GaMO: Geometry-aware Multi-view Diffusion Outpainting for Sparse-View 3D Reconstruction

Anonymous authors
Paper under double-blind review

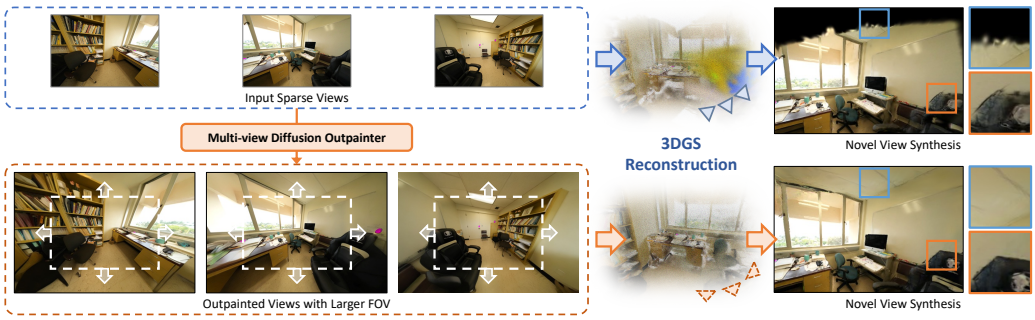

Figure 1: **GaMO: Geometry-Aware Multi-View Diffusion Outpainting for Sparse-View 3D Reconstruction.** GaMO employs zero-shot geometry-aware conditioning and denoising to expand sparse input views into geometrically consistent wide-FOV images, increasing scene coverage from the same camera viewpoints without fine-tuning. The expanded views enable high-quality 3DGS reconstruction and novel view synthesis.

Abstract

Recent 3D reconstruction methods achieve impressive results with dense multi-view imagery but struggle when only a few views are available. Various approaches, including regularization techniques, semantic priors, and geometric constraints, have been implemented to address this challenge. Recent diffusion-based approaches further improve performance by generating novel views to augment training data. Despite this progress, we identify three critical limitations in current state-of-the-art approaches: (i) inadequate coverage beyond known view peripheries, (ii) geometric inconsistencies across generated views, and (iii) computational inefficiency due to expensive pipelines. We introduce **GaMO** (**G**eometry-**a**ware **M**ulti-view **O**utpainter), a framework that reformulates sparse-view reconstruction through multi-view outpainting. Instead of generating new viewpoints, GaMO expands the field of view from existing camera poses, which inherently preserves geometric consistency while providing broader scene coverage. Our approach employs multi-view conditioning and geometry-aware denoising strategies in a zero-shot manner without training. Extensive experiments on Replica, ScanNet++, and Mip-NeRF 360 demonstrate strong reconstruction performance across sparse-view settings (3, 6, and 9 input views). Notably, our method is significantly more efficient than existing diffusion-based approaches, reducing the overall runtime to within 10 minutes.

1 Introduction

Reconstructing complete 3D scenes from limited input views is a fundamental problem with numerous tangible applications, ranging from virtual property tours to immersive telepresence. However, it remains notoriously difficult, often resulting in broken geometry and visible visual artifacts. Previous approaches attempted to address the sparsity of input views through regularization, semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

priors, or geometric constraints (Niemeyer et al., 2022; Jain et al., 2021; Yang et al., 2023; Zhu et al., 2024; Wang et al., 2023a; Li et al., 2024b). These methods remain limited in handling unobserved regions.

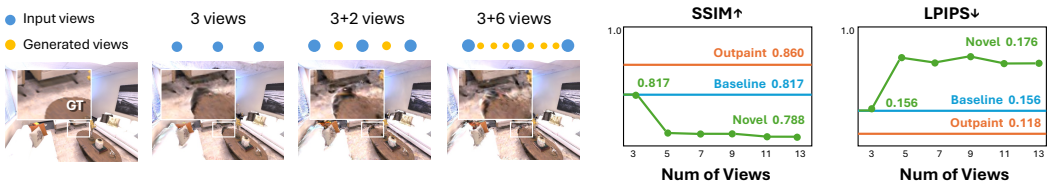

Figure 2: **Motivation: Outpainting vs. diffusion-generated novel views.** Using multi-view diffusion (Cao et al., 2025a), we train 3DGS (Kerbl et al., 2023) with three settings: interpolated views (green), a 3-view baseline (blue), and outpainting (orange). **Left:** Adding more diffusion-generated views (3–13) may introduce geometric inconsistencies under sparse-view conditions. **Right:** SSIM/LPIPS show a similar trend: increasing generated views can degrade reconstruction quality, while outpainting remains more stable.

Recently, diffusion-based approaches (Zhong et al., 2025a; Wu et al., 2025c; 2024b; 2025b; Ancikevičius et al., 2023) have generated novel views to improve the reconstruction quality for sparse observations. Nevertheless, these methods show three fundamental limitations: (1) novel view generation mainly focuses on enhancing angular coverage of existing geometry and often overlooks the extension beyond the periphery, leaving persistent **holes** and **ghostings** in the reconstruction; (2) geometric and photometric **inconsistencies** across novel and input views inevitably become prominent as view overlap increases due to internal diffusion variations; (3) novel view generation requires elaborate trajectory planning and camera pose sampling, making the process **time-consuming**.

Recent multi-view diffusion models incorporate geometry-aware priors and achieve strong geometric consistency across generated views. However, we find that directly using them to expand camera viewpoints is not suitable under extremely sparse-view settings. Fig. 2 shows an experiment where three input views are used to generate interpolated novel views via diffusion, which are then used to train 3DGS (Kerbl et al., 2023). Increasing the number of generated views (from 3 to 13 total views) degrades reconstruction quality, leading to lower SSIM and higher LPIPS. This observation suggests that generating additional viewpoints is not always beneficial under sparse-view settings.

Instead, we observe that *outpainting*, rather than novel view generation, better leverages the geometry-aware priors of multi-view diffusion models for sparse-view 3D reconstruction. By extending content around existing views, outpainting expands coverage while preserving geometric consistency. We therefore propose **GaMO** (**G**eometry-**a**ware **M**ulti-view **O**utpainting), which (1) expands the field of view (FOV) to cover unobserved regions, (2) avoids fusing multiple hallucinated views in 3D space, reducing misalignment artifacts, and (3) reconstructs a scene in a single outpainting pass, tens of times faster than video diffusion-based methods. On Replica (Straub et al., 2019), ScanNet++ (Yeshwanth et al., 2023), and Mip-NeRF 360 (Barron et al., 2022), GaMO achieves competitive performance with improved perceptual quality.

We summarize our contributions as follows:

- We establish outpainting as a superior paradigm for sparse-view reconstruction, eliminating common issues including holes, ghosting artifacts, and geometric inconsistencies.
- We develop a geometry-aware outpainting approach with novel conditioning and denoising strategies in a zero-shot manner, without fine-tuning.
- We achieve competitive performance across Replica, ScanNet++, and Mip-NeRF 360, demonstrating strong geometric accuracy and perceptual quality with 3, 6, and 9 input views while reducing reconstruction time to under 10 minutes.

2 Related Work

Sparse-view 3D Gaussian Splatting. Sparse-view 3DGS (Kerbl et al., 2023) remains difficult (Younis & Cheng, 2025), especially indoors (Li et al., 2024c). Optimization-based methods

add depth, geometric, or large-model priors (Li et al., 2024b; Chung et al., 2023b; Zhu et al., 2024; Zhang et al., 2024d; Park et al., 2025; Peng et al., 2024; Lin et al., 2025a; Turkulainen et al., 2025; Mihajlovic et al., 2024; Yu et al., 2024b; Wang et al., 2024d; He et al., 2025b; Shih et al., 2025), or diffusion priors that synthesize pseudo-views, repair renderings, or inpaint unseen regions (Kong et al., 2025; Yang et al., 2024a; Paul et al., 2024; Paliwal et al., 2025; Fischer et al., 2025; Yin et al., 2025; Liu et al., 2026a; Ni et al., 2026); feed-forward methods range from cost volumes and transformers to 3D foundation, ray-conditioned, and generative scene models (Chen et al., 2024b; Charatan et al., 2024; Fan et al., 2024; Xu et al., 2025a; Liu et al., 2024f; Wang et al., 2024f; Xu et al., 2025b; Tang et al., 2024b; Min et al., 2024; Szymanowicz et al., 2024; Leroy et al., 2024; Ye et al., 2025; Wang et al., 2025c; Jiang et al., 2025b; Jin et al., 2024; Jiang et al., 2025a; Wang et al., 2025b; Szymanowicz et al., 2025; Liang et al., 2025; Bahmani et al., 2026). Quality still drops sharply outside the input coverage (Han et al., 2025): these methods *regularize* the representation or *inpaint* at novel poses, whereas we *augment* training data by outpainting the input views themselves.

Multi-view diffusion models for 3D. Building on view-conditioned diffusion (Liu et al., 2023a), multi-view diffusion enforces cross-view consistency through attention, geometric, and camera-ray conditioning (Shi et al., 2023c; Wang & Shi, 2023; Liu et al., 2023b; Shi et al., 2023b; Long et al., 2024; Kong et al., 2024; Yang et al., 2024b; Huang et al., 2024; Kuang et al., 2024; Hu et al., 2024; Lu et al., 2024a; Tang et al., 2024c; Li et al., 2024d; Liu et al., 2024d; Tang et al., 2023; Chan et al., 2023; Chen et al., 2023; Zheng & Vedaldi, 2024; Sitzmann et al., 2021), and lifts its outputs to meshes or Gaussians (Xu et al., 2024f; Tang et al., 2024a). Video diffusion adds temporal consistency, extends to pose-free videos, unposed inputs, and 4D scenes, and gains 3D-aware camera control or zero-shot steering (Voleti et al., 2024; Gao et al., 2024; Kwak et al., 2024; Zhang et al., 2024b; Chen et al., 2024c;d; Yu et al., 2024e; Ma et al., 2025; Li et al., 2025; Zhao et al., 2025; Xu et al., 2024c; Li et al., 2024a; Bahmani et al., 2025a; Yu et al., 2025b; Bai et al., 2025; He et al., 2025a; Cao et al., 2025b; Ren et al., 2025; You et al., 2024). All synthesize *novel views* at new poses; GaMO instead *outpaints* existing views, which better preserves geometric consistency.

Diffusion inpainting, outpainting, and scene extrapolation. Pre-trained diffusion models (Ho et al., 2020; Rombach et al., 2022) can be steered at sampling time via noised-reference guidance, masked replacement, resampling, manifold constraints, or window fusion (Meng et al., 2021; Choi et al., 2021; Avrahami et al., 2022; 2023; Lugmayr et al., 2022; Chung et al., 2022b;a; Bar-Tal et al., 2023; Lee et al., 2023), or trained for inpainting directly (Saharia et al., 2022; Suvorov et al., 2022; Ju et al., 2024; Zhuang et al., 2024); GaMO applies blended-latent replacement and RePaint-style resampling inside a multi-view diffusion model with a timestep-scheduled mask and no inpainting-specific weights. Outpainting predates diffusion (Teterwak et al., 2019; Yang et al., 2019; Wang et al., 2019; Lin et al., 2021; Yao et al., 2022); in 3D, perpetual view generation and single-view scene generation grow scenes outward (Liu et al., 2021; Li et al., 2022; Fridman et al., 2023; Höllein et al., 2023; Chung et al., 2023a; Shriram et al., 2024; Yu et al., 2024c; Engstler et al., 2025; Pu et al., 2024; Zhou et al., 2024; Seo et al., 2024; Sargent et al., 2024; Wang et al., 2025a; Chen et al., 2026; Huang et al., 2025), and 3D inpainting fills masked regions across views (Mirzaei et al., 2023; Liu et al., 2026b; Salimi et al., 2025; Weber et al., 2026; Chen et al., 2025). Closest to us, ExtraNeRF (Shih et al., 2024), OGGSplat (Wang et al., 2025d), MVInpainter (Cao et al., 2024), GEN3C (Ren et al., 2025), and concurrent 3DGS extrapolation (Zhou et al., 2026) also confine generation to unseen regions, but target novel poses or object edits and mostly rely on trained or per-scene fine-tuned models; GaMO keeps the input poses, enlarges only their field of view, and outpaints all views jointly and zero-shot. See appendix for more related work.

3 METHOD

To address challenges in sparse-view 3D reconstruction, we perform geometry-aware outpainting with multi-view diffusion models. Expanding the field-of-view (FOV) of the input images fills holes, fixes blurred boundaries, and preserves geometric consistency without modifying existing content, yielding a simpler and faster reconstruction process. As illustrated in Fig. 3, the pipeline has three stages: coarse 3D initialization for geometry priors (Sec. 3.1), geometry-aware multi-view outpainting to generate enlarged-FOV views (Sec. 3.2), and refined 3D reconstruction from the outpainted views (Sec. 3.3).

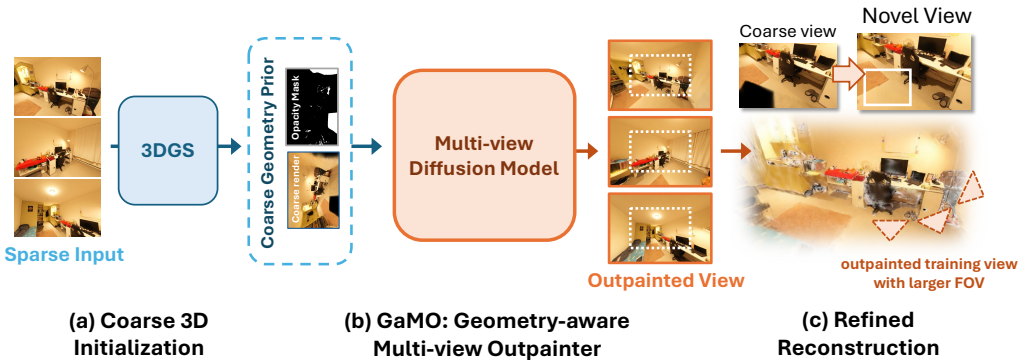

Figure 3: **Overview of Our Pipeline.** Given sparse input views, our method follows a three-stage process. **(a) Coarse 3D Initialization:** We obtain geometry priors from initial 3D reconstruction, including an opacity mask and coarse render that provide essential structural cues. **(b) Geometry-aware Multi-view Outpainter:** Using the geometry priors, GaMO generates outpainted views with enlarged FOV via a multi-view diffusion model. **(c) Refined Reconstruction:** The outpainted views are used to refine the 3D reconstruction, resulting in improved completeness and consistency.

3.1 Coarse 3D Initialization

To ensure geometric consistency in the diffusion model, we use DUSt3R (Wang et al., 2024e) to generate an initial point cloud and train a coarse 3DGS model. Using this coarse model, we identify outpainting regions by rendering an opacity mask with a FOV wider than the input views. We also render a coarse color image to provide appearance priors for the outpainting in Sec. 3.2.

Opacity Mask. We enlarge the FOV by reducing the focal lengths with a scaling ratio $S_k < 1$ (i.e., $f'_x = f_x \times S_k$, $f'_y = f_y \times S_k$). For each target outpainted view, we first render an opacity map $\mathcal{O}$ by α-blending the opacity values of the Gaussians:

$$\mathcal{O}(\mathbf{u}) = \sum_{i \in \mathcal{N}} \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \tag{1}$$

where $\sigma_i = \alpha_i \mathcal{G}'_i(\mathbf{u})$ denotes the opacity contribution of the i-th Gaussian at pixel $\mathbf{u}$. The opacity mask $\mathcal{M}$ is then obtained by thresholding the opacity map with $\mathcal{M} = \mathbb{I}(\mathcal{O} < \eta_{\text{mask}})$, where η_{mask} is a threshold value and $\mathbb{I}(\cdot)$ is the indicator function. Regions where $\mathcal{M} = 1$ correspond to areas with low opacity that require outpainting.

Coarse Rendering. We render a color image I_{coarse} with the enlarged FOV from the coarse 3DGS model. This coarse rendering serves as a reference that provides geometric and appearance priors to the diffusion model, maintaining consistency between outpainted and existing scene content.

3.2 GaMO: Geometry-aware Multi-view Diffusion Outpainter

Our geometry-aware outpainting method operates through three key components: (1) multi-view conditioning that provides structural and appearance guidance; (2) mask latent blending that integrates coarse geometry priors during denoising; and (3) iterative mask scheduling with noise resampling that ensure smooth transitions. The model operates in latent space using DDIM sampling (Song et al., 2021) for efficient denoising.

Multi-View Conditioning. Given a set of sparse input RGB images $\{I_i\}_{i=1}^N$ and their corresponding camera parameters $\{\Pi_i\}_{i=1}^N$, our model generates outpainted views conditioned on camera representations, geometric correspondences, and appearance features, as illustrated in Fig. 4(a).

For camera representation, we employ Plücker ray embeddings (Xu et al., 2024g) that provide dense 6D ray parameterizations for each pixel, compactly encoding both ray origin and direction for geometry-aware reasoning. The embedding of each input view $\mathcal{P}_r$ is derived from its corresponding camera parameters Π_r, while the embedding of the outpainted view $\mathcal{P}_t^*$ uses the same camera parameters with scaled focal lengths (f'_x, f'_y) to align with the enlarged FOV.

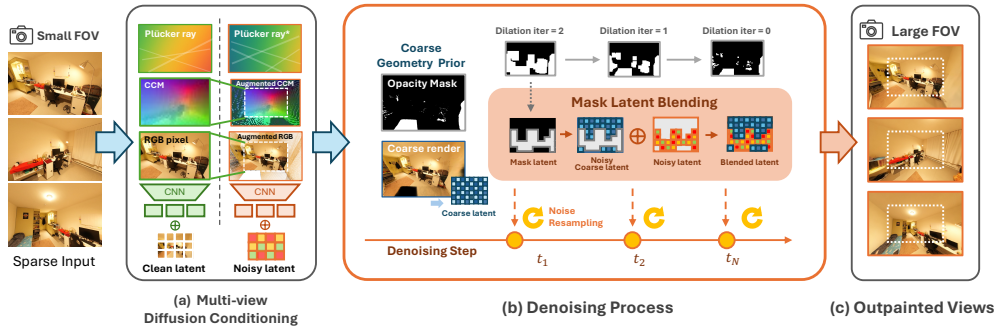

Figure 4: **Overview of GaMO (Geometry-aware Multi-view Diffusion Outpainter).** (a) Multi-view Diffusion Conditioning: Sparse input views are encoded into clean latents and combined with multi-view conditions, including Plücker ray embeddings for input views ($\mathcal{P}_r$) and the target view with enlarged FOV ($\mathcal{P}_t^*$), along with original and augmented Canonical Coordinate Map (CCM) and RGB, to provide both geometric and appearance cues for diffusion model conditioning. (b) Denoising Process: Coarse geometry priors (opacity mask and coarse render) guide the denoising through mask latent blending performed at multiple timesteps $(t_1, t_2, ..., t_N)$ with progressive dilation and noise resampling, generating outpainted views with enlarged FOV (c).

For geometric correspondence, we warp input RGB images and Canonical Coordinate Maps (CCM) (Li et al., 2024e) to align with the expanded FOV by unprojecting pixels to 3D and reprojecting onto the outpainted camera plane, producing $\mathcal{C}_{r\to t}^{\text{warp}}$ and $I_{r\to t}^{\text{warp}}$. We then downsample the original inputs by factor S_k and place them at the center of the warped features, creating augmented signals $I_{r\to t}^{\text{aug}}$ and $\mathcal{C}_{r\to t}^{\text{aug}}$ where the center preserves exact input information while the periphery retains warped geometric structure to guide outpainting.

For appearance features, we encode the input RGB images through a variational autoencoder (VAE) to obtain clean latent features $\mathbf{z}_r$. The noisy latent features $\mathbf{z}_t$ are randomly generated and will be denoised to generate the outpainted views.

Algorithm 1 Geometry-aware Multi-view Outpainter

Input: Coarse render I_{coarse}, opacity mask $\mathcal{M}$
Output: Outpainted views $\{S_j^{\text{out}}\}_{j=1}^{M}$
Setup: Noise schedule $\Sigma = \{\sigma_1, \ldots, \sigma_T\}$
Setup: latent blending iterations $\{t_1, \ldots, t_N\}$
Setup: resampling iterations R
$\mathbf{z}^{\text{coarse}} \leftarrow \text{Encode}(I_{\text{coarse}})$
$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $s = T, \ldots, 1$ **do**
  $\mathbf{z}_{s-1} = \text{Denoise}(\mathbf{z}_s, \text{conditions})$ ▷ equation 2
  if $s \in \{t_1, \ldots, t_N\}$ **then**
    $\mathbf{z}_{s-1}^{\text{coarse}} = \text{AddNoise}(\mathbf{z}^{\text{coarse}}, \sigma_{s-1})$
    $\mathbf{z}_{s-1}^{\text{blend}} = \text{IMS}(\mathcal{M}, \mathbf{z}_{s-1}, \mathbf{z}_{s-1}^{\text{coarse}})$ ▷ equation 3
    $\mathbf{z}_{s-1} = \mathbf{z}_{s-1}^{\text{blend}}$
    for r in R **do**
      $\hat{\mathbf{z}}_0 = \text{Predict}(\mathbf{z}_{s-1})$
      $\mathbf{z}_s^{\text{resamp}} \leftarrow \text{AddNoise}(\hat{\mathbf{z}}_0, \sigma_s)$ ▷ equation 4
      $\mathbf{z}_{s-1} \leftarrow \text{Denoise}(\mathbf{z}_s^{\text{resamp}}, \text{conditions})$
    end for
  end if
end for
$\{S_j^{\text{out}}\}_{j=1}^{M} = \text{Decode}(\mathbf{z}_0)$

Lightweight convolutional encoders process all conditioning signals. We add the Plücker ray embeddings $\mathcal{P}_r$, CCM $\mathcal{C}_r$, and RGB images I_r of the input views to the clean latent features $\mathbf{z}_r$, and $\mathcal{P}_t^*$, $\mathcal{C}_{r\to t}^{\text{aug}}$, and $I_{r\to t}^{\text{aug}}$ to the noisy latent features $\mathbf{z}_t$ of the target outpainted view. We then condition the pre-trained diffusion model on the fused features to generate outpainted view latents zero-shot,

$$p_\theta(\mathbf{z}_t | \mathbf{z}_r, \mathcal{P}_r, \mathcal{C}_r, I_r, \mathcal{P}_t^*, \mathcal{C}_{r\to t}^{\text{aug}}, I_{r\to t}^{\text{aug}}), \tag{2}$$

where θ denotes the pre-trained model (Cao et al., 2025a) parameters. These multi-view conditions keep the diffusion process geometrically consistent across views, even under an enlarged FOV.

Denoising Process with Mask Latent Blending. As the central component of our geometry-aware framework, mask latent blending integrates coarse geometry priors from the coarse 3D initialization (Sec. 3.1) into the diffusion loop. As outlined in Alg. 1, this process ensures that outpainted content respects existing scene structures while generating plausible peripheral regions. Fig. 4(b) shows that the opacity mask $\mathcal{M}$ and coarse rendering I_{coarse} provide consistent structural guidance throughout denoising.

At selected denoising timesteps $\{t_1, t_2, ..., t_N\}$, we perform *mask latent blending* between the denoised latent and the coarse geometry prior. To ensure both latents share the same noise level, we add noise to the coarse latent, which is obtained by encoding the coarse rendering into latent space, before blending them using a latent-space mask $\mathcal{M}_{\text{latent}}$. The mask evolution is controlled by

Table 1: **Quantitative comparison on Replica (Straub et al., 2019) and ScanNet++ (Yeshwanth et al., 2023) with 3 and 6 input views.** Red, orange, and yellow mark the best, second-best, and third-best result in each column. † numbers are those reported in the original paper, shown in gray and excluded from the ranking, with dashes where the setting is not reported; ‡ is our reproduction under the evaluation protocol used here. Run time is not ranked.

	Replica (3 views)			Replica (6 views)			ScanNet++ (3 views)			ScanNet++ (6 views)			
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Run time
FSGS	20.84	0.815	0.172	23.91	0.846	0.145	16.58	0.690	0.359	21.69	0.801	0.298	12 min
InstantSplat	20.65	0.815	0.142	23.09	0.849	0.140	16.72	0.720	0.315	21.19	0.811	0.193	1 min
Difix3D+	19.47	0.783	0.194	21.86	0.810	0.188	15.64	0.659	0.346	20.62	0.764	0.244	31 min
GenFusion	22.34	0.833	0.172	23.98	0.855	0.142	17.97	0.725	0.354	21.96	0.808	0.218	22 min
GuidedVD-3DGS†	23.98	0.848	0.136	26.35	0.872	0.122	-	-	-	23.89	0.850	0.182	-
GuidedVD-3DGS‡	25.26	0.864	0.138	26.68	0.880	0.133	18.82	0.720	0.328	22.98	0.815	0.204	3h 20 min
Ours	24.40	0.865	0.117	26.40	0.882	0.104	20.06	0.759	0.265	23.41	0.835	0.181	8 min

iterative mask scheduling:

$$\mathbf{z}_{t_k}^{\text{blend}} = (1 - \mathcal{M}_{\text{latent}}^{(k)}) \odot \mathbf{z}_{t_k}^{\text{coarse}} + \mathcal{M}_{\text{latent}}^{(k)} \odot \mathbf{z}_{t_k}, \tag{3}$$

where $\mathbf{z}_{t_k}$ is the denoised latent, $\mathbf{z}_{t_k}^{\text{coarse}}$ is the coarse latent with matching noise level (Alg. 1, line 11), $\mathcal{M}_{\text{latent}}^{(k)}$ is the dilated mask at iteration k, and $\odot$ denotes element-wise multiplication.

Iterative Mask Scheduling and Noise Resampling. To gradually integrate generated content with the existing geometric structure, Iterative Mask Scheduling progressively adjusts $\mathcal{M}_{\text{latent}}^{(k)}$ over iterations k to control the ratio between outpainting and known coarse regions. The mask dilation is progressively reduced as denoising proceeds, allowing the model to first explore peripheral content and later refine geometry within coarse regions.

To maintain smooth transitions across blended regions, we perform noise resampling after each blending operation. After blending, we perform noise resampling R times on the blended latent to eliminate boundary artifacts and ensure smooth integration between the coarse geometry and generated content (Alg. 1, lines 14–17). Specifically, we first predict the clean latent $\hat{\mathbf{z}}_0$ from the blended latent, then add noise back to the current timestep t_k:

$$\mathbf{z}_{t_k}^{\text{resamp}} = \sqrt{\bar{\alpha}_{t_k}}\hat{\mathbf{z}}_0 + \sqrt{1 - \bar{\alpha}_{t_k}}\boldsymbol{\epsilon}, \tag{4}$$

where $\hat{\mathbf{z}}_0$ is the predicted clean latent from $\mathbf{z}_{t_k}^{\text{blend}}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes sampled Gaussian noise. This resampling prevents boundary artifacts and ensures smooth blending.

This framework ensures that outpainted regions seamlessly blend with known content while maintaining geometric plausibility, with the coarse 3DGS geometry providing structural guidance throughout the generation process. Importantly, it requires only inference without fine-tuning the backbone diffusion model.

3.3 3DGS Refinement with Outpainted Views

Given the original input views $\{I_i^{gt}\}_{i=1}^N$ and the generated outpainted views $\{S_j^{\text{out}}\}_{j=1}^M$ from Sec. 3.2, we refine the 3DGS model by jointly optimizing with both sets of views. During training, we sample either an input view or an outpainted view for supervision at each iteration.

Loss for Input Views. We employ the standard 3DGS reconstruction loss (Kerbl et al., 2023) to ensure accurate reconstruction of the observed regions:

$$\mathcal{L}_{\text{input}} = (1 - \lambda_s)\mathcal{L}_1(I_i, I_i^{gt}) + \lambda_s\mathcal{L}_{\text{D-SSIM}}(I_i, I_i^{gt}), \tag{5}$$

where I_i denotes the rendered image from input viewpoint, I_i^{gt} is the ground truth input view, and λ_s is a weighting factor that balances the $\mathcal{L}_1$ loss and structural similarity loss.

Loss for Outpainted Views. Relying solely on reconstruction loss fails to fill unobserved regions and causes artifacts. We incorporate perceptual loss (Johnson et al., 2016) $\mathcal{L}_{\text{LPIPS}}$ to provide balanced gradients across outpainted and original regions, effectively guiding training while maintaining perceptual consistency. The loss is:

$$\begin{aligned} \mathcal{L}_{\text{recon}} &= (1 - \lambda_s)\mathcal{L}_1(S_j, S_j^{\text{out}}) + \lambda_s\mathcal{L}_{\text{D-SSIM}}(S_j, S_j^{\text{out}}), \\ \mathcal{L}_{\text{outpainted}} &= \mathcal{L}_{\text{recon}}(S_j, S_j^{\text{out}}) + \lambda_{\text{perc}}\mathcal{L}_{\text{LPIPS}}(S_j, S_j^{\text{out}}), \end{aligned} \tag{6}$$

where S_j denotes the rendered wide-FOV image and S_j^{out} is the generated outpainted image.

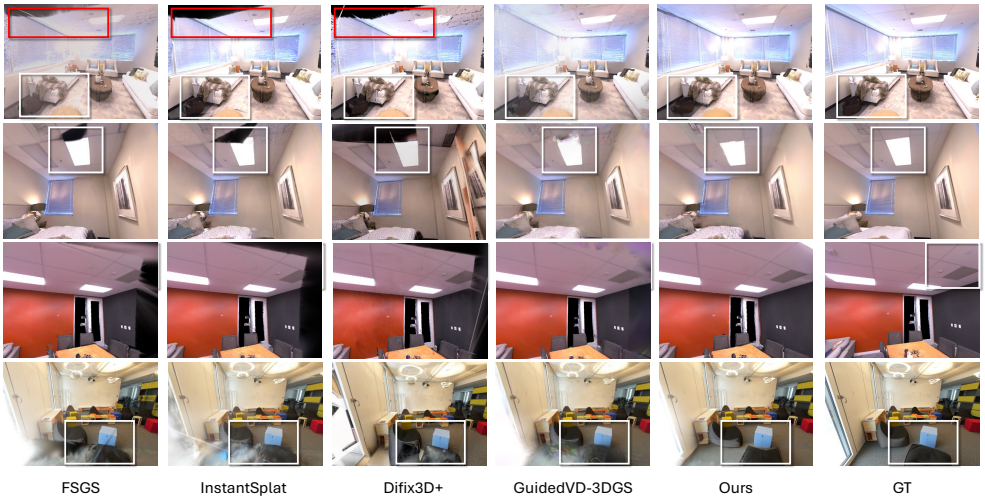

Figure 5: **Qualitative comparison on Replica and ScanNet++ under sparse 3-view and 6-view settings.** Our method produces better coverage, improved geometric consistency, and fewer artifacts. White boxes highlight challenging regions.

4 Experiments

4.1 Experimental Setups

Datasets and Evaluation Protocol. We evaluate on **Replica** (Straub et al., 2019), **ScanNet++** (Yeshwanth et al., 2023), and **Mip-NeRF 360** (Barron et al., 2022). Following (Zhong et al., 2025a;b), we adopt sparse-view settings with 3, 6, and 9 views for indoor datasets (**Replica** and **ScanNet++**), where the 9-view results are reported in the supplementary material, and 6 and 9 views for the unbounded **Mip-NeRF 360** dataset. For Replica and ScanNet++, we follow the view splits used in (Zhong et al., 2025a;b), while for Mip-NeRF 360 we use the original 6- and 9-view configurations. For ScanNet++, we additionally construct the 3- and 9-view setups to ensure suitable coverage for training and evaluation. Performance is evaluated using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018).

Implementation Details. We adopt **MVGenMaster** (Cao et al., 2025a) as the diffusion backbone, which supplies geometry-aware priors zero-shot. For fair comparison, we standardize resolutions to 512×384 on **Replica** and 576×384 on **ScanNet++** and **Mip-NeRF 360**. GaMO reconstructs a scene in about 8 minutes on a single RTX 4090, including diffusion outpainting and final 3DGS optimization. We provide further implementation details, hyperparameters, runtime breakdown, and extended results in the supplementary material.

Baselines. We compare against several state-of-the-art sparse-view reconstruction methods: (1) **FSGS** (Zhu et al., 2024), using depth-guided Gaussian unpooling; (2) **InstantSplat** (Fan et al., 2024), employing MASt3R priors and self-supervised bundle adjustment; (3) **Difix3D** (Wu et al., 2025b), applying single-step diffusion refinement; (4) **GenFusion** (Wu et al., 2025c), integrating reconstruction and video diffusion via cyclical fusion; and (5) **GuidedVD-3DGS** (Zhong et al., 2025a), leveraging video diffusion and evaluated using the authors' official implementation and settings. For fair comparison, all methods except InstantSplat use DUSt3R (Wang et al., 2024e) for initialization.

4.2 Comparisons

Quantitative Results. Tab. 1 and Tab. 2 compare methods on Replica, ScanNet++, and Mip-NeRF 360 under sparse views. On Replica and ScanNet++, GaMO obtains the best SSIM and lowest LPIPS in most settings. On the challenging Mip-NeRF 360 benchmark, it achieves the best PSNR and SSIM at both 6 and 9 views with competitive LPIPS. GaMO is also up to $25\times$ faster than the diffusion-based GuidedVD-3DGS (Zhong et al., 2025a), reconstructing a scene in under 10 minutes.

Table 2: **Quantitative comparison on Mip-NeRF 360 (Barron et al., 2022) with 6 and 9 input views.** Red, orange, and yellow mark the best, second-best, and third-best result in each column.

Method	MIP-NERF 360 (6 VIEWS)			MIP-NERF 360 (9 VIEWS)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FSGS	14.76	0.326	0.532	15.36	0.373	0.502
InstantSplat	15.91	0.388	0.443	16.32	0.433	0.419
Difix3D+	14.94	0.308	0.419	16.04	0.368	0.371
GenFusion	16.40	0.384	0.487	17.55	0.435	0.409
GuidedVD-3DGS	13.89	0.273	0.640	15.77	0.386	0.418
Ours	16.74	0.393	0.436	17.56	0.448	0.381

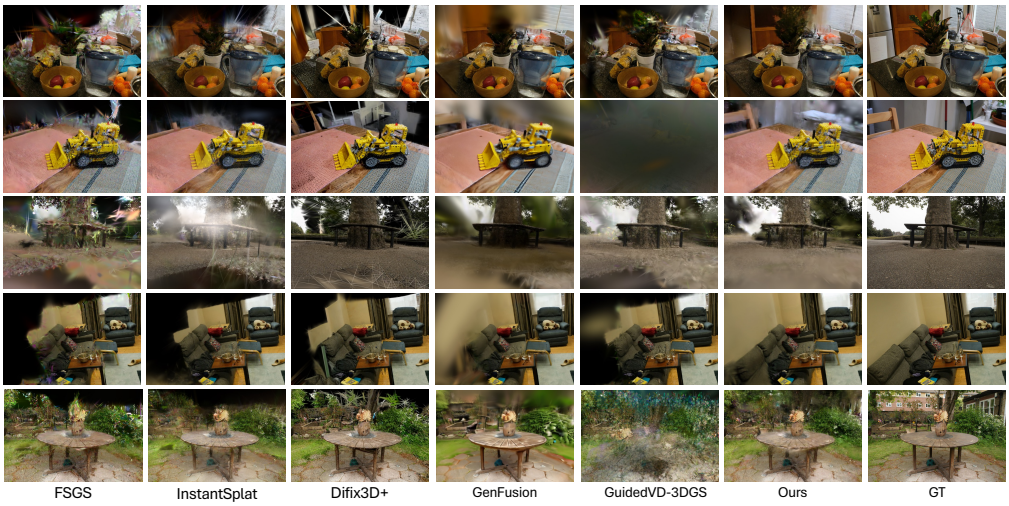

Figure 6: **Qualitative comparison on Mip-NeRF 360 (Barron et al., 2022) under sparse 6-view and 9-view settings.** Large-scale outdoor scenes with wide baselines make sparse-view reconstruction challenging. Our method achieves better coverage, improved geometric consistency, and fewer artifacts than prior methods. White boxes highlight challenging regions.

Qualitative Results. Fig. 5 and Fig. 6 show visual comparisons on representative scenes. Our method produces more complete reconstructions by effectively addressing the key challenges in sparse-view reconstruction: reducing black holes in unobserved regions, minimizing rendering artifacts, and improving geometric consistency. These results demonstrate the effectiveness of GaMO in generating high-quality outpainted views that enhance 3D reconstruction.

4.3 ABLATION STUDIES

We conduct comprehensive ablation studies on Replica and ScanNet++ datasets with 6 input views. To separately assess the outpainted view quality and novel view synthesis performance, we center

Table 3: **Quantitative ablation of blending components.** We evaluate the impact of augmented conditioning (Aug.), hard (H.)/soft (S.) mask blending, and noise resampling (N.). P, S, L denote PSNR↑, SSIM↑, and LPIPS↓, respectively. Row 5 represents our full model.

	Components				Outpainted			Novel View		
#	Aug.	H.	S.	N.	P	S	L	P	S	L
1					18.97	.776	.210	22.37	.821	.197
2	✓				19.11	.779	.207	22.53	.822	.192
3	✓	✓			19.77	.797	.199	23.52	.839	.174
4	✓		✓	✓	19.37	.797	.196	23.22	**.840**	.173
5*	✓	✓		✓	**20.01**	**.800**	**.190**	**23.53**	.839	**.172**

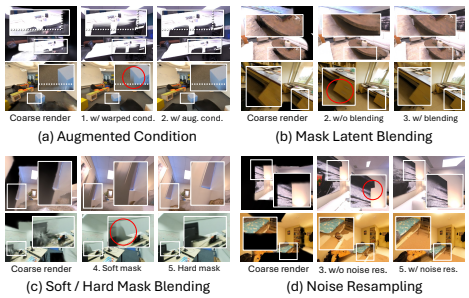

Figure 7: **Qualitative ablation.** (a) Aug. cond. aligns content. (b-c) Mask blending improves geometry. (d) Noise resampling removes seams.

Table 4: **Ablation on mask blending scheduling.** Comparison of blending at different timesteps: t_k (single-step), $t_{1\to t_N}$ (multi-step), All (every step), and IMS (Iterative Mask Scheduling). Row 4* is our full method. *Time (s)* is total generation time.

	Scheduling				Outpainted				Novel View		
#	t_k	$t_{1\to t_N}$	All	IMS	P↑	S↑	L↓	T(s)↓	P↑	S↑	L↓
1	✓				20.09	.804	.198	**85**	23.38	.837	.176
2		✓			19.85	.799	.173	93	23.53	.839	.179
3			✓		**20.31**	**.809**	.201	167	**23.67**	**.842**	.178
4*		✓		✓	20.07	.801	**.169**	93	23.65	.839	**.171**

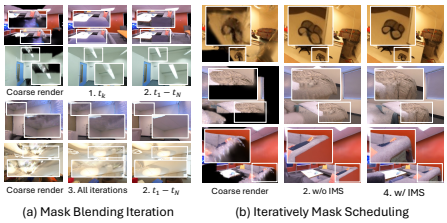

Figure 8: **Qualitative ablation.** (a) Blending timesteps impact. (b) IMS progressively improves detail coherence and alignment.

Table 5: **Ablation on 3DGS refinement.** Impact of point re-init. and perceptual loss on reconstruction quality. Row 3* is full method.

	Components		Novel View		
#	Re-init.	Percep.	P↑	S↑	L↓
1		✓	24.80	.860	.140
2	✓		**25.14**	.857	.139
3*	✓	✓	24.93	**.861**	**.135**



Figure 9: **Qualitative ablation.** (a) Re-init. enables Gaussian generation in new regions. (b) Perceptual loss recovers sharp details.

crop the input images to 0.6× of their original size. For each ablation, we provide quantitative results and visual comparisons, where numbered configurations (e.g., "1.") in figures correspond to table rows, and letters (e.g., (a), (b)) denote different visual comparison aspects.

Latent Blending Strategies. Tab. 3 and Fig. 7 present ablation results on our latent blending design. Augmenting the warped features with downscaled reference RGB and CCM (rows 1-2) prevents incorrect hallucinations in known regions (a). Mask latent blending (rows 2-3) prevents severe geometric misalignment (b, red circle) and improves PSNR by 0.66 dB. Hard masking (rows 4-5) produces sharper boundaries (c) with 0.64 dB gain over soft masking. Finally, noise resampling (rows 3 vs. 5) reduces blending artifacts by 0.24 dB, generating more coherent results (d).

Mask Blending Scheduling. Tab. 4 and Fig. 8 ablate the blending schedule. Single-step blending (row 1) is insufficient, since denoising washes out the coarse geometry, while multi-step blending (row 2) better preserves geometric cues (a-top). Blending at every step (row 3) gives slightly higher PSNR/SSIM but blurs boundaries (a-bottom) and increases denoising time, so we keep the multi-step schedule. Iterative Mask Scheduling (rows 2 vs. 4) substantially improves perceptual quality, as progressive mask dilation gives smoother transitions and more coherent details (b).

3DGS Refinement Components. Tab. 5 and Fig. 9 present ablation results on 3DGS refinement components. Point cloud re-initialization using outpainted views (rows 1 vs. 3) enables the successful generation of Gaussian points in outpainted regions (a). Perceptual loss (rows 2 vs. 3) effectively fills holes and reduces artifacts by providing better gradient guidance for outpainted regions (b), producing cleaner and more realistic renderings.

5 Conclusion

We present *GaMO*, a zero-shot outpainting framework for sparse-view 3D reconstruction. By expanding existing views rather than synthesizing new perspectives, GaMO increases scene coverage while preserving geometric consistency, reducing artifacts and improving reconstruction quality at significantly lower runtime.

Limitations. GaMO cannot recover fully occluded regions, and clustered or misaligned views may degrade performance. Future work could explore adaptive outpainting and hybrid strategies.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with writing, proofreading, and improving the clarity and organization of the manuscript. We have not used generative AI tools to generate experimental results, conduct experiments, or produce scientific claims, and these uses do not apply to this work. All AI-assisted work was reviewed and verified by the authors, and generated code was tested for correctness before use. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We release our code and provide implementation details, experimental configurations, and evaluation protocols to ensure full reproducibility of our results.

REFERENCES

Titas Ancikevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18187–18197. IEEE, 2022.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. VD3D: Taming large video diffusion transformers for 3D camera control. In *International Conference on Learning Representations (ICLR)*, 2025b.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, volume 2026, pp. 82850–82880, 2026.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5470–5479, 2022.

Chenjie Cao, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *arXiv preprint arXiv:2408.08000*, 2024.

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6045–6056, 2025a.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025b.

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4217–4229, 2023.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23185–23194. IEEE, 2023.

Honghua Chen, Chen Change Loy, and Xingang Pan. Mvip-nerf: Multi-view 3d inpainting on nerf scenes via diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5344–5353, 2024a.

Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexible-view exploration. *Advances in Neural Information Processing Systems*, 38:172270–172299, 2026.

Weiliang Chen, Jiayi Bi, Yuanhui Huang, Wenzhao Zheng, and Yueqi Duan. Scenecompleter: Dense 3d scene completion for generative novel view synthesis. *arXiv preprint arXiv:2506.10981*, 2025.

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 *European Conference on Computer Vision*, pp. 370–386. Springer, 2024b.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024c.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024d.

Zixuan Chen, Ruijie Su, Jiahao Zhu, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Vividdreamer: Towards high-fidelity and efficient text-to-3d generation. *arXiv preprint arXiv:2406.14964*, 2024e.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022a.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in neural information processing systems*, 35:25683–25696, 2022b.

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

Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d Gaussian Splatting in few-shot images. *arXiv preprint arXiv:2311.13398*, 2023b.

Yiquan Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *European Conference on Computer Vision*, pp. 432–449. Springer, 2024.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision 2025*, 2025. URL https://openreview.net/forum?id=5uw1GRBFoT.

Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. In *2025 International Conference on 3D Vision (3DV)*, pp. 457–468. IEEE, 2025.

Zhiwen Fan, Wenyan Cong, Kairun Wen, 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. *arXiv preprint arXiv:2403.20309*, 2024.

Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023.

Tobias Fischer, Samuel Rota Bulò, Yung-Hsu Yang, Nikhil Keetha, Lorenzo Porzi, Norman Müller, Katja Schwarz, Jonathon Luiten, Marc Pollefeys, and Peter Kontschieder. Flowr: Flowing from sparse to dense 3d reconstructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27702–27712, 2025.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36:39897–39914, 2023.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

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

Vitor Guizilini, Muhammad Zubair Irshad, Dian Chen, Greg Shakhnarovich, and Rares Ambrus. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 764–776, 2025.

Xiangyu Han, Zhen Jia, Boyi Li, Yan Wang, Boris Ivanovic, Yurong You, Lingjie Liu, Yue Wang, Marco Pavone, Chen Feng, et al. Extrapolated urban view synthesis benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28718–28728. IEEE, 2025.

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

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025a.

Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. In *European Conference on Computer Vision*, pp. 463–479. Springer, 2024b.

Zongqi He, Zhe Xiao, Kin-Chung Chan, Yushen Zuo, Jun Xiao, and Kin-Man Lam. See in detail: Enhancing sparse-view 3d gaussian splatting with local depth and semantic regularization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7875–7886. IEEE, 2023.

Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9698–9707, 2024.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025.

Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9784–9794, 2024.

Mude Hui, Zihao Wei, Hongru Zhu, Fei Xia, and Yuyin Zhou. Microdiffusion: Implicit representation-guided diffusion for 3d reconstruction from limited 2d microscopy projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11460–11469, 2024.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021.

Chenhao Ji, Chaohui Yu, Junyao Gao, Fan Wang, and Cairong Zhao. Campvg: Camera-controlled panoramic video generation with epipolar-aware diffusion. *arXiv preprint arXiv:2509.19979*, 2025.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4918–4929. IEEE, 2025a.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025b.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pp. 694–711. Springer, 2016.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pp. 150–168. Springer, 2024.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10026–10038, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

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.

Hanyang Kong, Xingyi Yang, and Xinchao Wang. Generative sparse-view gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26745–26755, 2025.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9503–9513. IEEE, 2024.

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024.

Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6775–6785, 2024.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pp. 112–130. Springer, 2024.

Kyungmin Lee, Kihyuk Sohn, and Jinwoo Shin. Dreamflow: High-quality text-to-3d generation by approximating probability flow. *arXiv preprint arXiv:2403.14966*, 2024.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36:50648–50660, 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *Advances in Neural Information Processing Systems*, 37:62189–62222, 2024a.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20775–20785, 2024b.

Lingen Li, Zhaoyang Zhang, Yaowei Li, Jiale Xu, Wenbo Hu, Xiaoyu Li, Weihao Cheng, Jinwei Gu, Tianfan Xue, and Ying Shan. Nvcomposer: Boosting generative novel view synthesis with multiple sparse and unposed images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 777–787. IEEE, 2025.

Ming-Feng Li, Yueh-Feng Ku, Hong-Xuan Yen, Chi Liu, Yu-Lun Liu, Albert YC Chen, Cheng-Hao Kuo, and Min Sun. Genrc: Generative 3d room completion from sparse image collections. In *European Conference on Computer Vision*, pp. 146–163. Springer, 2024c.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3D: High-resolution multiview diffusion using efficient row-wise attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024d.

Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. In *International Conference on Learning Representations (ICLR)*, 2024e.

Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European conference on computer vision*, pp. 515–534. Springer, 2022.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 798–810. IEEE, 2025.

Yixun 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6517–6526, 2024.

Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. *arXiv preprint arXiv:2104.03963*, 2021.

Chin-Yang Lin, Cheng Sun, Fu-En Yang, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Longsplat: Robust unposed 3d gaussian splatting for casual long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27412–27422, 2025a.

Chin-Yang Lin, Chung-Ho Wu, Chang-Han Yeh, Shih-Han Yen, Cheng Sun, and Yu-Lun Liu. Frugalnerf: Fast convergence for extreme few-shot novel view synthesis without learned priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11227–11238, 2025b.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 14458–14467, 2021.

Aoming Liu, Zhong Li, Zhang Chen, Nannan Li, Yi Xu, and Bryan A Plummer. Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance. In *European Conference on Computer Vision*, pp. 146–164. Springer, 2024a.

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

Kuan-Hung Liu, Cheng-Kun Yang, Min-Hung Chen, Yu-Lun Liu, and Yen-Yu Lin. Corrfill: Enhancing faithfulness in reference-based inpainting with correspondence guidance in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1618–1627. IEEE, 2025.

Kunhao Liu, Ling Shao, and Shijian Lu. Novel view extrapolation with video diffusion priors. *arXiv preprint arXiv:2411.14208*, 2024c.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10072–10083, 2024d.

Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. MeshFormer: High-quality mesh generation with 3D-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024e.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275. IEEE, 2023a.

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 *European Conference on Computer Vision*, pp. 37–53. Springer, 2024f.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024g.

Xinhang Liu, Jiaben Chen, Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-NeRF/3DGS: Diffusion-generated pseudo-observations for high-quality sparse-view reconstruction. In *European Conference on Computer Vision (ECCV)*, 2024h.

Xinhui Liu, Can Wang, Wei Jiang, Wei Wang, and Dong Xu. Gaussvid: Sparse-view gaussian splatting with 3d-aware video diffusion priors. *arXiv preprint arXiv:2608.21849*, 2026a.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b.

Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Ping Luo, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5726–5738, 2026b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8744–8753, 2024a.

Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14211–14219, 2024b.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Campagnolo Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin Solomon. Score distillation via reparametrized DDIM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2016–2029. IEEE, 2025.

David McAllister, Songwei Ge, Jia-Bin Huang, David Jacobs, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. *Advances in Neural Information Processing Systems*, 37:33779–33804, 2024.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. Splatfields: Neural gaussian splats for sparse 3d and 4d reconstruction. In *European Conference on Computer Vision*, pp. 313–332. Springer, 2024.

Zhiyuan Min, Yawei Luo, Jianwen Sun, and Yi Yang. Epipolar-free 3d gaussian splatting for generalizable novel view synthesis. *Advances in Neural Information Processing Systems*, 37:39573–39596, 2024.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20669–20679. IEEE, 2023.

Ashkan Mirzaei, Riccardo De Lutio, Seung Wook Kim, David Acuna, Jonathan Kelly, Sanja Fidler, Igor Gilitschenski, and Zan Gojcic. Reffusion: Reference adapted diffusion models for 3d scene inpainting. *arXiv preprint arXiv:2404.10765*, 2024.

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulo, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10258–10268, 2024.

Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6022–6033, 2025.

Junfeng Ni, Yixin Chen, Zhifei Yang, Yu Liu, Ruijie Lu, Song-Chun Zhu, and Siyuan Huang. G4splat: Geometry-guided gaussian splatting with generative prior. In *International Conference on Learning Representations*, volume 2026, pp. 138924–138951, 2026.

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.

Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Avinash Paliwal, Xilong Zhou, Wei Ye, Jinhui Xiong, Rakesh Ranjan, and Nima Khademi Kalantari. Ri3d: Few-shot gaussian splatting with repair and inpainting diffusion priors. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25094–25103. IEEE, 2025.

Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the computer vision and pattern recognition conference*, pp. 21600–21609, 2025.

Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28285–28295, 2024.

Soumava Paul, Christopher Wewer, Bernt Schiele, and Jan Eric Lenssen. Sp2360: Sparse-view 360○ scene reconstruction using cascaded 2d diffusion priors. In *ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2024.

Rui Peng, Wangze Xu, Luyang Tang, Liwei Liao, Jianbo Jiao, and Ronggang Wang. SCGaussian: Structure consistent Gaussian Splatting with matching prior for few-shot novel view synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9914–9925, 2024.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Ahmad Salimi, Tristan Aumentado-Armstrong, Marcus A Brubaker, and Konstantinos G Derpanis. Geometry-aware diffusion models for multiview scene inpainting. *arXiv preprint arXiv:2502.13335*, 2025.

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

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 37:80220–80243, 2024.

Hao Shi, Yu Li, Kailun Yang, Jiaming Zhang, Kunyu Peng, Alina Roitberg, Yaozu Ye, Huajian Ni, Kaiwei Wang, and Rainer Stiefelhagen. Fishdreamer: Towards fisheye semantic completion via unified image outpainting and segmentation. *arXiv preprint arXiv:2303.13842*, 2023a.

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

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*, 2023c.

Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, and Janne Kontkanen. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20385–20395. IEEE, 2024.

Meng-Li Shih, Ying-Huan Chen, Yu-Lun Liu, and Brian Curless. Prior-enhanced gaussian splatting for dynamic scene reconstruction from casual video. *arXiv preprint arXiv:2512.11356*, 2025.

Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in neural information processing systems*, 34:19313–19325, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. Replica: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Chih-Hai Su, Chih-Yao Hu, Shr-Ruei Tsai, Jie-Ying Lee, Chin-Yang Lin, and Yu-Lun Liu. Boost-mvsnerfs: Boosting mvs-based nerfs to generalizable view synthesis in large-scale scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

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 *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3172–3182. IEEE, 2022.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024.

Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24846–24857. IEEE, 2025.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024a.

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

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *Advances in Neural Information Processing Systems*, 2023.

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. In *European Conference on Computer Vision*, pp. 175–191. Springer, 2024c.

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. *arXiv preprint arXiv:1908.07007*, 2019.

Shr-Ruei Tsai, Wei-Cheng Chang, Jie-Ying Lee, Chih-Hai Su, and Yu-Lun Liu. Lightsout: Diffusion-based outpainting for enhanced lens flare removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6353–6363, 2025.

Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6996–7007, 2023.

Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2421–2431. IEEE, 2025.

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

Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. In *European Conference on Computer Vision*, pp. 153–168. Springer, 2024a.

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

Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multiview consistent images for single-view scene reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26772–26782. IEEE, 2025a.

Hanyang Wang, Fangfu Liu, Jiawei Chi, and Yueqi Duan. Videoscene: Distilling video diffusion model to generate 3d scenes in one step. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16475–16485. IEEE, 2025b.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025c.

Ning-Hsu Wang and Yu-Lun Liu. Depth Anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, and Vikas Chandra. Taming mode collapse in score distillation for text-to-3D generation. *arXiv preprint arXiv:2401.00909*, 2024b.

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

Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6913–6923, 2024c.

Qisen Wang, Yifan Zhao, Jiawei Ma, and Jia Li. How to use diffusion priors under sparse views? *Advances in Neural Information Processing Systems*, 37:30394–30424, 2024d.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024e.

Yanbo Wang, Ziyi Wang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Oggsplat: Open gaussian growing for generalizable reconstruction with expanded field-of-view. *arXiv preprint arXiv:2506.05204*, 2025d.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1399–1408. IEEE, 2019.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *Advances in Neural Information Processing Systems*, 37:107326–107349, 2024f.

Zhen Wang, Qiangeng Xu, Feitong Tan, Menglei Chai, Shichen Liu, Rohit Pandey, Sean Fanello, Achuta Kadambi, and Yinda Zhang. Mvdd: Multi-view depth diffusion models. In *European Conference on Computer Vision*, pp. 236–253. Springer, 2024g.

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. *Advances in neural information processing systems*, 36:8406–8441, 2023b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

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

Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20731–20741, 2024.

Ethan Weber, Norman Müller, Yash Kant, Vasu Agrawal, Michael Zollhöfer, Angjoo Kanazawa, and Christian Richardt. Fillerbuster: Unified generative scene completion model for casual captures. In *2026 International Conference on 3D Vision (3DV)*, pp. 532–542. IEEE, 2026.

Chung-Ho Wu, Yang-Jung Chen, Ying-Huan Chen, Jie-Ying Lee, Bo-Hsu Ke, Chun-Wei Tuan Mu, Yi-Chuan Huang, Chin-Yang Lin, Min-Hung Chen, Yen-Yu Lin, et al. Aurafusion360: Augmented unseen region alignment for reference-based 360deg unbounded scene inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16366–16376, 2025a.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26024–26035, 2025b.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024a.

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

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024c.

Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6078–6088, 2025c.

Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. *arXiv preprint arXiv:2307.03177*, 2023.

Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9892–9902, 2024d.

Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E Kaufman. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6369–6379, 2024.

Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pp. 143–163. Springer, 2024a.

Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3difftection: 3d object detection with geometry-aware diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10617–10627, 2024b.

Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024c.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024d.

Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, and Zhuowen Tu. Bayesian diffusion models for 3d shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10628–10638, 2024e.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025a.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024f.

Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25442–25452, 2025b.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In *International Conference on Learning Representations*, 2024g.

Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human-3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. *Advances in Neural Information Processing Systems*, 37:99601–99645, 2024.

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. *arXiv preprint arXiv:2402.10259*, 2024a.

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

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2024b.

Xiaofeng Yang, Yiwen Chen, Cheng Chen, Chi Zhang, Yi Xu, Xulei Yang, Fayao Liu, and Guosheng Lin. Learn to optimize denoising scores: A unified and improved diffusion prior for 3d generation. In *European Conference on Computer Vision*, pp. 136–152. Springer, 2024c.

Xiuyu Yang, Yunze Man, Jun-Kun Chen, and Yu-Xiong Wang. SceneCraft: Layout-guided 3D scene generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024d.

Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10560–10569. IEEE, 2019.

Kai Yao, Penglei Gao, Xi Yang, Jie Sun, Rui Zhang, and Kaizhu Huang. Outpainting by queries. In *European conference on computer vision*, pp. 153–169. Springer, 2022.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. *Advances in Neural Information Processing Systems*, 37:1304–1332, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6796–6807, 2024.

Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. Gsfixer: Improving 3d gaussian splatting with reference-guided video diffusion priors. *arXiv preprint arXiv:2508.09667*, 2025.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

Tanveer Younis and Zhanglin Cheng. Sparse-view 3d reconstruction: Recent advances and open challenges. *arXiv preprint arXiv:2507.16406*, 2025.

Hang Yu, Ruilin Li, Shaorong Xie, and Jiayan Qiu. Shadow-enlightened image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7850–7860, 2024a.

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6658–6667. IEEE, 2024c.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5916–5926, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025b.

Rui Yu, Jiachen Liu, Zihan Zhou, and Sharon X Huang. Nerf-enhanced outpainting for faithful field-of-view extrapolation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16826–16833. IEEE, 2024d.

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*, 2024e.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023.

Zhongrui Yu, Martina Megaro-Boldini, Robert W Sumner, and Abdelaziz Djelouah. Unboxed: Geometrically and temporally consistent video outpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7309–7319, 2025c.

Xiaoding Yuan, Shitao Tang, Kejie Li, and Peng Wang. Camfreediff: Camera-free image to panorama generation with diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16408–16417, 2025.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6347–6357, 2024a.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37:15272–15295, 2024b.

Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024c.

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 *European Conference on Computer Vision*, pp. 335–352. Springer, 2024d.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Shaofeng Zhang, Jinfa Huang, Qiang Zhou, Zhibin Wang, Fan Wang, Jiebo Luo, and Junchi Yan. Continuous-multiple image outpainting in one-step via positional query and a diffusion-based approach. *arXiv preprint arXiv:2401.15652*, 2024e.

Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7162–7172, 2024f.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim H Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. In *International Conference on Learning Representations*, volume 2025, pp. 95479–95499, 2025.

Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9720–9731. IEEE, 2024.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Yingji Zhong, Zhihao Li, Dave Zhenyu Chen, Lanqing Hong, and Dan Xu. Taming video diffusion prior with scene-grounding guidance for 3d gaussian splatting from sparse inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6133–6143, 2025a.

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*, 2025b.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

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

Yunlai Zhou, Yiren Lu, Tuo Liang, Disheng Liu, Vipin Chaudhary, and Yu Yin. Filling the unseen: Scene extrapolation via 3d gaussian splatting. *arXiv preprint arXiv:2609.13262*, 2026.

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, pp. 145–163. Springer, 2024.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.

APPENDIX

A GENERATION-BASED COMPARISON

Table 6: **Comparison with the diffusion-based novel view generation method SEVA.** We report PSNR↑, SSIM↑, and LPIPS↓.

	Replica (3v)			ScanNet++ (3v)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
SEVA	18.52	0.677	0.193	15.79	0.631	0.343
Ours	**24.40**	**0.865**	**0.117**	**20.06**	**0.759**	**0.265**
	Replica (6v)			**ScanNet++ (6v)**		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
SEVA	18.64	0.691	0.206	16.93	0.640	0.276
Ours	**26.40**	**0.882**	**0.104**	**23.41**	**0.835**	**0.181**



Figure 10: **Qualitative comparison with SEVA.** While SEVA produces visually plausible results, it often lacks pixel-level accuracy and geometric consistency.

To further compare with approaches designed for direct novel view generation, we evaluate our method against the diffusion-based method SEVA Zhou et al. (2025). As shown in Table 6, our approach consistently outperformed SEVA across both datasets and view settings. In particular, our method achieved significantly higher PSNR and SSIM while reducing LPIPS, indicating more accurate reconstruction and better perceptual quality.

Qualitative results in Fig. 10 further illustrated these differences. While SEVA produced visually plausible results, it often exhibited inaccurate pixel-level details and geometrically inconsistent structures. In contrast, our method preserved accurate details and more coherent scene geometry.

A.1 INTERPOLATION VS. OUTPAINTING

To isolate the effect of the generation strategy, we further compare novel-view interpolation and FOV outpainting using the same SEVA backbone. As shown in Fig. 11, increasing the number of interpolated views progressively degrades reconstruction quality. In contrast, outpainting expands scene coverage while preserving the original viewpoints, reducing potential inconsistencies introduced by generated novel views.

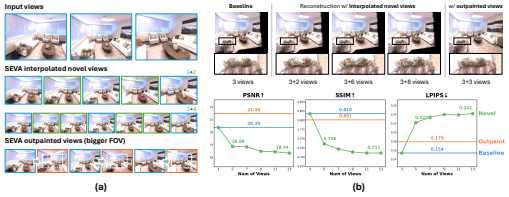

Figure 11: **SEVA interpolation vs. outpainting.** Increasing interpolated views degrades reconstruction quality.

B EFFECT OF NOVEL-VIEW SUPERVISION FREQUENCY

We investigate whether increasing diffusion-generated novel-view supervision improves reconstruction. Using GenFusion Wu et al. (2025c), we increase its supervision frequency by reducing the diffusion generation interval. As shown in Fig. 12, higher frequencies consistently degrade PSNR, SSIM, and LPIPS, introducing blurrier reconstructions. These results highlight the potential inconsistencies of excessive diffusion-generated supervision.

Freq.	PSNR↑	SSIM↑	LPIPS↓
Official (1×)	**23.98**	**0.855**	**0.142**
2×	23.71	0.849	0.154
4×	23.57	0.846	0.162

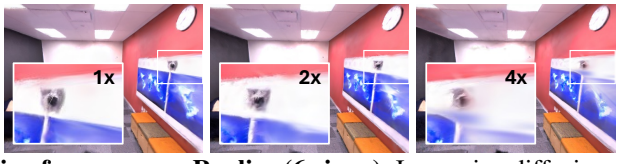

Figure 12: **Effect of GenFusion supervision frequency on Replica (6 views).** Increasing diffusion-generated supervision from 1× to 4× degrades reconstruction quality and introduces blurring.

C Outpainting as Complementary Supervision

We investigate whether GaMO's outpainting complements diffusion-generated novel-view supervision by incorporating our outpainted views into GuidedVD-3DGS Zhong et al. (2025a). As shown in Table 7, combining both supervision strategies improves PSNR, SSIM, and LPIPS under both 3-view and 6-view settings on Replica.

Table 7: **Complementary supervision on Replica.** Results under 3-view and 6-view settings.

Views	Method	P↑	S↑	L↓
3	GuidedVD	25.26	0.864	0.138
	+ GaMO	**25.88**	**0.877**	**0.117**
6	GuidedVD	26.68	0.880	0.133
	+ GaMO	**27.08**	**0.887**	**0.112**

D Implementation Details

For coarse initialization, we train 3DGS for 10,000 iterations with $\lambda_s = 0.2$ and opacity threshold $\eta_{\text{mask}} = 0.6$.

For outpainting, we use the multi-view diffusion model Cao et al. (2025a) with focal-length scaling $S_k \in [0.5, 0.7]$, adjusted per scene depending on scene scale. We use DDIM sampling Song et al. (2021) with $T = 50$ steps, and perform latent blending at timesteps $t_1 = 0.7T$, $t_2 = 0.5T$, $t_3 = 0.3T$ with noise resampling $R = 3$. Input and outpainted views share the same resolution (differing only in FOV), with dimensions set as multiples of 64. Before refinement, we alpha-blend downscaled inputs at the center.

For refinement, we optimize 3DGS for 3,000 iterations (3 views) or 7,000 iterations (6/9 views) with $\lambda_{\text{perc}} = 0.1$, alternating supervision between input and outpainted views. In addition, we make minor adjustments to several 3DGS refinement hyperparameters based on scene characteristics to ensure stable optimization. We will release all code, configurations, and scripts used in our experiments. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

E Outpainting Comparison Using Multi-View Diffusion Models



Figure 13: **Comparison of outpainting using adapted multi-view diffusion models.** Top: input views. Middle: outpainted views generated by adapted SEVA Zhou et al. (2025), MVGenMaster Cao et al. (2025a), and our GaMO. Bottom: novel views after 3DGS refinement using the generated outpainted views. Adapted multi-view diffusion models suffer from multi-view inconsistency, resulting in noisy reconstructions, while our method produces consistent outpainted views that improve reconstruction quality.

We compare our method against adapted multi-view diffusion models for outpainting. Specifically, we adapt SEVA Zhou et al. (2025) and MVGenMaster Cao et al. (2025a) by modifying the camera intrinsics to generate outpainted versions of the input views with extended FOV. These outpainted input views are then used to train 3DGS for improved novel view synthesis.

As shown in Fig. 13, SEVA produces highly noisy novel views after 3DGS refinement due to severe multi-view inconsistency caused by lack of geometric constraints. While MVGenMaster incorporates additional geometric mechanisms (e.g., multi-view conditioning), it still suffers from inconsistency issues that introduce artifacts in the refined reconstruction. In contrast, our GaMO effectively addresses the multi-view inconsistency problem, providing consistent outpainted views across

multiple viewpoints that successfully refine 3DGS quality without introducing additional noise or artifacts.

F Iterative Mask Scheduling Implementation

To maximize the utilization of coarse geometry during outpainting while preserving the generative diversity of the diffusion model, we introduce an **Iterative Mask Scheduling (IMS)** strategy. IMS dynamically adjusts the mask region throughout the denoising process, allowing the diffusion model to first freely hallucinate missing regions and later progressively align the generated content with the coarse 3D initialization.

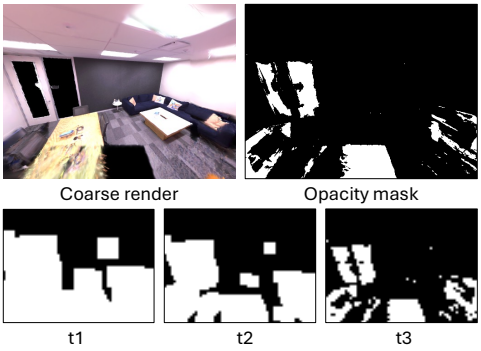

Figure 14: **Iterative Mask Scheduling visualization.** Top: coarse render and opacity mask derived from the coarse 3D initialization. Bottom: progressive mask shrinking during denoising at timesteps $t = 35, 25, 15$, with 2, 1, and 0 dilation iterations, respectively.

Design Rationale. As demonstrated in the ablation studies in the main paper (Tab. 4), we found that applying mask latent blending at specific denoising steps yields significantly better results than continuous blending throughout the entire denoising process. Based on these findings, we strategically select three representative timesteps corresponding to the early, middle, and late stages of denoising to participate in the latent blending process. At each stage, we employ progressively shrinking mask sizes to control the degree of interference with the denoising process: larger masks in early stages allow more freedom for generation, while smaller masks in later stages enforce stronger alignment with coarse geometry.

Implementation Details. As illustrated in Fig. 14, we generate three mask levels through morphological dilation:

$$\mathcal{M}_{\text{latent}}^{(k)} = \text{Dilate}(\mathcal{M}_{\text{base}}^{\downarrow}, \text{kernel} = 5, \text{iterations} = \frac{k-15}{10}), \tag{7}$$

where $\mathcal{M}_{\text{base}}^{\downarrow}$ denotes the downsampled base mask from the coarse geometry opacity map, aligned to the latent space resolution of 64×48 via adaptive max pooling. The $\text{Dilate}(\cdot)$ operation applies iterative max pooling with 5×5 kernel to expand the masked region. During the denoising process from $t = 50$ to $t = 0$, we apply $\mathcal{M}_{\text{latent}}^{(35)}$ at $t = 35$, $\mathcal{M}_{\text{latent}}^{(25)}$ at $t = 25$, and $\mathcal{M}_{\text{latent}}^{(15)}$ at $t = 15$, as visualized in Fig. 14. This staged approach balances generative freedom with geometric consistency, as validated by our ablation experiments.

G FOV Expansion and Multi-FOV Ablation

We evaluate different FOV expansion ratios and multi-FOV supervision on a representative Replica scene under the 6-view setting. As shown in Fig. 15, excessive expansion (0.2× focal-length scaling) substantially degrades reconstruction quality, while 0.6× provides the best overall single-FOV performance. Combining multiple FOVs (0.4/0.6/0.8×) slightly improves PSNR and LPIPS, at the cost of lower SSIM.

Scale	P↑	S↑	L↓
0.2×	15.51	0.643	0.539
0.4×	25.16	0.849	0.140
0.6×	26.00	**0.878**	0.128
0.8×	25.64	**0.878**	0.133
Multi-FOV	**26.01**	0.871	**0.125**

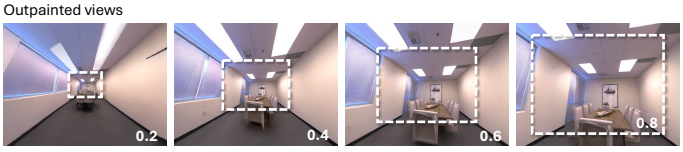

Figure 15: **FOV expansion and multi-FOV ablation.** Quantitative and qualitative results under different focal-length scales. Multi-FOV combines 0.4, 0.6, and 0.8×.

H More Quantitative Comparison

We provide additional quantitative results on Replica Straub et al. (2019) and ScanNet++ Yeshwanth et al. (2023) datasets with varying numbers of input views (3, 6, and 9 views), as shown in Tab. 8 and Tab. 9. We focus our comparison GuidedVD-3DGS Zhong et al. (2025a), a competitive state-of-the-art diffusion-based method.

Evaluation Protocol. For Replica, we follow the evaluation protocol from Zhong et al. (2025a) for all three view settings. For ScanNet++, the 6-view setting follows Zhong et al. (2025a), while the 3-view and 9-view settings use manually selected views to maximize spatial coverage. All methods use DUSt3R Wang et al. (2024e) for point cloud initialization.

Results. Our method consistently outperforms baselines across most metrics and view settings. On Replica, we achieve the best SSIM and LPIPS scores across all view counts. On ScanNet++, we obtain superior performance across all metrics in all view settings. Notably, our method maintains competitive quality with GuidedVD-3DGS Zhong et al. (2025a) while being significantly faster (approximately 6-9 minutes vs. 3+ hours).

Table 8: **Quantitative comparison on Replica Straub et al. (2019) with 3, 6, and 9 input views.**

Method	Replica (3 views)			Replica (6 views)			Replica (9 views)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
GuidedVD-3DGS Zhong et al. (2025a)	25.26	0.864	0.138	26.68	0.880	0.133	28.08	0.901	0.108
Ours	24.40	0.865	0.117	26.40	0.882	0.104	27.58	0.902	0.096

Table 9: **Quantitative comparison on ScanNet++ Yeshwanth et al. (2023) with 3, 6, and 9 input views.**

Method	ScanNet++ (3 views)			ScanNet++ (6 views)			ScanNet++ (9 views)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
GuidedVD-3DGS Zhong et al. (2025a)	18.82	0.720	0.328	22.98	0.815	0.204	24.65	0.843	0.159
Ours	20.06	0.759	0.265	23.41	0.835	0.181	25.17	0.860	0.152

I More Qualitative Comparison

We provide additional qualitative results across 3-, 6-, and 9-view settings on Replica Straub et al. (2019) and ScanNet++ Yeshwanth et al. (2023) datasets, as shown in Fig. 16. We compare against FSGS Zhu et al. (2024), InstantSplat Fan et al. (2024), DiFix3D Wu et al. (2025b), GenFusion Wu et al. (2025c), and GuidedVD-3DGS Zhong et al. (2025a) using the same baseline configurations as described in the main paper.

As illustrated in Fig. 16, even with extremely sparse inputs (3 views), our method produces reasonable content and geometry while maintaining consistency. Compared to baselines, our approach demonstrates better scene coverage with fewer missing regions (black holes), improved geometric consistency with reduced ghosting artifacts, and overall higher visual quality. These improvements are particularly evident in challenging regions highlighted by white boxes.

J Runtime Analysis

We report the end-to-end runtime of our pipeline on a representative indoor scene (Replica_6, *office_2*) with 6 input views at 512×384 resolution, evaluated on a single NVIDIA RTX 4090 GPU. The pipeline consists of three main stages: coarse 3DGS reconstruction and rendering, multi-view diffusion outpainting, and the final 3DGS refinement stage that incorporates DUSt3R point cloud initialization and refined 3DGS training.

Since GuidedVD-3DGS requires significantly larger GPU memory, it is evaluated on an NVIDIA RTX A6000 GPU. For

Table 10: **Runtime breakdown of our method on RTX 4090.**

Stage	Time (s)	Time (min)
Coarse 3DGS init. & render	118	1.97
Multi-view outpainting	93	1.55
3DGS refine (train + render)	280	4.67
Total	491	8.18

Table 11: **End-to-end runtime comparison RTX A6000.**

Method	Total Time
GuidedVD-3DGS Zhong et al. (2025a)	∼3h 20min
Ours	**9min 54s**

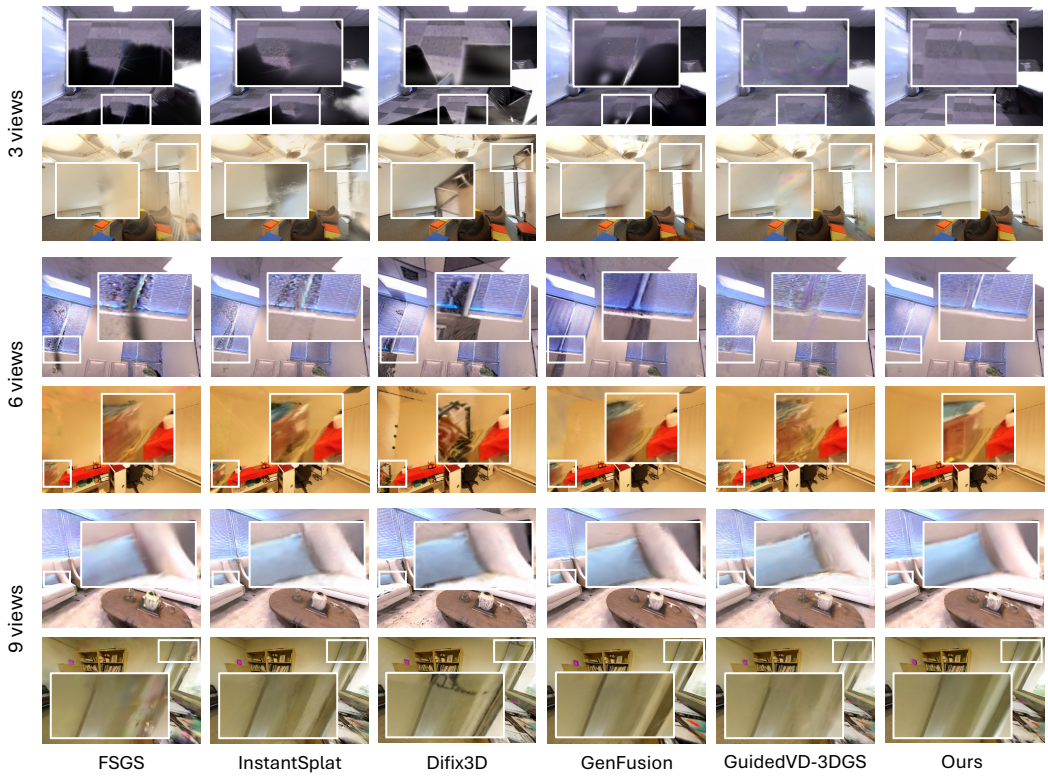

Figure 16: **Qualitative comparison on Replica Straub et al. (2019) and ScanNet++ Yeshwanth et al. (2023) with 3, 6, and 9 sparse views.** Our method produces better coverage, geometric consistency, and fewer artifacts compared to baselines. White boxes highlight challenging regions. Best viewed zoomed in.

a fair comparison, we additionally report the runtime of our method on the same hardware. On the Replica *office_2* scene, our pipeline takes **9 min 54 s**, while GuidedVD-3DGS requires approximately **3 hours 20 minutes**. This demonstrates that our method achieves substantially faster reconstruction while maintaining competitive reconstruction quality.

K Failure Cases

While our method demonstrates strong performance across a wide range of scenarios, it remains limited in scenes containing severe occlusions. This limitation is inherent to all methods that generate novel views for 3D reconstruction; current multi-view diffusion models face the same challenge, as even densely sampled novel viewpoints struggle to reconstruct regions that are severely occluded by obstacles. As shown in Fig. 17, such heavily occluded areas remain challenging for both geometry-aware outpainting and multi-view diffusion methods.

Potential Solutions. A promising direction to address viewpoint-specific occlusions is to generate outpainted views from alternative camera perspectives with geometry-aware mechanisms, such as bird's-eye or top-down views with a larger FOV. By generating content from drastically different viewing angles, these views could potentially observe regions that are occluded from the original camera poses, thereby providing complementary supervision for the occluded areas.

L Baseline Implementation Details

In this section, we describe the implementation details of all baseline methods used in our experiments. Unless otherwise stated, we follow the official implementations and recommended settings



Figure 17: **Failure cases in heavily occluded regions.** Due to severe occlusions in the scene, certain regions are never observed across all input views. Both outpainting (ours) and novel view generation methods struggle to reconstruct these completely unobserved areas. Red boxes highlight the occluded regions where reconstruction fails.

provided by the authors. For most methods, the input camera poses are provided and the reconstruction is initialized using point clouds estimated by DUSt3R Wang et al. (2024e). InstantSplat follows its original pipeline and uses MASt3R Duisterhof et al. (2025) for initialization. Method-specific differences (e.g., initialization strategies) are described in the corresponding subsections.

FSGS Zhu et al. (2024). For FSGS, we follow the official implementation and training configuration provided by the authors Zhu et al. (2024). In particular, we adopt the version that incorporates the depth supervision term, as suggested in the original implementation. All other training parameters remain consistent with the default configuration.

InstantSplat Fan et al. (2024). InstantSplat is implemented using the official codebase Fan et al. (2024) with its default settings. The Gaussian initialization is obtained from point clouds reconstructed using MASt3R Duisterhof et al. (2025), following the standard pipeline described in the original work.

Difix3D+ Wu et al. (2025b). The Difix3D framework contains multiple variants. In the original implementation, *DiFix3D* and *DiFix3D+* correspond to different variants within the same framework Wu et al. (2025b), differing in whether a time neural enhancer is used during inference. In our experiments, we adopt the version without the time neural enhancer to maintain consistency with the base reconstruction pipeline.

GenFusion Wu et al. (2025c). GenFusion is evaluated using the official implementation and recommended hyperparameters Wu et al. (2025c). We follow the standard training and inference pipeline without modifying any of the default settings.

GuidedVD-3DGS Zhong et al. (2025a). GuidedVD-3DGS is implemented using the official codebase Zhong et al. (2025a). However, the results reported in the original paper use a different rendering resolution compared to our evaluation protocol. This difference leads to minor numerical discrepancies between the reported numbers in the paper and those reproduced in our environment. For transparency and fair comparison, we report both the original results from the paper and the results reproduced under our experimental settings.

M Related Work

Diffusion priors for 3D reconstruction. Diffusion models provide learned priors through Score Distillation Sampling Poole et al. (2022); Wang et al. (2023b); Liang et al. (2024). Improvements address over-smoothing Lukoianov et al. (2024), mode collapse Wang et al. (2024b), provide unified frameworks McAllister et al. (2024), classifier-based distillation Yu et al. (2023), ODE-based sampling Wu et al. (2024d); Lee et al. (2024), and optimize both 3D models and priors Yang et al. (2024c); Chen et al. (2024e). Video diffusion serves as powerful priors Liu et al. (2024b); Melas-Kyriazi et al. (2024); Yi et al. (2024). Reconstruction methods use multi-view conditioning Wu et al. (2024b), pseudo-observation enhancement Liu et al. (2024h), scene-grounding guidance Zhong et al.

(2025a), iterative refinement Wu et al. (2025c;b); Liu et al. (2024g), inline prior-guided score matching under sparse views Wang et al. (2024d), visibility-guided decompositional reconstruction Ni et al. (2025), and various coupling strategies Xu et al. (2024e); Hui et al. (2024); Müller et al. (2024); Liu et al. (2024e); Xue et al. (2024). Native 3D diffusion includes latent approaches Lan et al. (2024); He et al. (2024b); Wu et al. (2024c;a); Yang et al. (2024d) and RL finetuning Xie et al. (2024). While these methods generate additional views or provide guidance, they face multi-view *inconsistency*. Our insight: diffusion models suit *outpainting known views* better than hallucinating novel perspectives, maintaining stronger geometric grounding.

Geometry-aware generation. Geometric consistency leverages Plücker coordinates for camera conditioning Xu et al. (2023); Kant et al. (2024); Zhang et al. (2024c); Xu et al. (2024d); Ji et al. (2025); He et al. (2024a); Bahmani et al. (2025b) and epipolar constraints or voxel representations Tu et al. (2023) for multi-view consistency Huang et al. (2024); Ye et al. (2024); Wang et al. (2024g); Kupyn et al. (2025); Müller et al. (2024); Xu et al. (2024b); Zheng et al. (2024). Camera motion control in video diffusion extends these representations to temporal settings Wang et al. (2024h). Joint pixel-level image and depth synthesis further improves geometric fidelity Guizilini et al. (2025). Depth and normal conditioning proves critical Lu et al. (2024a); Long et al. (2024); Ke et al. (2024); Fu et al. (2024); Patni et al. (2024); Hu et al. (2024); Duan et al. (2024); Wang & Liu (2024); Qiu et al. (2024). Recent panoramic generation Zhang et al. (2024e); Wang et al. (2024c) and video outpainting Yu et al. (2025c) typically operate in 2D or single-view scenarios. Our approach combines *multi-view outpainting* with *geometry awareness* through coarse 3DGS rendering, opacity-based masking, and noise resampling for consistent, geometrically plausible FOV expansion.

Outpainting and FOV expansion. Diffusion-based outpainting includes panoramic methods Zhang et al. (2024e); Wu et al. (2023); Kalischek et al. (2025); Yuan et al. (2025); Feng et al. (2023); Shi et al. (2023a); Yu et al. (2024a); Zhang et al. (2024a); Lu et al. (2024b); Liu et al. (2024a) and restoration tasks Tsai et al. (2025); Liu et al. (2025). Video outpainting methods leverage input-specific adaptation Wang et al. (2024a) and temporal diffusion Zhang et al. (2024f). For 3D scenarios, methods employ visibility-aware inpainting Liu et al. (2024c); Wu et al. (2025a), generative scene completion via grid priors Weber et al. (2024) or interactive extrapolation Yu et al. (2025a), reference-adapted diffusion for 3D inpainting Mirzaei et al. (2024), video diffusion priors Liu et al. (2024c), NeRF-guided training Yu et al. (2024d), iterative 3DGS updates Yu et al. (2025c), and multi-view SDS Chen et al. (2024a). General sparse-view baselines include feed-forward prediction Charatan et al. (2024); Chen et al. (2024b), regularized optimization Paliwal et al. (2024); Xu et al. (2024a), and NeRF-based methods Wang et al. (2023a); Yang et al. (2023); Niemeyer et al. (2022); Lin et al. (2025b); Su et al. (2024). These works require per-scene fine-tuning or focus on single-view outpainting. Our method performs *zero-shot multi-view outpainting* using pre-trained MVGenMaster Cao et al. (2025a) with geometry-aware mechanisms ensuring cross-view consistency without scene-specific training.

N Per-Scene Quantitative Results

We provide detailed per-scene quantitative results for every method across all datasets and view settings. Tab. 12 reports results on ScanNet++ and Replica with 3 input views, Tab. 13 with 6 input views, and Tab. 14 reports results on Mip-NeRF360 with 6 and 9 input views. Each entry reports PSNR, SSIM, and LPIPS from top to bottom. Our method achieves consistently strong performance across scenes and datasets, ranking first or second in the majority of per-scene metrics while maintaining competitive results throughout, all while being **20× faster** than the strongest competing method.

Table 12: **Per-scene quantitative results on ScanNet++ and Replica (3 views).** Each entry reports PSNR↑, SSIM↑, and LPIPS↓ from top to bottom.

	ScanNet++ (3 Views)					Replica (3 Views)						
Method	8a20	94ee	7831	a29c	avg	office_2	office_3	office_4	room_0	room_1	room_2	avg
FSGS	18.09	15.97	17.44	14.82	16.58	21.07	22.31	20.24	18.40	21.81	21.20	20.84
	0.689	0.718	0.666	0.688	0.690	0.872	0.864	0.846	0.719	0.801	0.789	0.815
	0.299	0.355	0.412	0.369	0.359	0.177	0.117	0.140	0.213	0.167	0.220	0.172
InstantSplat	17.27	17.42	17.79	14.38	16.72	19.78	22.97	22.43	18.05	22.62	18.05	20.65
	0.733	0.754	0.694	0.698	0.720	0.862	0.879	0.872	0.729	0.822	0.726	0.815
	0.232	0.311	0.399	0.317	0.315	0.124	0.096	0.103	0.177	0.130	0.219	0.142
Difix3D+	16.05	16.05	16.38	14.09	15.64	19.58	20.77	20.65	17.89	18.79	19.14	19.47
	0.656	0.692	0.650	0.636	0.659	0.834	0.838	0.836	0.712	0.737	0.742	0.783
	0.297	0.364	0.355	0.366	0.346	0.203	0.145	0.138	0.232	0.201	0.245	0.194
GenFusion	20.58	16.63	17.46	17.22	17.97	22.43	23.53	23.46	19.40	24.18	21.02	22.34
	0.747	0.725	0.693	0.734	0.725	0.882	0.871	0.861	0.745	0.835	0.803	0.833
	0.356	0.344	0.376	0.340	0.354	0.165	0.138	0.156	0.215	0.142	0.216	0.172
GuidedVD-3DGS	17.32	**19.04**	19.23	19.70	18.82	25.07	**26.80**	**25.63**	**24.14**	**26.84**	**23.08**	**25.26**
	0.642	0.756	0.716	0.764	0.720	0.905	0.910	**0.898**	**0.791**	**0.854**	0.825	0.864
	0.342	0.316	0.359	0.306	0.328	0.114	0.097	0.109	0.171	0.126	0.208	0.138
Ours	**20.77**	18.81	**20.35**	**20.31**	**20.06**	**25.13**	25.68	24.57	23.71	25.17	22.14	24.40
	0.753	**0.770**	**0.748**	**0.766**	**0.759**	**0.913**	**0.911**	0.897	0.789	0.851	**0.827**	**0.865**
	0.225	**0.273**	**0.299**	**0.263**	**0.265**	**0.088**	**0.075**	**0.095**	**0.150**	**0.114**	**0.179**	**0.117**

Table 13: **Per-scene quantitative results on ScanNet++ and Replica (6views).** Each entry reports PSNR↑, SSIM↑, and LPIPS↓ from top to bottom.

	ScanNet++ (6 Views)					Replica (6 Views)						
Method	8a20	94ee	7831	a29c	avg	office_2	office_3	office_4	room_0	room_1	room_2	avg
FSGS	23.56	20.72	20.86	21.60	21.69	25.99	24.24	23.51	21.65	22.89	23.18	23.91
	0.831	0.755	0.806	0.812	0.801	0.903	0.877	0.866	0.772	0.820	0.836	0.846
	0.234	0.402	0.309	0.246	0.298	0.092	0.114	0.148	0.171	0.171	0.175	0.145
InstantSplat	22.84	20.27	21.16	20.50	21.19	26.62	20.04	22.72	21.35	23.65	24.14	23.09
	0.844	0.815	0.761	0.822	0.811	0.913	0.850	0.864	0.782	0.835	0.850	0.849
	0.121	0.209	0.275	0.165	0.193	0.071	0.187	0.150	0.138	0.142	0.150	0.140
Difix3D+	23.86	19.42	18.86	20.34	20.62	23.40	21.70	22.79	20.42	21.22	21.62	21.86
	0.811	0.765	0.695	0.784	0.764	0.862	0.845	0.831	0.747	0.776	0.800	0.810
	0.175	0.269	0.312	0.219	0.244	0.160	0.149	0.176	0.204	0.219	0.218	0.188
GenFusion	24.69	21.01	20.21	21.91	21.96	25.82	23.42	24.91	23.57	23.04	23.14	23.98
	0.860	0.818	0.750	0.805	0.808	0.909	0.869	0.878	0.808	0.830	0.837	0.855
	0.127	0.211	0.312	0.221	0.218	0.085	0.127	0.139	0.149	0.165	0.187	0.142
GuidedVD-3DGS†	25.10	23.10	22.16	25.21	23.89	27.46	26.81	27.43	24.85	26.00	25.53	26.35
	0.882	0.860	0.803	0.857	0.850	0.916	0.902	0.897	0.796	0.851	0.872	0.872
	0.118	0.201	0.269	0.157	0.182	0.083	0.099	0.122	0.145	0.142	0.142	0.122
GuidedVD-3DGS‡	24.18	22.11	**21.84**	23.79	22.98	**28.04**	**26.46**	**27.63**	25.55	**26.18**	**26.24**	**26.68**
	0.844	0.824	0.768	0.824	0.815	0.925	**0.901**	0.900	0.823	**0.858**	0.871	0.880
	0.143	0.219	0.283	0.172	0.204	0.087	0.112	0.130	0.143	0.157	0.167	0.133
Ours	**24.98**	**22.38**	21.56	**24.70**	**23.41**	28.18	26.23	26.85	**25.65**	25.47	26.00	26.40
	0.877	**0.840**	**0.783**	**0.841**	**0.835**	**0.927**	0.900	**0.900**	**0.831**	0.857	**0.878**	**0.882**
	0.104	**0.186**	**0.269**	**0.164**	**0.181**	**0.062**	**0.087**	**0.104**	**0.117**	**0.127**	**0.128**	**0.104**

† Reported in paper, ‡ Our reproduction.

Table 14: **Per-scene quantitative results on Mip-NeRF360 with 6 and 9 input views.** Each entry reports PSNR↑, SSIM↑, and LPIPS↓ from top to bottom.

	MIP-NERF360 (6 VIEWS)										MIP-NERF360 (9 VIEWS)									
Method	bicycle	bonsai	counter	flowers	garden	kitchen	room	stump	treehill	avg	bicycle	bonsai	counter	flowers	garden	kitchen	room	stump	treehill	avg
FSGS	14.18	13.39	15.02	12.49	17.17	17.14	14.58	15.75	13.16	14.76	14.14	14.36	16.18	13.60	17.75	15.02	15.99	17.37	13.84	15.36
	0.193	0.322	0.409	0.161	0.397	0.520	0.472	0.183	0.278	0.326	0.251	0.387	0.486	0.190	0.419	0.495	0.554	0.266	0.307	0.373
	0.604	0.588	0.452	0.653	0.354	0.355	0.487	0.556	0.736	0.532	0.603	0.508	0.387	0.636	0.375	0.448	0.393	0.463	0.703	0.502
InstantSplat	16.93	13.89	16.14	**13.65**	17.37	18.26	15.51	**18.11**	13.33	15.91	12.86	14.84	17.30	13.38	18.41	**19.82**	17.79	**19.19**	13.29	16.32
	0.334	0.369	0.463	0.186	0.405	**0.606**	0.511	**0.277**	0.342	0.388	0.291	0.424	0.526	0.198	0.446	**0.666**	0.601	**0.358**	0.385	0.433
	0.457	0.550	0.410	0.546	0.296	**0.259**	0.432	0.459	0.577	0.443	0.617	0.482	0.362	0.571	0.257	**0.211**	0.331	0.405	0.532	0.419
Difix3D+	15.04	13.47	14.96	12.95	17.13	17.37	14.62	15.39	13.57	14.94	15.38	14.55	16.65	13.83	18.52	17.98	16.51	16.71	14.25	16.04
	0.187	0.323	0.397	0.141	0.374	0.500	0.450	0.141	0.263	0.309	0.219	0.387	0.484	0.182	0.437	0.546	0.536	0.215	0.308	0.368
	0.468	**0.494**	**0.380**	**0.493**	**0.291**	0.281	0.431	**0.445**	**0.491**	**0.419**	**0.451**	**0.427**	**0.323**	**0.434**	0.244	0.259	0.355	**0.395**	**0.451**	**0.371**
GenFusion	16.29	14.43	**16.54**	13.90	18.23	18.09	17.24	17.68	**15.18**	16.40	**16.53**	**15.29**	**17.76**	**14.91**	19.38	19.88	19.03	18.53	**16.62**	17.55
	0.292	0.389	0.474	**0.200**	0.413	0.521	0.579	0.226	**0.363**	0.384	0.309	**0.429**	**0.547**	**0.234**	0.475	0.586	0.651	0.290	**0.393**	0.435
	0.570	0.534	0.420	0.595	0.379	0.352	0.407	0.604	0.613	0.487	0.537	0.448	0.364	0.496	0.302	0.254	0.307	0.486	0.488	0.409
GuidedVD-3DGS	14.80	12.75	12.05	12.48	15.20	14.32	13.57	15.76	14.12	13.89	14.49	14.56	16.81	13.49	18.52	15.81	17.57	17.27	13.44	15.77
	0.189	0.293	0.305	0.135	0.210	0.359	0.462	0.173	0.333	0.273	0.238	0.383	0.506	0.194	0.456	0.542	0.598	0.246	0.313	0.386
	0.670	0.722	0.714	0.717	0.541	0.658	0.566	0.602	0.569	0.640	0.519	0.483	0.346	0.492	0.253	0.373	0.327	0.432	0.538	0.418
Ours	**17.33**	**14.74**	16.45	13.55	**18.84**	**18.68**	**17.93**	17.72	15.39	**16.74**	16.21	16.15	17.34	14.98	**19.89**	19.38	**20.55**	18.66	14.87	**17.56**
	0.285	**0.392**	**0.481**	0.192	**0.430**	0.567	**0.615**	0.231	0.341	**0.393**	**0.314**	0.461	0.539	0.229	**0.492**	0.626	**0.700**	0.289	0.384	**0.448**
	0.476	0.524	0.398	0.547	0.294	0.304	**0.370**	0.488	0.522	0.436	**0.450**	0.444	0.365	0.463	0.244	0.266	**0.282**	0.423	0.488	0.381