

WorldExplorer: Towards Generating Fully Navigable 3D Scenes

MANUEL-ANDREAS SCHNEIDER*, Technical University of Munich, Germany

LUKAS HÖLLEIN*, Technical University of Munich, Germany

MATTHIAS NIESSNER, Technical University of Munich, Germany

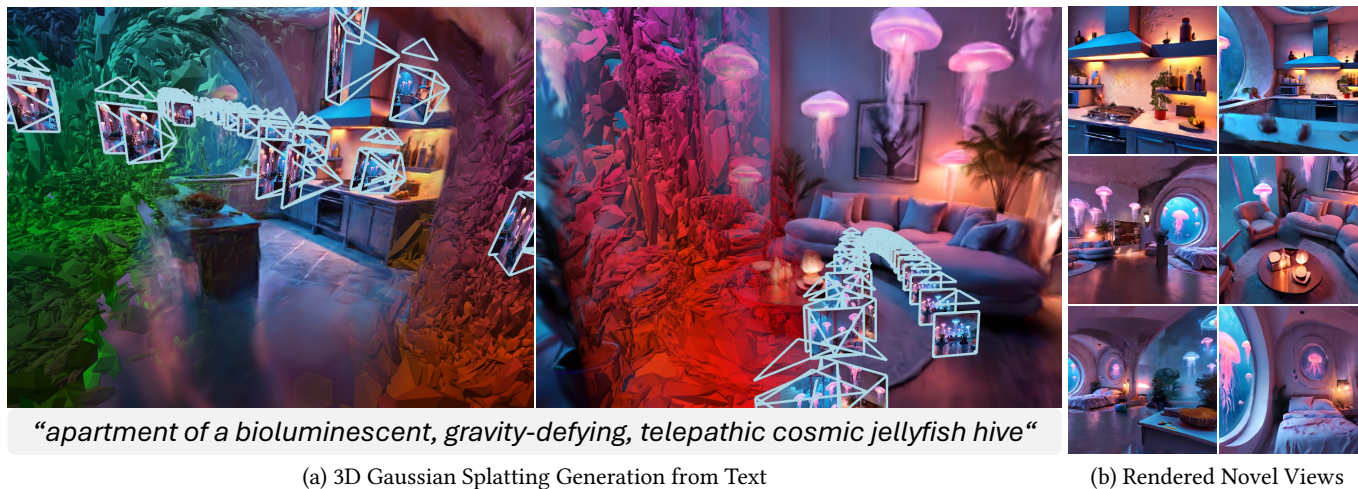


Fig. 1. **3D world generation from text input.** We exploit camera-guided video diffusion models [Zhou et al. 2025] to iteratively generate multi-view observations of a scene. To this end, we define multiple trajectories (cameras marked in blue) that explore different areas. Finally, we reconstruct a 3D Gaussian Splatting scene [Kerbl et al. 2023] from all generated images (we visualize the individual Gaussians on top of the renderings). Our scenes can be rendered under a large variety of viewpoints, i.e., they can be fully navigated beyond centered perspectives.

Generating 3D worlds from text is a highly anticipated goal in computer vision. Existing works are limited by the degree of exploration they allow inside of a scene, i.e., produce stretched-out and noisy artifacts when moving beyond central or panoramic perspectives. To this end, we propose WorldExplorer, a novel method based on autoregressive video trajectory generation, which builds fully navigable 3D scenes with consistent visual quality across a wide range of viewpoints. We initialize our scenes by creating multi-view consistent images corresponding to a 360 degree panorama. Then, we expand it by leveraging video diffusion models in an iterative scene generation pipeline. Concretely, we generate multiple videos along short, pre-defined trajectories, that explore the scene in depth, including motion around objects. Our novel scene memory conditions each video on the most relevant prior views, while a collision-detection mechanism prevents degenerate results, like moving into objects. Finally, we fuse all generated views into a unified 3D representation via 3D Gaussian Splatting optimization. Compared to prior approaches, WorldExplorer produces high-quality scenes that remain stable under large camera motion, enabling for the first time realistic and

unrestricted exploration. We believe this marks a significant step toward generating immersive and truly explorable virtual 3D environments.

CCS Concepts: • **Computing methodologies** → **Rendering; Volumetric models; Reconstruction.**

Additional Key Words and Phrases: scene generation, video diffusion, 3D Gaussian Splatting

ACM Reference Format:

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. 2025. WorldExplorer: Towards Generating Fully Navigable 3D Scenes. In *SIG-GRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*, December 15–18, 2025, Hong Kong, Hong Kong. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3757377.3763946>

1 Introduction

High-quality and diverse 3D environments are a key requirement for applications ranging from AR/VR asset creation to computer graphics in gaming and film. As creating these worlds manually requires significant time and expertise, automated generation offers a scalable and accessible alternative for producing rich 3D content. The advent of text-to-image (T2I) and text-to-video (T2V) models [Labs 2023; Yang et al. 2024c] enables users to synthesize varied and visually compelling images directly from text input. Recent methods have begun leveraging these advances in 2D generation for 3D tasks such as object synthesis [Poole et al. 2023; Siddiqui et al. 2024] and scene construction [Sun et al. 2024; Yu et al. 2024c]. However, existing approaches to 3D scene generation still exhibit several limitations that hinder their practicality for world creation.

*denotes equal contribution. Website: <https://mschneider456.github.io/world-explorer>

Authors' Contact Information: Manuel-Andreas Schneider, Technical University of Munich, Germany, manuel.schneider@tum.de; Lukas Höllein, Technical University of Munich, Germany, lukas.hoellein@tum.de; Matthias Nießner, Technical University of Munich, Germany, niessner@tum.de.



This work is licensed under a Creative Commons Attribution 4.0 International License. SA Conference Papers '25, Hong Kong, Hong Kong
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2137-3/2025/12
<https://doi.org/10.1145/3757377.3763946>

One of the core challenges in generating 3D environments is enabling users to *freely explore* them in an interactive fashion. That is, to create immersive surroundings, users must be able to navigate the scene and look around objects, all while maintaining realistic view synthesis. One line of existing methods leverages 360° panoramas and converts them into 3D scenes [Yang et al. 2024b; Zhou et al. 2024]. Despite high-quality view synthesis near the scene’s center, their ability to support broader exploration is very limited; occlusions and missing multi-view information result in distorted geometry and inconsistent appearance under larger view-point changes. Alternative approaches generate scenes through iterative *render-refine-repeat* pipelines [Höllein et al. 2023; Shriram et al. 2024; Yu et al. 2024a], often relying on T2I models and monocular depth estimators [Yang et al. 2024a]. While this enables them to produce entire scenes beyond panoramas, their iterative nature often leads to stretching or distortion artifacts, since the object geometry is expanded frame-by-frame. More recently, methods using video diffusion models have demonstrated locally plausible object geometry [Chen et al. 2025b; Ren et al. 2025; Sun et al. 2024], i.e., their world prior over continuous motions helps to mitigate such artifacts. However, the generated worlds cannot yet be interactively explored, as they either lack full 360° scene coverage or exhibit distortion artifacts under novel views beyond centered perspectives.

To address these challenges, we propose a novel scene generation method leveraging camera-guided video generation, that enables free exploration of text-generated 3D environments (Figure 1). Our approach combines the strengths of panorama-based initialization, video diffusion models, and iterative scene generation. First, we construct a 360° panorama from multiple images, describing separate areas of a diverse environment. By utilizing depth estimation [Yang et al. 2024a] and T2I inpainting [Labs 2023], we create this initial scene scaffold. Next, we leverage video diffusion models in an iterative generation scheme to explore and expand the areas described by these initial observations. Concretely, we generate multiple small videos, that move into different scene parts and thus provide more multi-view information about their contained objects. We define the video trajectories *a-priori*, i.e., a diverse set of directions and perspectives iteratively explores the entire 360° scene. To ensure view-consistency between the separate videos, we propose a scene memory mechanism. It builds a conditional input for the video model from our scene scaffold and the most suitable views among all previously generated images. We further develop a collision detection mechanism, that dynamically adjusts the trajectory length based on the actual generated content, i.e., we avoid degenerate outputs like moving the camera into walls or objects. Finally, all generated views are fused into a single 3D scene representation via 3D Gaussian Splatting optimization [Kerbl et al. 2023]. We initialize this reconstruction by estimating a point cloud from all generated images using [Wang et al. 2025a] and rigidly align it with our generated cameras. This enables interactive exploration of generated 3D worlds from arbitrary viewpoints in real-time based on rasterization. To summarize, our contributions are:

- We introduce the first method for generating 3D scenes from text that supports high-quality view synthesis while enabling exploration across a wide range of camera poses.

- We propose an autoregressive scene expansion via video diffusion, driven by trajectory sampling and adaptive collision detection.
- We design a scene memory to condition each generation step on key prior frames, ensuring view consistency and scene coherence.

2 Related Work

Image- and Video-Generation Models. Text-to-image (T2I) models can create high-quality, diverse images from text [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022]. Their visual quality continuously improved [Labs 2023; Podell et al. 2023; Xie et al. 2024; Zhang et al. 2023b]. As they are trained on billions of images [Schuhmann et al. 2022], they represent strong 2D world priors, that can be adopted for many downstream tasks like inpainting/editing [Brooks et al. 2023; Lugmayr et al. 2022], semantic controls [Mou et al. 2024; Ye et al. 2023; Zhang et al. 2023a], or personalization [Hu et al. 2022; Ruiz et al. 2023]. We leverage T2I models in our panorama initialization stage to create large and diverse environments. Video models can generate worlds under continuously changing perspectives [Blattmann et al. 2023; Seaweed et al. 2025; Yang et al. 2024c]. A popular approach trains diffusion models [Ho et al. 2020] via flow-matching [Lipman et al. 2022] with U-Net [Ronneberger et al. 2015] or DiT [Peebles and Xie 2023; Vaswani et al. 2017] backbones. We leverage these models to build entire 3D scenes autoregressively.

3D Scene Generation. Several approaches use 2D models as priors for 3D generation. Score distillation sampling [Poole et al. 2023] optimizes for diverse objects [Chen et al. 2023; Siddiqui et al. 2024; Wang et al. 2023] or scenes [Cohen-Bar et al. 2023; Li et al. 2024b; Zhang et al. 2024]. Others create multi-view consistent images for 3D reconstruction [Gao et al. 2024; Höllein et al. 2024; Szymanowicz et al. 2025]. In order to create entire scenes, [Höllein et al. 2023; Schult et al. 2024; Shriram et al. 2024; Yu et al. 2024b] utilize depth estimation [Ke et al. 2024; Yang et al. 2024a] and inpainting in an iterative generation pipeline [Liu et al. 2021]. Alternatively, [Li et al. 2024a; Schwarz et al. 2025; Wang et al. 2024] turn 360° panoramas into 3D Gaussian Splatting [Kerbl et al. 2023] scenes. Recently, camera-guided video diffusion models are utilized in similar pipelines [Liu et al. 2024b; Ren et al. 2025; Sun et al. 2024; Yu et al. 2024c]. We leverage such video models in a novel iterative pipeline to generate large-scale 3D scenes that allow interactive exploration.

Camera-Controlled Video Generation. Camera control in video generation is beneficial for world generation. Recent approaches finetune models with additional pose input [Bahmani et al. 2025, 2024; Bai et al. 2025; He et al. 2025; Liang et al. 2024]. Others render point clouds into novel views to condition the next frame generation [Liu et al. 2024a; Wang et al. 2025b; Zhang et al. 2025]. While the motion of the videos can be precisely controlled, their length is often limited (e.g., small rotations around objects). By using T2V models in a sliding-window fashion, [Song et al. 2025; Zhou et al. 2025] generate longer and consistent videos. However, to create entire 3D scenes, it is crucial to generate consistent outputs even when revisiting previously generated areas or when viewing them from different perspectives. To this end, we exploit these models in a novel iterative scheme. Backed by a scene memory mechanism, we generate multiple small videos in a 3D-consistent fashion.

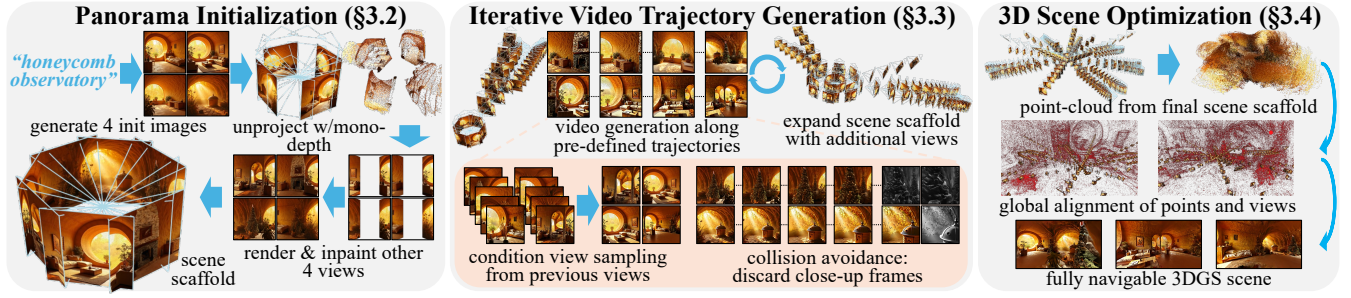


Fig. 2. **Method overview.** We generate a fully navigable 3D scene from text input in three stages. **Left:** we create four initial images with a T2I model [Labs 2023], and we define that they look outwards, distributed uniformly around the scene center (without overlap). We leverage monocular depth estimators [Yang et al. 2024a] and inpainting to create the remaining four images of our scene scaffold. **Mid:** we propose a novel iterative video trajectory generation pipeline. It exploits camera-guided video diffusion models [Zhou et al. 2025] to generate additional images of the scene along pre-defined trajectories. For each video generation, we select optimal conditioning views from all previously generated frames, which encourages 3D-consistent content generation. We further discard frames (marked as grey), that are too close to (running into) objects to avoid degenerate outputs. We repeat this trajectory generation multiple times to obtain a final set of posed images. **Right:** we optimize a 3D Gaussian-Splatting scene [Kerbl et al. 2023] from our generated images. To this end, we first obtain a point-cloud initialization using [Wang et al. 2025a]. Then we align the predicted points to our camera poses via rigid transformation and scaling. Finally, we can explore the scene, i.e., synthesize novel views from perspectives beyond the scene center, in real-time using standard computer graphics rasterization.

3 Method

Our method creates a 3D scene from text as input. It is divided into three stages (see Figure 2). First, we generate a panorama initialization to define a rough scaffold of the scene layout (Section 3.2). Then, we utilize camera-guided video diffusion models (Section 3.1) to expand the scene. The core idea of our approach is a scene generation scheme, which creates multiple small videos that progressively explore the scene (Section 3.3). Each video contributes multi-view information about individual objects, i.e., we move “inside and around” of different parts of our scene scaffold. To ensure 3D-consistency between videos, we propose a novel scene memory mechanism, that selects optimal conditioning images for the next video, and a collision detection mechanism. Finally, we reconstruct the entire 3D scene from all generated images (Section 3.4).

3.1 Review of Video Diffusion Models

Video diffusion [Blattmann et al. 2023; Seawead et al. 2025; Yang et al. 2024c] models the conditional probability distribution $p_\theta(\mathbf{I}^{\text{tgt}}|\mathbf{I}^{\text{src}})$ over data $(\mathbf{I}^{\text{tgt}}, \mathbf{I}^{\text{src}}) \sim q(\mathbf{V})$, where $\mathbf{V} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N\}$ is a dataset of continuous videos with $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$ being the individual images. To achieve more control over the generated video samples, a camera condition $(\pi^{\text{src}}, \pi^{\text{tgt}})$ is added to each image, which typically is encoded into dense Plücker embeddings [Liang et al. 2024; Zhou et al. 2025]. The network θ then parameterizes the conditional diffusion denoising probability $p_\theta(\mathbf{I}^{\text{tgt}}|\mathbf{c}) = \int p_\phi(\mathbf{I}_{0:T}^{\text{tgt}}|\mathbf{c}) d\mathbf{I}_{1:T}^{\text{tgt}}$ with $\mathbf{c} = (\mathbf{I}^{\text{src}}, \pi^{\text{src}}, \pi^{\text{tgt}})$ and we progressively add more Gaussian noise to the latent variables $\mathbf{I}_{1:T}^{\text{tgt}}$ following [Ho et al. 2020; Lipman et al. 2022]. We leverage the recent camera-guided video diffusion model from [Zhou et al. 2025] as a strong prior over the real world. In other words, we exploit the model to generate entire 3D scenes autoregressively, i.e., we generate a large set of images expressed as multiple shorter videos $\{\hat{\mathbf{V}}_1, \dots, \hat{\mathbf{V}}_M\}$. To ensure high-quality and 3D-consistent video generation, we develop a novel conditioning

mechanism that selects the optimal $\mathbf{c}_k$ from all previously generated data $\{\hat{\mathbf{V}}_1, \dots, \hat{\mathbf{V}}_{k-1}\}$ (see Section 3.3).

3.2 Panorama Initialization

In the first step of our method, we create a panoramic overview of the scene (see Figure 2 left). We utilize the T2I model Flux [Labs 2023] to generate four initial images, denoted as $\{\mathbf{I}_i\}_{i=1}^4$. Each of these images describes separate areas of the scene, e.g., a kitchen or a living room. Then, we define four fixed camera poses, each consisting of a rotation $R_i \in \text{SO}(3)$ and a translation $t_i \in \mathbb{R}^3$, with $t_i = \mathbf{0}$ for all i . The rotations correspond to uniformly distributed yaw angles, i.e.,

$$R_i = R_y(\theta_i), \quad \theta_i \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}, \quad (1)$$

where $R_y(\theta)$ denotes a rotation of angle θ about the global y -axis. In other words, each camera is oriented outward from the origin. Next, we utilize these poses together with a pretrained monocular depth estimator [Yang et al. 2024a] to unproject each image into a point cloud as $\mathcal{P}_i = R_i K^{-1} [u, v, 1]^\top \cdot \mathbf{D}_i(u, v)$, where $\mathbf{K}$ are the pinhole camera intrinsics with a field of view of 60° and $\mathbf{D}_i$ is the estimated depth map for image $\mathbf{I}_i$. We now similarly define four additional camera poses at $\theta_i \in \{45^\circ, 135^\circ, 225^\circ, 315^\circ\}$ and project the point clouds into these views using point splatting [Ravi et al. 2020; Wiles et al. 2020]. Finally, we inpaint the unobserved regions using [Labs 2023] to obtain our panoramic scene overview $\{\mathbf{I}_i\}_{i=1}^8$.

In contrast to existing text-to-panorama generators like [Tang et al. 2023; Yang et al. 2024b], this approach can create more diverse scenes. For example, the four initial images can represent drastically different areas (e.g., an indoor kitchen and outdoor forest), whereas panorama generators tend to produce 360° views of a single scene type only (see Figure 4 and the supplementary material). It also offers users an easy interface for interactive generation of diverse scenes by only having to generate and arrange four images according to personal preferences.

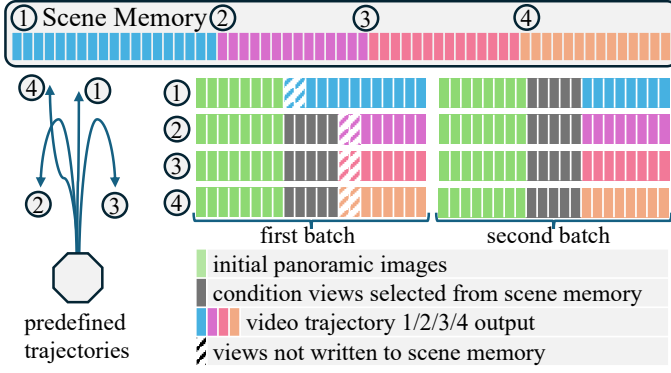


Fig. 3. **Conditional video generation via scene memory.** We sequentially generate videos in multiple batches of 21 frames along 4 trajectories (blue, purple, red, orange). These trajectories start from each of our initial panoramic images (bottom left). To ensure 3D-consistent outputs, we prefill the first 8 frames with our initial scene scaffold (marked in green). The remaining 5 are sampled based on rotational similarity from our scene memory, which stores the images and poses of all previously generated frames (top). After each trajectory is generated, we append all frames starting from the third image to our scene memory for conditioning of later frames.

3.3 Iterative Video Trajectory Generation

Given the initial scene scaffold $\{I_i\}_{i=1}^8$, we create additional observations with the video diffusion model. Concretely, we propose an iterative generation scheme: starting from an initial image, we utilize four pre-defined trajectories $\{\pi_k^{\text{tgt}}\}_{k=1}^4$ that explore the scene by moving “into and around” that area. In other words, we sequentially generate the videos $\hat{V}_{4i+k} \sim p_\theta(I_k^{\text{tgt}} | I^{\text{src}}, \pi_k^{\text{src}}, \pi_k^{\text{tgt}})$ for all i, k , where suitable condition frames I^{src} are selected per video from all previous frames. We iterate this principle multiple times to create in total 32 videos. Next, we explain the generation in more detail.

Trajectory Sampling. Camera-guided video diffusion models such as [Zhou et al. 2025] are powerful priors for generating *locally plausible* multi-view observations (e.g., rotating around objects). In contrast, generating entire 3D scenes along a single, *large* camera path remains challenging due to several reasons. First, defining a continuous path a-priori, that covers an entire 3D scene is ambiguous. In other words, we have to adapt the trajectory on-the-fly based on the generated content (e.g., prevent running into walls, resolve occlusions, discover novel generated areas). Second, to achieve 3D-consistency in large 3D worlds, the model must memorize previously generated objects/structures across a large history of previous frames. While approaches like [Song et al. 2025; Zhou et al. 2025] show promising results, their models still suffer from inconsistencies, collisions, and catastrophic forgetting, when re-visiting previously observed areas (also see Figure 4 and the supplementary material).

To this end, we rely on a different strategy to turn video diffusion models into 3D-consistent generators for large 3D scenes. Inspired by iterative generation pipelines like [Höllein et al. 2023; Liu et al. 2021], we split scene generation into multiple smaller trajectories. Concretely, our four pre-defined trajectories exploit video diffusion

models to generate additional observations within a specific local scene area, starting from one of the 8 initial panoramic images. As can be seen in Figure 3, we create camera paths that zoom-in along a straight line (1), rotate to the left/right (2,3) and elevate up/look down (4). These trajectories allow us to rotate around arbitrary generated content in a brute-force fashion, i.e., we use the same camera paths starting from each initial image.

Crucially, resolving collisions becomes a much easier task this way: since we already split scene generation into discontinuous and small camera paths, we only need to prevent such degenerate cases for each trajectory separately. In other words, the next trajectory is not impacted by a collision in previous trajectories: it can still be generated starting from the scene scaffold. Concretely, we first generate the entire video and then utilize a video depth estimator [Chen et al. 2025a] to predict normalized depth maps $\hat{D} \in [0, 1]^{N \times H \times W}$ for all generated frames. We define a fixed area of interest as a binary mask $M \in \{0, 1\}^{H \times W}$, where pixels (u, v) within a center crop of extent (w, h) are set to 1:

$$M(u, v) = \begin{cases} 1 & \text{if } u \in \left[\frac{W-w}{2}, \frac{W+w}{2}\right], v \in \left[\frac{H-h}{2}, \frac{H+h}{2}\right] \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We then compute a masked average depth per frame $t \in \{1 \dots N\}$ as

$$\bar{d}_t = \frac{\sum_{u,v} M(u, v) \cdot \hat{D}_t(u, v)}{\sum_{u,v} M(u, v)}. \quad (3)$$

We discard the trajectory from the first frame t^* onward for which $\bar{d}_{t^*} < 0.4$, and retain only the generated frames before that point. Empirically, we define the center crop to extend to 20% of the image resolution. This avoids ending trajectories too early if they move closely besides objects/walls, which usually leads to low depth values on the edges of images.

Scene Memory. Our scene generation is divided into multiple short video segments, denoted as $\{\hat{V}_{4i+k}\}_{i,k}$, where each segment corresponds to a separate camera trajectory. To ensure 3D-consistency across videos, we introduce a novel scene memory mechanism (summarized in Figure 3). Assuming each trajectory produces N frames, we further divide them into video batches of 21 frames. Similar to [Zhou et al. 2025], we condition each batch on a set of 13 previously generated frames. The first 8 are fixed and correspond to the initial panoramic images $\{I_i\}_{i=1}^8$, which ensures global scene layout awareness at all times. The remaining 5 context frames are dynamically selected based on rotational similarity to all previously generated frames across earlier trajectories. Concretely, we select the top-5 nearest camera poses w.r.t. all 21 frames in a manner similar to [Zhou et al. 2025]. Different from their approach, we consider all previously generated images across all trajectories as potential candidates. Additionally, we leverage our known trajectory layout to discard cameras, that are directly opposite to the initial image of the current trajectory. After the video generation completed, we add the newly generated images and their associated cameras into our scene memory, i.e., the next trajectory can be conditioned on them. For improved diversity, we avoid saving the first two generated frames of any trajectory, as their images are still similar to the initial panoramic images.

Our method leverages the two-stage video generation approach of [Zhou et al. 2025], where anchor frames are first generated uniformly along the trajectory, followed by interpolation between these anchors. Our scene memory mechanism ensures consistency of the anchor frames across trajectories. The subsequent interpolation then naturally produces scene-consistent intermediate frames.

3.4 3D Scene Optimization

The final step of our method is to reconstruct the 3D scene from all generated images. To this end, we utilize 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] as our scene representation, since it offers real-time, high-quality novel-view-synthesis (NVS) capabilities on standard computer graphics pipelines. This makes it possible to create immersive and fully navigable 3D scene experiences.

First, we obtain a point cloud initialization from our generated images. In traditional NVS, Structure-from-Motion typically provides such a point cloud when regressing camera poses from the input images. However, we generate more than 1K images, which makes running methods like COLMAP [Schonberger and Frahm 2016] prohibitively slow. To this end, we utilize the recent approach VGGT [Wang et al. 2025a] to obtain a sparse point cloud from our generated images. We downsample the prediction to 200K points and use those as initialization for 3DGS training.

We perform an additional alignment step to our known camera poses, since VGGT predicts the point cloud $\hat{\mathcal{P}}_{GS}$ and the corresponding camera poses in its own world coordinate system. First, we obtain the rigid transformation between VGGT and our coordinate system by comparing the predicted camera pose of $\mathbf{I}_1$ with our known pose: $\pi_t = \pi_1 \hat{\pi}_1^{-1}$. Next, we calculate a scaling factor as $s = \frac{\text{hull}(\{\pi_i\}_{i=1}^N)}{\text{hull}(\{\hat{\pi}_i\}_{i=1}^N)}$, where *hull* denotes the length of the bounding box around all camera centers. Finally, we transform the VGGT point cloud into our known world coordinate system as $\mathcal{P}_{GS} = s \cdot \pi_t \cdot \hat{\mathcal{P}}_{GS}$.

We optimize the 3DGS scene following the loss function introduced in [Kerbl et al. 2023]:

$$\mathcal{L}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (\lambda_1 |\hat{\mathbf{I}}_i - \mathbf{I}_i| + \lambda_2 (1 - \text{SSIM}(\hat{\mathbf{I}}_i, \mathbf{I}_i))) \quad (4)$$

where $\hat{\mathbf{I}}_i$ is the rendered image from the 3DGS scene, $\mathbf{I}_i$ our generated frames from the second stage, and λ_1, λ_2 the loss weights.

4 Results

Implementation Details. We generate all results with our default trajectory sampling scheme of 8 initial start images and 4 pre-defined trajectories. All 32 videos are generated sequentially in a clockwise order of starting images. Within each we first generate the “zoom-in” trajectory (1) followed by “left” (2), “right” (3), and “up” (4). We generate 44 frames per video for the trajectory types (1) and (4), and 134 frames for (2) and (3), respectively. Depending on the collision detection, we generate between 1.5K-2.5K images per scene at resolution 576×576 . The complete three-stage generation process takes on average 7h on a single RTX3090 GPU. Our runtime is dominated by the second stage (~10min/trajectory), while the first stage (~5min) and third stage (~11min) are comparatively faster.

We experiment with multiple different prompts and scene types in our main results. Concretely, we generate prompts with LLMs and use a different room type for each of the four start images (e.g., living room, bathroom, bedroom, office, kitchen) for our indoor scene results. We also demonstrate the capability of our method to generate outdoor scenes in the supplementary material.

Baselines. We compare against multiple types of text-to-3D scene generation methods. We provide additional details about baselines in the supplementary material.

- **Panorama-To-3D:** we compare against DreamScene360 [Zhou et al. 2024] and LayerPano3D [Yang et al. 2024b] as state-of-the-art methods that first generate a 360° panorama from text and then turn it into a 3DGS [Kerbl et al. 2023] scene.
- **Iterative T2I Lifting:** we compare against Text2Room [Höllein et al. 2023] and WonderWorld [Yu et al. 2024a], which generate 3D scenes through a render-refine-repeat pattern based on T2I models and monocular depth estimators. We provide one of our four start images as input to these methods for fair comparisons.
- **Camera-Guided Video Diffusion:** most related to our approach, we compare against FlexWorld [Chen et al. 2025b] and SEVA [Zhou et al. 2025]. We choose them over other existing models, as they have demonstrated *complete* 360° scene generation with video diffusion (instead of only locally looking around objects or scene parts). We provide one of our four start images as input to both methods and add our 3DGS post-optimization step to SEVA.

Metrics. We evaluate our generated 3D scenes both qualitatively and quantitatively. We calculate CLIP Score (CS) [Radford et al. 2021] and Inception Score (IS) [Salimans et al. 2016] on RGB renderings from novel perspectives. For each scene, we sample cameras *beyond* panoramic/centered perspectives, to assess the quality under arbitrary viewpoints. Additionally, we conduct a user study and ask n=64 participants about the Perceptual Quality (PQ) and the 3D-Consistency (3DC). We present them with rendered videos that move throughout the whole 3D scene from novel (non-centered) perspectives and ask to score the scenes on a scale of 1–5.

4.1 Qualitative Comparisons

We compare the quality of generated scenes against all baselines in Figure 4. Also, we provide additional results of our method in Figure 6 and the supplementary material. In both cases, we split the visualization into two parts. First, we visualize a scene overview located at the center of the scene to evaluate the quality and diversity of generated 3D scenes. Second, we render novel views that explore the scene, i.e., they are rotated around objects and moved away from the center. This allows us to showcase the ability of each method to fully navigate their generated scenes.

The panorama-based methods DreamScene360 [Zhou et al. 2024] and LayerPano3D [Yang et al. 2024b] create compelling 360° views, but they do not hallucinate unobserved regions beyond the panoramic viewpoint. Thus, their novel views contain visible distortions and occlusions. Text2Room [Höllein et al. 2023] and WonderWorld [Yu et al. 2024a] create high-detail textures due to their iterative unprojection steps, but suffer from error accumulation, i.e., their object geometry appears stretched out due to erroneous depth estimation.



Fig. 4. **Qualitative comparison of our method and baselines.** We generate a scene given the text prompt “star-written, constellation-binding, midnight celestial ink library”. We visualize a scene overview from a centered camera position and 4 novel views that explore the scene from non-centered perspectives. Our method retains compelling visual quality and generates plausible continuations of the scene, even under these extreme camera positions. Please see the supplementary material for animated comparisons against baselines.

Table 1. **Quantitative comparison.** We report 2D metrics and user study results, including: Clip Score (CS), Inception Score (IS), Perceptual Quality (PQ), and 3D-Consistency (3DC). We evaluate all methods on novel views rendered from perspectives beyond the scene center, i.e., we quantify their ability to explore the scene arbitrarily. Our method creates scenes with the highest text-prompt similarity and is preferred in our user study.

Method	2D Metrics		User Study	
	CS ↑	IS ↑	PQ ↑	3DC ↑
DreamScene360 [Zhou et al. 2024]	23.51	2.06	3.01	3.04
LayerPano3D [Yang et al. 2024b]	24.30	2.04	2.16	2.14
Text2Room [Höllein et al. 2023]	24.37	1.92	2.68	2.32
WonderWorld [Yu et al. 2024a]	19.74	2.23	2.34	1.75
FlexWorld [Chen et al. 2025b]	21.61	2.31	2.89	2.71
SEVA [Zhou et al. 2025]	21.63	2.13	2.51	1.91
Ours (1 image)	21.42	1.88	2.42	2.03
Ours (4 images)	25.65	2.29	2.48	2.14
Ours (w/o scene-mem)	25.91	2.21	2.61	2.36
Ours (w/o coll-det)	25.37	1.89	3.77	3.42
Ours	25.94	2.27	4.04	4.02

FlexWorld [Chen et al. 2025b] creates complete 3D scenes with compelling hallucinations (e.g., the furnace). However, their trajectory-based generation does not cover large viewpoint changes, and as such novel views rotated around objects still exhibit missing and distorted geometry (e.g., behind the couch/table). SEVA [Zhou et al. 2025] uses a single, continuous trajectory to generate an entire 3D scene given a single image as input. This enables them to generate novel geometry, that matches the style of the input image. However, the generated video frames contain noticeable flickering artifacts (see suppl. material), which become visible in the 3DGS renderings as duplicated objects (e.g., multiple overlaid couches/tables) or floating artifacts (e.g., the bottom-left of the overview rendering).

This showcases the advantage of our three-stage approach: by conditioning the video diffusion model on 8 panoramic images, we generate high-quality and fitting novel views for the entire 360° scene layout. Similar to [Chen et al. 2025b; Zhou et al. 2025] our video diffusion stage generates locally plausible multi-view information. Additionally, our novel scene memory and iterative trajectory generation lift this to a complete 3D scene along multiple, discontinuous trajectories. Overall, this leads to high-quality renderings even for novel views far beyond the scene center. This enables users to interactively explore our environments from arbitrary camera viewpoints. Please see the suppl. material for more scenes, animated video results, and intermediate baseline outputs (e.g., video frames from SEVA [Zhou et al. 2025] and panoramas from DreamScene360 [Li et al. 2024b] and LayerPano3d [Yang et al. 2024b]).

4.2 Quantitative Comparisons

We present the quantitative results averaged across the novel view renderings of multiple scenes in Table 1. Our approach achieves the highest CLIP score, since our generated geometry remains consistent towards the scene prompt even beyond panoramic perspectives.

This highlights the advantage of our panorama initialization: the scene scaffold is already created from the beginning, which makes inpainting around it easier than for the iterative T2I and video diffusion baselines, that have to hallucinate the novel content completely from scratch. We obtain competitive Inception scores, which signals that our generated images are as sharp as baseline methods. Users give the highest ratings to our method, which highlights the quality of our generated 3D scenes under challenging, novel perspectives. We also provide a runtime and memory comparison against baselines in the supplementary material.

4.3 Ablations

The key components of our method are the panorama initialization (Section 3.2) and iterative video trajectory generation, utilizing our novel scene memory and collision detection mechanisms (Section 3.3). We demonstrate their importance in Figure 5 and Table 1.

How important is the panorama initialization? We generate multiple small videos along pre-defined trajectories, that start from each of the 8 initial panoramic images. We ablate the importance of having 8 initial images by only providing 4 or 1 of them as condition during video generation (but still generate all 32 trajectories). Scene diversity and sharpness drop noticeably, i.e., generating entire scene parts from scratch is harder than inpainting around our initial scene scaffold. Concretely, when providing only a single panoramic image, the video diffusion model needs to hallucinate entire new areas in a 3D-consistent fashion. This oftentimes leads to oversimplified outputs (if using a low guidance scale) or oversaturated frames (if using a high guidance scale). While providing 4 images helps to retain large scenes, their 3D-consistency severely drops, as the individual trajectories are less well-conditioned. This results in blurriness and floating artifacts in the 3D scene reconstruction.

How important is the scene memory? We propose a novel scene memory mechanism, that conditions each video generation on the most suitable previous frames across all trajectories. We ablate this choice by replacing it with the scene memory of [Zhou et al. 2025], i.e., we only select condition frames across the individual batches *within* each trajectory. Since we design multiple small camera paths instead of one large trajectory, the amount of multi-view information beyond the local areas of each trajectory is very limited under this setup. As a result, each video generates slightly different object types/positions. We can observe severe degradation of our scenes, as the lack of 3D-consistency results in blurry reconstructions.

How important is the collision detection? Our collision detection mechanism avoids degenerate videos due to running into generated content like walls. It is a necessary component, as we define our trajectories a-priori, i.e., we always use the same trajectories regardless of the actual generated content. Without it, collisions may occur, which lead to blurry edges around those areas. For example, the video generation ran into the archway in front of the kitchen, which makes its edges fuzzy when observed from up-close. Since collisions do not necessarily appear for every trajectory, other areas like the kitchen remain unaffected of these local degenerations. This further signals the robustness of our iterative generation: later trajectories are not adversely impacted by collisions of previous ones.



Fig. 5. **Ablation study on the key components of our method.** When conditioning video generation on only 1 initial image, generated scenes become smaller and less diverse: it is harder to generate novel areas from scratch than to inpaint around existing panoramic observations. While 4 initial panoramic images enable the creation of larger scenes, the frames of trajectories without a starting image are less 3D-consistent. This results in floating artifacts during rendering. Our proposed scene memory mechanism is necessary, as each individual trajectory only covers a local area of the scene. Thus, conditioning only within that trajectory similarly creates 3D-inconsistent scenes. Our collision detection ensures, that the pre-defined trajectories do not intersect with generated scene content like the archway towards the kitchen area. Without it, these areas become more fuzzy due to degenerate videos running into these structures.

4.4 Limitations

Our method generates high-quality 3D scenes from text as input, that can be fully explored, i.e., we can interactively move into scene parts and rotate around objects without quality degradation. Nevertheless, there are remaining limitations (see supplementary material). Our video diffusion model sometimes produces high-frequency flickering artifacts across trajectories, which manifest in blurry renderings. Extending our 3D scene optimization with uncertainty [Goli et al. 2024] or robustness [Sabour et al. 2023] weighting schemes specifically tailored for these common problems of video generators may further detect and ignore such outliers. Furthermore,

our method uses a set of pre-defined trajectories, which explore and expand the initial scene scaffold. While it proved robust towards a large variety of room and prompt types, it limits the layout of the generated scenes. Incorporating semantic layouts as condition [Schult et al. 2024; Zhang et al. 2023a], enabling interactive and unlimited scene generations, and generating dynamic 4D worlds are interesting future directions. Lastly, our approach generates 3D scenes with baked-in material and lighting effects, since our base diffusion models operate in RGB space. Recent text-to-x generators like [Ke et al. 2025; Kocsis et al. 2025] have the potential to replace RGB-space models in downstream tasks such as scene generation.



Fig. 6. **World generation results of our method.** We visualize scene overview renderings close to the scene center facing in two opposite directions. Additionally, we render novel views far beyond the scene center, that explore the different generated areas in more detail. This showcases the possibility to interactively explore our worlds from arbitrary viewpoints. We also demonstrate the diversity of our scenes by mixing indoor/outdoor elements in a single 3D scene. Please see the supplementary material for animated flythroughs of our generated scenes.

5 Conclusion

We have presented WorldExplorer, a novel method for text-to-3D scene generation. By utilizing video diffusion models in a scene generation scheme, we iteratively expand the generated scene from a panoramic scaffold towards fully navigable 3D environments. The core insight of our method is a tailored scene memory mechanism for our pre-defined video trajectories. It allows us to condition the

video diffusion model on the most relevant previous frames and thus ensures 3D-consistent image generations. We further demonstrate, our panorama stage as a helpful initializer for creating diverse environments. Our output 3DGS scenes can be rendered from arbitrary novel viewpoints in real-time with classical rasterization pipelines. We believe our method demonstrates an important step towards generating immersive 3D environments from only text input.

Acknowledgments

This project was funded by the ERC Consolidator Grant Gen3D (101171131) as well as the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing”. We also thank Angela Dai for the video voice-over and Simon Giebenhain for help on the Gaussian overlay visualization.

References

- Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. 2025. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.
- Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. 2024. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781* (2024).
- Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuoqiu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. *arXiv preprint arXiv:2503.11647* (2025).
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.
- Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. 2025b. FlexWorld: Progressively Expanding 3D Scenes for Flexible-View Synthesis. *arXiv preprint arXiv:2503.13265* (2025).
- Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22246–22256.
- Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2025a. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. *arXiv:2501.12375* (2025).
- Dana Cohen-Bar, Elad Richardson, Gal Metzger, Raja Giryes, and Daniel Cohen-Or. 2023. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2920–2929.
- Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *Advances in Neural Information Processing Systems* (2024).
- Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. 2024. Bayes’ rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20061–20070.
- Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangkue Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 2025. CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models. *arXiv preprint arXiv:2503.10592* (2025).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- Lukas Höllein, Aljaž Božić, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. 2024. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5043–5052.
- Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7909–7920.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZevKeeFY9>
- Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. 2025. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis. *arXiv preprint arXiv:2505.09358* (2025).
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- Peter Kocsis, Lukas Höllein, and Matthias Nießner. 2025. IntrinsicX: High-Quality PBR Generation using Image Priors. *arXiv preprint arXiv:2504.01008* (2025).
- Black Forest Labs. 2023. FLUX. <https://github.com/black-forest-labs/flux>.
- Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Peng Yuan Zhou. 2024b. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision*. Springer, 214–230.
- Wenrui Li, Fucheng Cai, Yapeng Mi, Zhe Yang, Wangmeng Zuo, Xingtao Wang, and Xiaopeng Fan. 2024a. Scenedreamer360: Text-driven 3d-consistent scene generation with panoramic gaussian splatting. *arXiv preprint arXiv:2408.13711* (2024).
- Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. 2024. Wonderland: Navigating 3D Scenes from a Single Image. *arXiv preprint arXiv:2412.12091* (2024).
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
- Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. 2021. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14458–14467.
- Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. 2024b. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. *arXiv:2408.16767* [cs.CV] <https://arxiv.org/abs/2408.16767>
- Kunhao Liu, Ling Shao, and Shijian Lu. 2024a. Novel View Extrapolation with Video Diffusion Priors. *arXiv preprint arXiv:2411.14208* (2024).
- Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adaptor: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR abs/2307.01952* (2023). <https://doi.org/10.48550/ARXIV.2307.01952> *arXiv:2307.01952*
- Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net. <https://openreview.net/forum?id=FjNys5c7VyY>
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- Aditya Ramesh, Pratul P. Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501* (2020).
- Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751* (2025).
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 234–241.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023*. IEEE, 22500–22510. <https://doi.org/10.1109/CVPR52729.2023.02155>
- Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. 2023. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20626–20636.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language

- Understanding.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).
- Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
- Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, et al. 2024. Controlroom3d: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6201–6210.
- Katja Schwarz, Denys Rozumnyi, Samuel Rota Bulò, Lorenzo Porzi, and Peter Kotschieder. 2025. A Recipe for Generating 3D Worlds From a Single Image. *arXiv preprint arXiv:2503.16611* (2025).
- Team Seaweed, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. 2025. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685* (2025).
- Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. 2024. Realdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199* (2024).
- Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotný. 2024. Meta 3D AssetGen: Text-to-Mesh Generation with High-Quality Geometry, Texture, and PBR Materials. *CoRR* abs/2407.02445 (2024). <https://doi.org/10.48550/ARXIV.2407.02445> arXiv:2407.02445
- Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. 2025. History-Guided Video Diffusion. *arXiv preprint arXiv:2502.06764* (2025).
- Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. 2024. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928* (2024).
- Stanislaw Szymonowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. 2025. Bolt3D: Generating 3D Scenes in Seconds. *arXiv:2503.14445* (2025).
- Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. *arXiv* (2023).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. 2024. Perf: Panoramic neural radiance field from a single panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- Hanyang Wang, Fangfu Liu, Jiawei Chi, and Yueqi Duan. 2025b. VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step. *arXiv preprint arXiv:2504.01956* (2025).
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025a. Vgg: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651* (2025).
- Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems* 36 (2023), 8406–8441.
- Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. 2020. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7467–7477.
- Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629* (2024).
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024a. Depth anything v2. *Advances in Neural Information Processing Systems* 37 (2024), 21875–21911.
- Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. 2024b. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv preprint arXiv:2408.13252* (2024).
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024c. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
- Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
- Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. 2024a. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394* (2024).
- Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snively, Jiajun Wu, et al. 2024b. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6658–6667.
- Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024c. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).
- Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023b. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909* (2023).
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
- Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. 2024. Towards text-guided 3d scene composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6829–6838.
- Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. 2025. Scene Splatter: Momentum 3D Scene Generation from Single Image with Video Diffusion Model. *arXiv preprint arXiv:2504.02764* (2025).
- Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishtha, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. 2025. STABLE VIRTUAL CAMERA: Generative View Synthesis with Diffusion Models. *arXiv preprint arXiv:2503.14489* (2025).
- Shijie Zhou, Zhiwen Fan, Dejie Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suyu You, Zhangyang Wang, and Achuta Kadambi. 2024. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*. Springer, 324–342.

WorldExplorer: Towards Generating Fully Navigable 3D Scenes

Supplementary Material

A Supplemental Video

Please watch our attached video ¹ for a comprehensive evaluation of the proposed method. We include rendered videos of multiple generated scenes from novel trajectories, that showcase the quality of our large-scale 3D scenes under various non-centered perspectives. We also compare against baselines and ablations of our method by showing rendered videos.

B User Study Details

We conduct a user study to determine Perceptual Quality (PQ) and 3D Consistency (3DC) of our 3D scenes. In order to determine these scores, we asked participants to rate our scenes and baselines based on overall quality and 3D-consistency on a scale of 1–5. Figure 7 shows an example of how we asked the users to score the above two metrics. We show the users short videos of trajectories rendered from our final scene, as well as from baselines, both displaying large camera motion. In total, we received 2432 datapoints from $n=64$ participants.

C Additional Baseline Details

We compare against several different types of methods for 3D scene generation. Here, we provide additional details about the chosen baselines.

Panorama-To-3D Generators. Since DreamScene360 [Zhou et al. 2024] and LayerPano3D [Yang et al. 2024b] can only be initialized from text input or 360 degree panorama input, we cannot directly provide our initial panoramic images to them. To ensure a fair comparison, we instead provide these methods the identical text prompt. Since their text-to-panorama backbone is built on a different T2I model than ours, we still obtain initialization with slightly different style. For consistency between the baselines, we generate one panorama using LayerPano3D [Yang et al. 2024b] and use it for both baselines. We visualize the panoramas for two of our comparison scenes in Figure 8 and also compare them against our panoramic scene initialization.

Text2Room [Höllein et al. 2023]. We provide one of our four initial images as input to the method. We run it with their default hyperparameters.

FlexWorld [Chen et al. 2025b]. We provide one of our four initial images as input to the method. We run it with their default hyperparameters.

WonderWorld [Yu et al. 2024a]. We provide one of our four initial images as input to the method. Their interactive scene generation scheme allows for fine-grained control over the next viewpoint for inpainting and its corresponding text prompt. We carefully select viewpoints in a manner similar to the original presented results.

SEVA [Zhou et al. 2025]. We follow the original codebase to generate an entire 3D scene starting from a single image. Concretely,

¹<https://youtu.be/N6NJsNyiv6I>

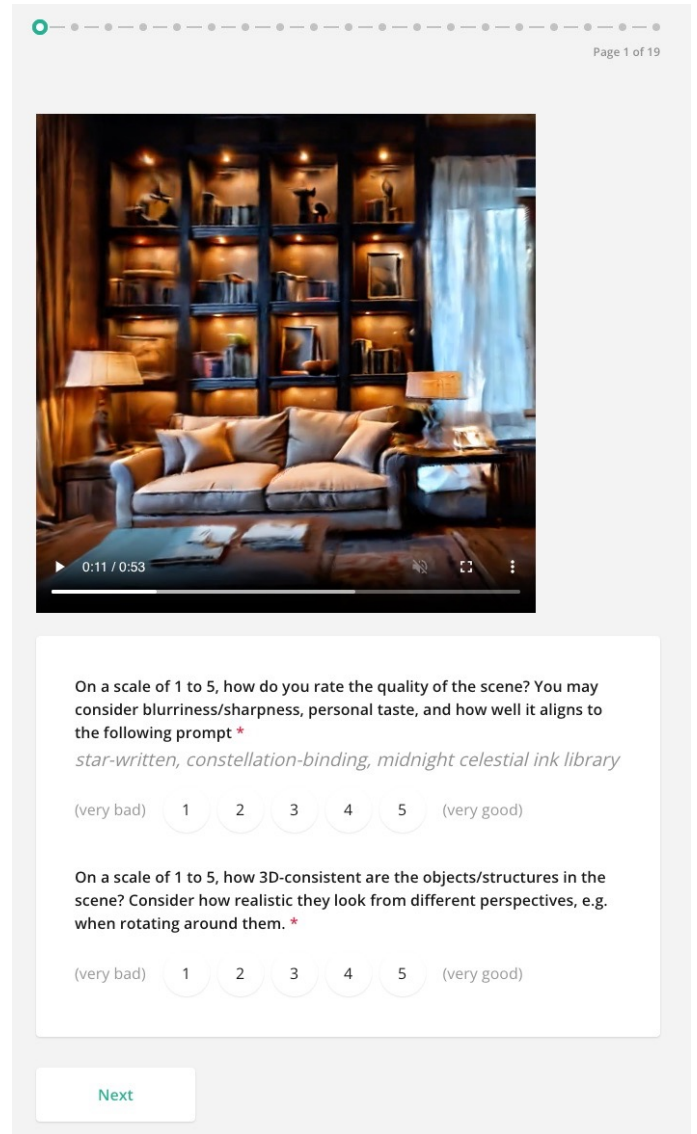


Fig. 7. **User study interface.** (Top) We provide users with videos along trajectories with large camera motion for our and baseline scenes. (Left) We ask users to scale the Perceptual Quality (PC) and 3D Consistency (3DC) on a scale from 1 to 5.

we manually define a single, continuous trajectory starting from one of our four initial panoramic images. This trajectory covers the entire 360 degree viewpoints around that start image by a mostly horizontal and outward-looking trajectory. However, coming up with a suitable trajectory a-priori from a single start image is hard, as the generated scene content is unknown. We visualize the generated video output in Figure 9. As can be seen, the video suffers



(a) LayerPano3D [Yang et al. 2024b] for “celestial ink library”



(b) LayerPano3D [Yang et al. 2024b] for “gothic revival mansion”



The 8 initial panoramic images of our method for “celestial ink library”



The 8 initial panoramic images of our method for “gothic revival mansion”

Fig. 8. **Intermediate baseline outputs.** We compare the panorama generation of [Yang et al. 2024b] against our first stage panorama initialization (Section 3.2). Since we generate four images independently from each other, we can obtain environments with higher diversity than the baselines (e.g., it is possible to specify certain room types like “living room” or “kitchen” separately for each of the four images).



Fig. 9. **Intermediate video generation results of [Zhou et al. 2025].** The above frames represents the video output generated along a single, continuous trajectory from the SEVA baseline [Zhou et al. 2025] initialized from a single input image on the “celestial ink library” scene. Even though we have chosen a high guidance scale of 6.0 for the generation, we still obtain inconsistent results (e.g. multiple different tables are generated in the bottom row). This leads to visible blurriness and duplicated object results when optimizing a 3DGS [Kerbl et al. 2023] representation from these images. In contrast, our full pipeline divides scene generation into multiple smaller videos that are further conditioned on 8 initial panoramic images. The resulting scenes from our approach exhibit 3D-consistent and high quality renderings.

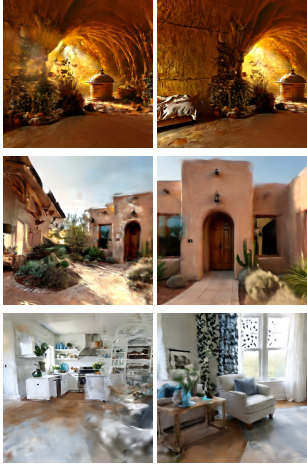


Fig. 10. **Limitations of our method.** Top and mid: high-frequency flickering in generated video frames lead to floating artifacts and blurry textures for certain parts of our 3D reconstruction. Bottom: objects placed at the very edge of our generated scene are rarely observed from all sides during video trajectory generation. Thus, the reconstructed scene is underconstrained at these areas, which results in degenerate rendering results.

from high-frequency flickering and inconsistencies across views, especially when moving far away from the initial image. As a result, the optimized 3DGS [Kerbl et al. 2023] scene in our main results suffers from blurriness and duplicated content. This highlights the necessity for an iterative scene generation scheme, that we propose on top of this camera-guided video diffusion model. In other words, our pre-defined trajectories provide a robust way to split scene generation into multiple local parts and our panorama initialization provides the necessary global conditioning for this iterative generation.

D Limitations

Our method generates high-quality 3D scenes from text as input, that can be fully explored, i.e., we can interactively move into scene parts and rotate around objects without quality degradation. Nevertheless, there are remaining limitations (see Figure 10). As shown in Figure 10 top and mid, our method is still susceptible to floaters and blurriness in the final 3D reconstruction from certain perspectives. This is partially the case since synthetically generated images are typically never fully 3D consistent, especially across thousands of individual images.

Additionally, our pre-defined trajectory scheme may not cover all areas of a generated scene. We visualize in Figure 10 bottom the front and back view of an armchair. Since the backside of the armchair is very close to the wall behind it, our pre-defined trajectories did never observe that part of the object. As a result, the 3D reconstruction remains inconsistent in these areas.

E Runtime/Memory Comparison

Our method generates fully-explorable scenes, as opposed to baseline-scenes viewable from limited camera-poses. We generate more

observations in order to resolve occlusions and increase the 3D completeness. This yields higher-quality scenes, as measured by Perceptual-Quality (PQ) and 3D-Consistency (3DC) of our user-study in Table 1. Therefore, we report runtime amortized per generated frame on the “celestial library” scene in Table 2. We achieve comparable per-frame generation rates and memory consumption as baseline approaches. Our runtime is dominated by the second stage (~10min/trajectory), while the first stage (~5min) and third stage (~11min) are comparatively faster. We believe real-time video-diffusion will reduce runtime towards interactive generation.

F Scene Layout Diversity

Our panoramic scaffold (Section 3.2) determines the global scene layout, i.e., which room types are combined. In our main results, we create the scaffold by using the same text prompt, but four different room prefixes (e.g. “kitchen/office of XYZ”). This creates scenes with a “cross-shaped” layout, where each of the four rooms becomes an explorable area.

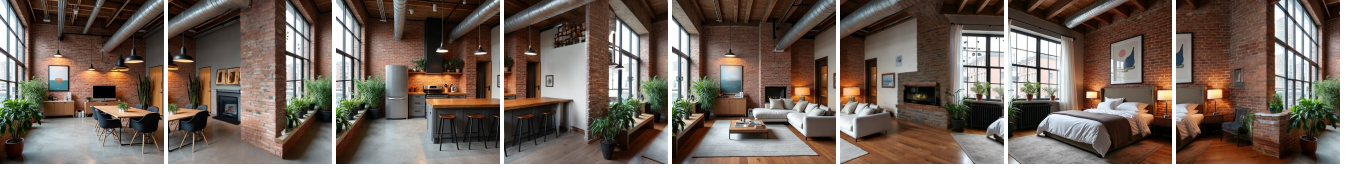
We showcase in Figure 11, that our method can also generate other layouts (e.g., “L-shaped” scenes) by adjusting the panoramic scaffold as desired. Concretely, we replace two room prefixes with “wall/window of XYZ” to generate flat, non-explorable areas. This demonstrates the diversity of our approach: it is possible to customize the generated scene layouts by adjusting the panoramic scene scaffold accordingly.

G Additional Qualitative Results

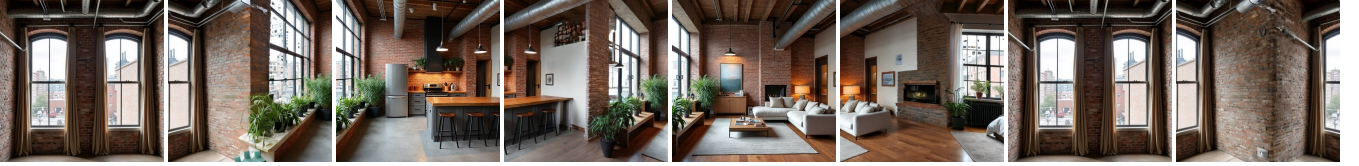
We show additional qualitative results of our method in Figure 12, Figure 13, and Figure 14.

Table 2. **Runtime/Memory comparison.** Our method generates fully-explorable scenes, as opposed to baseline-scenes viewable from limited camera-poses. We generate more observations in order to resolve occlusions and increase the 3D completeness. The runtime amortized per generated frame on the “celestial library” scene is comparable to baselines.

Method	Total runtime	Generated frames	Runtime per frame	Memory
DreamScene360 [Zhou et al. 2024]	0.5h	one panorama	N/A	16GB
LayerPano3D [Yang et al. 2024b]	0.2h	one panorama	N/A	5GB
Text2Room [Höllein et al. 2023]	0.9h	200	16.2s	15GB
WonderWorld [Yu et al. 2024a]	user-specific interactive generation		~10s	45GB
FlexWorld [Chen et al. 2025b]	0.5h	144	12.9s	48GB
SEVA [Zhou et al. 2025]	0.9h	270	13.5s	22GB
Ours	7h	2251	14.1s	22GB



The 8 initial panoramic images corresponding to the prefixes: dining room → kitchen → living room → bedroom



The 8 initial panoramic images corresponding to the prefixes: wall → kitchen → living room → wall



Scene rendering: kitchen → living room

Scene rendering: dining room → kitchen

Scene rendering: wall → kitchen

Fig. 11. **Scene layout diversity.** We generate an “urban industrial loft” with two different layouts by modifying the text prompts in the first stage. Top: generated panoramic scene scaffold with the room prefixes “dining room, kitchen, living room, bedroom”. Mid: replacing “dining room” and “bedroom” prefixes with “wall”. Bottom left: in both cases, we obtain scenes that transition from a kitchen to a living room. Bottom mid: in the first case (top) the generated scene contains another transition from a dining room area to the kitchen. Bottom right: in the second case (mid) we instead create a smaller scene without another room, but instead a flat wall transitioning to the kitchen.

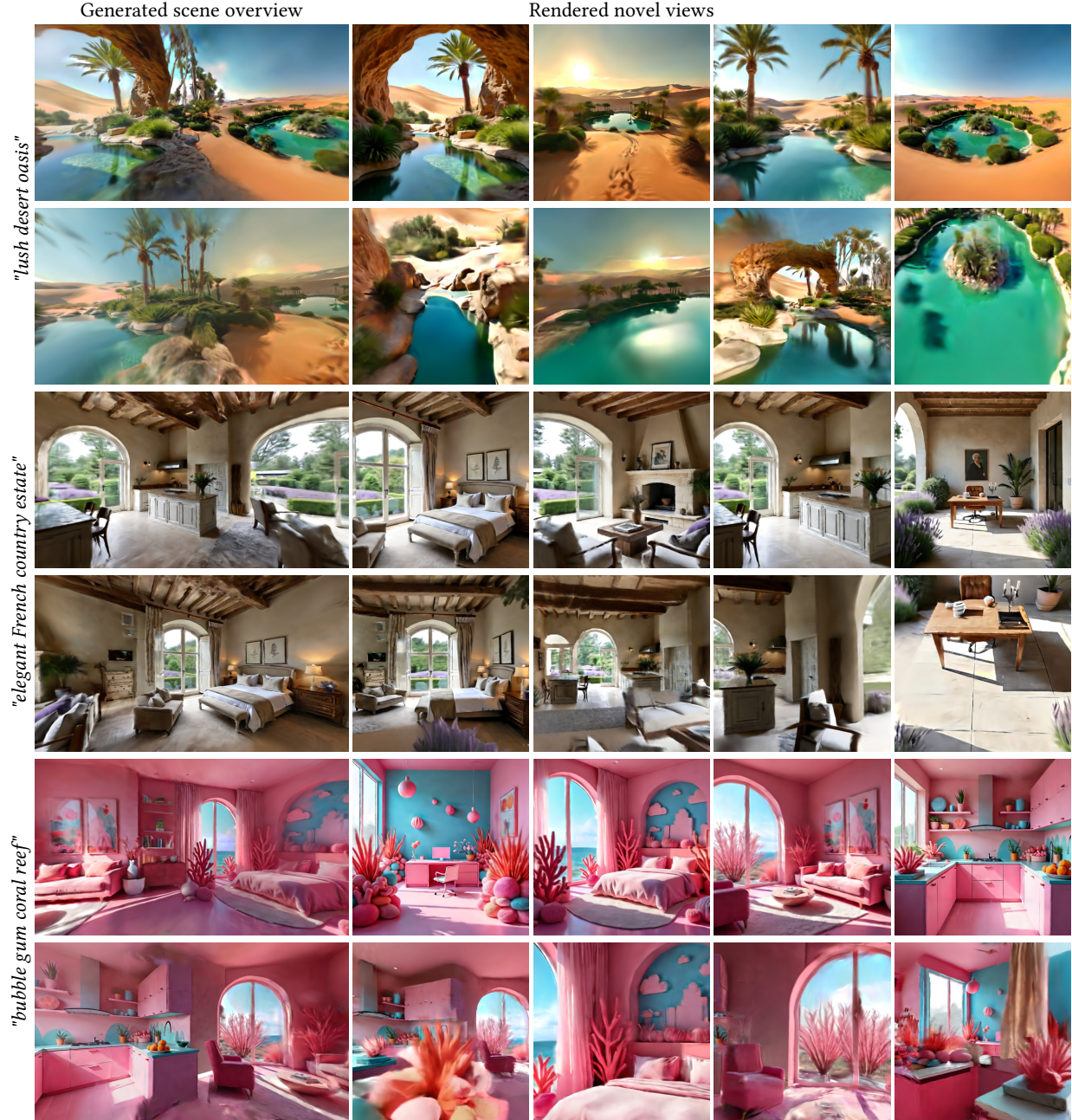


Fig. 12. **Additional world generation results of our method.** We visualize scene overview renderings close to the scene center facing in two opposite directions. Additionally, we render novel views far beyond the scene center, that explore the different generated areas in more detail. This showcases the possibility to interactively explore our worlds from arbitrary viewpoints. Please see the supplementary video for animated flythroughs of our generated scenes.



Fig. 13. **Additional world generation results of our method.** We visualize scene overview renderings close to the scene center facing in two opposite directions. Additionally, we render novel views far beyond the scene center, that explore the different generated areas in more detail. This showcases the possibility to interactively explore our worlds from arbitrary viewpoints. Please see the supplementary video for animated flythroughs of our generated scenes.

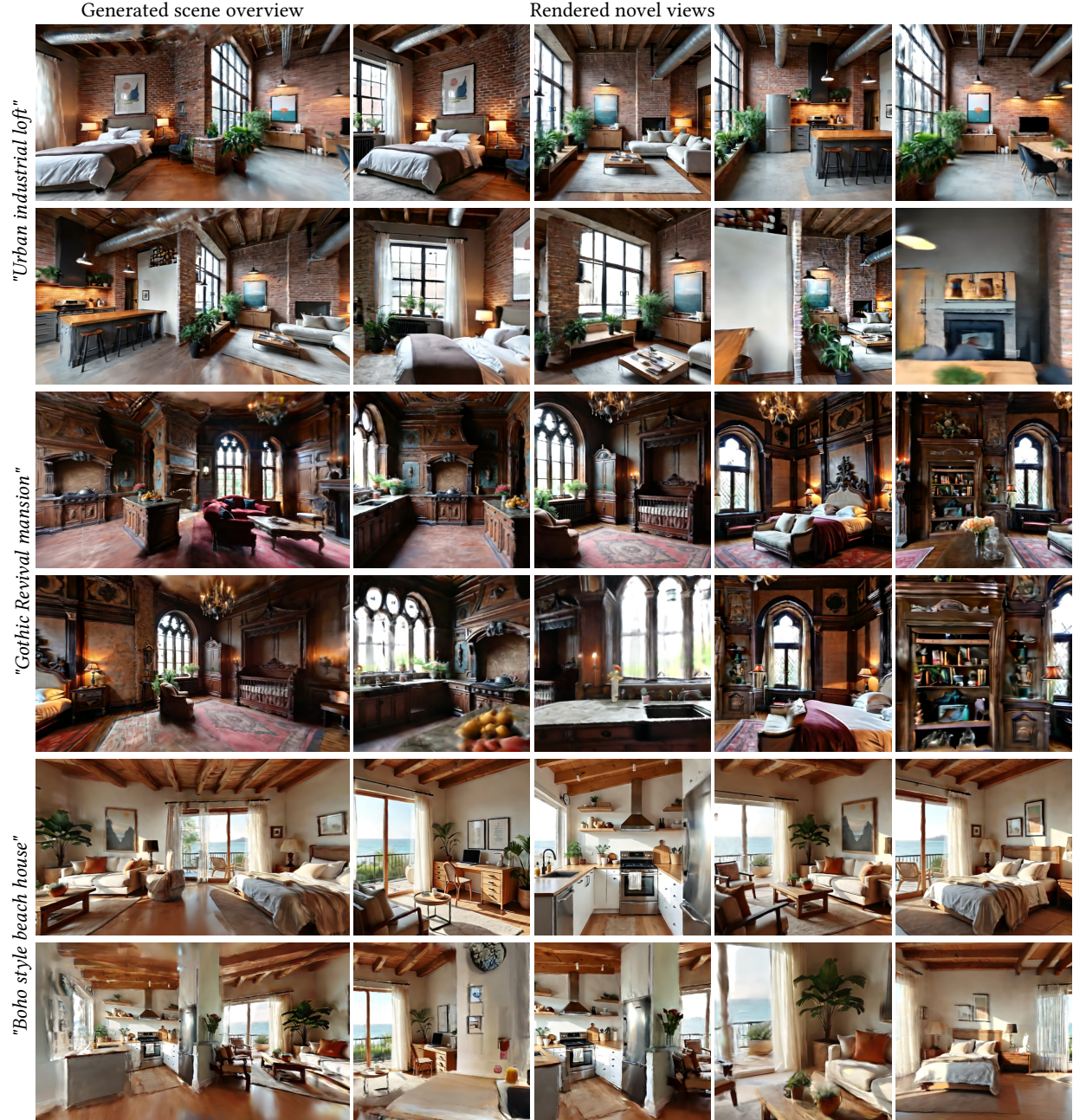


Fig. 14. **Additional world generation results of our method.** We visualize scene overview renderings close to the scene center facing in two opposite directions. Additionally, we render novel views far beyond the scene center, that explore the different generated areas in more detail. This showcases the possibility to interactively explore our worlds from arbitrary viewpoints. Please see the supplementary video for animated flythroughs of our generated scenes.