One2Scene: Geometric Consistent Explorable 3D Scene Generation from a Single Image

Anonymous authors
Paper under double-blind review

Abstract

Generating explorable 3D scenes from a single image is a highly challenging problem in 3D vision. Existing methods struggle to support free exploration, often producing severe geometric distortions and noisy artifacts when the viewpoint moves far from the original perspective. We introduce **One2Scene**, an effective framework that decomposes this ill-posed problem into three tractable sub-tasks to enable immersive explorable scene generation. We first use a panorama generator to produce anchor views from a single input image as initialization. Then, we lift these 2D anchors into an explicit 3D geometric scaffold via a generalizable, feed-forward Gaussian Splatting network. Instead of treating the panorama as a single image for reconstruction, we project it into multiple sparse anchor views and reformulate the reconstruction task as multi-view stereo matching, which allows us to leverage robust geometric priors learned from large-scale multi-view datasets. A bidirectional feature fusion module is used to enforce cross-view consistency, yielding an efficient and geometrically reliable scaffold. Finally, the scaffold serves as a strong prior for a novel view generator to produce photorealistic and geometrically accurate views at arbitrary cameras. By explicitly conditioning on a 3D-consistent scaffold to perform reconstruction, One2Scene works stably under large camera motions, supporting immersive scene exploration. Extensive experiments show that One2Scene substantially outperforms state-of-the-art methods in panorama depth estimation, feed-forward 360° reconstruction, and explorable 3D scene generation. Code and models will be released. Anonymous project page can be found at: `https://one2scene5406.github.io/`.

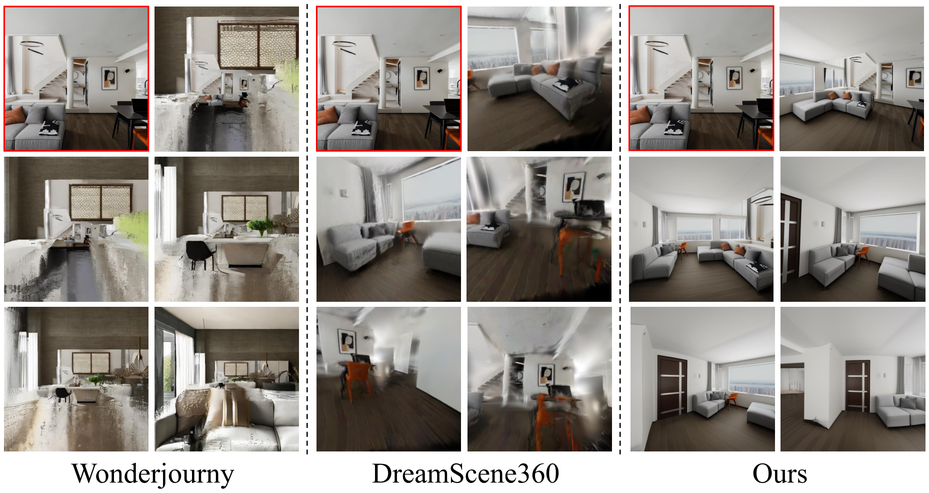

Figure 1: **Comparison on large-viewpoint novel view synthesis.** Existing methods such as Wonderjourny (Yu et al., 2023) and Dreamscene360 (Zhou et al., 2024) exhibit clear geometric distortions and artifacts, while our method generates photorealistic and geometrically accurate novel views. The input image is highlighted by a red bounding box. The other images represent the novel views.

1 INTRODUCTION

The increasing demand for high-quality 3D content is reshaping the landscape of video games, visual effects, and mixed reality, making 3D generation a highly active research topicm (Valevski et al., 2024; Adamkiewicz et al., 2022; Martin-Brualla et al., 2021; Ye et al., 2024b). Reconstruction-based methods like Neural Radiance Fields (NeRF) (Mildenhall et al., 2020) and Gaussian Splatting (GS) (Kerbl et al., 2023) have achieved remarkable results, but they typically require hundreds or even thousands of input images. Although sparse-view reconstruction approaches alleviate this requirement (Wang et al., 2023; Yang et al., 2023; Yu et al., 2024a; Charatan et al., 2024; Liu et al., 2024c;b; Wu et al., 2024a; Szymanowicz et al., 2024b), these methods struggle with large viewpoint extrapolation and fail to generalize to unseen regions. In stark contrast, generative view synthesis (Liu et al., 2023; Sargent et al., 2024; Liu et al., 2024a; Yu et al., 2024b) is emerging as a significant advancement in 3D content creation, as it can generate plausible content in unobserved regions (Shi et al., 2024; Zhou et al., 2025; Szymanowicz et al., 2025).

Although object-level 3D generation (Liu et al., 2023; Sargent et al., 2024; Ye et al., 2024b) has achieved rapid progress, generating an explorable 3D scene from a single image remains a significant challenge. One of the key challenges is how to maintain 3D geometric consistency and visual quality under large viewpoint changes and long-term generation. Some methods leverage pre-trained video generation models (Brooks et al., 2024; Xing et al., 2024; Hong et al., 2022; Yang et al., 2024) to create 3D-aware sequences (Liu et al., 2024a;d; Yu et al., 2024b; Chen et al., 2024; Sun et al., 2024; Liang et al., 2024), but they often suffer from geometric inconsistency and loop-closure consistency. Panorama-based pipelines such as Dreamscene360 (Zhou et al., 2024) and DreamCube (Huang et al., 2025) attempt to convert panoramas into 3D scenes, but their ability to support broader exploration is very limited, as shown in Figure 1 (a). Although navigation and inpainting-based methods (Chung et al., 2023; Yu et al., 2023; Höllein et al., 2023) enable the generation of more expansive scenes, their iterative nature often causes global semantic drift. Furthermore, cumulative errors often result in stretched or distorted geometry, as shown in Figure 1 (b). These limitations highlight the need for a new approach that can produce geometrically accurate and photorealistic scenes from a single image while supporting broad exploration.

To achieve the goal mentioned above, in this paper we introduce **One2Scene**, a novel framework that systematically decomposes explorable 3D scene generation into three distinct, yet more manageable subtasks. First, to overcome the profound information deficit of a single image, we generate a set of anchor views for global coverage using a panoramic cubemap representation. Note that these anchor views alone are insufficient to create a truly explorable scene, as shown in Figure 1 (a). Full exploration requires synthesizing high-quality novel views from arbitrary viewpoints, while how to ensure 3D consistency presents a significant hurdle. To this end, we introduce a powerful and efficient prior that encodes both geometry and appearance to stably constrain the generative process. Specifically, we reformulate the problem of monocular panoramic depth estimation as a multi-view stereo matching problem across extremely sparse anchor views, and lift the 2D anchor views into an explicit 3D geometric scaffold using a feed-forward 3D GS model. Such a design not only ensures the high efficiency of our feed-forward model but also critically enables us to leverage robust geometric priors learned from large-scale multi-view datasets. To further enforce geometric consistency across anchor-view boundaries, we introduce a bidirectional fusion module. As a result, our feed-forward model can reconstruct a geometrically accurate, high-quality 3D scaffold in 0.5 seconds.

The constructed explicit geometric scaffold provides strong priors for both geometry and appearance to guide the final novel view synthesis. To effectively utilize this scaffold, we introduce a novel Dual-LoRA training strategy. Unlike common refinement models that use channel-wise conditional injection (Wu et al., 2025), our strategy effectively fuses information from the high-quality input view with the coarse yet geometrically-rich views rendered from our scaffold. These combined conditions then guide the generation process at arbitrary camera views via a global 3D-aware attention mechanism. Our experiments demonstrate that this design significantly enhances the model's ability to leverage the priors provided. By grounding the generation process in a consistent 3D representation, the final results of our One2Scene model are not only photorealistic but also exhibit superior multi-view consistency, as demonstrated in Figure 1 (c).

Our contributions can be summarized as follows. First, we introduce a powerful feed-forward 3D GS model with a bidirectional fusion module to construct a high-quality 3D scaffold by reformulating the monocular panoramic depth estimation into a multi-view stereo problem. Second, we present a scaffold-guided synthesis method to utilize explicit geometric and appearance priors from any target view, which robustly grounds the final rendering and resolves the geometric ambiguities inherent in single-image generation. Finally, we demonstrate that our proposed One2Scene sets a new state-of-the-art on explorable 3D scene generation, achieving superior photorealism and geometric accuracy, particularly under significant viewpoint shifts.

2 RELATED WORK

3D Scene Reconstruction. Differentiable rendering techniques, such as NeRF (Mildenhall et al., 2020) and 3DGS (Kerbl et al., 2023), are primarily designed for per-scene optimization and require dense input views, which limit their practical applications in the real world. To reduce the need for dense images, the research community has proposed various sparse-view reconstruction methods (Wang et al., 2023; Yang et al., 2023; Yu et al., 2024a; Charatan et al., 2024; Liu et al., 2024c;b; Wu et al., 2024a; Szymanowicz et al., 2024b). Concurrently, generalizable feed-forward models (Charatan et al., 2024; Chen et al., 2025; Szymanowicz et al., 2024b;a; Wewer et al., 2024; Xu et al., 2025; Ye et al., 2024a; Hong et al., 2024; Tang et al., 2024), which can directly produce 3D representations from sparse inputs without per-instance optimization, have garnered significant attention. However, a fundamental challenge shared by these sparse-view approaches is their limited extrapolation capability, as they are unable to render unobserved regions.

Video Diffusion-based 3D Scene Generation. Recent video generation models (Brooks et al., 2024; Xing et al., 2024; Hong et al., 2022; Yang et al., 2024) have shown great potential to generate 3D-aware sequences. These models can naturally serve as 3D scene generators when camera poses are controllable (Guo et al., 2024; Wang et al., 2024b; Melas-Kyriazi et al., 2024; Voleti et al., 2024; Liang et al., 2024). To enhance 3D consistency, recent works such as ReconX (Liu et al., 2024a), ViewCrafter (Yu et al., 2024b) and VMem (Li et al., 2025) have integrated 3D geometric priors into their frameworks by leveraging reconstruction models such as DUSt3R (Wang et al., 2024a) and CUT3R (Wang et al., 2025b).

Image Diffusion-based 3D Scene Generation.

Several innovative investigations (Liu et al., 2023; Wu et al., 2024b; Sargent et al., 2024; Höllein et al., 2024; Seo et al., 2024; Shi et al., 2024; Wang & Shi, 2023; Shi et al., 2023; Liu et al., 2024e) have incorporated camera pose information into pre-trained T2I models to generate novel views. Within this category, two key strategies have emerged for generating explorable scenes from a single image. The first strategy employs **pose-conditioned view synthesis**. Methods such as SEVA (Zhou et al., 2025) and CAT3D (Gao et al., 2024) leverage camera pose information to guide the generation of novel views, demonstrating impressive scene-level results. However, when applied to single-image inputs over extended camera trajectories, these methods struggle to maintain long-range geometric consistency and visual coherence, often resulting in accumulated errors and semantic drift that compromise global scene structure. The second strategy relies on **iterative navigation and inpainting**(Pu et al., 2024; Chung et al., 2023; Yu et al., 2023; Höllein et al., 2023). One notable example, Pano2Room(Pu et al., 2024), builds the scene sequentially by navigating through space and inpainting unseen areas. Although it can produce plausible indoor results, this iterative framework is inherently prone to accumulating geometric and appearance errors over time, compromising global scene consistency. A second limitation is its design, which incorporates strong indoor priors that restrict its generalization to outdoor scenes and diverse visual styles.

In contrast to these sequential approaches, our One2Scene framework introduces a novel scaffold-guided paradigm. It decomposes the ill-posed single-image-to-scene problem into more manageable subtasks, achieving superior geometric fidelity and photorealistic quality. By first generating a globally consistent 3D scaffold in a single, feed-forward pass, our One2Scene method establishes a robust geometric and semantic foundation for the entire scene. This holistic global prior directly counteracts the error accumulation inherent in sequential methods like pose-conditioned synthesis and iterative inpainting. Consequently, our approach is not only more geometrically consistent but also significantly more general than specialized methods like Pano2Room, demonstrating superior performance across both indoor and outdoor environments.

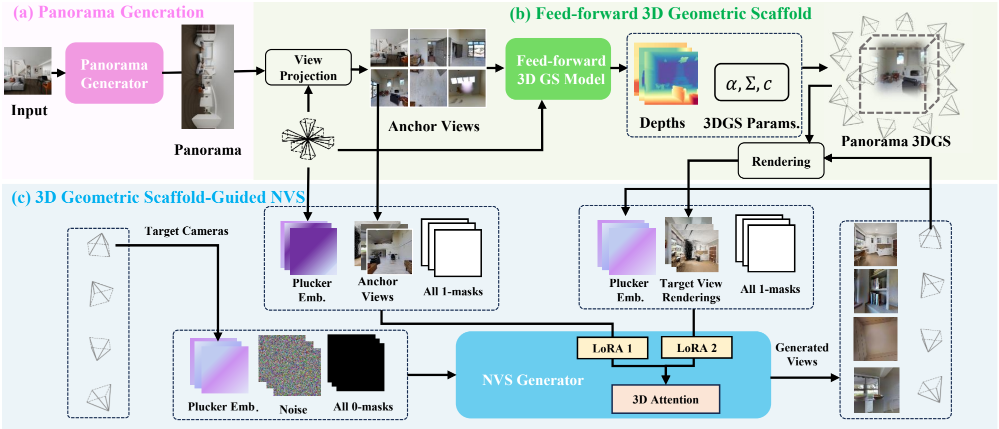

Figure 2: **Overview of One2Scene.** Our method consists of three stages: (a) an anchor view generation stage to establish an initial 360-degree representation, (b) a feed-forward 3D Gaussian Splatting stage to construct an explicit 3D geometric scaffold, and (c) a synthesis stage that leverages the scaffold information to produce high-quality novel views. The pipeline enables geometrically consistent and photorealistic novel view synthesis from a single input image.

3 Methodology

This section details our One2Scene framework, which can generate an explorable 3D scene from a single image by decomposing this ill-posed problem into a sequence of manageable sub-tasks, as illustrated in Figure 2. First, to overcome the severe information deficit, we generate a panorama to cover the global scene. Second, we obtain a set of anchor views from the panorama and introduce a feed-forward 3D GS model to lift these 2D anchor views into an explicit 3D geometric scaffold. Finally, with the strong geometric and appearance priors provided by the 3D scaffold, a synthesis network is used to generate photorealistic and consistent novel views from arbitrary camera poses.

3.1 Panorama Generation

Generating explorable 3D scenes from a single image is a highly challenging problem, often resulting in pronounced semantic drift and geometric inconsistency across long-range novel views. To address this challenge, we adopt a progressive approach that first expands visual information content and subsequently establishes a robust geometric foundation. We employ a specialized image-to-panorama generation model to transform the limited input view into a 360° panoramic representation. This representational choice is motivated by two primary considerations. First, the comprehensive field of view provides more visual cues that facilitate subsequent globally consistent scene generation. Second, compared to direct arbitrary novel view synthesis, panoramic image generation with a single image as input is a more well-posed computational task. In particular, we employ Hunyuan-Pano-DiT (Wang et al., 2025c), which demonstrates exceptional generalization capabilities acquired through training on extensive large-scale datasets, to generate the panoramic image.

3.2 Feed-forward 3D Geometric Scaffold

Although the panorama generated from the initial stage provides global coverage, it remains a 2D representation estimated from a single viewpoint and lacks explicit 3D information. Maintaining geometric consistency when synthesizing with large viewpoint changes and long sequences remains a fundamental challenge in explorable scene generation. To this end, we introduce a novel feed-forward 3D GS model to predict a set of 3D Gaussian parameters $(\boldsymbol{\mu}_i, \alpha_i, \boldsymbol{\Sigma}_i, \boldsymbol{c}_i), i = 1^{H \times W \times N}$ for each pixel in the generated panorama. This process provides the scene with explicit 3D information, thereby ensuring global geometric consistency.

Anchor View Projection. Accurate depth estimation is the cornerstone of this model, as inaccurate depth can introduce severe rendering artifacts. Although significant progress has been made in depth estimation from a single panoramic image (Ai et al., 2023; Wang & Liu, 2024; Pintore et al., 2023), this task remains highly challenging. A key difficulty lies in the lack of large-scale datasets comparable to those available for perspective images, limiting the generalization ability of panoramic depth estimators. To achieve robust depth estimation, we propose to reformulate the problem of monocular panoramic depth estimation as a multi-view stereo matching problem. Specifically, we first project the 360° panorama into a set of six perspective cubemap views, which serve as the input anchor views for our model. This strategy allows us to leverage powerful geometric priors learned from large-scale multi-view datasets. We choose to use cubemaps because they provide the most compact perspective representation of the panoramic scene, ensuring high efficiency. To facilitate correspondence matching across views, we expand each cubemap's Field of View (FoV) to 95°, creating a 2.5° overlap at adjacent view boundaries. For further details, please refer to appendix A.2.

Bidirectional Fusion Module. Although a 2.5-degree overlap is established between adjacent anchor views, the correspondence remains extremely sparse. Existing multi-view stereo models like VGGT (Wang et al., 2025a), which rely on substantial inter-view overlap, suffer from significant performance degradation in such scenarios. To address this limitation, we propose novel architectural modifications to VGGT to explicitly enforce cross-view consistency and improve the robustness of depth estimation. Specifically, we integrate a bidirectional fusion mechanism into the pre-trained DPT head of VGGT to promote cross-view depth consistency. This mechanism establishes geometric correspondence across views while preserving view-specific details.

To effectively handle overlapped regions, we introduce a Cube-to-Equirectangular (C2E) transformation module that projects the dense feature maps $\mathbf{F}_i$ from the six anchor views into a unified equirectangular latent. Subsequently, these equirectangular features are fused using a convolutional layer $\mathbf{H}_c$. Then, the fused features $\mathbf{F}_e$ are transformed back to the cubic space via an Equirectangular-to-Cube (E2C) module and merged with the original anchor view features through a residual connection. The finally updated feature for each view, $\mathbf{F}'_i$, is computed as follows:

$$\mathbf{F}_e = \mathbf{H}_c(\text{C2E}(\{\mathbf{F}_i\}_{i=1}^{6})), \quad \mathbf{F}'_i = \mathbf{F}_i + \text{E2C}(\mathbf{F}_e). \tag{1}$$

This bidirectional transformation and fusion mechanism aligns features in overlapped regions to achieve geometric consistency via C2E/E2C transformations, while using residual connections to maintain view-specific details simultaneously. For further details, please refer to appendix A.3.

Gaussian Parameter Prediction Heads. For each pixel, the Gaussian center $\boldsymbol{\mu}$ is computed by unprojecting the predicted depth into 3D space using the camera intrinsics: $\boldsymbol{\mu} = \mathbf{K}^{-1}\boldsymbol{u}d + \Delta$, where $\mathbf{K}$ denotes the camera intrinsic matrix, $\boldsymbol{u} = (u_x, u_y, 1)$ represents the pixel coordinates, and $\Delta \in \mathbb{R}^3$ indicates the predicted positional offset. To predict the remaining Gaussian parameters (opacity, covariance, and color), we employ an additional prediction head based on the DPT architecture. Following NoPosplat (Ye et al., 2024a), this prediction head takes both VGGT features and the RGB image as inputs. The direct pathway from RGB images complements VGGT's high-level semantic-focused features by preserving essential fine textural details.

Training. The feed-forward 3DGS model is trained using a composite loss function, which includes a rendering loss and a depth loss. The rendering loss is a combination of the Mean Squared Error (MSE) and the LPIPS perceptual loss (Johnson et al., 2016), while the depth loss is the Scale-Invariant Logarithmic (SILog) loss (Eigen et al., 2014). The model is trained on a collection of four datasets: two synthetic datasets, Structured3D (Zheng et al., 2020) and Deep360 (Li et al., 2022), and two real-world datasets, Matterport3D (Chang et al., 2017) and Stanford2D3D (Armeni et al., 2017). Through this training regimen, our feed-forward 3DGS model demonstrates precise geometric modeling capabilities and robust generalization across indoor, outdoor, and even stylized scenes.

3.3 3D Scaffold Guided Novel View Synthesis

In the final stage of our pipeline, we leverage the 3D geometric scaffold to generate a fully explorable 3D scene. In particular, we propose to transform the task of novel view synthesis from a single view to the problem of synthesis conditioned on the set of anchor views:

$$p\left(\mathbf{I}^{\text{tgt}} \mid \mathbf{I}^{\text{anchor}}, \mathbf{p}^{\text{anchor}}, \mathbf{p}^{\text{tgt}}\right). \tag{2}$$

However, the above formulation remains limited since the anchor views are all observations from a single point in the space, and they lack the explicit scale and geometric information required for robust 3D understanding. Our 3D geometric scaffold, with its precise geometric modeling capabilities, overcomes this limitation by enabling the rendering of novel views from arbitrary viewpoint. These rendered views contain rich geometric and appearance information. Therefore, they can serve as powerful conditions to guide the synthesis of novel views, significantly enhancing their realism and consistency. Although these rendered views may exhibit artifacts or occlusions (e.g., black holes) for large viewpoint changes, they still retain a substantial amount of useful structural information, owing to our model's accurate depth estimation. This insight allows us to further reformulate the synthesis problem as follows:

$$p\left(\mathbf{I}^{\text{tgt}} \mid \mathbf{I}^{\text{anchor}}, \mathbf{p}^{\text{anchor}}, \mathbf{I}^{\text{render}}, \mathbf{p}^{\text{tgt}}\right), \tag{3}$$

where view $\mathbf{I}^{\text{render}}$ is rendered from the scaffold in the camera pose of the target view $\mathbf{I}^{\text{tgt}}$.

Dual-LoRA Training. It is a challenging task to manage two distinct types of conditions in the synthesis process: the high-quality anchor views, which offer pristine appearance but are geometrically ambiguous, and the rendered views, which provide strong geometric priors but may contain artifacts. To effectively guide the synthesis using both conditions, we need to process these heterogeneous signals. Inspired by MMDiT (Esser et al., 2024), which uses separate encoders for different modalities, such as text and images, before fusing their features for self-attention, we propose a Dual-LoRA training strategy. Built upon the SEVA architecture (Zhou et al., 2025), our approach employs two different LoRA modules to process the anchor view and the rendered view independently, as shown in Figure 2 (c). The features from both conditions are then integrated with the noisy latent representation through a 3D attention mechanism. Our experiments confirm that this method demonstrates significantly stronger learning capabilities compared to a naive approach of simply concatenating the rendered view with the noise latent.

Memory Condition. To ensure temporal and spatial consistency when generating a large number of frames for a continuous 3D scene, we introduce an additional memory condition during inference. This condition is a previously generated frame selected from a memory bank, which has the closest average camera pose to the current target frame. The synthesis problem is thus further refined to:

$$p\left(\mathbf{I}^{\text{tgt}} \mid \mathbf{I}^{\text{anchor}}, \mathbf{p}^{\text{anchor}}, \mathbf{I}^{\text{render}}, \mathbf{I}^{\text{mem}}, \mathbf{p}^{\text{mem}}, \mathbf{p}^{\text{tgt}}\right). \tag{4}$$

This memory-guided approach effectively preserves visual consistency, particularly when synthesizing content in occluded regions.

Training Data Construction. To assemble a dataset for supervised training, we perform sparse 3D reconstructions on the DL3DV (Ling et al., 2023) and RealEstate10K (Zhou et al., 2018) datasets using the pre-trained feed-forward 3DGS model MVSplat (Chen et al., 2025). This strategy is intentionally employed to simulate the artifacts and holes that arise in rendered views when the reconstruction is based on sparse input viewpoints. By using the camera trajectories inherent to these datasets, we sample novel views that exhibit significant viewpoint deviations. Training pairs are subsequently formed, each comprising a ground truth image and its corresponding view rendered from the sparse 3D reconstruction at the identical camera pose.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation Details. In the panorama generation stage, we employ Hunyuan-Pano-DiT (Wang et al., 2025c) as the generator. The feed-forward 3DGS model is trained for 80,000 iterations using the AdamW optimizer. We set the learning rate of the VGGT backbone to 2e-5, and set the learning rate to 2e-4 for all other modules. In the final stage, the 3D scaffold-guided novel view synthesis model is trained for 40,000 iterations using the Adam optimizer based on SEVA (Zhou et al., 2025), with a batch size of 16 and a learning rate of 1.25e-5.

Experiments Setup. To more comprehensively evaluate our proposed One2Scene model and demonstrate its effectiveness and advantages, we conduct the following experiments. (1) First, we benchmark our One2Scene model against the SOTA 3D scene generation models in producing high-quality, explorable 3D scenes. (2) Second, we evaluate the key component of our One2Scene model, i.e., the feed-forward 360° reconstruction network, by comparing its quality, efficiency, and geometric accuracy with the SOTA methods. Its depth estimation performance is also evaluated on standard panorama depth estimation benchmarks. (3) Third, we conduct a series of ablation studies to dissect the effectiveness of our design of One2Scene.

Evaluation Metrics. We evaluate the quality of our generated scenes across three key dimensions. (1) Visual Fidelity. We measure visual quality using two no-reference image quality assessment metrics: NIQE (Mittal et al., 2012) and Q-Align (Wu et al., 2023). (2) Semantic Consistency. We measure the semantic consistency between the initial image and the novel views using CLIP-I score (Hessel et al., 2021). (3) Geometric Consistency. We evaluate geometric stability by first estimating the camera poses of the generated views with a pre-trained VGGT model. These estimated poses are then benchmarked against the ground-truth camera trajectories to compute Rotation Error (RotError) (He et al., 2024), Camera Motion Consistency (CamMC) (Wang et al., 2024b), and Translation Error (TransError) (He et al., 2024). More details of our evaluation protocol are provided in Appendix A.1.

4.2 Main Results

4.2.1 Explorable 3D Scene Generation

To establish a rigorous evaluation protocol in the absence of a standard benchmark for explorable 3D scene generation, we adapt the WorldScore benchmark (Duan et al., 2025), which is originally proposed for short-sequence 3D scene evaluation. To ensure a comprehensive assessment, we sample 40 scenes spanning four diverse static scene categories: indoor-real, indoor-stylized, outdoor-real, and outdoor-stylized (10 per category). This diverse benchmark allows us to thoroughly test the robustness and quality of the generated 3D scenes from single-view inputs.

Results. We compare One2Scene with DreamScene360 (Zhou et al., 2024), WonderJourney (Yu et al., 2023), VMem (Li et al., 2025) and SEVA (Zhou et al., 2025). Quantitative results are reported in Table 1. For methods that accept camera-conditioned novel view synthesis, we additionally evaluate geometric consistency. Since DreamScene360 and WonderJourney do not produce fully explorable scenes (as shown in Figure 1), we can only perform qualitative comparisons with VMem and SEVAin, as shown in Figure 3. We also condition VMem and SEVA on the anchor views produced in our One2Scene method, and denote the corresponding methods as VMem+ and SEVA+.

Semantic and Appearance Consistency. As demonstrated in Figure 3, SEVA and VMem often hallucinate content in unobserved regions, leading to semantic inconsistencies. Our 3D scaffold, however, preserves global semantic coherence. This advantage is validated by our quantitative results in Table 1: our One2Scene achieves superior NIQE (4.43) and Q-Align (4.13) scores, and its CLIP-I score (89.95) markedly surpasses those of SEVA (87.82) and VMem (75.80).

Scale Ambiguity and Drift. As noted by Zhou et al. (2025) in SEVA, the single input image makes SEVA suffer from scale ambiguity issues. This manifests the distortion of object size and physically implausible geometric artifacts, such as cameras penetrating through walls (see Figure 3). Even conditioned on our anchor views, SEVA+ and VMem+ remain unable to effectively resolve the scale drift problem. This fundamental limitation stems from the lack of relative translation information in anchor views, which prevents the model from inferring a unified global scale. In contrast, our method explicitly constructs a 3D scaffold that provides robust scale constraints, effectively mitigating the scale ambiguity issue and producing physically plausible results.

Geometric Stability. Existing methods often struggle to maintain long-term geometric stability. SEVA, for example, lacks a persistent geometric representation, causing inconsistent reconstructions in loop-closure scenarios (e.g., frame 78 vs. 255 in Figure 3). VMem attempts to enforce consistency via online reconstruction with CUT3R, but this strategy is highly susceptible to a vicious cycle of error accumulation: generated low-quality frames destroy the geometry, which in turn provide wrong guidance for subsequent frames, leading to catastrophic failure. In contrast, our pre-built 3D scaffold provides a stable geometric prior, effectively preventing error propagation. This advantage is substantiated by the quantitative results: our method achieves a score of 0.389 in CamMC, significantly outperforming VMem (0.998, see Table 1).

The above results highlight the superiority of our three-stage design of One2Scene, which systematically addresses the global semantic inconsistency, scale ambiguity, and geometric instability. More results can be found in Appendix A.6 and our anonymous project page.

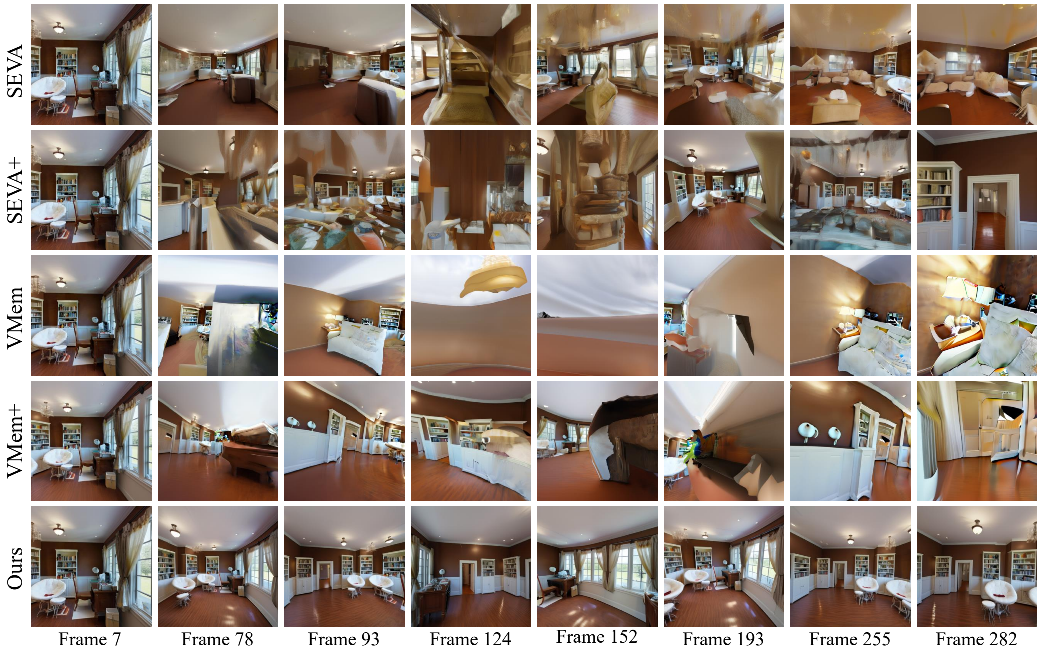

Figure 3: **Qualitative comparison.** Our method retains compelling visual quality and generates plausible continuations of the scene, even under large viewpoint change.

Table 1: Quantitative comparisons for 3D scene generation.

Methods	NIQE↓	Q-Align↑	CLIP-I↑	TransErr ↓	RotErr ↓	CamMC ↓
DreamScene360 (Zhou et al. (2024))	8.40	1.91	74.24	-	-	-
WonderJourney (Yu et al. (2023))	4.97	3.02	77.92	-	-	-
SEVA (Zhou et al. (2025))	4.53	3.20	87.82	0.460	0.165	0.558
SEVA (Zhou et al. (2025)) + Anchor	4.45	3.45	88.70	0.422	0.116	0.460
VMem (Li et al. (2025))	6.86	2.95	75.80	0.573	0.569	0.998
VMem (Li et al. (2025)) + Anchor	5.23	3.04	81.33	0.613	0.426	0.887
One2Scene (Ours)	**4.43**	**4.13**	**89.95**	**0.326**	**0.107**	**0.389**

4.2.2 FEED-FORWARD 360° RECONSTRUCTION

This section validates the core advantages of our feed-forward 3DGS network, a cornerstone of our pipeline. We demonstrate its superiority in reconstruction quality, computational efficiency, and geometric accuracy compared to SOTA methods.

Reconstruction Quality. We conduct a direct comparison with the SOTA method, AnySplat (Jiang et al., 2025). Since both methods are extensions of the VGGT model, this shared foundation ensures a fair evaluation. As shown in Figure 4, AnySplat's reconstruction fails with only 6 sparse views. This is because it predicts an erroneous depth map, which results in a distorted geometric scene. Even when 20 densely tangent patches with substantial overlap are projected from a panorama, its performance remains sub-par, suffering from severe artifacts in drastic viewpoint changes. In stark contrast, our model constructs a high-quality and robust 3D geometric scaffold even from sparse inputs. Although large rotations can introduce minor local artifacts due to occlusion, the underlying geometric foundation remains stable, providing crucial priors for the subsequent generation task. The importance of our scaffold is further confirmed by the experiment in Table 2: replacing our reconstruction module with AnySplat causes a significant degradation in final generation quality.

Table 2: Comparison on the 3D scene generation performance by replacing our feed-forward 360° reconstruction network with AnySplat.

Methods	NIQE↓	Q-Align↑	CLIP-I↑	TransErr ↓	RotErr ↓	CamMC ↓
AnySplat (Jiang et al., 2025)	4.96	3.61	81.96	0.332	0.367	0.616
Ours	**4.43**	**4.13**	**89.95**	**0.326**	**0.107**	**0.389**

Table 3: Comparison of depth estimation on Matterport3D and Stanford2D3D datasets.

Methods	Matterport3D				Stanford2D3D			
	AbsRel↓	δ_1↑	δ_2↑	δ_3↑	*AbsRel*↓	δ_1↑	δ_2↑	δ_3↑
BiFuse (Wang et al., 2020)	0.2048	84.52	93.19	96.32	0.1209	86.60	95.80	98.60
UniFuse (Jiang et al., 2021)	0.1063	88.97	96.23	98.31	0.1114	87.11	96.64	98.82
HoHoNet (Sun et al., 2020))	0.1488	87.86	95.19	97.71	0.1014	90.54	96.93	98.86
BiFuse++ (Wang et al., 2022)	–	87.90	95.17	97.72	–	87.83	96.49	98.84
ACDNet (Zhuang et al., 2022)	0.1010	90.00	96.78	98.76	0.0984	88.72	97.04	98.95
PanoFormer (Shen et al., 2022)	0.0904	88.16	96.61	98.78	0.1131	88.08	96.23	98.55
HRDFuse (Ai et al., 2023)	0.0967	91.62	96.69	98.44	0.0935	91.40	97.98	99.27
EGFormer (Yun et al., 2023)	0.1473	81.58	93.90	97.35	0.1528	81.85	93.38	97.36
Elite360D (Ai & Wang, 2024)	0.1115	88.15	96.46	98.74	0.1182	88.72	96.84	98.92
Depth Anywhere (Wang & Liu, 2024)	0.0850	91.70	97.60	99.10	0.1180	91.00	97.10	98.70
Ours (Zero-shot)	0.1070	88.97	96.51	98.61	0.0675	95.20	98.53	99.30
Ours (Finetune)	**0.0391**	**98.09**	**99.41**	**99.74**	**0.0444**	**96.95**	**98.85**	**99.44**

Computational Efficiency. Using six sparse views, our model reconstructs a high-quality scaffold in 0.5 seconds on an H20 GPU, marking a 5.6× speedup over AnySplat, which relies on a dense view set and requires 2.8 seconds. The inference time is further slashed to only 0.1 seconds when using a more powerful NVIDIA H100 GPU.

Accurate Depth Estimation. To quantitatively assess the geometric accuracy of our model, we evaluate its depth estimation performance against SOTA methods on the Matterport3D and Stanford2D3D datasets. As detailed in Table 3, the results are compelling: our model, when applied in a zero-shot setting, surpasses all compared approaches on the Stanford2D3D dataset. This result indicates that our method effectively inherits and transfers geometric priors from the foundational VGGT model. Furthermore, when our model is fine-tuned on the Matterport3D and Stanford2D3D datasets, it demonstrates exceptional performance, boosting the AbsRel metric by over 50%. This further underscores the powerful geometric modeling capabilities of our reconstruction model.

4.3 Ablations and Analysis

Given limited space, we provide comprehensive ablation studies in the Appendix, featuring in-depth analyses of our Dual-LoRA training methodology, memory condition mechanism, and bidirectional fusion module (see Appendix A.4). We also provide detailed quantitative evaluation results for our generation model on the DL3DV dataset (see Appendix A.5).

5 Conclusion and Limitations

In this paper, we introduced One2Scene, a novel and effective framework for generating fully explorable 3D scenes from a single image. We addressed the critical challenge of geometric distortion and artifact generation in existing methods when there were large viewpoint changes. Our core contribution lied in the decomposition of this ill-posed problem into three tractable subtasks: initializing sparse anchor views via a panorama generator, lifting them into an explicit and geometrically reliable 3D scaffold by a feed-forward GS network, and finally, leveraging the scaffold as a strong prior for photorealistic novel view synthesis. Our extensive experiments validated that One2Scene substantially outperformed state-of-the-art methods in explorable 3D scene generation.

Limitations. While our approach significantly improves 3D consistency across long sequences and large viewpoint changes, the generated views may contain subtle inconsistencies. Similar to CAT3D (Gao et al., 2024), we can further enhance geometric consistency through post-reconstruction processing. Please see the "Result Gallery" on our anonymous project page. In future work, we plan to construct larger-scale datasets to further improve our model's performance and robustness.

6 Ethics Statement

This research does not involve human participants or the collection of sensitive personal information. All datasets utilized in this study are employed in strict accordance with their respective licensing agreements and terms of use.

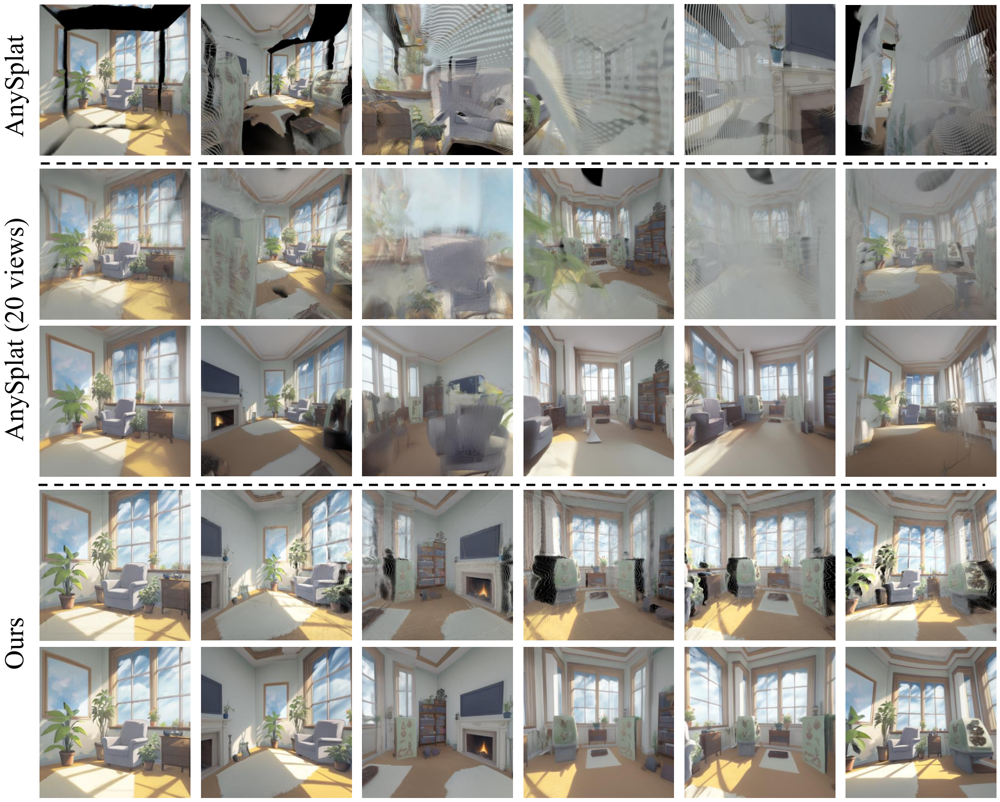

Figure 4: Ablation study on reconstruction performance. We compare the 3D scene generation quality by replacing our feedforward network with AnySplat. Top row: reconstruction results. Bottom row: generation results using our model.

The proposed methodology is designed exclusively for academic research and scientific advancement. While we do not anticipate direct harmful applications, we recognize the potential for misuse if deployed without appropriate ethical considerations and safety measures. We advocate for the responsible application of our research contributions, emphasizing the importance of fairness, transparency, and adherence to applicable legal frameworks.

7 Reproducibility Statement

We have implemented comprehensive measures to facilitate the reproducibility of our research findings. The main manuscript provides thorough documentation of our proposed framework, including detailed descriptions of the model architecture, dataset preprocessing methodologies, and algorithmic implementations. Complete hyperparameter configurations and training protocols are explicitly specified to enable independent replication of our results.

References

Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *CVPR*, 2024.

Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13273–13282, 2023.

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pp. 19457–19467, 2024.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS (NeurIPS)*, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, pp. 370–386. Springer, 2025.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *CoRR*, abs/2311.13384, 2023.

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

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989*, 2023.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5043–5052, 2024.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Yukun Huang, Yanning Zhou, Jianan Wang, Kaiyi Huang, and Xihui Liu. DreamCube: 3D Panorama Generation via Multi-plane Synchronization. 2025.

Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 6:1519–1526, 2021.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

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

Ming Li, Xueqian Jin, Xuejiao Hu, Jingzhao Dai, Sidan Du, and Yang Li. Mode: Multi-view omnidirectional depth estimation with 360 cameras. In *European Conference on Computer Vision*, pp. 197–213. Springer, 2022.

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

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. *arXiv preprint arXiv:2412.12091*, 2024.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *arXiv preprint arXiv:2312.16256*, 2023.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024a.

Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan. Make-your-3d: Fast and consistent subject-driven 3d content generation. *arXiv preprint arXiv:2403.09625*, 2024b.

Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *CVPR*, pp. 20763–20774, 2024c.

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

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *arXiv preprint arXiv:2410.16266*, 2024d.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024e.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pp. 7210–7219, 2021.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024.

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020.

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

Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. Deep scene synthesis of atlanta-world interiors from a single omnidirectional image. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4708–4718, 2023.

Guo Pu, Yiming Zhao, and Zhouhui Lian. Pano2room: Novel view synthesis from a single indoor panorama. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9420–9429, 2024.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 2024.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations*, 2024.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2573–2582, 2020.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion, 2024. URL `https://arxiv.org/abs/2411.04928`.

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024a.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, pp. 10208–10217, 2024b.

Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. Bolt3D: Generating 3D Scenes in Seconds. *arXiv:2503.14445*, 2025.

Shengji Tang, Weicai Ye, Peng Ye, Weihao Lin, Yang Zhou, Tao Chen, and Wanli Ouyang. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. *arXiv preprint arXiv:2410.06245*, 2024.

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

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *CVPR*, pp. 459–468. Computer Vision Foundation / IEEE, 2020.

Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5448–5460, 2022.

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, pp. 9065–9076, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Ning-Hsu Albert Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Advances in Neural Information Processing Systems*, 37:127739–127764, 2024.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024a.

Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025c.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024.

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

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. *arXiv preprint arXiv: 2503.01774*, 2025.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024a.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024b.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, pp. 399–417. Springer, 2024.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025.

Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, pp. 8254–8263, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024a.

Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. *arXiv preprint arXiv:2403.14613*, 2024b.

Hanyang Yu, Xiaoxiao Long, and Ping Tan. Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. *arXiv preprint arXiv:2409.03456*, 2024a.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. Wonderjourney: Going from anywhere to everywhere. *CoRR*, abs/2312.03884, 2023.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024b.

Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. *arXiv preprint arXiv:2304.07803*, 2023.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535. Springer, 2020.

Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv e-prints*, pp. arXiv–2503, 2025.

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

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph*, 37, 2018.

Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3653–3661, 2022.

A Appendix

We provide the following materials in this appendix:

- Appendix A.1: Detailed evaluation protocol.
- Appendix A.2: Details about cube projection.
- Appendix A.3: Details about bidirectional fusion module.
- Appendix A.4: Ablation study and analysis.
- Appendix A.5: More NVS results on DL3DV.
- Appendix A.6: More qualitative results.
- Appendix A.7: Declaration of LLM assistance.

A.1 Evaluation Protocol

To assess the quality of our generated scenes, we evaluate them across three key aspects: visual quality, input-output alignment, and geometric consistency.

For visual quality, we use two no-reference image quality assessment (NR-IQA) metrics. The first is NIQE (Mittal et al., 2012), where a lower score indicates that the image's statistics are more similar to a natural image. The second is Q-Align (Wu et al., 2023), a state-of-the-art model where a higher score signifies better perceptual quality.

For input-output alignment, we use the CLIP-I score (Hessel et al., 2021) to measure the semantic similarity between the generated images and the single input image. A higher score means the content and style are better preserved.

For geometric consistency, we evaluate how accurately the generated camera trajectory matches the ground truth. Our process is as follows: we sample a frame for every 10 frames from the generated sequence, estimate their camera poses using a pre-trained VGGT model (Wang et al., 2025a), and then compare these estimated poses to the ground-truth poses used for generation. This comparison is quantified using three metrics: RotError, TransError, and CamMC. To ensure a fair comparison, all methods are tested on the same set of camera trajectories, which are combinations of linear movements (move forward/backward/left/right) and curvilinear movements (orbit, lemniscate). These metrics are defined as follows:

RotError (He et al., 2024). It measures the average per-frame rotation error between the estimated rotation $\tilde{R}_i$ and the ground-truth rotation R_i:

$$\text{RotErr} = \frac{1}{n}\sum_{i=1}^{n} \arccos \frac{\operatorname{tr}(\tilde{R}_i R_i^{\mathrm{T}}) - 1}{2}.$$

TransError (He et al., 2024). It measures the average per-frame position error, calculated as the L2 distance between the estimated translation $\tilde{T}_i$ and the ground-truth translation T_i:

$$\text{TransErr} = \frac{1}{n}\sum_{i=1}^{n} \left\| \tilde{T}_i - T_i \right\|_2.$$

CamMC (Wang et al., 2024b). It provides a single score for the average overall pose error by computing the Frobenius norm of the difference between the estimated and ground-truth 3x4 pose matrices:

$$\text{CamMC} = \frac{1}{n}\sum_{i=1}^{n} \left\| [\tilde{R}_i|\tilde{T}_i] - [R_i|T_i] \right\|_F.$$

For all geometric error metrics, a lower value indicates better performance.

A.2 Details about Cube Projection

For equirectangular to cube (E2C) projection, the field-of-view (FoV) of each cube face is equal to 90 degrees; each face can be considered as a perspective camera whose focal length is $w/2$, and all faces share the same center point in the world coordinate. Since the six cube faces share the same center point, the extrinsic matrix of each camera can be defined by a rotation matrix R_i. p is then the pixel on the cube face:

$$p = K \cdot R_i^T \cdot q, \tag{5}$$

where

$$q = \begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} sin(\theta) \cdot \cos(\phi) \\ \sin(\phi) \\ \cos\theta \cdot \cos\phi \end{bmatrix}, K = \begin{bmatrix} w/2 & 0 & w/2 \\ 0 & w/2 & w/2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

where θ and ϕ are longitude and latitude in equirectangular projection and q is the position in Euclidean space coordinates.

While the 90° FoV model is mathematically exact for a perfect cube, it can introduce rendering artifacts at the seams between adjacent faces. To resolve this, we expand the field-of-view slightly, for instance to 95°. This modification ensures that each cube face captures a small, overlapped region from its neighbors. The projection methodology remains the same, but the camera's intrinsic matrix must be recalculated.

The relationship between focal length f, image width w, and FoV is given by $f = (w/2)/\tan(\text{FoV}/2)$. For a 95° FoV, the new focal length, denoted by f', is:

$$f' = \frac{w/2}{\tan(95°/2)} = \frac{w/2}{\tan(47.5°)}. \tag{7}$$

This results in a modified intrinsic matrix, K', where the focal length term $w/2$ is replaced by f':

$$K' = \begin{bmatrix} \frac{w/2}{\tan(47.5°)} & 0 & w/2 \\ 0 & \frac{w/2}{\tan(47.5°)} & w/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{8}$$

The final projection equation using the improved model is:

$$p = K' \cdot R_i^T \cdot q. \tag{9}$$

This adjustment, while minor, is critical for producing high-quality, artifact-free cubemaps suitable for production rendering environments. The definitions of q and R_i remain unchanged.

The inverse transformation, Cube to Equirectangular (C2E) projection, which is used to project features from the cube faces back to the panoramic view, is achieved by mathematically reversing this projection process. This robust projection method is essential for the bidirectional feature exchange in our model.

A.3 Details about Bidirectional Fusion Module

The performance of traditional multi-view models, such as VGGT that relies on dense overlap, degrades significantly when faced with extremely sparse correspondences resulting from a mere 2.5-degree overlap between anchor views. To address this issue, we introduce an innovative modification to the VGGT architecture, which aims to explicitly enhance cross-view consistency, thereby improving the robustness of depth estimation. Specifically, we integrate a Bidirectional Fusion Module into the pre-trained DPT head to promote cross-view depth consistency. The core principle of this module is to establish geometric correspondences across views while preserving the unique, high-fidelity details inherent to each individual view.

The module commences with the feature maps $\{\mathbf{F}_i\}_{i=1}^6$ extracted from the six anchor views. To effectively process the overlapping regions, we first introduce a C2E transformation module. As detailed in Appendix A.2, the C2E transformation leverages strict geometric projection principles to seamlessly project and aggregate the features from the six discrete cube views into a unified equirectangular latent space via differentiable bilinear sampling.

Subsequently, a lightweight convolutional layer, $\mathbf{H}_c$, is applied to this aggregated global feature map. Its purpose is to smooth the boundaries between the projected views and fuse their information, forming a globally consistent feature representation, $\mathbf{F}_e$. This step can be conceptualized as a process that information from all views is aggregated to build a consensus representation. This forward fusion process is formulated as:

$$\mathbf{F}_e = \mathbf{H}_c(\text{C2E}(\{\mathbf{F}_i\}_{i=1}^{6})). \tag{10}$$

Next, to propagate this global consistency information back to each individual view, we perform an inverse process. Through an E2C transformation, the fused global feature $\mathbf{F}_e$ is re-projected into the coordinate spaces of the six original anchor views.

Finally and crucially, rather than directly replacing the original features with this global information, we employ a residual connection to add it to the original feature map $\mathbf{F}_i$, yielding the updated view-specific feature $\mathbf{F}'_i$:

$$\mathbf{F}'_i = \mathbf{F}_i + \text{E2C}(\mathbf{F}_e). \tag{11}$$

The elegance of this “local-to-global-to-local” bidirectional mechanism lies in its dual function: the C2E/E2C transformations are responsible for aligning features in overlapping regions to enforce geometric consistency, while the residual connection ensures that the model retains and utilizes the original, high-fidelity details from each view. In this manner, our module effectively strengthens cross-view constraints while preventing the loss of view-specific information that can occur with forced fusion.

A.4 Ablation Study and Analysis

Effectiveness of Dual-LoRA Training. We first compare our Dual-LoRA training against the common channel-wise concatenation method. As shown in Figure A1, our model exhibits superior generation quality, no matter with and without the memory condition. This is because our Dual-LoRA approach can better leverage the two conditions of varying quality. The results in Table A1 further confirm that Dual-LoRA achieves better visual quality and geometric consistency.

Effectiveness of Memory Condition. We then analyze the impact of incorporating an additional memory condition at inference time. Although the quantitative results in Table A1 do not show a significant improvement, we observe a clear qualitative benefit. As highlighted by the colored boxes in Figure A1, this condition helps our model maintain better multi-view consistency, especially in occluded regions requiring significant content synthesis.

Effectiveness of Bidirectional Fusion Module. Our baseline approach directly applies VGGT for multi-view consistent depth estimation. However, due to the extremely sparse overlap between anchor views in panoramic scenarios, VGGT struggles to handle such conditions, resulting in significant performance degradation compared to geometric estimation tasks with larger overlaps. We fine-tune VGGT on panoramic images without any architectural modifications, which leads to noticeable performance improvements but still exhibits seaming artifacts at view boundaries.

Our proposed Bidirectional Fusion (BF) module substantially alleviates the geometric inconsistencies at edges. The BF module leverages complementary Cubemap-to-Equirectangular (C2E) and Equirectangular-to-Cubemap (E2C) transformations to establish robust geometric correspondences through residual connections. This bidirectional information flow enables the model to better handle the sparse overlap challenge inherent in panoramic depth estimation. As demonstrated in Table A2, the integration of the BF module yields significant performance improvements across both datasets, with notable gains in accuracy metrics such as reduced AbsRel error and increased δ_1, δ_2 and δ_3, confirming the effectiveness of our approach in addressing multi-view consistency challenges in panoramic depth estimation.

A.5 NVS Results on DL3DV

Competing Method. Our primary competing method is MVSplat360 (Chen et al., 2024), a state-of-the-art method capable of refining rendered views. To ensure a direct and fair comparison, we strictly

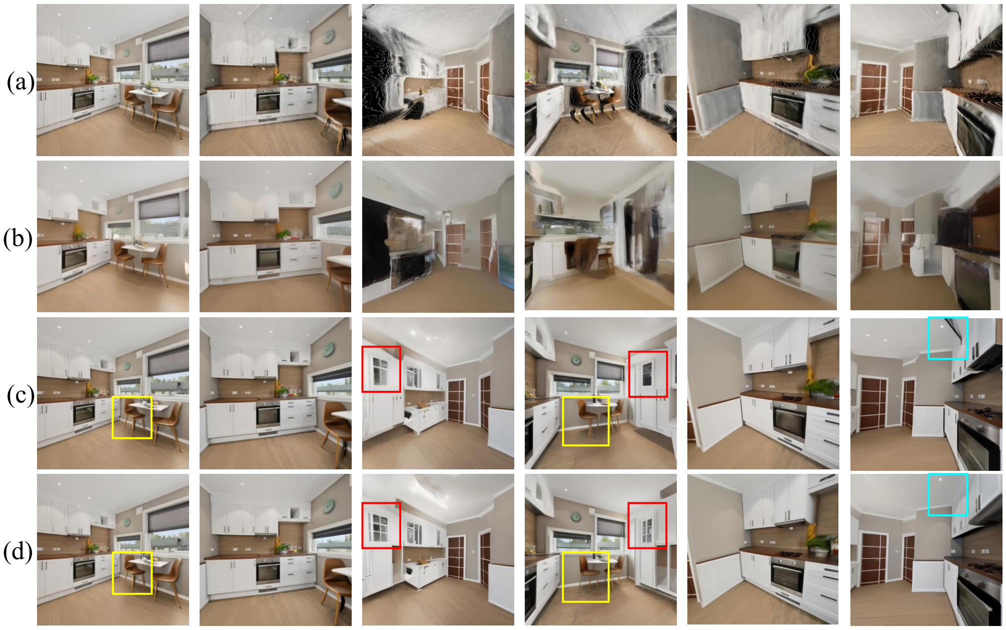

Figure A1: Qualitative comparison for the ablation study. (a) Render views from our 3D scaffold. (b) Naive concatenation baseline. (c) Ours (Dual-LoRA training only). (d) Ours (Full model with memory condition).

Table A1: Ablation study on 3D scaffold guided novel view synthesis.

Methods	NIQE↓	Q-Align↑	CLIP-I↑	TransErr ↓	RotErr ↓	CamMC ↓
Naive Concat.	5.04	3.41	85.30	0.481	0.260	0.655
Dual-LoRA Training	4.42	4.10	89.51	0.326	0.119	0.401
+ Memory Condition	4.43	4.13	89.95	0.326	0.107	0.389

Table A2: Effectiveness of BF module. Zero-shot quantitative comparison on Matterport3D and Stanford2D3D datasets.

Methods	Matterport3D				Stanford2D3D			
	AbsRel↓	δ_1↑	δ_2↑	δ_3↑	*AbsRel*↓	δ_1↑	δ_2↑	δ_3↑
Baseline	0.1576	78.82	93.20	96.15	0.1497	81.99	93.53	97.88
w/o BF	0.1204	86.28	95.36	97.45	0.0797	94.31	97.42	98.85
w BF	0.1070	88.97	96.51	98.61	0.0675	95.20	98.53	99.30

adhere to the evaluation protocol established for the DL3DV (Ling et al., 2023) dataset, as utilized by the competing method.

Quantitative Results. As detailed in Table A3, our method demonstrates superior performance over MVSplat360 across all evaluation metrics. Specifically, our method achieves a PSNR of 17.35 (+0.98) and an FID of 116.84 (-1.48). Furthermore, we observe substantial reductions in both LPIPS (0.343) and DIST (0.181) indices, indicating superior perceptual similarity and geometric accuracy, respectively. Collectively, these quantitative improvements underscore our method's enhanced effectiveness in leveraging auxiliary views to synthesize more accurate and high-fidelity novel views.

Qualitative Results. The qualitative comparisons presented in Figure A2 visually corroborate our quantitative findings. Our method consistently generates sharper and more structurally coherent scenes, showcasing an effective use of information from auxiliary views. In contrast, the results from MVSplat360 frequently exhibit noticeable artifacts and structural distortions, particularly when synthesizing views with large camera pose changes.

Table A3: The NVS numerical comparison on the DL3DV (Ling et al., 2023) dataset.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)	DIST (↓)	FID (↓)
PixelSplat	15.32	0.422	0.517	0.374	139.75
MVSplat	15.94	0.441	0.459	0.282	73.91
MVSplat360	16.37	0.453	0.439	0.238	18.32
Ours	**17.35**	**0.506**	**0.343**	**0.181**	**16.84**

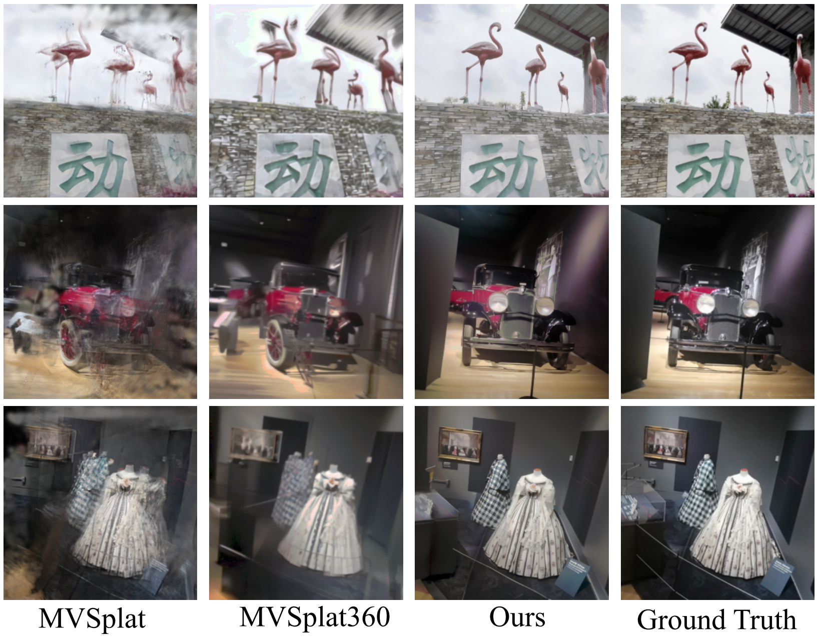

Figure A2: Visual comparison with existing SOTA methods on DL3DV.

A.6 More Qualitative Results

In this section, we provide more qualitative results to further support the claims presented in the main paper. We showcase a broader range of visual comparisons against baseline methods across diverse and challenging scenes, including indoor, outdoor, and stylized scenes. These examples serve to visually corroborate the quantitative improvements reported in the main paper, highlighting our method's superior performance in generating explorable 3D scenes.

We present side-by-side visualizations to compare our method, One2Scene, against key competitors: VMem and SEVA. Consistent with the main paper, we also include results for their '+' variants (VMem+ and SEVA+), which are conditioned on our generated anchor views. These comparisons, as shown from Figure A3 to Figure A7, further demonstrate the superior performance of our method in terms of visual fidelity, 3D geometric consistency, and the effective mitigation of scale ambiguity artifacts in previous methods.

A.7 Declaration of Generative AI Assistance

During the preparation of this manuscript, we utilized Gemini-2.5-Pro to assist in improving its linguistic quality. Specifically, after completing the initial draft, we provided the model with selected passages to obtain suggestions for grammar, clarity, and conciseness. All AI-assisted revisions were rigorously reviewed and edited by the authors, who assume full responsibility for the final accuracy and scholarly appropriateness of the content.

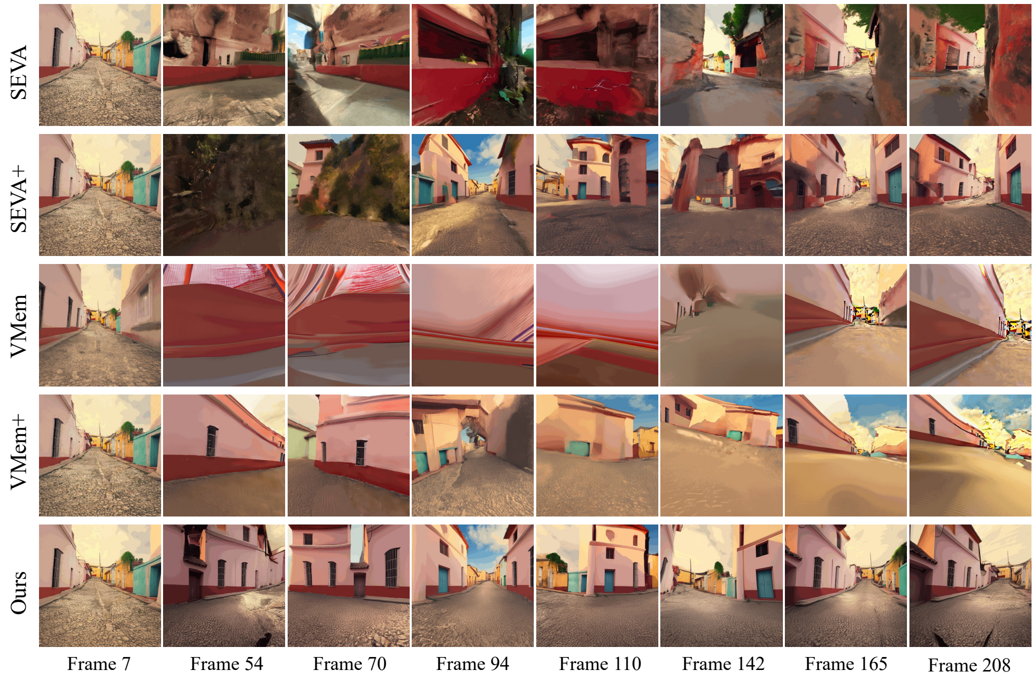

Figure A3: Qualitative comparison between One2Scene and SOTA methods.

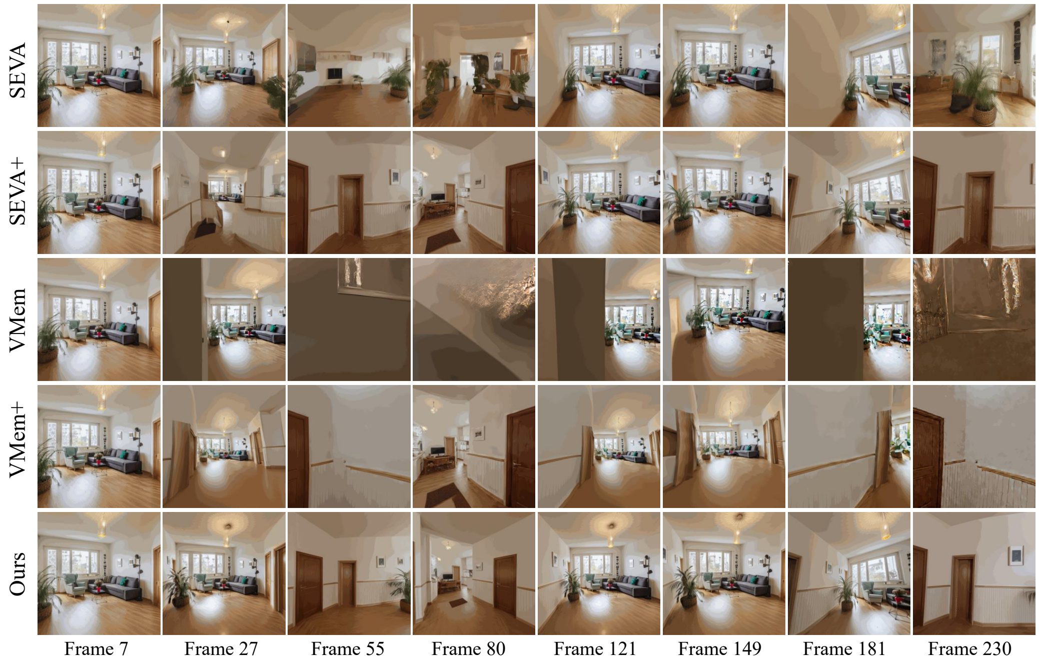

Figure A4: Qualitative comparison between One2Scene and SOTA methods.

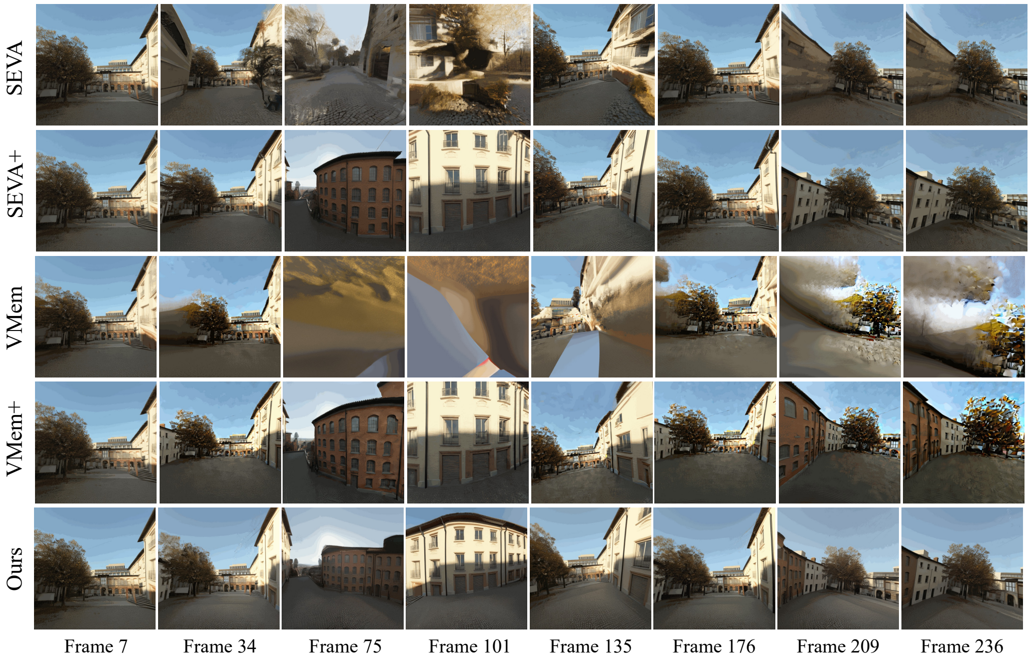

Figure A5: Qualitative comparison between One2Scene and SOTA methods.

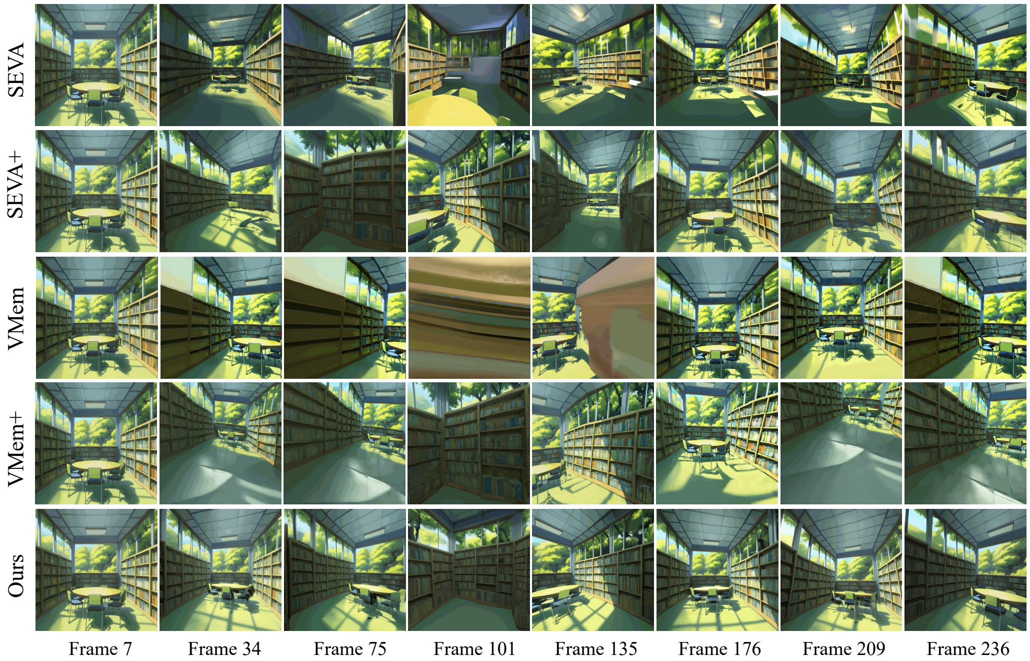

Figure A6: Qualitative comparison between One2Scene and SOTA methods.

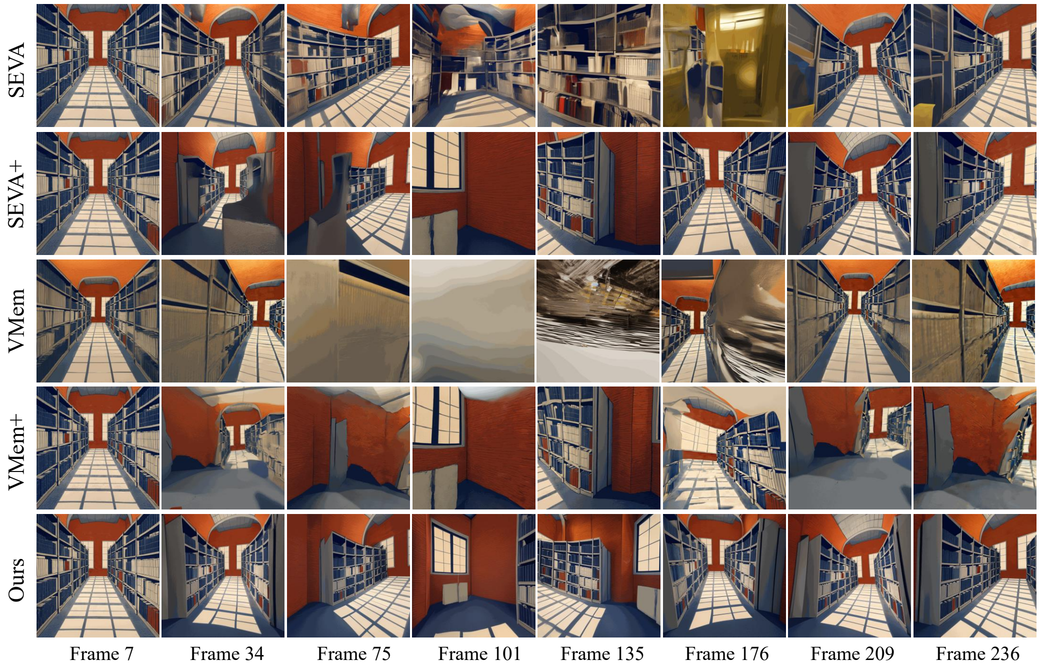

Figure A7: Qualitative comparison between One2Scene and SOTA methods.