

Taming Video Diffusion Prior with Scene-Grounding Guidance for 3D Gaussian Splatting from Sparse Inputs

Yingji Zhong¹ Zhihao Li² Dave Zhenyu Chen² Lanqing Hong² Dan Xu¹

¹The Hong Kong University of Science and Technology ²Huawei Noah's Ark Lab

{yzhongbn, danxu}@cse.ust.hk, {zhihao.li, dave.zhenyuchen, honglanqing}@huawei.com



Figure 1. We tackle the critical issues of (a) extrapolation and (b) occlusion in sparse-input 3DGS by leveraging a video diffusion model. Vanilla generation often suffers from inconsistencies within the generated sequences (as highlighted by the yellow arrows), leading to black shadows in the rendered images. In contrast, our scene-grounding generation produces consistent sequences, effectively addressing these issues and enhancing overall quality (c), as indicated by the blue boxes. The numbers refer to PSNR values. Zoom in for better visualization.

Abstract

Despite recent successes in novel view synthesis using 3D Gaussian Splatting (3DGS), modeling scenes with sparse inputs remains a challenge. In this work, we address two critical yet overlooked issues in real-world sparse-input modeling: extrapolation and occlusion. To tackle these issues, we propose to use a reconstruction by generation pipeline that leverages learned priors from video diffusion models to provide plausible interpretations for regions outside the field of view or occluded. However, the generated sequences exhibit inconsistencies that do not fully benefit subsequent 3DGS modeling. To address the challenge of inconsistencies, we introduce a novel scene-grounding guidance based on rendered sequences from an optimized 3DGS, which tames the diffusion model to generate consistent sequences. This guidance is training-free and does not require any fine-tuning of the diffusion

model. To facilitate holistic scene modeling, we also propose a trajectory initialization method. It effectively identifies regions that are outside the field of view and occluded. We further design a scheme tailored for 3DGS optimization with generated sequences. Experiments demonstrate that our method significantly improves upon the baseline and achieves state-of-the-art performance on challenging benchmarks. The project page is available at <https://zhongyingji.github.io/guidevd-3dgs>.

1. Introduction

Recent advances in 3D scene representation such as Neural Radiance Fields (NeRF) [2–4, 32, 34, 47] have greatly boosted the performance of Novel View Synthesis (NVS). NeRF represents the scene with a Multi-Layer Perceptron (MLP) and renders high-fidelity images with vol-

umetric rendering. More recently, 3D Gaussian Splatting (3DGS) [19, 28, 29, 62] emerges as a powerful explicit representation that models the scene with a set of 3D gaussian primitives and renders images via differentiable splatting. 3DGS achieves comparable performance to NeRF while requiring significantly less training time and offering higher inference speeds.

Despite recent advances in scene representations based on 3DGS, modeling scenes with sparse inputs remains a significant challenge. The sparse supervision often leads radiance fields to learn degenerate representations due to shape-radiance ambiguity [63], regardless of whether the representation is NeRF or 3DGS. While there have been promising improvements [21, 35, 48, 67], the commonly used face-forwarding [17, 31] and object-oriented ‘outside-in’ viewing [32] settings oversimplify real-world sparse-input modeling, causing many methods to overlook two critical issues: **(i) extrapolation** - while the sparse inputs typically cover the scene as much as possible, there may still exist regions that are outside the field of view, as shown in Fig. 1 (a); **(ii) occlusion** - occlusion frequently occurs for novel views that deviate even slightly from the training input views, as illustrated in Fig. 1 (b). When rendering with an optimized 3DGS, these issues can cause severe artifacts, such as black holes, significantly degrading image quality. Therefore, handling these two issues, i.e., extrapolation and occlusion, is critical for real-world sparse-input modeling.

To address the above-discussed issues, we propose a novel reconstruction by generation pipeline based on 3DGS. Intuitively, we use video diffusion models [8, 15, 55, 61] to generate multi-view sequences, which provide plausible interpretations of the scene based on priors learned from large-scale datasets. These sequences significantly enlarge the viewing instances, offering a high potential to address the extrapolation and occlusion issues. With the sparse inputs and the generated sequences, we can optimize a 3DGS to model the scene. However, as shown in Fig. 1, this vanilla pipeline brings little improvement or may even degrade the performance. The main reason can be attributed to the multi-view inconsistency within the generated sequences. The inconsistency manifests in two aspects: **(i)** the appearance inconsistency between frames within a sequence; **(ii)** the generated sequence may contain hallucinated elements not present in the scene.

To fully leverage the learned prior from video diffusion models for sparse-input 3DGS, we further explore addressing the challenges of inconsistencies within the generated sequences. Unlike existing methods that resolve appearance inconsistencies by assigning per-frame learnable appearance embeddings [24, 30], we focus on taming video diffusion models to *directly generate sequences with consistency*. Inspired by training-free guidance methods for diffusion models [1, 42, 57, 60] that enable controllable

generation through external guidance, we introduce a novel strategy called *scene-grounding guidance* to ensure consistent generation without requiring further fine-tuning of the diffusion models. Specifically, the scene-grounding guidance is based on a rendered sequence from an optimized 3DGS. During each step of the denoising process, the noisy sequence receives gradients from the supervision of the rendered sequence. Although the rendered sequence does not provide perfect guidance, our key insight in employing it to address the inconsistency is twofold: **(i)** adjacent frames within the rendered sequence are highly consistent due to limited camera movement between them; **(ii)** the rendered sequence is scene-grounding, which can guide the diffusion model to avoid generating elements that do not exist in the scene. In addition to addressing the issues of extrapolation and occlusion, our method also enhances the overall quality of the rendered images, as demonstrated in Fig. 1 (c). To effectively identify regions that are outside the field of view or occluded, we propose a trajectory initialization strategy to determine the camera trajectory during sequence generation, which is also based on an optimized 3DGS. With the proposed method, we can perform a holistic modeling of the scene. Additionally, we introduce a scheme for optimizing 3DGS with generated sequences, focusing on loss and sampling designs, which further enhance overall performance. Following [68], we evaluate our method on two challenging indoor datasets: Replica [44] and ScanNet++ [58], where the issues of extrapolation and occlusion are pronounced. The experiments demonstrate that our method achieves notable improvements and establishes state-of-the-art performance. Our contributions are summarized as:

- This paper is the first to explicitly address the challenges of extrapolation and occlusion in 3DGS modeling from sparse inputs.
- We propose a novel reconstruction by generation pipeline with a designed scene-grounding guidance, which tames the video diffusion models to generate consistent and plausible sequences to effectively tackle the issues of extrapolation and occlusion.
- We present a trajectory initialization strategy that effectively identifies regions that are outside the field of view and occluded, facilitating holistic scene modeling. We also introduce a scheme for optimizing 3DGS with generated sequences, further improving the performance.
- Our method demonstrates significant improvements over the baseline, achieving over 3.5 dB and 2.5 dB PSNR enhancements on the Replica [44] and ScanNet++ [58] datasets, respectively, thereby establishing state-of-the-art performance.

2. Related Works

Radiance Fields from Sparse Inputs. Although improvements have been made in scene representation [2–4, 19, 28,

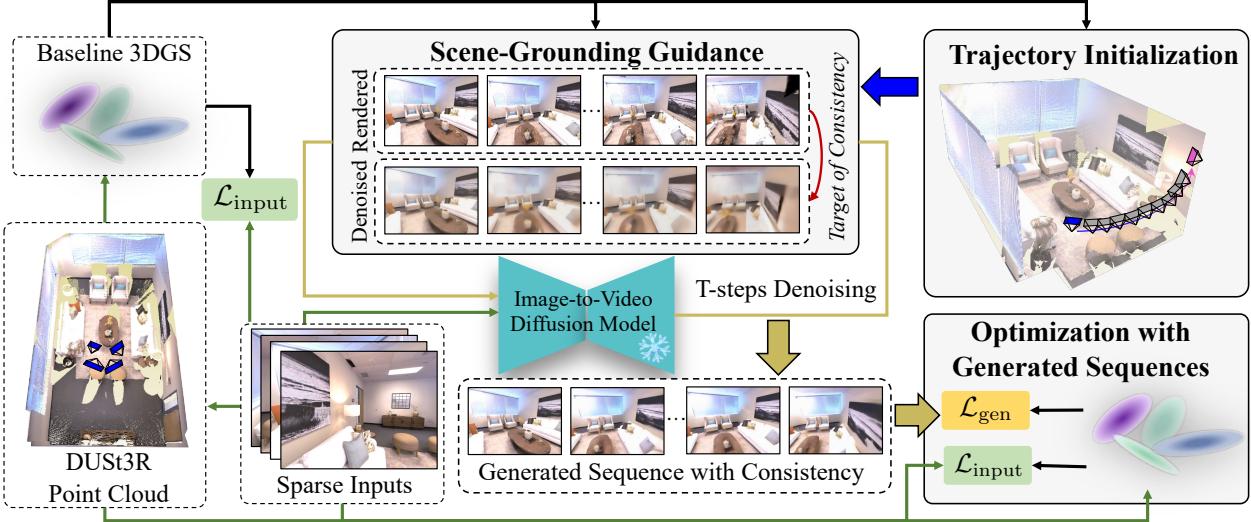


Figure 2. **Framework overview of our proposed method.** It consists of three parts: scene-grounding guidance, trajectory initialization, and optimization scheme with generated sequences. Initially, a baseline 3DGS is trained using sparse inputs and initialized with the point cloud from DUS3R [50]. Yellow regions denote uncovered areas, e.g., those outside the field of view or occluded. The trajectory initialization determines the paths for sequence generation based on renderings from the baseline 3DGS, facilitating holistic scene modeling. The video diffusion model receives an input image along with the trajectory for sequence generation, incorporating scene-grounding guidance during the denoising process to ensure consistent output. The guidance is based on the rendered sequences. Finally, the generated sequences are utilized to optimize the final 3DGS through a tailored optimization scheme.

[32, 34, 62], learning a robust radiance field typically requires dense inputs as supervision due to the radiance-shape ambiguity [63]. Current works can be roughly categorized into two lines of research. The first line of research focuses on pre-training generalizable radiance fields using multi-view datasets. Generalizable NeRF [6, 27, 49, 59] typically projects the ray points to reference images and integrates 2D features as the auxiliary input of MLP. In contrast, generalizable 3DGS [5, 9, 25, 64] commonly perform dense depth prediction associated with gaussian properties. The second line of research leverages regularization techniques for optimization, which are applicable to both NeRF and 3DGS, including hand-crafted constraints [20, 35, 67], and those derived from pre-trained models [11, 21, 36, 38, 48, 54, 69]. In this paper, we explore the priors from video diffusion models [8, 15, 55, 61] for sparse-input 3DGS modeling. Unlike previous works, our method trains sparse-input 3DGS using augmented sequences from video diffusion models, which provide interpolation or extrapolation around the input views. We design a guidance strategy that tames the video diffusion model to generate more plausible and scene-grounded sequences, greatly enhancing the performance.

Diffusion Prior for Radiance Fields. Diffusion models [14, 39, 43] have shown remarkable generation capabilities. The strong prior knowledge embedded in diffusion models can facilitate the training of radiance fields. Specifically, several works leverage Score Distillation Sampling (SDS) [7, 22, 23, 37, 52] using a frozen diffusion model to

train a 3D consistent representation based on text prompts in a zero-shot manner. Other studies focus on training view-consistent [40, 46, 53] or quality-enhanced [26] diffusion models, where the generated results can be directly applied to train a radiance field. Our method also uses the generated results to train a radiance field. However, our primary contribution lies in a novel guidance strategy that controls the generation to be consistent, which is crucial for sparse-input modeling augmented with generation.

Controllable Generation for Diffusion Models. Current works of controllable generation can be categorized into training-required and training-free methods. Training-required methods fine-tune the diffusion models with additional conditions [33, 65], or train an additional noise-conditioned external guidance function, e.g., classifier guidance [10] for denoising sampler. Training-free methods freeze the foundation diffusion model, and modify the denoising process with the control signal from the external guidance functions [1, 42, 57, 60]. These methods do not require training additional guidance functions or fine-tuning diffusion models; instead, they enable controllable generation in a plug-and-play manner. Our work is inspired by training-free methods. However, we concentrate on multi-view modeling, which necessitates high consistency control over the generated results. Besides, unlike these methods that typically rely on pre-trained models for guidance, we utilize rendered results to provide guidance.

3. The Proposed Method

In this paper, we utilize video diffusion models to tackle two critical issues in real-world sparse-input modeling: extrapolation and occlusion, as illustrated in Fig. 1. The overview of our method is illustrated in Fig. 2, which consists of three proposed components: a scene-grounding guidance (Sec. 3.2), a trajectory initialization strategy (Sec. 3.3), and a scheme for 3DGS optimization with generated sequences (Sec. 3.4). We will detail these components following a preliminary review of 3DGS and diffusion models.

3.1. Preliminary

3D Gaussian Splatting (3DGS) [19] represents a scene with a set of anisotropic 3D Gaussian primitives. Each Gaussian primitive is parametrized by a set of attributes: a center μ , a scaling factor s , a quaternion q , an opacity value α , and a feature vector f . The basis function of each Gaussian primitive is formulated as $\mathcal{G}(x) = \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu))$, where Σ is the covariance matrix derived from s and q . 3DGS renders the scene through a differentiable splatting, which firstly transforms the 3D Gaussian $\mathcal{G}(x)$ into 2D Gaussian $\mathcal{G}'(x)$ on the image plane via projection [70], and applies a tile-based rasterizer for rendering, which sorts the 2D Gaussians by depth and employs the α -blending as follows:

$$C(x_p) = \sum_{i \in K} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \quad \sigma_i = \alpha_i \mathcal{G}'_i(x_p), \quad (1)$$

where x_p is the pixel position, K refers to the number of 2D Gaussians associated with the pixel, and c represents the decoded color of feature f .

Diffusion Models [14, 43] are a family of generative models that progressively perturb data with intensifying Gaussian noises (i.e., forward noising), and then learn to reverse this process for sample generation (i.e., reverse denoising). The key of the diffusion model is a U-Net ϵ_θ which is trained to predict the noise that is injected in the current sample \mathbf{x}_t . The sampling is conducted by iterative denoising for T steps [60] as follows:

$$\mathbf{x}_{t-1} = (1 + \beta_t/2)\mathbf{x}_t + \beta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \sqrt{\beta_t} \mathbf{z} \quad (2)$$

where $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ is the estimated score function which can be derived from $\epsilon_\theta(\mathbf{x}_t, t)$; β_t is pre-defined parameters; $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. In this work, we leverage a camera-controlled image-to-video diffusion model [61], whose condition includes an image for the first frame, and the camera trajectory for the path of the generated sequence. The model is operated in a latent space of dimension d , supporting the sequence length of L , thus $\mathbf{x}_t \in \mathbb{R}^{L \times h \times w \times d}$.

3.2. Generation via Scene-Grounding Guidance

Applying the generated sequences from the video diffusion model can provide plausible interpretations of regions not

Algorithm 1 Generation with Scene-Grounding Guidance

```

1: Function GENERATOR( $\mathcal{R}, I, \{\phi_j\}_{j=1}^L$ )
2: Input: Optimized 3DGS model  $\mathcal{R}$ , input image  $I$ , camera trajectory of a sequence  $\{\phi_j\}_{j=1}^L$ .
3: Given: Latent image-to-video diffusion model  $\epsilon_\theta$ , VAE decoder  $\mathcal{D}$ , pre-defined  $\beta_t, \bar{\alpha}_t$  and guidance scale  $\gamma_t$ .
4: Abbreviate  $\epsilon_\theta(\mathbf{x}_t, t, I, \{\phi_j\}_{j=1}^L)$  as  $\epsilon_\theta(\mathbf{x}_t, t)$ 
5:  $\mathbf{S}, \mathbf{M} = \text{rasterize}(\{\phi_j\}_{j=1}^L, \mathcal{R}) \quad \triangleright \text{Eq. (1)& (5)}$ 
6:  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ 
7: for  $t = T, \dots, 1$  do
8:    $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$  if  $t > 1$ , else  $\mathbf{z} = \mathbf{0}$ 
9:    $\hat{\mathbf{x}}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t) + \sqrt{\beta_t}\mathbf{z}$ 
10:   $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t))$ 
11:   $\mathbf{X}_{0|t} = \mathcal{D}(\mathbf{x}_{0|t})$ 
12:   $\mathbf{g}_t = \nabla_{\mathbf{x}_t} \mathcal{L}(\mathbf{S}, \mathbf{M}, \mathbf{X}_{0|t}) \quad \triangleright \text{Eq. (6)}$ 
13:   $\mathbf{x}_{t-1} = \hat{\mathbf{x}}_{t-1} - \gamma_t \mathbf{g}_t \quad \triangleright \text{Eq. (2)& (4)}$ 
14: end for
15: return  $\mathcal{D}(\mathbf{x}_0)$ 

```

covered by the sparse inputs. However, as illustrated in Fig. 1, the inconsistency within the generated sequences manifests as: **(i)** appearance inconsistencies across frames and **(ii)** the occurrence of non-existent elements, which can negatively impact the 3DGS modeling. In this section, we propose an innovative scene-grounding guidance method that directs the video diffusion model to generate consistent sequences, significantly enhancing the performance of sparse-input 3DGS.

Inspired by previous training-free guidance methods [1, 60] that achieve their objectives by modifying the sampler in Eq. (2), we adopt a similar approach to attain the goal of consistency. Specifically, we firstly replace $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ with a conditional score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{Q}|\mathbf{x}_t)$, where \mathbf{Q} refers to *the target of consistency*. The conditional score function can be expanded by the Bayesian rule as:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{Q}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{Q}|\mathbf{x}_t), \quad (3)$$

where $\nabla_{\mathbf{x}_t} \log p(\mathbf{Q}|\mathbf{x}_t)$ can be considered as a guidance term that injects the consistency constraint into Eq. (2). We further formulate $p(\mathbf{Q}|\mathbf{x}_t)$ as: $p(\mathbf{Q}|\mathbf{x}_t) = \exp(-\lambda \mathcal{L}(\mathbf{Q}, \mathbf{x}_t)/Z)$, where $\mathcal{L}(\mathbf{Q}, \mathbf{x}_t)$ indicates how well the current sample \mathbf{x}_t is aligned with the target, and Z is a normalization term. The guidance term can thus be implemented using the gradient of the following loss function:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{Q}|\mathbf{x}_t) \propto -\nabla_{\mathbf{x}_t} \mathcal{L}(\mathbf{Q}, \mathbf{x}_t), \quad (4)$$

which is appended to Eq. (2) to achieve the target of consistency during the denoising sampling.

The remaining problem lies in how to define the consistency target \mathbf{Q} . Unlike previous works [1, 42, 60] that define the target based on external pre-trained models, we

establish the target using a rendered sequence from an optimized 3DGS model \mathcal{R} . Though the rendered sequence is not perfect, our key insights are as follows: **(i)** the rendered images of adjacent frames are highly consistent, as the camera movement between them is typically minor; **(ii)** the rendered frames provide scene grounding, clearly indicating which elements exist in the scene. Therefore, the rendered sequence can serve as an effective guidance for the generated sequence to achieve the target of consistency.

Given a camera trajectory $\{\phi_j\}_{j=1}^L$ for the sequence generation, we first utilize the optimized 3DGS to render a sequence $\{S_j\}_{j=1}^L$, along with a mask sequence $\{M_j\}_{j=1}^L$ that indicates the regions not covered by the sparse inputs. To get the mask, we first render a transmittance map, which is obtained by α -blending (as Eq. (1)) on the opacity. For each pixel x_p , the α -blending is formulated as:

$$O(x_p) = \sum_{i \in K} \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \quad (5)$$

where σ and O refer to the opacity of the gaussian and the transmittance map, respectively. The mask is then acquired by thresholding the transmittance map with a value η_{mask} : $M = (O < \eta_{\text{mask}})$. For convenience, we stack $\{S_j\}_{j=1}^L$ and $\{M_j\}_{j=1}^L$ to \mathbf{S} and \mathbf{M} , which are of shape $L \times H \times W \times 3$ and $L \times H \times W \times 1$, respectively. Since the target of consistency is based on the rendered sequence in clean data space, to receive the guidance, we transform the noisy latent \mathbf{x}_t into a latent $\mathbf{x}_{0|t}$ in the clean data space, based on prediction from the model ϵ_θ : $\mathbf{x}_{0|t} = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$. With the consistency target \mathcal{Q} that is based on the rendered sequence \mathbf{S} , we formulate the function \mathcal{L} in the guidance term (Eq. (4)) as:

$$\begin{aligned} \mathcal{L}(\mathbf{S}, \mathbf{M}, \mathbf{X}_{0|t}) = & \|\mathbf{M} \odot (\mathbf{S} - \mathbf{X}_{0|t})\|_1 \\ & + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}}(\mathbf{M} \odot \mathbf{S}, \mathbf{M} \odot \mathbf{X}_{0|t}), \end{aligned} \quad (6)$$

where $\mathbf{X}_{0|t}$ is decoded from the latent $\mathbf{x}_{0|t}$ by a VAE decoder, \odot is the Hadamard product, and $\mathcal{L}_{\text{perc}}$ is a perceptual loss [18] with its corresponding weight as λ_{perc} .

With the guidance from Eq. (6), the denoising process balances the consistency constraint and the prior from the diffusion model, integrating them into plausible generation results. This guidance does not involve any fine-tuning of the diffusion model, thereby preserving its generative capabilities. The detailed pipeline is outlined in Alg. 1.

3.3. Trajectory Initialization Strategy

To enable holistic modeling of the scene, the camera trajectories for the video diffusion model should cover regions that are outside the field of view or occluded as much as possible. The generated sequences can thus provide plausible interpretations for these regions, which serve as the basis for optimizing the subsequent 3DGS model. Similar to the scene-grounding guidance discussed in Sec. 3.2,

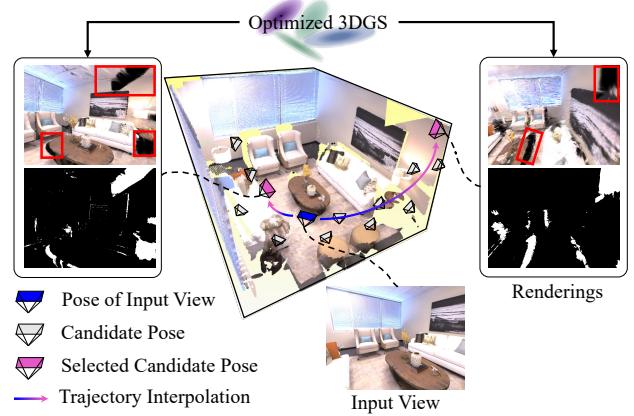


Figure 3. Illustration of the proposed trajectory initialization strategy. The yellow parts represent unobserved regions. For each input view, we sample a set of candidate poses around it, and render at these poses using an optimized 3DGS. We select candidate poses whose renderings exhibit significant holes (highlighted by red boxes), and interpolate trajectories between these candidate poses and the input view’s pose.

the proposed trajectory initialization method is also based on an optimized 3DGS model. For the i -th sparse input view with camera pose φ_i , we first sample a set of candidate poses around it, as depicted in Fig. 3. Suppose there are a total of m candidate poses, we use the optimized 3DGS model \mathcal{R} to render images for these poses as $\{\hat{S}_c^{(i)}, \hat{M}_c^{(i)}\}_{c=1}^m = \text{rasterize}(\{\hat{\phi}_c^{(i)}\}_{c=1}^m, \mathcal{R})$. For poses where the rendered images exhibit significant black holes, as indicated by the mask \hat{M} calculated from Eq. (5), we interpolate a trajectory of length L (matching the length of video diffusion model) between the input camera pose and these poses as follows: $\{\phi_j^{(i,c)}\}_{j=1}^L = \text{interp}(\varphi_i, \hat{\phi}_c^{(i)})$, where $\hat{\phi}_c^{(i)}$ refers to one selected candidate pose from the i -th input. In practice, we select the top- k candidate poses based on the sizes of their corresponding masks. Then, we build a trajectory pool by traversing all input views and their respective selected candidate poses as:

$$\Phi = \{\{\phi_j^{(i,c)}\}_{j=1}^L | i, c\}, \quad (7)$$

where each element in the pool is sampled for the sequence generation.

3.4. 3DGS Optimization with Generation

Given sparse inputs of N images along with their poses, i.e., $\{C_i^{\text{gt}}, \varphi_i\}_{i=1}^N$, we aim at optimizing a 3DGS model with the auxiliary generated sequences. For simplicity, we refer to the input images paired with their poses as ‘input views’, and we term the generated images with their associated poses as ‘generated views’. During each iteration, we sample an input view and a generated view for supervision. Specifically, for the input view, we employ the default

Algorithm 2 3DGS Optimization with Generation

```

1: Input: Sparse inputs of N images  $\{C_i^{\text{gt}}, \varphi_i\}_{i=1}^N$ .
2: Given: Number of iterations  $N_{\text{iter}}$ , generation interval
    $N_{\text{gen}}$ , ratio of samples from other sequences  $\eta$ .
3: Variable: Global list of generated views  $\mathbf{G} = []$ .
4: Baseline 3DGS model optimization  $\Rightarrow \mathcal{R}$ 
5: Trajectory initialization  $\Rightarrow \Phi$   $\triangleright$  Eq. (7)
6: for  $t = 0, \dots, N_{\text{iter}} - 1$  do
7:   If  $t \% N_{\text{gen}} = 0$  then
8:     Sample an input view  $I$ 
9:     Sample a trajectory around  $I$  from  $\Phi \Rightarrow \{\phi_j\}_{j=1}^L$ 
10:     $\mathbf{S} = \text{GENERATOR}(\mathcal{R}, I, \{\phi_j\}_{j=1}^L)$ 
11:    Append  $\mathbf{S}$  to  $\mathbf{G}$ 
12:   End If
13:   Sample an input view to get  $\mathcal{L}^{\text{input}}$   $\triangleright$  Eq. (8)
14:   If  $\text{rand}() \geq \eta$  then
15:     Sample a generated view from  $\mathbf{S}$ 
16:   Else Sample a generated view from  $\mathbf{G}$ 
17:   End If
18:   Use the generated view to get  $\mathcal{L}^{\text{gen}}$   $\triangleright$  Eq. (9)
19:    $(\mathcal{L}^{\text{input}} + \mathcal{L}^{\text{gen}}).\text{backward}()$ 
20:   # Densification and opacity reset
21: end for

```

reconstruction loss [19] written as:

$$\mathcal{L}^{\text{input}} = (1 - \lambda)\mathcal{L}_1(C_i, C_i^{\text{gt}}) + \lambda\mathcal{L}_{\text{D-SSIM}}(C_i, C_i^{\text{gt}}), \quad (8)$$

where C_i refers to the rendered image and λ is a weighting factor. For supervision of generated views, we find that the reconstruction loss does not effectively fill the hole regions, and increasing its weight leads to performance degradation due to flaws in the generated images. To address this issue, we propose using perceptual loss [18]. The perceptual loss is calculated over the entire image, allowing those hole regions to significantly influence the gradients, thereby effectively driving the model to fill those holes. Thus, the loss on the generated views is formulated as:

$$\mathcal{L}^{\text{gen}} = \lambda_{\text{gen1}}\mathcal{L}_1(C_j, S_j) + \lambda_{\text{gen2}}\mathcal{L}_{\text{perc}}(C_j, S_j), \quad (9)$$

where S_j refers to the generated image, λ_{gen1} and λ_{gen2} are two balancing factors, respectively.

We empirically find that conducting local sampling within a specific optimization interval, where a substantial portion of generated views is sampled from the same sequence of local regions, enhances visual quality. However, sampling exclusively from a single sequence can lead to a forgetting issue, where optimized information about holes in other regions becomes diluted. Therefore, within each interval of local sampling, we also include generated views from other sequences with a ratio η . The optimization pipeline is presented in Alg. 2.



Figure 4. Sequences from the vanilla generation suffer from inconsistencies. A 3DGS model optimized with these sequences renders images with black shadows, highlighted by red boxes, while our method solves this issue with the scene-grounding guidance.

4. Experiments

4.1. Experimental Setups

Datasets and Metrics. We target addressing the issues of extrapolation and occlusion for sparse-input 3DGS scene modeling, which are overlooked by current benchmarks. To evaluate the effectiveness of our method, we conduct experiments on a benchmark [68] created from two indoor datasets, i.e., the synthetic Replica [44] and the realistic ScanNet++ [58], which consists of 6 and 4 scenes, respectively. Although the selected six input views for each scene can cover most regions, there are still areas outside the field of view. Moreover, the ‘inside-out’ viewing directions make occlusion common in this benchmark. For quantitative comparisons, we report PSNR, SSIM [51], and LPIPS [66] scores.

Baseline. We train a baseline 3DGS model initialized with the point cloud from DUS3R [50], incorporating the gaussian unpooling in FSGS [69], which makes the optimized model a strong baseline. Based on this we conduct experiments to verify the effectiveness of our method. The model is denoted as ‘Baseline 3DGS’ in the following.

Implementation Details. The baseline model described above serves as the model \mathcal{R} for scene-grounding guidance and trajectory initialization. For sequence generation, we employ the camera-controlled image-to-video diffusion model [61] which supports the generation of $L = 25$ frames. The weighting factors, λ , λ_{perc} , λ_{gen1} , and λ_{gen2} are set to 0.2, 10^{-4} , 0.1, and 0.01, respectively. The threshold of η_{mask} is set to 0.9, while η is set to 0.5. The generation interval N_{gen} is set to 260 and N_{iter} is set to 10,000.

4.2. Comparisons

Comparison on Replica. As shown in Tab. 1, our method achieves the highest performance on the Replica dataset, outperforming DNGaussian [21] and FSGS [69] by a significant margin of over 3.0 dB in PSNR. Fig. 5 illustrates that our method effectively addresses occlusion and extrapolation, while other 3DGS-based methods struggle with these challenges. Additionally, their depth regularization often compromise thin structures, such as the flower in the vase in the second row. FreeNeRF [56] exhibits severe artifacts because it cannot effectively utilize the strong prior from

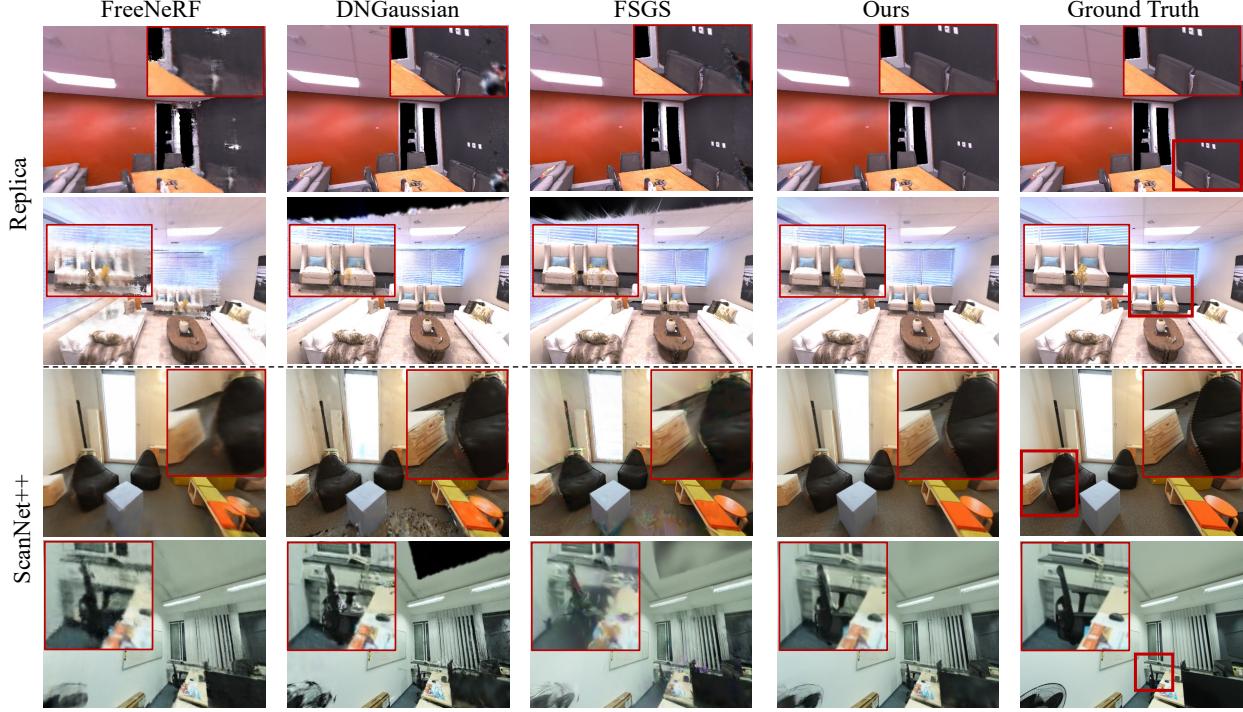


Figure 5. Qualitative comparisons on the Replica and ScanNet++ datasets. All 3DGS-based methods are optimized using the initialized point cloud from DUSt3R [50]. Our method effectively addresses the issues of extrapolation and occlusion while preserving finer details and reducing artifacts. For better visualization, please zoom in on the results.

Method	Replica [44]			ScanNet++ [58]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Mip-NeRF [2]	18.12	0.707	0.391	19.58	0.755	0.389
InfoNeRF [20]	13.07	0.598	0.552	14.54	0.646	0.495
DietNeRF [16]	18.99	0.676	0.444	19.76	0.719	0.431
FreeNeRF [56]	20.99	0.765	0.324	20.17	0.756	0.368
S ³ NeRF [68]	22.54	0.800	0.287	22.21	0.787	0.364
3DGS [‡] [19]	22.80	0.818	0.179	21.41	0.817	0.211
DNGaussian [21]	17.63	0.718	0.435	19.01	0.754	0.367
DNGaussian [†] [21]	22.71	0.821	0.189	20.68	0.788	0.281
FSGS [69]	20.22	0.760	0.304	17.95	0.730	0.373
FSGS [†] [69]	22.99	0.833	0.205	21.23	0.813	0.257
Ours	26.35	0.872	0.122	23.89	0.850	0.182

Table 1. Quantitative comparisons on the Replica and ScanNet++ datasets. Including our approach, 3DGS-based methods marked with \ddagger are initialized with the point cloud from DUSt3R [50].

the DUSt3R point cloud. Although FreeNeRF can fill hole regions through neighboring interpolation (e.g., the wall behind the chair in the first row and the ceiling in the second row), the results frequently exhibit blurring or artifacts.

Comparison on ScanNet++. ScanNet++ is a dataset captured in realistic scenes, so it is more complicated and challenging than the synthetic Replica dataset. The results in Tab. 1 demonstrate that our method has a clear advantage over current approaches, surpassing FSGS by more than 2.5 dB in PSNR. As depicted in Fig. 5, our method effectively addresses the extrapolation issue (e.g., the ceiling in the fourth row) and mitigates needle-like artifacts observed

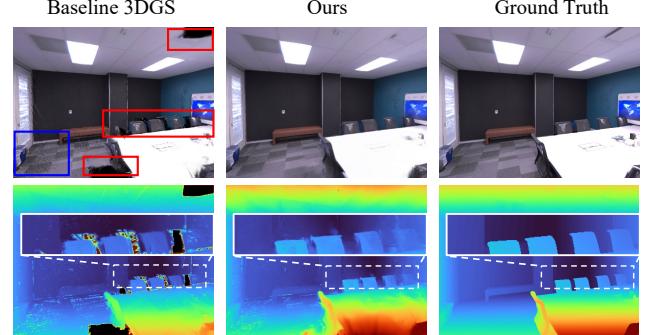


Figure 6. Our method not only effectively addresses extrapolation and occlusion (red boxes), improving the overall quality (blue boxes), but also predicts more plausible geometry.

in the rendered images of DNGaussian [21] and FSGS [69] (the third row). Furthermore, the comparisons in the third row highlight our method’s superiority in preserving finer details compared to all other methods.

4.3. Ablation Studies

Our technical contributions consist of three key components. We analyze their effects on the Replica dataset.

Generation with Scene-Grounding Guidance. Optimizing a 3DGS with sequences from vanilla generation results in quality degradation. In Tab. 2 (a), while the full image metrics are enhanced due to slightly improved modeling at occluded regions, the visual quality degrades, as indicated by PSNR of observable regions dropping from 25.45 dB

(a)	Gen.	Guide.	Traj.	Full Image			Observable Regions			(b)	PSNR↑	SSIM↑	LPIPS↓
				PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓				
Baseline 3DGS				22.80	0.818	0.179	25.45	0.860	0.129	Baseline 3DGS	22.80	0.818	0.179
w/ Vanilla Generation	✓			23.69	0.840	0.160	25.00	0.870	0.119	w/ Guided Generation&Traj.	25.58	0.859	0.138
w/ Guided Generation	✓	✓		25.03	0.852	0.139	26.52	0.881	0.101	w/ perceptual loss	26.35	0.872	0.122
w/ Guided Generation&Traj.	✓	✓	✓	25.58	0.859	0.138	26.53	0.883	0.100	w/o local sampling	26.28	0.871	0.127
										w/o global list	26.01	0.867	0.122

Table 2. Ablation experiments on the Replica dataset. (a) Effectiveness of the proposed scene-grounding guidance (*Guide.*) for generation, and the trajectory initialization strategy (*Traj.*). (*Gen.*) indicates utilizing generated sequences for modeling. Metrics of observable regions mask out regions outside the field of view or occluded. (b) Effectiveness of the proposed scheme for 3DGS optimization.



Figure 7. The perceptual loss for generated views greatly increases the modeling capability at hole regions.

	PSNR↑	SSIM↑	LPIPS↓
Baseline 3DGS+LaMa [45]	24.56	0.833	0.167
Baseline 3DGS+SDS [37] with SDInpaint	24.33	0.844	0.153
Baseline 3DGS+SDS [37] with SDInpaint*	25.15	0.853	0.141
Ours	26.35	0.872	0.122

Table 3. Comparisons with inpainting methods on the Replica dataset. * indicates the usage of our trajectory initialization.

to 25.00 dB. This degradation is attributed to inconsistencies within generated sequences, which can result in black shadows in rendered images as illustrated in Fig. 4. In contrast, our scene-grounding guidance ensures that the generated sequences remain consistent, significantly enhancing the modeling capability in regions outside the field of view and occluded, while also improving the overall quality, evidenced by the ‘w/ Guided Generation’ results in Tab. 2 (a).

Trajectory Initialization Strategy. Tab. 2 (a) further demonstrates that the proposed trajectory initialization strategy significantly boosts the performance, notated as ‘w/ Guided Generation&Traj’. The improvement mainly arises from enhanced modeling of the regions outside the field of view or the occluded areas, as the metrics of visible regions plateau while the overall image metrics improve by over 0.5 dB in PSNR. This indicates that the initialization strategy effectively identifies hole regions for holistic modeling.

Scheme for 3DGS Optimization with Generation. We verify the effectiveness of the proposed scheme in Tab. 2 (b). Specifically, the perceptual loss of Eq. (9) increases PSNR by over 0.5 dB, which is crucial for the model to fill the hole regions, as shown in Fig. 7. We empirically find that local sampling brings improvement in Sec. 3.4. This is evidenced by the performance decrease of ‘w/o local sampling’, which randomly samples generated views from all generated sequences. Alg. 2 shows that we use a global list to avoid the forgetting problem, and its necessity is verified by an over 0.3 dB PSNR drop observed with ‘w/o global list’. Combining these contributions, our full model effec-

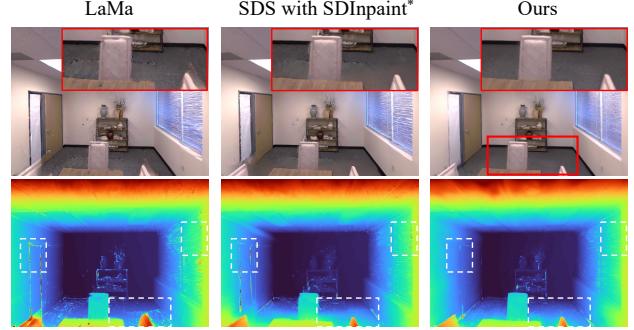


Figure 8. Qualitative comparisons with other inpainting methods. Our approach not only produces more plausible appearances around the inpainting regions but also predicts more consistent geometries in fine-grained local areas.

tively addresses extrapolation and occlusion while enhancing overall image quality, meanwhile exhibiting much better geometry, as shown in Fig. 6.

4.4. Further Comparisons with Inpainting Methods

Extrapolation and occlusion can also be addressed using inpainting methods. We thus compare our approach with two inpainting-based methods. One method applies LaMa [45] inpainting on hole regions, while the other optimizes a 3DGS by Score Distillation Sampling (SDS) [37] based on a SDInpaint model [39]. We also incorporate our proposed trajectory initialization into the SDS method to enhance the optimization of inpainting regions. Results in Tab. 3 show that our method outperforms these two methods by more than 1.0 dB in PSNR. Qualitative results in Fig. 8 indicate that, under certain conditions, the SDS-based method produces inpainted areas with strange appearances, while LaMa tends to create blurring artifacts interpolated from neighboring regions. Besides, Fig. 8 also shows that our method predicts better geometry in regions with local details. The inpainting results from our approach are more plausible due to the well-designed guidance, which effectively exploits prior knowledge from the diffusion model.

5. Conclusion

In this paper, we have explored to address the critical issues of extrapolation and occlusion in sparse-input 3DGS modeling. We propose using video diffusion models that provide plausible interpretations for regions that are outside

the field of view and occluded. To resolve inconsistencies within generated sequences, we introduce a novel scene-grounding guidance that controls the diffusion model to generate consistent sequences without any fine-tuning. Additionally, we propose a trajectory initialization strategy to enhance holistic modeling and develop a scheme for optimizing 3DGSS with generated sequences. Extensive experiments validate our approach, demonstrating that it outperforms current methods by a significant margin.

References

- [1] 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](#)
- [2] 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](#), [2](#), [7](#), [3](#)
- [3] 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.
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. [1](#), [2](#)
- [5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. [3](#)
- [6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 2021. [3](#)
- [7] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *CVPR*, 2024. [3](#)
- [8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. [2](#), [3](#)
- [9] 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*, 2024. [3](#)
- [10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. [3](#)
- [11] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantssplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. [3](#), [1](#), [2](#)
- [12] 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. In *NeurIPS*, 2024. [1](#), [3](#)
- [13] Liang Han, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. In *NeurIPS*, 2024. [1](#)
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. [3](#), [4](#)
- [15] 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 preprint arXiv:2210.02303*, 2022. [2](#), [3](#)
- [16] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. [7](#), [3](#)
- [17] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. [2](#)
- [18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. [5](#), [6](#)
- [19] 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. [2](#), [4](#), [6](#), [7](#), [3](#)
- [20] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. [3](#), [7](#)
- [21] 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. [2](#), [3](#), [6](#), [7](#)
- [22] Yixin Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *CVPR*, 2024. [3](#)
- [23] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. [3](#)
- [24] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024. [2](#)
- [25] 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. [3](#)
- [26] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In *NeurIPS*, 2024. [3](#)
- [27] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *CVPR*, 2022. [3](#)
- [28] 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. [2](#)

- [29] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Francisco Vicente Carrasco, Markus Steinberger, and Fernando De La Torre. Taming 3dgs: High-quality radiance fields with limited resources. *arXiv preprint arXiv:2406.15643*, 2024. 2
- [30] 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*, 2021. 2
- [31] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM TOG*, 38(4):1–14, 2019. 2
- [32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3
- [33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 3
- [34] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 3
- [35] 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. 2, 3
- [36] Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. Coherentngs: Sparse novel view synthesis with coherent 3d gaussians. In *ECCV*, 2024. 3
- [37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3, 8
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
- [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 8
- [40] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1
- [42] 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
- [43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3, 4
- [44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6, 7, 3
- [45] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 8, 3
- [46] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024. 3
- [47] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 1
- [48] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, 2023. 2, 3
- [49] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 3
- [50] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3, 6, 7, 1, 2, 4
- [51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6
- [52] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2024. 3
- [53] 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 *CVPR*, 2024. 3, 1
- [54] Jamie Wynn and Daniyar Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *CVPR*, 2023. 3
- [55] 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*, 2024. 2, 3
- [56] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, 2023. 6, 7, 3
- [57] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024. 2, 3

- [58] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2, 6, 7, 3
- [59] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 3
- [60] Jiwen Yu, Yinhui Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *ICCV*, 2023. 2, 3, 4
- [61] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3, 4, 6, 1
- [62] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024. 2, 3
- [63] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2, 3
- [64] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024. 3
- [65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3
- [66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6
- [67] Yingji Zhong, Lanqing Hong, Zhenguo Li, and Dan Xu. Cvt-xrf: Contrastive in-voxel transformer for 3d consistent radiance fields from sparse inputs. In *CVPR*, 2024. 2, 3
- [68] Yingji Zhong, Kaichen Zhou, Zhihao Li, Lanqing Hong, Zhenguo Li, and Dan Xu. Empowering sparse-input neural radiance fields with dual-level semantic guidance from dense novel views. *arXiv preprint arXiv:2503.02230*, 2025. 2, 6, 7
- [69] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *ECCV*, 2024. 3, 6, 7
- [70] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002. 4

Taming Video Diffusion Prior with Scene-Grounding Guidance for 3D Gaussian Splatting from Sparse Inputs

Supplementary Material

A. Implementation Details

During the denoising sampling process, we employ the DDIM sampler [41] combined with our proposed guidance, setting the number of sampling steps to 50. Regarding the trajectory initialization strategy, for each input view in its camera space, we sample views by changing the polar/azimuth angle to $[-30^\circ, -15^\circ, 0^\circ, 15^\circ, 30^\circ]$, and setting the radial distance to $[1, \frac{1}{3}, \frac{1}{10}]$ of the depth of the center pixel (from the prediction of ViewCrafter [61]). Out of 75 sampled views, we discard those whose renderings exhibit holes larger than 10% of the image size (to filter out uncommon viewpoints), then select the top 6 views with the largest holes from the remaining. To obtain the point cloud used for initialization, we follow the standard pipeline provided on the DUSt3R [50] webpage. Since our focus is sparse-input radiance fields reconstruction, the groundtruth camera poses and intrinsics are provided. During DUSt3R optimization, we fix both the poses and intrinsics to their groundtruth values.

B. More Results

Our method focuses on holistic modeling of an indoor scene of a moderate size, and we conduct the experiments in the main paper with 6 input views, since 6 input views are basically sufficient to cover the entire room. To validate the effectiveness of our method, we also test our method with different number of views following the common 3/6/9-view settings of sparse-input modeling. Tab. A1 validates that, our method is effective given different number of input views, with consistent improvements over our baseline. InstantSplat [11] is a strong baseline of sparse-input pose-free modeling, leveraging DUSt3R [50] point cloud for 3DGS initialization. Our method also consistently outperforms InstantSplat as shown in Tab. A1.

To obtain a thorough understanding of the source of the performance improvement, we show some quantitative results regarding performances of observable and the unobservable regions respectively in Tab. A2. The results show that our method brings improvement in both observable and unobservable regions.

We further compare our method with two representative methods that leverage diffusion models for sparse-input modeling, ReconFusion [53] and CAT3D [12] on the datasets of RealEstate10K and LLFF. We adhere to their settings for fair comparisons and the results are shown in Tab. A3. On the LLFF dataset, our method is based on the strong baseline of binocular-guided 3DGS [13]. The

results show that our method achieves comparable performance with both ReconFusion and CAT3D.

We provide per-scene comparisons in Table A4, demonstrating that our method consistently achieves superior performance across all scenes. Additional qualitative results are shown in Fig. A3. These results highlight the effectiveness of our approach in addressing issues such as extrapolation and occlusion, as seen in examples like the wall behind the chair (second row) and the ceiling (third row). Furthermore, our method preserves more intact structures with finer details, such as the edges in the fifth and sixth rows.

We present a comparison of the generated sequences from the video diffusion model with and without the proposed guidance in Fig. A2. The results clearly show that our proposed guidance enhances the plausibility of the generated sequences by maintaining consistent appearances and ensuring that only elements present in the scene are generated. Consistency in the generated video is crucial for effective 3DGS optimization. Using inconsistent sequences for 3DGS optimization often leads to artifacts, such as black shadows in the renderings, which significantly degrade visual quality, as demonstrated on the demo page.

C. Discussion

While our approach significantly improves overall quality by addressing extrapolation and occlusion challenges, we observe that it occasionally produces over-smoothed results. We hypothesize that this is due to the limited resolution supported by the video diffusion model during generation. On a 32GB V100 GPU, we are constrained to generating sequences at resolutions of 320×448 for the Replica dataset and 320×512 for the ScanNet++ dataset, which are subsequently upsampled to rendering resolutions of 480×640 and 480×720, respectively, for supervision during 3DGS optimization. This upsampling process introduces undersampling, which can smooth out certain regions and result in over-smoothed effects. Addressing the challenge of preserving high-frequency details during 3DGS optimization under resource-limited sequence generation remains an open problem and is a direction for future work.



Figure A1. Point clouds from DUStr3R [50] optimized with sparse input views on the Replica dataset. The yellow parts represent unobserved regions, e.g., regions that are outside the field of view or occluded. Note that the ceilings are removed for better visualization.

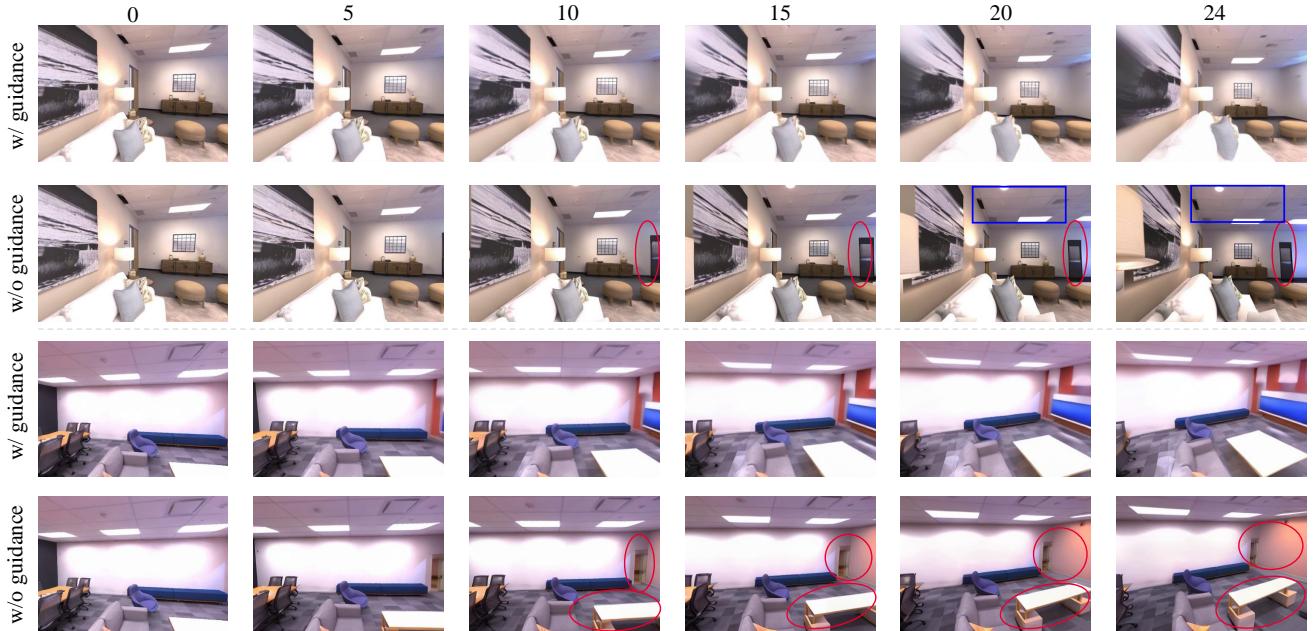


Figure A2. Generated frames from the video diffusion model with and without the proposed guidance. The numbers at the top indicate the frame IDs. The first frame corresponds to an image from the sparse input views, while other frames are generated. Without guidance, the generated sequences exhibit significant inconsistencies: (i) appearance inconsistencies, highlighted by the blue boxes; and (ii) hallucinated elements that do not exist in the scene, highlighted by the red boxes. In contrast, with the proposed guidance, the generated sequences are more plausible and consistent.

	3-view			6-view			9-view		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Replica									
Baseline 3DGS	19.87	0.794	0.178	22.80	0.818	0.179	24.81	0.863	0.124
InstantSplat [11]	20.49	0.766	0.226	20.35	0.760	0.290	18.44	0.708	0.373
Ours	23.98	0.848	0.136	26.35	0.872	0.122	27.42	0.891	0.111

Table A1. Our method brings performance improvement over the baseline with different number of input views, and consistently outperforms another strong sparse-input modeling baseline InstantSplat [11].

Replica 6-view	Full Image			Observable Regions			Unobservable Regions		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline 3DGS w/ Vanilla Generation Ours	22.80	0.818	0.179	25.45	0.860	0.129	14.27	0.967	0.029
	23.69	0.840	0.160	25.00	0.870	0.119	17.11	0.977	0.025
	26.35	0.872	0.122	27.12	0.894	0.091	20.85	0.985	0.020
Baseline 3DGS+LaMa [45]	24.56	0.833	0.167	25.45	0.860	0.129	17.80	0.981	0.021
Baseline 3DGS+SDInpaint [39]*	25.15	0.853	0.141	26.13	0.878	0.104	19.25	0.982	0.022

Table A2. Analysis of performance regarding observable and unobservable regions. * refers to incorporating our trajectory initialization strategy. The methods in the second block utilize inpainting models.

	3-view			6-view			9-view		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RealEstate10K									
ReconFusion [53]	25.84	0.910	0.144	29.99	0.951	0.103	31.82	0.961	0.092
CAT3D [12]	26.78	0.917	0.132	31.07	0.954	0.092	32.20	0.963	0.082
Ours	25.03	0.871	0.136	30.62	0.944	0.069	32.45	0.955	0.062
LLFF									
ReconFusion [53]	21.34	0.724	0.203	24.25	0.815	0.152	25.21	0.848	0.134
CAT3D [12]	21.58	0.731	0.181	24.71	0.833	0.121	25.63	0.860	0.107
Ours	21.35	0.746	0.173	25.13	0.851	0.102	26.29	0.880	0.084

Table A3. Comparisons with ReconFusion [53] and CAT3D [12] on the RealEstate10K and LLFF datasets.

	ScanNet++ [58]					Replica [44]						
	a2ccc	8a20d	94ee1	78318	avg	office2	office3	office4	room0	room1	room2	avg
Mip-NeRF [2]	18.28	23.48	16.93	19.63	19.58	17.43	19.04	19.08	17.46	16.57	19.16	18.12
	0.759	0.799	0.725	0.735	0.755	0.539	0.685	0.727	0.762	0.721	0.808	0.707
	0.351	0.321	0.431	0.451	0.389	0.486	0.421	0.393	0.342	0.386	0.317	0.391
InfoNeRF [20]	13.90	17.69	14.34	12.21	14.54	13.66	12.53	11.51	12.58	14.11	14.00	13.07
	0.662	0.691	0.627	0.605	0.646	0.463	0.545	0.592	0.618	0.689	0.678	0.598
	0.468	0.437	0.516	0.558	0.495	0.612	0.623	0.624	0.542	0.435	0.477	0.552
DietNeRF [16]	20.67	23.00	15.34	20.02	19.76	19.12	19.35	18.97	19.84	17.18	19.46	18.99
	0.751	0.776	0.627	0.725	0.719	0.612	0.695	0.419	0.783	0.749	0.797	0.676
	0.385	0.363	0.516	0.459	0.431	0.458	0.417	0.721	0.34	0.386	0.343	0.444
FreeNeRF [56]	19.93	22.37	19.42	18.94	20.17	20.89	21.06	20.25	22.55	19.69	21.43	20.99
	0.759	0.791	0.762	0.711	0.756	0.688	0.735	0.750	0.831	0.781	0.807	0.765
	0.307	0.299	0.417	0.449	0.368	0.359	0.340	0.364	0.234	0.325	0.321	0.324
S ³ NeRF [?]	21.81	25.60	20.05	21.36	22.21	22.79	23.83	23.08	24.01	19.66	21.87	22.54
	0.801	0.811	0.784	0.753	0.787	0.728	0.773	0.801	0.862	0.808	0.825	0.800
	0.324	0.330	0.357	0.444	0.364	0.326	0.309	0.301	0.213	0.277	0.293	0.287
3DGS [†] [19]	20.65	23.49	20.38	21.11	21.41	25.03	23.60	22.14	20.32	22.68	23.07	22.80
	0.824	0.857	0.821	0.764	0.817	0.873	0.858	0.834	0.720	0.802	0.824	0.818
	0.193	0.136	0.218	0.298	0.211	0.141	0.147	0.180	0.204	0.203	0.196	0.179
DNGaussian [21]	19.10	21.21	17.55	18.20	19.01	22.68	18.40	12.31	12.60	18.87	20.91	17.63
	0.765	0.781	0.743	0.730	0.755	0.843	0.789	0.644	0.534	0.708	0.790	0.718
	0.343	0.292	0.382	0.450	0.367	0.233	0.291	0.628	0.722	0.397	0.338	0.435
DNGaussian [†] [21]	20.47	23.73	18.90	19.61	20.68	25.31	23.34	21.83	20.33	22.59	22.88	22.71
	0.805	0.842	0.784	0.722	0.788	0.890	0.853	0.837	0.729	0.800	0.820	0.821
	0.213	0.183	0.287	0.357	0.281	0.124	0.161	0.197	0.226	0.208	0.219	0.189
FSGS [69]	19.19	18.98	15.77	17.87	17.95	20.70	20.26	21.62	19.65	19.23	19.89	20.22
	0.760	0.735	0.719	0.708	0.730	0.802	0.790	0.825	0.654	0.712	0.779	0.760
	0.321	0.316	0.415	0.442	0.373	0.266	0.255	0.271	0.315	0.374	0.342	0.304
FSGS [†] [69]	21.28	22.56	20.28	20.79	21.23	24.37	23.41	23.45	21.02	23.56	22.14	22.99
	0.826	0.844	0.815	0.767	0.813	0.873	0.856	0.862	0.759	0.823	0.822	0.833
	0.219	0.193	0.267	0.350	0.257	0.194	0.174	0.189	0.198	0.205	0.270	0.205
Ours	25.21	25.10	23.10	22.16	23.89	27.46	26.81	27.43	24.85	26.00	25.53	26.35
	0.857	0.882	0.860	0.803	0.850	0.916	0.902	0.897	0.796	0.851	0.872	0.872
	0.157	0.118	0.201	0.269	0.182	0.083	0.099	0.122	0.145	0.142	0.122	0.122

Table A4. Per-scene performance of various models on the ScanNet++ and Replica datasets. For each method, the three rows represent PSNR, SSIM, and LPIPS, respectively. avg indicates the average performance across all scenes in each dataset. Including our approach, 3DGS-based methods marked with [†] are initialized with the point cloud from DUS3R [50].

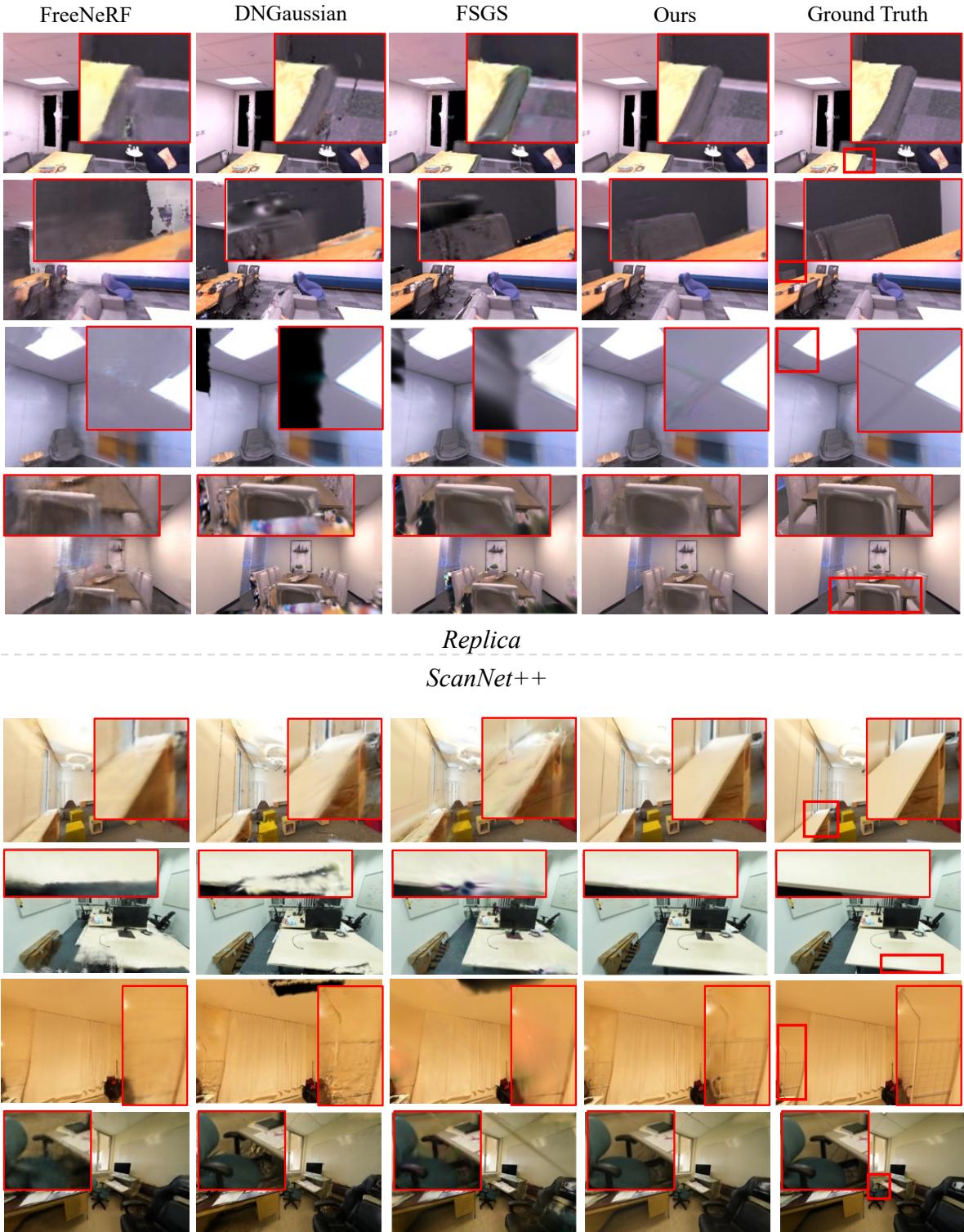


Figure A3. Qualitative comparisons between other works on Replica and ScanNet++ datasets. All 3DGS-based methods are optimized using the initialized point cloud from DUST3R [50].