# SplatFusion: Training-Free 3D Scene Completion from Sparse Views Using Temporal Diffusion Priors and Gaussian Splatting

Tanveer Younis[1] Dawar Khan[2] Zhanglin Cheng[1*]
younis@siat.ac.cn dawar.khan@kaust.edu.sa zl.cheng@siat.ac.cn
[1]Shenzhen VisuCA Key Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]King Abdullah University of Science and Technology, Saudi Arabia

## Abstract

*Reconstructing complete 3D scenes from extremely sparse viewpoints (e.g., 2–3 wide-baseline images) remains a core yet unsolved challenge. Existing 3D Gaussian Splatting (3DGS) and neural rendering methods degrade severely when view overlap is limited, often producing incomplete or geometrically distorted results. We introduce SplatFusion, a reconstruction framework that requires no training or fine-tuning of diffusion models, instead relying solely on pretrained video diffusion priors to synthesize missing scene content plausibly. Our core idea is a Scene-Consistent Temporal Guidance (SCTG) mechanism that tightly couples 3D structure with generative diffusion models. Specifically, SCTG conditions video diffusion on sequences rendered from the evolving 3DGS representation, enforcing both spatial alignment with geometry and temporal coherence across synthesized frames. These refined views are back-projected to densify and correct the 3D scene iteratively. Extensive experiments on diverse real-world datasets demonstrate that SplatFusion consistently outperforms existing sparse-view reconstruction methods. Evaluations using VLM-based perceptual scores and the MEt3R metric for geometric consistency show clear gains in visual fidelity and temporal coherence, even in scenarios where previous approaches fail. Our training-free framework opens new possibilities for practical 3D reconstruction applications where dense view acquisition is impractical.*

## 1. Introduction

Recent breakthroughs in neural 3D scene representation, particularly through Neural Radiance Fields (NeRF) [3–5, 14, 27, 28, 41, 43] and 3D Gaussian Splatting (3DGS) [13, 16, 21–24, 49], have revolutionized novel view synthesis and 3D reconstruction. Although these methods excel under dense-view conditions, the fundamental challenge of reconstructing accurate 3D scenes from sparse viewpoints remains largely unsolved, which is a critical requirement for practical applications. Sparse-view reconstruction becomes especially difficult under *extreme sparsity* when using only 2–3 input images with wide baselines (30°–60°) and minimal overlap. In this setting, structure-from-motion fails to recover correspondences, neural rendering methods exhibit severe view-dependent artifacts, and geometric priors alone cannot resolve the highly ill-posed reconstruction task.

Current state-of-the-art approaches, including pose-free methods such as InstantSplat [7], CoR-GS [48], and optimization-based techniques such as FSGS [49] and DNGaussian [16], demonstrate promising results under moderate sparsity but fail catastrophically when confronted with extreme baseline separations and minimal overlap, precisely the conditions most relevant for practical deployment. The core challenge lies in the fundamental shape-radiance ambiguity, where insufficient constraints lead to reconstructions with significant geometric inaccuracies in the unobserved regions. Without sufficient view coverage, reconstruction methods struggle to separate the true surface geometry from appearance or lighting changes, often producing results that appear plausible from training viewpoints but contain large geometric errors in unseen regions. In wide-baseline settings, the absence of reliable correspondences further limits geometric constraints, forcing methods to rely heavily on learned priors that may fail to generalize across diverse scenes.

To address these fundamental limitations, we propose SplatFusion, a method that uses pretrained diffusion models without any training or fine-tuning, and that to our knowledge is the first to complete 3D scenes from only 2–3 wide-baseline input images under extremely sparse-view conditions. Note that we define "training-free" as avoiding the expensive fine-tuning of the diffusion backbone or train-

*Corresponding author

ing a feed-forward generalizable model. However, consistent with the nature of 3D Gaussian Splatting, our method still involves per-scene optimization of the Gaussian primitives. Unlike existing methods that rely on dense supervision or scene-specific training, our approach leverages rich generative priors embedded in pretrained video diffusion models to hallucinate plausible scene content in unobserved regions while maintaining strict geometric and temporal consistency. Central to our contribution is the Scene-Consistent Temporal Guidance (SCTG) mechanism, which bridges the gap between 2D generative modeling and 3D geometric reasoning by conditioning the diffusion process on rendered sequences from an evolving 3DGS representation. Our framework operates through an iterative refinement process that alternates between diffusion-based view synthesis and 3D geometry updating. Starting from a coarse 3DGS initialization obtained using geometric priors from recent correspondence models [37], we generate temporally coherent novel views through guided diffusion, ensuring that each frame maintains spatial alignment with the current 3D scene structure while benefiting from the hallucination capabilities of pretrained generative models. These refined views are then back-projected into a 3D representation, enabling progressive geometry densification and quality improvement without requiring scene-specific training.

The key insight driving our approach is that, while individual 2D diffusion models may lack explicit 3D awareness, their temporal variants can be effectively guided to maintain geometric consistency across time through carefully designed conditioning mechanisms. By treating the sparse-view reconstruction problem as a temporally coherent view synthesis task, we leveraged the powerful generative capabilities of video diffusion models while constraining their outputs to remain geometrically plausible through our novel guidance strategy. Our contributions are summarized as follows:

- This paper presents the first training-free method for extreme sparse-view 3D scene reconstruction, handling as few as 2–3 wide-baseline images with minimal overlap.
- We introduce a novel SCTG mechanism that enforces spatial and temporal consistency by conditioning a diffusion model on rendered sequences from an optimized 3DGS representation.
- We propose a novel 3D-aware guidance loss that uses a visibility mask to fuse 2D generative priors with an evolving 3DGS representation, enabling robust structural densification in sparse-view settings.
- Our method demonstrates significant improvements over baselines on challenging benchmarks (Tanks and Temples, MipNeRF360, DL3DV-10k), achieving higher perceptual quality and geometric accuracy, particularly under extreme sparsity where prior approaches fail.

## 2. Related Work

Our work intersects three main areas of research: sparse-view 3D reconstruction with 3DGS, diffusion-based view synthesis, and training-free guidance mechanisms for generative models.

### 2.1. Sparse-view 3D Reconstruction and 3DGS

Neural rendering methods such as NeRF [27] and 3D Gaussian Splatting (3DGS) [13] offer efficient view synthesis but degrade under sparse-view conditions. To address this, pose-free methods such as InstantSplat [7], CoR-GS [48], and FSGS [49] have been proposed. However, these methods still yield incomplete geometry and appearance inconsistencies under extreme sparsity. DNGaussian [16] adds depth normalization but cannot hallucinate unseen content. Our approach leverages pretrained video diffusion priors with scene-consistent temporal supervision to complete missing regions and improve fidelity without dense supervision or scene-specific training.

### 2.2. Diffusion-based View Synthesis

Diffusion priors are increasingly used for 3D reconstruction [29], yet methods like ReconFusion [39], CAT3D [10], and ReconX [18] typically rely on scene-specific fine-tuning, limiting zero-shot use. Likewise, post-processing pipelines [22, 38] and object-centric methods [40] often require optimization or denser inputs. In contrast, our method is fully training-free and tailored for sparse, wide-baseline, unbounded scenes. Its key component, Scene-Consistent Temporal Guidance (SCTG), integrates 2D generative priors with optimized 3DGS geometry to enforce spatial and temporal coherence, enabling iterative geometry densification and improved robustness to missing data.

### 2.3. Training-free Guidance for Diffusion Models

Diffusion models can be guided without retraining using external signals [2, 33, 42, 45] to improve spatial and temporal coherence. Building on this, we introduce Scene-Consistent Temporal Guidance (SCTG), which extends this paradigm to sparse-view 3D reconstruction by combining pretrained video diffusion priors with explicit 3DGS renderings for scene-consistent supervision. Training-free methods, including optimization-based approaches like Universal Guidance [2], Loss-Guided Diffusion [33], and FreeDoM [45], and structured methods like Semantic Diffusion Guidance [19] and Structured Diffusion [8], offer a general way to guide models toward new goals without retraining. These methods operate as gradient-based plug-in controllers, where pretrained discriminative models provide guidance signals [31]. SCTG uses 3DGS renderings to define spatiotemporal losses at inference, providing a robust, training-free signal for geometry correction under extreme sparsity.

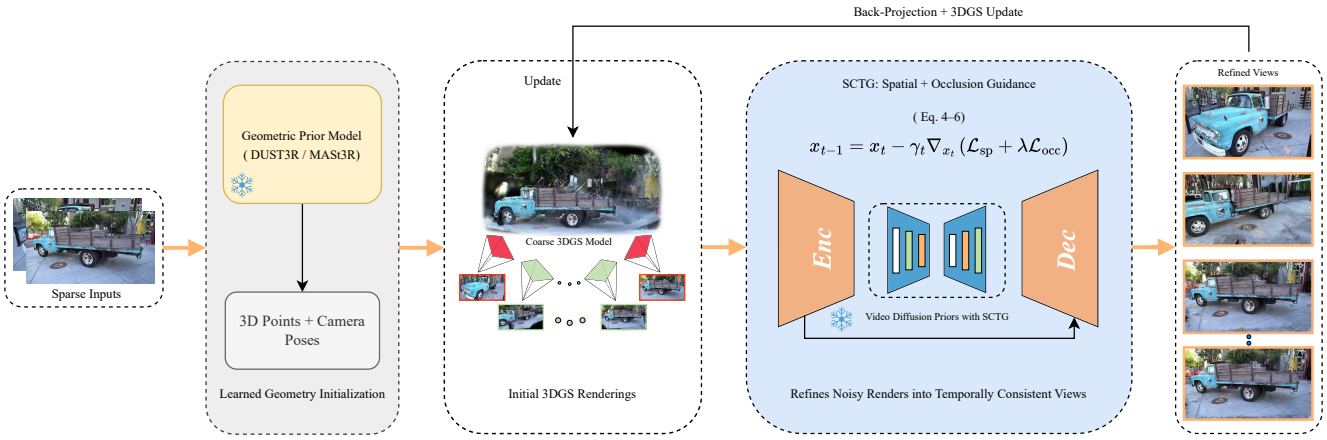


Figure 1. Overview of our *SplatFusion* method. We begin with a coarse 3DGS representation initialized using geometric priors from DUST3R/MASt3R [37] and camera poses from 2-3 unposed inputs. To address artifacts and inconsistencies from sparse data, our Scene-Consistent Temporal Guidance (SCTG) module guides a pretrained video diffusion model to generate spatially and temporally consistent novel views (see Algorithm 2). These refined views are then back-projected into the 3DGS scene, enabling iterative geometry densification and correction. The complete algorithmic workflow is presented in Algorithm 1.

## 3. Method

**Problem Statement.** Given $N \in \{2, 3\}$ input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ of a static scene with unknown camera parameters $\Theta = \{(R_i, t_i, K_i)\}_{i=1}^{N}$, the goal of *SplatFusion* is to reconstruct a complete *3DGS* representation $\mathcal{G} = \{g_m\}_{m=1}^{M}$ and refined camera parameters $\Theta$ for photorealistic novel view synthesis with spatial and temporal consistency. We began by leveraging geometric priors and pose propagation techniques [37] to estimate the initial camera poses. These poses are then used for coarse 3DGS initialization, following a geometric-prior-based approach for sparse inputs [7].

### 3.1. Method Overview

The overall *SplatFusion* pipeline is illustrated in Fig. 1 and Algorithm 1. SplatFusion is a training-free framework for sparse-view 3D reconstruction that leverages pretrained video diffusion priors to refine and complete 3DGS representations. Beginning with a coarse 3DGS scene initialized from sparse, unposed input images, our method addresses the resulting gaps and artifacts with a novel Scene-Consistent Temporal Guidance (SCTG) mechanism(see Fig. 2). The SCTG uses a pretrained video diffusion model to refine the rendered views in a spatially and temporally consistent manner. These refined frames are then back-projected into the 3DGS representation, improving the geometry and scene coverage in an iterative loop. This training-free approach, which requires no scene-specific fine-tuning of the diffusion model, tightly integrates diffusion-based refinement with 3D consistency, allowing the scene to evolve coherently, even under severe input sparsity.

### 3.2. Diffusion for Image and Video Generation

Diffusion models [11, 12] have emerged as powerful generative frameworks for images and videos by progressively corrupting data with Gaussian noise and learning to reverse this process through iterative denoising. The widely used Denoising Diffusion Probabilistic Model trains a U-Net denoiser $\epsilon_\theta$ to predict the noise at each timestep $t$. In the forward diffusion process, a clean sample $x_0$ is gradually transformed into a noisy version $x_t$ as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative product of the noise schedule. The model is optimized by minimizing the simplified variational loss:

$$L_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2}$$

At inference, denoising begins from pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and proceeds iteratively using an approximation of the reverse process as follows:

$$x_{t-1} = \left(1 + \frac{\beta_t}{2}\right) x_t + \beta_t \nabla_{x_t} \log p(x_t) + \sqrt{\beta_t} z, \quad z \sim \mathcal{N}(0, I), \tag{3}$$

where $\nabla_{x_t} \log p(x_t)$ represents the score function, that is, the gradient of the log-density of the noisy sample $x_t$, which is not explicitly known but is approximated by the neural network model trained to predict the noise component at each diffusion timestep [34]. This approximation guides the reverse denoising process by indicating how to adjust $x_t$ towards regions of higher data likelihood. Inspired by training-free guidance strategies [2, 33], external losses can be incorporated during reverse denoising to steer the generation toward the desired conditions. Our method leverages

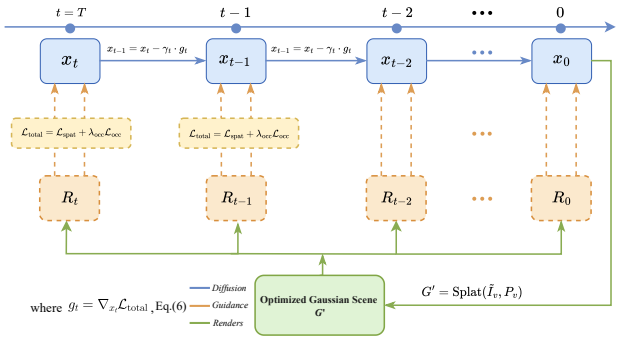


Figure 2. Overview of the proposed Scene-Consistent Temporal Guidance (SCTG) strategy. At each diffusion step $t$, the current image sample $x_t$ is aligned with the scene geometry by comparing it with a reference view $R_t$. Here, $R_t$ represents the view rendered from the current state of the optimized Gaussian scene $\mathcal{G}$ at the target pose $P_v$. The total guidance loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spatial}} + \lambda_{\text{occ}}\mathcal{L}_{\text{occ}}$, where $\mathcal{L}_{\text{spatial}}$ is a perceptual loss based on VGG features, and $\mathcal{L}_{\text{occ}}$ is an L2 loss weighted by visibility masks. This loss yields a gradient $g_t$, which is used in the update rule $x_{t-1} = x_t - \gamma_t \cdot g_t$ (Eq. (6)) to denoise the image in the next step. Once the final output $x_0$ is generated, high-confidence pixels are back-projected into the 3D scene via $\mathcal{G}' = \text{Splat}(\tilde{I}_v, P_v)$ (Eq. (9)).

a camera-conditioned image-to-video diffusion model [46], which operates in a latent space $\mathbb{R}^{L\times h\times w\times d}$, where $L$ is the video length, to generate temporally coherent and view-consistent sequences. This design is critical for the effective refinement of sparse-view 3D reconstructions by ensuring spatial and temporal coherence across the generated frames.

### 3.3. Scene-Consistent Temporal Guidance (SCTG)

Traditional 3DGS rendering under sparse views often produces structurally inconsistent outputs in unobserved regions. Although pretrained video diffusion models can hallucinate plausible content, they lack explicit scene awareness and geometric grounding. To address this, we introduced the SCTG mechanism, which injects a scene-specific structure into the generative process by guiding denoising using losses computed against 3D ground-truth scene (3DGS) renderings. This design allows the diffusion model to refine novel views in a way that aligns with the evolving 3D geometry while preserving a high perceptual quality. To enforce spatial alignment with the scene structure and maintain temporal consistency across generated frames, SCTG operates in the temporally conditioned latent space of a pretrained video diffusion model (see Fig. 2). At each time step, it leverages rendered reference frames from the current 3DGS model to compute external loss gradients that guide the denoising process toward structure-aware outputs.

Here, $x_t$ denotes the current noisy latent representation at time $t$, and $R_t$ is the corresponding rendered reference frame from the 3DGS model. The SCTG introduces a spatial consistency loss defined as:

$$L_{\text{spatial}} = \|f_{\text{feat}}(x_t) - f_{\text{feat}}(R_t)\|_2^2, \tag{4}$$

where $f_{\text{feat}}(\cdot)$ extracts perceptual features using the pretrained VGG network [32]. We use VGG-based perceptual features in our spatial consistency loss, as they capture higher-level structural and textural cues compared to simple pixel-wise $L_2$ differences. This loss encourages the structural alignment of the generated frame with the perceived structure of the 3DGS rendering represented by $R_t$, which is more robust for 3D alignment than low-level pixel losses. Because some regions in $R_t$ may be occluded or contain artifacts, we incorporate an occlusion-aware mask $M_t$, obtained via visibility estimation in 3DGS rendering. This mask restricts consistency enforcement to valid regions, and the occlusion loss is formulated as:

$$L_{\text{occ}} = \sum_{i,j} M_t(i,j) \cdot \|x_t(i,j) - R_t(i,j)\|_2^2, \tag{5}$$

where $M_t(i,j) \in \{0,1\}$ indicates the visibility of pixel $(i,j)$, and $x_t(i,j)$ and $R_t(i,j)$ are the pixel values in the generated and reference frames, respectively.

These losses are integrated into the diffusion denoising step via gradient guidance as follows:

$$x_{t-1} = x_t - \gamma_t \nabla_{x_t} \left(L_{\text{spatial}} + \lambda L_{\text{occ}}\right), \tag{6}$$

where $\gamma_t$ controls the guidance strength at timestep $t$, and $\lambda$ balances the spatial and occlusion losses.

Our method builds upon the insight that in diffusion models, the score function $\nabla_{x_t} \log p(x_t)$ is not explicitly known but can be approximated using external guidance signals, as demonstrated in Universal Guidance [2] and Loss-Guided Diffusion [33]. By treating spatial and occlusion losses as pseudo-energy functions, their gradients $-\nabla_{x_t}(\mathcal{L}_{\text{spatial}} + \lambda\mathcal{L}_{\text{occ}})$ act as surrogate score corrections. This biases the generative process toward scene-consistent and temporally aligned outputs while preserving the core diffusion model's generative capacity.

### 3.4. Camera Trajectory Generation

To generate temporally coherent intermediate views from sparse input poses, we constructed a smooth camera trajectory (see Fig. 3) by interpolating high-level geometric descriptors with a multidimensional B-spline inspired by the prior work in view synthesis and neural rendering [25, 27]. The camera path is generated by smoothly interpolating pose-derived elements, including the camera position, look-at direction, and up vector, to ensure coherent motion and orientation across views. We begin by decomposing each camera-to-world matrix $P_i \in \mathbb{R}^{3\times4}$ into a triplet of 3D points: the camera position $c_i = P_i[:,3]$, a look-at point $l_i = c_i - \delta \cdot d_i$, and an up-reference point $u_i = c_i + \delta \cdot v_i$.

**Algorithm 1:** *SplatFusion* (Full Pipeline)

**Input:** Sparse input images $\{I_o\}_{o=1}^{O}$
**Output:** Optimized Gaussian scene $G^\star$

**Geometric prior & pose estimation:**;
$\{P_o\} \leftarrow \textsc{Dust3R/MASt3R}(\{I_o\})$;
$G \leftarrow \textsc{InstantSplatInit}(\{I_o, P_o\})$;

**for** $k \leftarrow 1$ **to** $K$ **do**
  **Camera-path sampling:**;
  $\{P_v\}_{v=1}^{V} \leftarrow \textsc{Trajectory}(\{P_o\})$;
  **View refinement via SCTG:**;
  **foreach** $P_v$ **do**
    $\tilde{I}_v \leftarrow \text{SCTG}(G, P_v, T)$;
    `// Algorithm 2 call`
  **Back-projection & 3DGS update:**;
  $G \leftarrow \textsc{SplatOptimize}\Big(\{(I_o, P_o)\} \cup \{(\tilde{I}_v, P_v)\}\Big)$;

**return** $G^\star \leftarrow G$

**Algorithm 2:** SCTG

**Input:** Optimised scene $G$, target pose $P_v$, timesteps $T$
**Result:** Refined frame $\tilde{I}_v$
$x_T \sim \mathcal{N}(0, I)$; `// Initial noise`
**for** $t = T$ **to** $1$ **do**
  $R_t \leftarrow \textsc{Render}(G, P_v)$;
  $M_t \leftarrow \textsc{VisibilityMask}(G, P_v)$;
  $L_{\text{sp}} = \lambda_{\text{perc}} \| f_{\text{feat}}(x_t) - f_{\text{feat}}(R_t) \|_2^2$;
  $L_{\text{occ}} = \| M_t \odot (x_t - R_t) \|_2^2$;
  $g_t = \nabla_{x_t}(L_{\text{sp}} + \lambda_{\text{occ}} L_{\text{occ}})$;
  $\hat{x}_{t-1} = x_t - \gamma_t g_t$; `// Guided step`
$\tilde{I}_v \leftarrow x_0$;
**return** $\tilde{I}_v$

Here, $d_i$ and $v_i$ represent the forward and upward directions derived from the rotation component of $P_i$, and $\delta$ is a scene-dependent scale factor that controls the interpolation span. These triplets $\{(c_i, l_i, u_i)\}$ are interpolated using a B-spline of degree $k$, yielding a set of smooth control points that define the dense virtual trajectory. For each interpolated step, we reconstructed the camera pose $\tilde{P}_j$ using a look-at frame formulation:

$$\tilde{P}_j = [R(d_j, v_j) \mid c_j], \quad \text{where } R = [x, y, z], \tag{7}$$

with $z = \text{normalize}(d_j)$, $x = \text{normalize}(v_j \times z)$, and $y = z \times x$. This yields a dense sequence of virtual camera poses that are both spatially smooth and temporally coherent, serving as an effective scaffold for video diffusion and 3D Gaussian Splatting refinement.

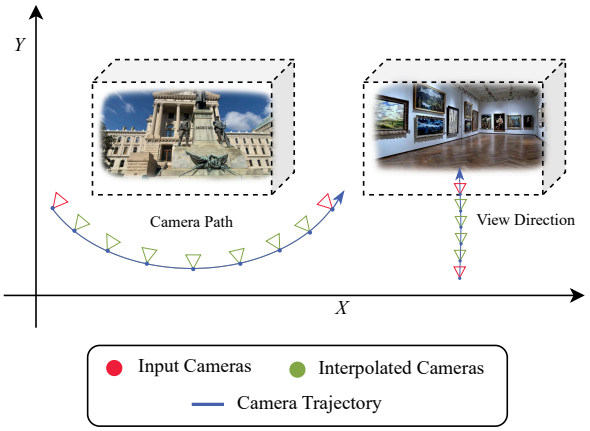


Figure 3. Illustration of our camera trajectory generation. Red markers indicate sparse input camera poses, while the blue curve shows the smooth trajectory of interpolated virtual cameras (green markers).

### 3.5. Back-Projection into 3D Gaussian Splatting

After refinement with the video diffusion model, the enhanced 2D frames are re-integrated into the 3D scene representation through back-projection. For each refined frame $\tilde{I}_v$, generated at a known or interpolated camera pose $P_v$, we treat the pair $(\tilde{I}_v, P_v)$ as a pseudo-observation of the scene. These pseudo-observations are combined with the original input images $(I_o, P_o)$ in a unified optimization step, yielding an updated set of Gaussian primitives, $G'$. Formally, we express this update as:

$$G' = \text{Splat}\left(\left\{(\tilde{I}_v, P_v)\right\}_{v=1}^{V} \cup \{(I_o, P_o)\}_{o=1}^{O}\right), \tag{8}$$

Here, $\text{Splat}(\cdot)$ refers to the differentiable optimization process that updates the Gaussian primitives $\{\mu_i, \Sigma_i, \alpha_i, c_i\}$ by minimizing the multi-view photometric loss function combined with spatial and perceptual regularization terms. Specifically, we solve the following objective:

$$\min_{G} \sum_{(I_v, P_v)} \|R(G, P_v) - I_v\|^2 + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(G), \tag{9}$$

where $R(G, P_v)$ denotes the differentiable 3DGS renderer at pose $P_v$, and $\mathcal{L}_{\text{reg}}$ includes smoothness and sparsity priors on Gaussian parameters. The optimization is performed via gradient descent using the backpropagation capability of the differentiable renderer. This process effectively integrates the refined pseudo-observations into a 3D representation, enabling geometric densification and correction of missing regions.

## 4. Experiments

All experiments were conducted using PyTorch on a single NVIDIA RTX 4090 GPU. For a full list of hyperparameters and settings, please refer to the supplementary material.

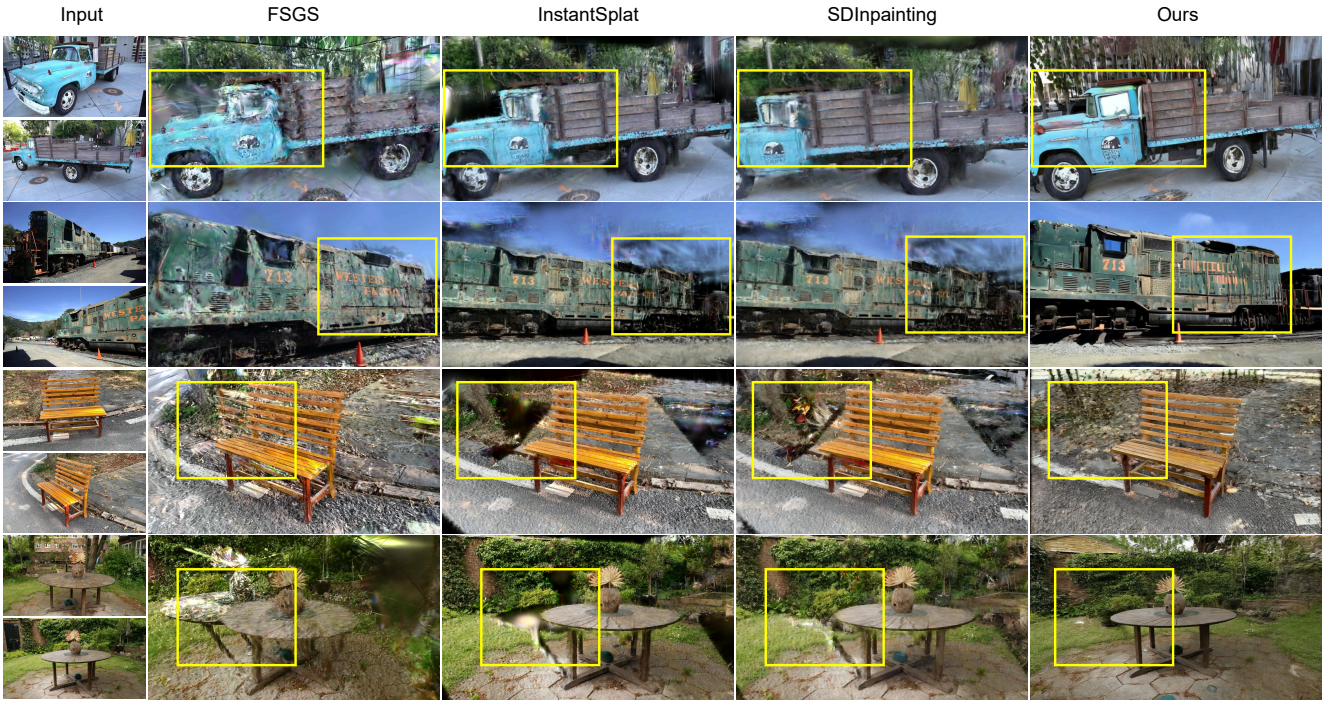


Figure 4. Qualitative comparison of novel view synthesis results on the *Tanks and Temples* [15], *MipNeRF360* [4], and *MVImgNet* [47] datasets. We compare our method against InstantSplat [7], Stable Diffusion Inpainting [30], and Few-Shot Gaussian Splatting (FSGS [49]). Under sparse-view conditions, prior methods often produce artifacts, incomplete regions, or structurally inconsistent renderings.

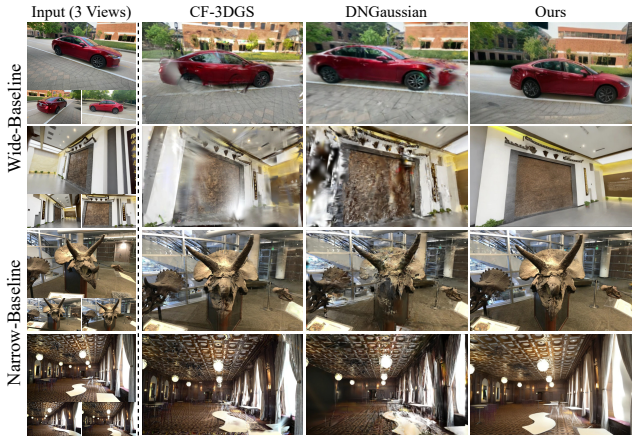


Figure 5. Qualitative comparison of novel view synthesis under narrow- and wide-baseline conditions. Our method preserves geometry, fine details, and coherence, while baselines such as CF-3DGS [6], DNGaussian [16] show distortions and missing regions, especially in wide-baseline cases on scenes from DL3DV-10k [17], Tanks and Temples [15], and LLFF datasets [26].

### 4.1. Datasets

We evaluated our method on diverse and challenging benchmarks that reflect the complexity of real-world 3D reconstruction. Following recent pose-free approaches [6, 9], we focused on datasets that stress sparse-view inputs and varied scene conditions. Specifically, we use *Tanks and Temples* [15] with only 2–3 wide-baseline views, *MVImgNet* [47] for diverse outdoor scenes, *DL3D* [17] for occlusion-heavy indoor and outdoor settings, and *MipNeRF360* [4] for large-scale 360° scenes with complex geometry and lighting. Together, these datasets form a comprehensive testbed for assessing robustness under extreme sparsity, occlusion, and scene complexity.

### 4.2. Evaluation Metrics

We use standard metrics (PSNR, SSIM, LPIPS) where ground truth is available, but these are unreliable under sparse-view conditions. For novel views without ground truth, we adopt a VLM-based evaluation inspired by recent work [35, 36, 44], using LLaVA [20] to assess perceptual quality and structural accuracy. This human-aligned protocol offers a more robust measure of visual fidelity in challenging settings.

### 4.3. Baseline Methods

We compared our method with recent state-of-the-art approaches in pose-free, sparse-view 3D reconstruction, including InstantSplat [7], which initializes 3DGS without known poses; FSGS [49], tailored for few-shot reconstruction; and CF-3DGS [6], which avoids traditional SfM-based pose estimation. We further evaluate against ReconX [18], a video diffusion model fine-tuned for sparse views, and

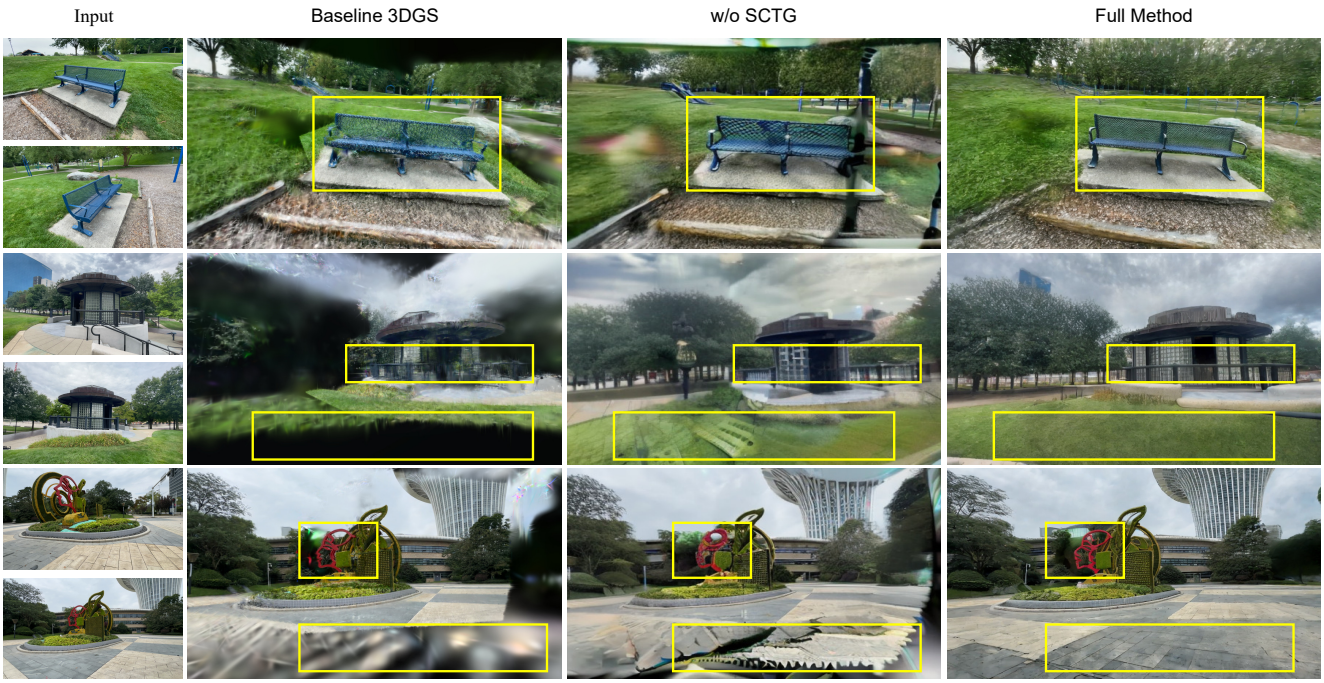


Figure 6. Qualitative ablation study on DL3DV-10k [17] showing the impact of SCTG. Sparse inputs (left) are compared against Baseline 3DGS and our method without SCTG (please zoom in to view the finer details).

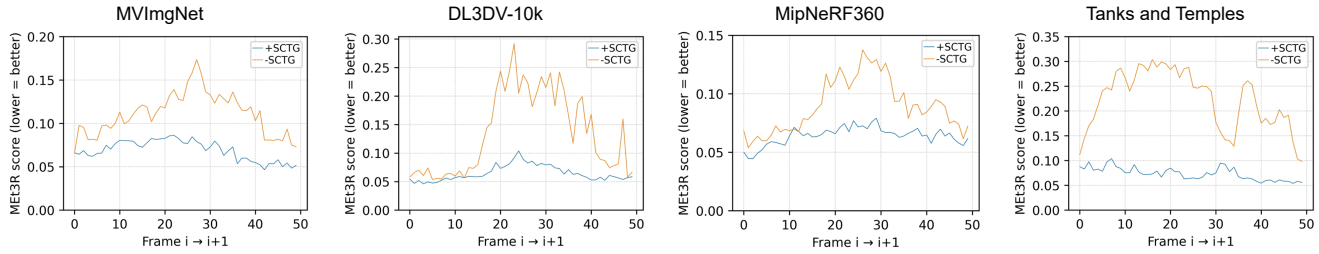


Figure 7. MEt3R [1] consistency comparison across multiple datasets. This metric quantifies the multi-view consistency by measuring the error between adjacent frames in a sequence. Our method (+SCTG) consistently achieves lower MEt3R scores than the baseline (-SCTG), indicating superior temporal coherence and fewer visual artifacts in the generated sequences (Lower scores are better).

DNGaussian [16], which employs global-local depth normalization for 3DGS optimization. Together, these methods provide a solid foundation for evaluating our approach.

### 4.4. Ablation Studies

To validate the contribution of each component, we conducted comprehensive ablation studies on the **DL3DV-10K** dataset [17]. Quantitative (Table 3) and qualitative (Fig. 6) results confirm the importance of each module. **Baseline 3DGS** yields a low PSNR of 22.15. Removing the entire **SCTG module** (w/o SCTG) causes a dramatic performance drop to a PSNR of 25.34, confirming that unguided diffusion priors are insufficient. Within SCTG, ablating the **Occlusion Mask** (w/o Occlusion Mask) degrades the PSNR to 28.91, as the model reinforces uncertain regions. Similarly, removing the **VGG Perceptual Loss** (w/o VGG Loss) results in a PSNR of 29.78 with a notable increase in LPIPS, highlighting its role in generating photorealistic details. Beyond single-image metrics, our **Full Method** significantly outperforms all variants in perceptual fidelity, as indicated by the VLM-based scores (Table 4). It also achieved the lowest MEt3R scores (Fig. 7), indicating superior geometric consistency.

### 4.5. Analysis

We evaluated our method using multiple benchmarks under sparse-view conditions, employing both standard metrics and VLM-based perceptual scores. As shown in Tab. 1, our approach consistently achieves the best LPIPS scores across all datasets, demonstrating superior perceptual fidelity and structural preservation. For novel view synthesis tasks without available ground truth, Tab. 2 reports the

| Method | MipNeRF360 | | | MVImgNet | | | DL3DV-10K | | | Tanks and Temples | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| SD-Inpainting | 21.93 | 0.477 | 0.188 | 20.50 | 0.522 | 0.186 | 23.31 | 0.740 | 0.105 | 23.31 | 0.730 | 0.105 |
| FSGS | 24.50 | 0.710 | 0.155 | 21.50 | 0.610 | 0.162 | 29.80 | 0.760 | 0.077 | 25.51 | 0.700 | 0.093 |
| InstantSplat | 24.85 | 0.720 | 0.148 | 21.58 | 0.618 | 0.158 | 30.15 | 0.825 | 0.062 | 25.63 | 0.715 | 0.091 |
| CF-3DGS | 19.75 | 0.520 | 0.235 | 18.20 | 0.485 | 0.285 | 18.99 | 0.737 | 0.271 | 15.95 | 0.557 | 0.503 |
| DNGaussian | 15.43 | 0.433 | 0.418 | 15.69 | 0.300 | 0.706 | 29.01 | 0.847 | 0.060 | 21.86 | 0.733 | 0.271 |
| **Ours** | 23.15 | 0.740 | 0.112 | 21.65 | 0.632 | 0.139 | 30.52 | 0.870 | 0.049 | 25.47 | 0.745 | 0.088 |

Table 1. Quantitative comparison of our method with SD-Inpainting, FSGS, InstantSplat, CF-3DGS, and DNGaussian across several benchmark datasets using PSNR, SSIM, and LPIPS metrics.

| Method | Tanks and Temples | | | | | MVImgNet | | | | | MipNeRF360 | | | | | DL3DV-10K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Noise-Free↑ | Edge↑ | Structure↑ | Detail↑ | Quality↑ | Noise-Free↑ | Edge↑ | Structure↑ | Detail↑ | Quality↑ | Noise-Free↑ | Edge↑ | Structure↑ | Detail↑ | Quality↑ | Noise-Free↑ | Edge↑ | Structure↑ | Detail↑ | Quality↑ |
| FSGS | 0.012 | 0.055 | 0.065 | 0.872 | 0.062 | 0.011 | 0.052 | 0.061 | 0.860 | 0.065 | 0.010 | 0.050 | 0.059 | 0.854 | 0.064 | 0.013 | 0.057 | 0.068 | 0.858 | 0.067 |
| InstantSplat | 0.027 | 0.194 | 0.178 | 0.971 | 0.141 | 0.024 | 0.183 | 0.166 | 0.976 | 0.137 | 0.017 | 0.158 | 0.152 | 0.968 | 0.125 | 0.021 | 0.181 | 0.163 | 0.973 | 0.130 |
| CF-3DGS | 0.032 | 0.210 | 0.190 | 0.963 | 0.148 | 0.031 | 0.205 | 0.188 | 0.958 | 0.140 | 0.026 | 0.192 | 0.174 | 0.957 | 0.133 | 0.027 | 0.189 | 0.182 | 0.954 | 0.135 |
| DNGaussian | 0.045 | 0.218 | 0.205 | 0.950 | 0.200 | 0.043 | 0.220 | 0.214 | 0.948 | 0.208 | 0.037 | 0.223 | 0.210 | 0.951 | 0.202 | 0.038 | 0.221 | 0.212 | 0.949 | 0.198 |
| SD-Inpainting | 0.023 | 0.215 | 0.228 | 0.960 | 0.212 | 0.022 | 0.210 | 0.223 | 0.951 | 0.204 | 0.020 | 0.208 | 0.216 | 0.947 | 0.193 | 0.019 | 0.205 | 0.214 | 0.952 | 0.199 |
| **Ours** | 0.174 | 0.702 | 0.545 | 0.991 | 0.401 | 0.179 | 0.711 | 0.552 | 0.989 | 0.409 | 0.165 | 0.698 | 0.537 | 0.988 | 0.395 | 0.172 | 0.705 | 0.542 | 0.990 | 0.403 |

Table 2. Quantitative comparison of perceptual quality for novel view synthesis where ground truth is not available. The VLM-based scores across multiple benchmark datasets indicate better performance with higher values in noise reduction, edge clarity, structural coherence, detail preservation, and overall visual quality.

| Variant | PSNR (↑) | SSIM (↑) | LPIPS (↓) |
|---|---|---|---|
| Baseline 3DGS | 22.15 | 0.751 | 0.182 |
| w/o SCTG | 25.34 | 0.795 | 0.115 |
| w/o Occlusion Mask | 28.91 | 0.842 | 0.068 |
| w/o VGG Loss | 29.78 | 0.861 | 0.059 |
| **Ours (Full Method)** | **30.52** | **0.870** | **0.049** |

Table 3. Ablation study of the core components of our method.

| Method | Noise-Free↑ | Edge↑ | Structure↑ | Detail↑ | Quality↑ |
|---|---|---|---|---|---|
| Baseline 3DGS | 0.015 | 0.185 | 0.168 | 0.955 | 0.132 |
| w/o SCTG | 0.025 | 0.215 | 0.220 | 0.968 | 0.205 |
| w/o Occlusion Mask | 0.121 | 0.595 | 0.435 | 0.981 | 0.315 |
| w/o VGG Loss | 0.158 | 0.681 | 0.515 | 0.986 | 0.380 |
| **Ours (Full)** | **0.172** | **0.705** | **0.542** | **0.990** | **0.403** |

Table 4. Ablation study across five LLaVA-IQA metrics.

VLM-based scores. Our method consistently outperformed all baselines, with scores in the "Structure" and "Edge" categories reaching up to three times higher than those of the next best method. Qualitative comparisons in Fig. 4 and 5, further support these findings. The SCTG-refined outputs preserved sharper edges and cleaner geometry, effectively restoring spatial fidelity in areas where the baselines exhibit artifacts. The combination of VLM-based metrics with our MEt3R Fig. 7 results provides a holistic validation, confirming that perceptual improvements are directly tied to enhanced geometric consistency. These improvements are consistent across diverse scenes, highlighting the robustness and strong generalization of our training-free strategy.

## 5. Conclusion

We introduce SplatFusion, a training-free framework that leverages pretrained video diffusion priors to enhance sparse-view 3D Gaussian Splatting (3DGS) reconstructions. Our central contribution, Scene-Consistent Temporal Guidance (SCTG), uses rendered 3DGS frames to enforce spatial and temporal coherence during the diffusion process. By reprojecting the refined 2D outputs back into the 3DGS representation, our method effectively fills occluded and underconstrained regions without requiring any scene-specific fine-tuning. Empirical evaluations show that SplatFusion consistently outperforms geometry-only baselines and achieves competitive visual fidelity. Quantitative metrics further confirmed that SCTG enhances both perceptual quality and geometric consistency across views. Ultimately, our work provides a promising paradigm for bridging the gap between 2D generative priors and 3D geometry. This approach also offers foundational insights into the future direction of the field, demonstrating that by carefully designing a guidance mechanism like SCTG, we can explicitly enforce geometric consistency to overcome the inherent limitations of 2D generative priors.

## Acknowledgments

This work was supported in part by NSFC (U21A20515), Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016), and the Shenzhen Science and Technology Program (CJGJZD20240729141906008).

## References


[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *CVPR*, 2025. 7

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPR*, 2023. 2, 3, 4

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

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 6

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 1

[6] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. COLMAP-Free 3D Gaussian splatting. In *CVPR*, 2024. 6

[7] Zhiwen Fan, Kairun Wen, Wenyan Cong, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Sparse-view gaussian splatting in seconds, 2024. arXiv:2403.20309. 1, 2, 3, 6

[8] Ruijia Feng, Wenhu Wang, Jiaming Song, Tongzhou Zhang, Yutong Chen, Yu Zeng, Luke Zettlemoyer, and Chuang Gan. Structured diffusion guidance for compositional text-to-image synthesis. In *NeurIPS*, 2023. 2

[9] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. NoPe-NeRF: Optimising neural radiance field with no pose prior. In *CVPR*, 2023. 6

[10] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, et al. CAT3D: Create anything in 3D with multi-view diffusion models. In *NeurIPS*, 2024. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen Video: High definition video generation with diffusion models. *arXiv*, 2022. arXiv:2210.02303. 3

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023. 1, 2

[14] Mijeong Kim, Seonguk Seo, and Bohyung Han. InfoNeRF: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. 1

[15] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 36(4):1–13, 2017. 6

[16] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. DNGaussian: Optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization. In *CVPR*, 2024. 1, 2, 6, 7

[17] Lu Ling, Yichen Sheng, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, 2024. 6, 7

[18] 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*, 2024. arXiv:2408.16767. 2, 6

[19] Haithem Liu, Yunfan Li, Ziyang Fang, Yongming He, Yuxuan Gao, Ziyu Zheng, Xiangyu Zhang, and Dahua Lin. More control: Leveraging semantic guidance for text-to-image diffusion models. *arXiv*, 2023. arXiv:2309.06726. 2

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 6

[21] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. MVSGaussian: Fast generalizable Gaussian splatting reconstruction from multi-view stereo. In *ECCV*, 2024. 1

[22] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3DGS-Enhancer: Enhancing unbounded 3D Gaussian splatting with view-consistent 2D diffusion priors. In *NeurIPS*, 2024. 2

[23] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-GS: Structured 3D Gaussians for view-adaptive rendering. In *CVPR*, 2024.

[24] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De La Torre. Taming 3DGS: High-quality radiance fields with limited resources. In *SIGGRAPH Asia*, 2024. 1

[25] Ricardo Martin-Brualla, Nazim Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Learning to reconstruct 3D manifolds from 2D supervision. In *ECCV*, 2018. 4

[26] Ben Mildenhall, Pratul P Srinivasan, Rafael Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Ayan Kar. Local Light Field Fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM TOG*, 38(4): 1–14, 2019. 6

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 4

[28] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 1

[29] Ben Poole et al. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 6

[31] Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *NeurIPS*, 2024. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[33] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *ICML*, 2023. 2, 3, 4

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3

[35] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. VistaDream: Sampling multiview consistent images for single-view scene reconstruction. In *ICCV*, 2025. 6

[36] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, 2023. 6

[37] Shuzhe Wang, Vincent Leroy, Yohann Cabon, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. 2, 3

[38] 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. In *CVPR*, 2025. 2

[39] Rundi Wu, Ben Mildenhall, Philipp Henzler, et al. ReconFusion: 3D reconstruction with diffusion priors. In *CVPR*, 2024. 2

[40] Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. GaussianObject: High-quality 3D object reconstruction from four views with Gaussian splatting. *ACM TOG*, 2024. 2

[41] Jiawei Yang, Marco Pavone, and Yue Wang. FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, 2023. 1

[42] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *NeurIPS*, 2024. 2

[43] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. PixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1

[44] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. WonderJourney: Going from anywhere to everywhere. In *CVPR*, 2024. 6

[45] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *ICCV*, 2023. 2

[46] 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*, 2024. arXiv:2409.02048. 4

[47] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. MVImgNet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 6

[48] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. CoR-GS: Sparse-view 3D Gaussian splatting via co-regularization. In *CVPR*, 2024. 1, 2

[49] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. FSGS: Real-time few-shot view synthesis using Gaussian splatting. In *ECCV*, 2024. 1, 2, 6