Fig. 4, with their corresponding input, generated multi-view panoramas, and obtained Gaussian Splatting reconstructions. As demonstrated, the generated novel-view panoramas retain all details from the initial panorama while correctly adjusting the geometry of the objects. For example, the highlighted chair is correctly translated despite its complex geometric structure. Furthermore, the previously occluded regions are well filled in, whilst respecting the overall context. This shows that our multi-view panorama generation model not only learns a strong geometric understanding of the environment, but also retains its powerful generative inpainting capability. The robust geometric reasoning capabilities are further underlined by the high-quality pointcloud, as produced by MapAnything. Finally, the 3DGS renderings are also high quality, even for viewpoints that are far away from the initial panorama.

**Comparisons with the state-of-the-art.** We also perform qualitative and quantitative evaluations against several state-of-the-art baselines: **LayerPano3D** [64] iteratively removes and inpaints layers from an input panorama to build a multi-plane panorama image. For detecting individual layers, they employ a depth and segmentation model. **WorldExplorer** [46] starts with a panoramic image and generates videos along several predefined trajectories using a camera-conditioned video diffusion model. To provide the scene context during generation, they sample previously generated frames and prepend them to the video. **Matrix-3D** [65], concurrent to us, fine-tunes a video diffusion model to generate panorama videos along a camera trajectory. However, due to video generation models being significantly more expensive than image generation models, they are restricted to a maximum resolution of $1440 \times 720$ despite significant computational resources.

We provide every method with the same initial panorama and a text prompt whenever applicable. After generation,

Figure 5. **Comparison with Baselines.** Given a high quality input panorama, we observe that our approach achieves consistent scene generation while showing significantly more details and sharpness in the rendered novel view images in comparison to the baselines.

we align the scenes to the ground-truth scale by comparing rendered and ground-truth depth maps. In Fig. 5, we show a qualitative comparison against the baselines. First, LayerPano3D provides fairly sharp renderings, however, quality degrades in occluded areas as the automatic layering of the scene oftentimes struggles, leading to artifacts and inconsistencies with the initial panorama. For example, the mountain background in the barbershop is replaced with a different door and the lowest layer (brown mist) also blends into the scene. The scenes generated by WorldExplorer are reasonable, but contain a fairly low level of detail. Especially when moving further away from the initial panorama, one experiences severe Gaussian defects. We attribute this to the fact that the generated videos experience drift over time, as well as local color inconsistencies. Finally, Matrix-3D generates very consistent and robust results. Nonetheless, the resulting scenes lack details and often appear blurry. We hypothesize that this is a result of the (for panoramic images) very low resolution of $1440 \times 720$. In contrast to them, our generated scenes are generally sharp and remain consistent even when moving further away. These findings are also consistent with our quantitative evaluation, which

| **Infinigen Indoors** | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| *WorldExplorer* | 11.864 | 0.674 | 0.739 |
| *LayerPano3D* | 18.305 | 0.783 | 0.509 |
| *Matrix-3D* | 18.532 | 0.753 | 0.502 |
| *Ours* | **21.775** | **0.797** | **0.430** |
| **Infinigen Outdoors** | PSNR↑ | SSIM↑ | LPIPS↓ |
| *WorldExplorer* | 13.912 | 0.561 | 0.594 |
| *LayerPano3D* | 17.364 | 0.590 | 0.537 |
| *Matrix-3D* | 17.970 | 0.581 | 0.529 |
| *Ours* | **20.507** | **0.646** | **0.384** |
| **Blender Scenes** | PSNR↑ | SSIM↑ | LPIPS↓ |
| *WorldExplorer* | 13.659 | 0.637 | 0.611 |
| *LayerPano3D* | 18.124 | 0.692 | 0.463 |
| *Matrix-3D* | 17.898 | 0.660 | 0.515 |
| *Ours* | **21.995** | **0.762** | **0.342** |
| **Average** | PSNR↑ | SSIM↑ | LPIPS↓ |
| *WorldExplorer* | 13.145 | 0.624 | 0.648 |
| *LayerPano3D* | 17.931 | 0.688 | 0.503 |
| *Matrix-3D* | 18.133 | 0.665 | 0.515 |
| *Ours* | **21.426** | **0.735** | **0.385** |

Table 1. **Quantitative Evaluation.** We compare our approach to the baseline methods for three different datasets on common image metrics. We observe that our approach yields significant improvements on all metrics and datasets.

we present in Tab. 1. Following common practice, we report the standard NVS metrics PSNR, SSIM, and LPIPS, comparing renderings from the generated Gaussians with the ground-truth views. As can be easily observed, our method clearly outperforms all baselines across all datasets and metrics. For example, we outperform the state of the art by at least 3.3 dB on average in PSNR. Similarly, for SSIM and LPIPS, Stepper achieves strong results of 0.735 and 0.385, clearly exceeding the second-best method, LayerPano3D, with 0.688 and 0.503, respectively.

### 4.3. Ablation Studies

**Auto-regressive expansion.** One of our main contributions is an auto-regressive scene expansion scheme via novel-view panorama synthesis, which enables an immersive experience. To further underline its necessity, we compare our final 3DGS scenes against scenes obtained from only the initial panorama and the respective scene geometry. As can be seen in Fig. 6, the scene obtained from only a single panorama has significantly more gaps. In contrast, our pipeline can retain all details from the initial scene whilst significantly improving completeness.

**Analysis of step size.** To test the effect of the fixed step size, we train two further model variants: **a)** A model for both forward and backward stepping, with the direction provided via a conditioning signal, and **b)** a model with a

Figure 6. **Single vs multiple panoramas to 3DGS.** The multi panorama input to the 3DGS reconstruction consistently fills in the unobserved regions in the initial panorama without sacrificing the quality of the input panorama.

Figure 7. **Effect of step size.** Novel panos. generated by a model with **a)** adjustable step direction, **b)** a larger step size $d = 0.5m$.

step size of $d = 0.5m$ (*vs* the default $d = 0.25m$). As can be seen in Fig. 7, the adjustable-step model sometimes produces wrong geometry and poor textures with artifacts, while our fixed-step model does not suffer from these issues. We attribute this to the learning task being easier when using a fixed step and thus selected this architecture. The $d = 0.5m$ model still generates high-quality novel panoramas, but is slightly worse at retaining details. The default $d = 0.25m$ model offers a good trade-off between step granularity, panorama quality, and exploration.

## 5. Conclusion

We introduced *Stepper*, a framework for text-driven immersive 3D scene generation that addresses the trade-off between visual fidelity and explorability. By combining a multi-view 360° diffusion model with a feed-forward reconstruction pipeline and a large-scale dataset of 230000 multi-view panoramas, Stepper generates high-quality, large-baseline explorable scenes without the context drift of prior methods. Our experiments demonstrate significant improvements over recent baselines, achieving an average PSNR improvement of 3.3dB, thereby establishing a new standard for immersive 3D scene synthesis.

## References


[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, pages 19697–19705, 2023. 3

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[4] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *ICCV*, pages 2139–2150, 2023. 1, 3

[5] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *IEEE TVCG*, 2025. 1, 3

[6] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *WACV*, pages 4334–4343, 2024. 2

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 2

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*. JMLR.org, 2024. 2

[9] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *CoRR*, 2023. 2

[10] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *NeurIPS*, 36:39897–39914, 2023. 1, 3

[11] Penglei Gao, Kai Yao, Tiandi Ye, Steven Wang, Yuan Yao, and Xiaofeng Wang. Opa-ma: Text guided mamba for 360-degree image out-painting. *arXiv preprint arXiv:2407.10923*, 2024. 2

[12] Ruiqi Gao, Aleksander Hołyński, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: create anything in 3d with multi-view diffusion models. In *NeurIPS*, pages 75468–75494, 2024. 3, 4

[13] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *ICLR*, 2025. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2

[16] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, pages 7909–7920, 2023. 1, 3

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[19] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE TPAMI*, 2024. 3

[20] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *ICLR*, 2025. 2, 3, 4, 5

[21] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024. 2

[22] 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, 5

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 2, 3, 4

[24] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *NeurIPS*, 37:80965–80986, 2024. 5

[25] 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

[26] Dmytro Kotovenko, Olga Grebenkova, and Björn Ommer. Edgs: Eliminating densification for efficient convergence of 3dgs, 2025. 5

[27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[28] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248. IEEE, 2016. 2

[29] 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 *CVPR*, pages 798–810, 2025. 3

[30] Guangben Lu, Yuzhen Du, Yizhe Tang, Zhimin Sun, Ran Yi, Yifan Qi, Tianyi Wang, Lizhuang Ma, and Fangyuan Zou. Pinco: Position-induced consistent adapter for diffusion transformer in foreground-conditioned inpainting. In *ICCV*, pages 15266–15276, 2025. 2

[31] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *AAAI*, pages 14211–14219, 2024. 2

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 2

[34] Michael Niemeyer, Fabian Manhardt, Marie-Julie Rakotosaona, Michael Oechsle, Daniel Duckworth, Rama Gosula, Keisuke Tateno, John Bates, Dominik Kaeser, and Federico Tombari. Radsplat: Radiance field-informed gaussian splatting for robust real-time rendering with 900+ fps. In *3DV*, 2025. 3, 5

[35] OpenAI. Sora: Video generation models as world simulators. https://openai.com/sora/, 2024. 2

[36] OpenAI. Sora 2. https://openai.com/index/sora-2/, 2025. 2

[37] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 3

[38] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *CVPR*, 2025.

[39] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. 3

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 2

[41] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *CVPR*, pages 12630–12641, 2023. 2, 5

[42] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *CVPR*, pages 21783–21794, 2024. 2, 5

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022. 2

[44] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 5

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4, 6

[46] Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. In *SIGGRAPH Asia*, 2025. 1, 2, 3, 6

[47] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. Controlroom3d: Room generation using semantic proxy rooms. In *CVPR*, 2024. 1, 3

[48] Katja Schwarz, Denis Rozumny, Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. A recipe for generating 3d worlds from a single image. In *ICCV*, pages 3520–3530, 2025. 2, 3

[49] 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. 4

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. pmlr, 2015. 2

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[52] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 2

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

[54] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *ACM MM*, pages 6811–6821, 2023. 2

[55] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 3

[56] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, pages 5261–5271, 2025. 3

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 3

[58] 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 *SIGGRAPH*, pages 1–11, 2024. 2

[59] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2

[60] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *ICLR*, 2024. 2

[61] Philipp Wulff, Felix Wimbauer, Dominik Muhle, and Daniel Cremers. Dream-to-recon: Monocular 3d reconstruction with diffusion-depth distillation from single images. In *ICCV*, pages 9352–9362, 2025. 3

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2

[64] 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 *SIGGRAPH*, pages 1–10, 2025. 1, 6

[65] 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. 1, 2, 3, 6

[66] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2

[67] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *JMLR*, 26(34): 1–17, 2025. 3

[68] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, pages 9043–9053, 2023. 3

[69] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *CVPR*, pages 6658–6667, 2024. 3

[70] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, pages 5916–5926, 2025. 3

[71] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025. 2

[72] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, 2025. 2

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[74] 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. 2

[75] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, pages 22490–22499, 2023. 2

[76] Haiyang Zhou, Xinhua Cheng, Wangbo Yu, Yonghong Tian, and Li Yuan. Holodreamer: Holistic 3d panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187*, 2024. 2, 3