Reflection-aware Generative Novel View Synthesis

GeonU Kim, Shin Dong-Yeon, and Tae-Hyun Oh

KAIST

{geonukim,shindy,taehyun.oh}@kaist.ac.kr

Fig. 1: Reflection-aware generative novel view synthesis on a single image containing a mirror. Given a single image and target novel view camera poses as input (left), both FlexWorld [10] and Stable Virtual Camera (SEVA) [60] (top two rows) fail to correctly interpret the presence of a mirror, resulting in baked mirror surfaces, leakage artifacts, or inconsistent reflections in the generated views. Ref-GeNVS (bottom row) generates both the mirror surface and the scene revealed in the reflection accurately, producing photorealistic and view-consistent results from single or sparse input images.

Abstract. We propose Ref-GeNVS, a training-free, reflection-aware method for generative novel view synthesis (NVS) in mirror scenes. Existing multi-view diffusion models often fail to recognize the mirror in the scene and cannot exploit reflected content for scene generation. To fix this issue without additional training, our key idea is to treat a mirror image as two complementary views. From input images, we estimate the mirror plane and reflect camera poses to form virtual views. Based on this virtual view setup, we propose a two-stage generation method consisting of Mirror-gated attention and Reflection injection, which enables reflection-consistent NVS by explicitly leveraging reflection relationships in a multi-view diffusion model. Ref-GeNVS inherits the strong generalizability of the multi-view diffusion backbone, while it does not require finetuning. On synthetic and real scenes including mirrors, Ref-GeNVS outperforms recent generative NVS methods by generating reflection-consistent and contextually coherent novel views, revealing scene structure visible only through mirrors.

Keywords: Generative Novel View Synthesis · Mirror Reflections

1 Introduction

Mirrors extend humans' limited field of view, allowing us to directly see regions that would otherwise be invisible. More specifically, humans treat the mirror as an additional viewpoint by separating the mirror from the background in their view, and extrapolate the rest of the scene from the observed evidence. This reflection-aware reasoning is critical in practice, from safety scenarios such as discovering hidden objects around road corners using mirrors, to immersive VR/AR experiences requiring faithful scene reconstruction. Despite this capability, we observe that state-of-the-art generative novel view synthesis (NVS) methods [10, 60] fall short of utilizing mirror cues and typically treat a mirror image as a single ordinary view. As a result, images with mirrors are handled as if the reflections provide no structural information, which leads to scene predictions that disagree with the true mirror content, baked mirror surface, and leakage from incorrect separation of reflective and non-reflective regions (see Fig. 1).

A straightforward approach for reflection-aware generative NVS would be to finetune large generative models on datasets containing mirrors. However, reflection-aware NVS is a physically constrained generation problem, and finetuning alone does not guarantee that models will learn the underlying reflection constraints correctly [12, 13, 22, 25]. Furthermore, collecting large-scale mirror datasets is challenging, and finetuning large models remains computationally expensive.

To circumvent these limitations, we propose Ref-GeNVS, a training-free method for reflection-aware generative NVS in scenes that contain mirrors. Ref-GeNVS uses a multi-view diffusion backbone [6,60] and is designed to incorporate reflected evidence even when only mirror-containing inputs are available. To explicitly inject reflected evidence into the backbone, we first estimate the plane equation of the mirror in the scene, and reflect camera poses of input views to the mirror to construct virtual reflected views. Then we generate the target views via a two-stage generation pipeline: (1) generate target views with mirror-masked input images and reflected virtual views while restricting conditioning to the mirror regions of the reflected views using *Mirror-gated attention*; and (2) complete the mirror surface via *Reflection injection* by injecting features from the reflected target poses into the mirror region at each denoising step. This approach closes the gap between human reflection-aware reasoning and current NVS models, enabling photorealistic and reflection-consistent novel views without finetuning. For evaluation, we construct a dataset of realistic 3D scenes with mirrors to probe reflection handling, and we also test on a real scene dataset [57]. Across both data, Ref-GeNVS generates photorealistic, reflection-consistent novel views from single and sparse inputs, surpassing recent generative NVS methods.

- We propose a training-free pipeline for *the first* reflection-aware generative NVS method, by treating a mirror image as two complementary views to generate the scene reflected on the mirror.
- We introduce Mirror-gated attention and Reflection injection, two techniques designed to explicitly leverage reflected cues during generation and improve reflection-consistent novel-view synthesis without model finetuning.

2 Related Work

Reflection-aware novel view synthesis. Novel view synthesis (NVS) renders photorealistic images from unseen viewpoints given calibrated inputs, typically via implicit radiance fields or explicit point-based primitives for volume rendering, with Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) as representative baselines [2,3,9,14,17,24,26,27,34,35,61]. Although these methods achieve impressive reconstruction quality, they often misinterpret reflective objects since strong view-dependent appearances violate the multi-view photometric consistency assumed by NeRF- and 3DGS-based pipelines. As a result, reflections are treated as outliers and the geometric correspondence induced by mirrors is not exploited. To overcome this limitation, prior reflection modeling separates diffuse and specular components with physically based decomposition and environment lighting estimation [23, 31], or renders real and reflected content using two radiance fields or reparameterized heads and blends the outputs [20, 45, 46]. Mirror-NeRF [57], MS-NeRF [54], TraM-NeRF [44], and MirrorGaussian [29] reconstruct mirrors by reflecting camera rays about the surface normal of the mirrors and rendering them, using volumetric ray marching for NeRF variants and point-based rasterization for Gaussian splatting. These methods only focus on rendering reflective objects themselves, whereas our main goal is to synthesize the remainder of the scene using a mirror and generate the surface of the mirror at the same time.

Our setting is therefore also related to OrCa [43] and World-from-Eyes [1], which can be framed as NVS applied to passive non-line-of-sight imaging. These methods use a multi-view captured glossy object as a proxy mirror to render the surrounding environment. Ref-GeNVS departs from these methods in two key aspects. Our method operates on extremely sparse input images (one or three input images) and assumes an ideal planar mirror.

Generative novel view synthesis. Generative novel view synthesis leverages strong priors from 2D/3D diffusion models to synthesize multi-view-consistent, photorealistic images from single or sparse input views [8, 19, 30, 39, 41, 42, 51, 60]. Another line of work adopts a warp-and-inpaint paradigm: they estimate coarse geometry, warp the input view into the target view, and then apply diffusion to inpaint holes from occlusions and out-of-frame regions [6, 10, 11, 32, 38, 40, 47, 55, 56, 58]. However, none of these methods are trained to handle mirrors. As a result, reflective regions are often misinterpreted as background, producing views inconsistent with the correct mirror reflections. In contrast, Ref-GeNVS explicitly incorporates mirror information into a multi-view diffusion backbone, enabling reflection-consistent novel view synthesis without additional finetuning.

3 Method

Given input images $\mathbf{I}_s$, our goal is twofold: to generate target views $\mathbf{I}_o$ which are consistent with the scene content revealed in the mirror, and to generate the mirror surface in $\mathbf{I}_o$ itself being globally consistent with the generated scene. The

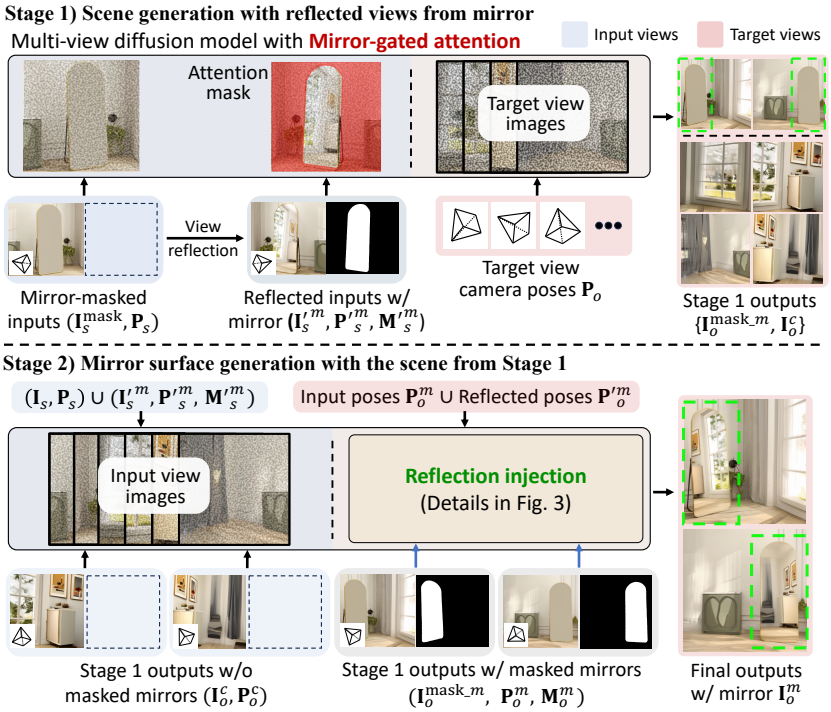

Fig. 2: Overview of Ref-GeNVS. In Stage 1, we first mask mirror surface to prevent interference between real-scene and reflected geometry. Then we generate the reflected scene by mirror-masked inputs and the reflected virtual views which consist of flipped input images, reflected camera poses, and the mirror masks. We propose *Mirror-gated attention* enabling only pixels inside the mirror influence the reflected views. In Stage 2, we perform *Reflection injection* process which generates the mirror surface using the images generated in Stage 1 as inputs. Finally, we obtain complete target images $\mathbf{I}_o$ from target views, which are consistent with the scene reflected on the mirror.

core idea of our method is providing additional information from the reflected region of the input images to the multi-view diffusion model, by obtaining the camera poses of reflected views.

In preprocess, we obtain reflected virtual views by reflecting the camera poses of input views containing mirrors across the estimated mirror plane. We then perform a two-stage generation process, as shown in Fig. 2. Stage 1 synthesizes the target scene from mirror-masked input views and reflected virtual views (Sec. 3.2), while Stage 2 restores the mirror appearance to be consistent with the generated scene (Sec. 3.3). We also introduce an automatic mirror plane estimation pipeline for preprocessing (Sec. 3.4).

Preprocess. Given input images $\mathbf{I}_s$ with binary mirror masks $\mathbf{M}_s$, camera poses $\mathbf{P}_s$ and a mirror plane $\mathbf{e} = [\mathbf{n}, d]$, we produce reflected view tuples $(\mathbf{I}_s'^m, \mathbf{P}_s'^m, \mathbf{M}_s'^m)$ for later stages. With the mirror plane $\mathbf{e}$, each input pose from view including

mirror $\mathbf{P}_s^m$ is reflected across the plane using Householder reflection [21] to obtain the pose of a virtual camera in the mirror. Then we apply horizontal flip to the both image/mask pair and the corresponding reflected camera pose for maintaining the same-hand coordinate convention as the original inputs. As a result, we obtain reflected views with the mirror masks $(\mathbf{I}_s'^m, \mathbf{P}_s'^m, \mathbf{M}_s'^m)$, where $(\mathbf{I}_s'^m, \mathbf{P}_s'^m)$ are additional inputs for multi-view diffusion model and the masks $\mathbf{M}_s'^m$ are used in *Mirror-gated attention* process.

3.1 Preliminary: Multi-view diffusion model

We adopt an "N-in M-out" multi-view diffusion model as our backbone, which aims to generate novel views sampled from:

$$p(\mathbf{I}_o \,|\, (\mathbf{I}_s, \mathbf{P}_s), \mathbf{P}_o)\,, \tag{1}$$

where $\mathbf{I}_s, \mathbf{I}_o$ denote the N input and M target images and $\mathbf{P}_s, \mathbf{P}_o$ are their camera poses, respectively. Multi-view diffusion models [6, 19, 60] use cross-view attention, which connects all different views in self attention by inflating the original 2D self-attention. During denoising process, the cross-view attention modules allow input and target frames to exchange information at every step, while an input/target indicator distinguishes encoded inputs from noisy targets. Building on the multi-view diffusion model, we develop a training-free reflection-aware NVS pipeline which introduces mirror-masked conditioning with *Mirror-gated attention* and novel *Reflection injection* algorithm. Our method can be generally incorporated into multi-view diffusion models using cross-view attention [6, 60] without finetuning.

3.2 Stage 1: Scene generation with Mirror-gated attention

Our goal is to generate target images $\mathbf{I}_o$ at target poses $\mathbf{P}_o$ using the reflected views $(\mathbf{I}_s'^m, \mathbf{P}_s'^m)$ obtained from the preprocess as additional inputs, which can be denoted as:

$$p_{\text{Ref-GeNVS}}\big(\mathbf{I}_o \,\big|\, (\mathbf{I}_s, \mathbf{P}_s), (\mathbf{I}_s'^m, \mathbf{P}_s'^m), \mathbf{P}_o\big)\,. \tag{2}$$

The concept is simple, however, naively feeding reflected views as additional views to a multi-view diffusion model presents three issues:

1. *Mirror recognition.* The model may not recognize the mirror separate it from the non-reflective region, leading to baked surfaces or leakage (see Fig. 1).
2. *Incorrect conditioning from non-mirror pixels.* If reflected images are used as inputs, pixels outside the mirror may also act as reflected evidence and incorrectly influence target-view generation.
3. *Reflection inconsistency in the generated mirror surface.* Since the mirror surface and the rest of the scene are generated independently, the synthesized target view can become inconsistent with the scene reflected in the mirror.

To address these issues, Stage 1 first masks mirror regions with a uniform color to explicitly separate mirrors from the background. The masked input views and reflected virtual views are then processed with *Mirror-gated attention*, which restricts attention to mirror-region tokens in the reflected inputs. In Stage 2, we perform *Reflection injection*, a guided inpainting process that uses the scene synthesized in Stage 1 to complete the mirror surface, ensuring that the synthesized reflections remain consistent with the generated scene.

Problem setup of Stage 1. As shown in Fig. 2 Stage 1, we use two input sets: (i) *mirror-masked original inputs* $(\mathbf{I}_s^{\text{mask}}, \mathbf{P}_s)$ formed by masking the mirror region $\mathbf{M}_s^m$ in $\mathbf{I}_s$, and (ii) the *reflected views with mirror mask* $(\mathbf{I}_s'^m, \mathbf{P}_s'^m, \mathbf{M}_s'^m)$ from the preprocess. The multi-view diffusion model receives $\{(\mathbf{I}_s^{\text{mask}}, \mathbf{P}_s)\} \cup \{(\mathbf{I}_s'^m, \mathbf{P}_s'^m, \mathbf{M}_s'^m)\}$ with $\mathbf{P}_o$ and generates $\mathbf{I}_o^{\text{mask}}$, which are target view images with masked mirror.

Mirror-gated attention for the reflected inputs. For reflected inputs with the binary mirror masks $(\mathbf{I}_s'^m, \mathbf{P}_s'^m, \mathbf{M}_s'^m)$, we use $\mathbf{M}_s'^m$ to restrict attention to tokens inside the mirror in the reflected views. Tokens outside the mirror region are masked during the cross-view attention process, ensuring that non-reflected content does not influence target-view generation.
Formally, let $Q \in \mathbb{R}^{N_q \times d}$ denote the query tokens of the target view and $K, V \in \mathbb{R}^{N_k \times d}$ denote the key and value tokens extracted from the reflected input views. Standard cross-view attention is computed as:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \tag{3}$$

To prevent tokens outside the mirror from contributing to the attention, we construct a binary mask $m \in \{0, 1\}^{N_k}$ derived from the reflected mirror mask $\mathbf{M}_s'^m$. The mask is obtained by downsampling and reshaping $\mathbf{M}_s'^m$ to align with the token layout of the attention module. The attention logits are then gated by:

$$S'_{ij} = \begin{cases} \frac{q_i^\top k_j}{\sqrt{d}}, & m_j = 1 \\ -\infty, & m_j = 0, \end{cases}, \qquad \text{Attn}_{\text{mirror}}(Q, K, V) = \text{softmax}(S')V. \tag{4}$$

This masking sets the attention weight of tokens outside the mirror region to zero after the softmax, ensuring that only reflected content contributes to the target-view synthesis. This Mirror-gated attention is applied consistently across all attention operators in the backbone, including the 3D attention and 1D attention modules [60] according to the model architecture. Restricting attention prevents artifacts caused by reflecting the non-mirror region of the input images.

3.3 Stage 2: Mirror surface generation with Reflection injection

In Stage 2, we complete the mirror surface in the outputs from Stage 1. To complete the mirror surface with reflection-consistent content, we introduce *Reflection injection* method that injects reflected-view evidence into the mirror region during the denoising process. To further stabilize generation near the

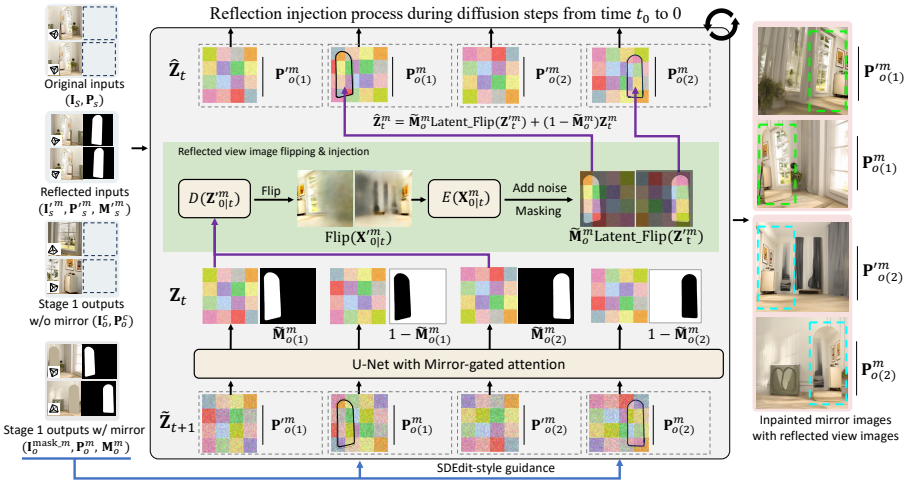

Fig. 3: Reflection injection process in Stage 2. We fill the mirror surface in the target views using three sources: the original inputs $(\mathbf{I}_s, \mathbf{P}_s)$, the reflected inputs from the preprocess $(\mathbf{I}'^m_s, \mathbf{P}'^m_s, \mathbf{M}'^m_s)$, and the Stage 1 outputs $(\mathbf{I}^{\text{mask}}_o, \mathbf{P}_o)$. To generate the mirror surface in the Stage 1 outputs $(\mathbf{I}^{\text{mask}\_m}_o, \mathbf{P}^m_o)$ consistent with the scene in reflection, we condition the model on two type of target-frame poses: the identity poses $\mathbf{P}^m_o$ and the corresponding reflected poses $\mathbf{P}'^m_o$. During each denoising step $t \to t-1$, the latent feature maps $\mathbf{Z}'^m_t$ predicted under $\mathbf{P}'^m_o$ are horizontally flipped and injected inside the target mirror mask $\tilde{\mathbf{M}}^m_o$ of the current target latent $\mathbf{Z}^m_t$. This latent injection is integrated with SDEdit-style guidance [33], enabling natural mirror-surface completion. The result is the inpainted mirror which is consistent with the scene generated in Stage 1 and with the reflected evidence from the original input images.

mirror boundary, we additionally employ an SDEdit-style guidance [33] which enables natural compositing of mirror and non-mirror region.

Input setup with the outputs from Stage 1. To inpaint the masked mirror in the Stage 1 outputs $\mathbf{I}^{\text{mask}}_o$, we first split $\mathbf{I}^{\text{mask}}_o$ into images containing mirrors and images without mirrors. Mirror masks $\mathbf{M}_o$ are obtained using SAM2 [37], guided by the source-view mirror masks $(\mathbf{I}^{\text{mask}}_s, \mathbf{M}^m_s)$. Since the mirror is already masked in Stage 1, SAM2 robustly tracks the mirror region in $\mathbf{I}^{\text{mask}}_o$. We therefore obtain $(\mathbf{I}^{\text{mask}}_o, \mathbf{M}_o) = \{(\mathbf{I}^{\text{mask}\_m}_o, \mathbf{M}^m_o), (\mathbf{I}^c_o, \mathbf{M}^c_o)\}$, where $\mathbf{I}^{\text{mask}\_m}_o$ contains mirrors and $\mathbf{I}^c_o$ contains no mirror regions.

Problem setup of Stage 2. As illustrated in Fig. 3, we use four types of inputs: the original views $(\mathbf{I}_s, \mathbf{P}_s)$, the reflected inputs $(\mathbf{I}'^m_s, \mathbf{P}'^m_s, \mathbf{M}'^m_s)$, the Stage 1 outputs with mirrors $(\mathbf{I}^{\text{mask}\_m}_o, \mathbf{P}^m_o, \mathbf{M}^m_o)$, and Stage 1 outputs without mirrors $(\mathbf{I}^c_o, \mathbf{M}^c_o)$. Our goal is to generate $\mathbf{I}^m_o$ at pose $\mathbf{P}^m_o$ such that the mirror region shows the reflection implied by the reflected virtual target pose $\mathbf{P}'^m_o$.

Reflection injection. Let $\mathbf{Z}^m_t$ be the latent at diffusion step t under pose $\mathbf{P}^m_o$ and $\mathbf{Z}'^m_t$ the latent under the reflected pose $\mathbf{P}'^m_o$. We inject reflected evidence

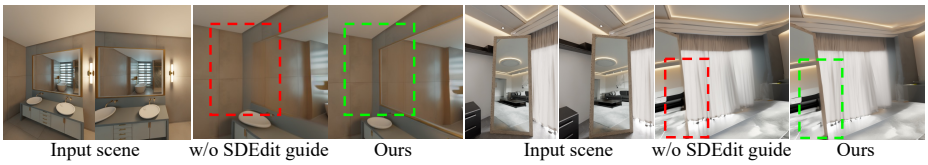

Fig. 4: Effect of SDEdit-style [33] guidance. Without SDEdit-style guidance, the model can produce noticeable artifacts near the boundaries of the mirror. The guidance enable natural compositing of mirror and non-mirror region.

into the mirror region using the binary mask $\mathbf{M}_o^m$:

$$\hat{\mathbf{Z}}_t^m = \tilde{\mathbf{M}}_o^m \, \text{Latent\_Flip}(\mathbf{Z}_t'^m) + (\mathbf{1} - \tilde{\mathbf{M}}_o^m) \, \mathbf{Z}_t^m, \tag{5}$$

where $\tilde{\mathbf{M}}_o^m$ denotes the mirror mask resized to the latent resolution. The operator Latent_Flip$(\cdot)$ aligns the reflected latent to the target frame via decode–flip–encode:

$$\text{Latent\_Flip}(\mathbf{Z}_t'^m) = E\Big(\text{Flip}\Big(D(\mathbf{Z}_{0|t}'^m)\Big)\Big) + \sigma_t \boldsymbol{\epsilon}_t, \tag{6}$$

where $D(\cdot)$ and $E(\cdot)$ denote the VAE decoder and encoder, $\mathbf{Z}_{0|t}'^m$ is the predicted clean latent from $\mathbf{Z}_t'^m$ using Tweedie's formula [15], and σ_t is the noise scale at step t. The decode–flip–encode route is necessary since the latent feature space is not inherently flip-equivariant with respect to the image space.

SDEdit-style guidance for mirror boundary stabilization. Although reflection injection enforces reflection consistency inside the mirror, directly injecting reflected latent into the mirror region may lead to unstable generation near the mirror boundary due to abrupt transitions between injected latent and surrounding regions, shown in Fig. 4. We therefore propose an SDEdit-style guidance that anchors the non-mirror region, which enables robust mirror boundary generation.

SDEdit generates images by guiding the diffusion process from an intermediate noisy state and has been applied to image editing and composition tasks (see [33] for further details). Inspired by this, we extend the idea to a multi-view diffusion setting to facilitate natural composition between mirror and non-mirror regions. We first encode the Stage 1 output $\mathbf{I}_o^{\text{mask\_}m}$ and apply forward diffusion to obtain, $\mathbf{Z}_t^{\text{init}} = \sqrt{\bar{\alpha}_t} E(\mathbf{I}_o^{\text{mask\_}m}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\bar{\alpha}_t$ denotes the cumulative noise schedule and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. During denoising, the non-mirror region is constrained by:

$$\tilde{\mathbf{Z}}_t^m = (\mathbf{1} - \tilde{\mathbf{M}}_o^m)\mathbf{Z}_t^{\text{init}} + \tilde{\mathbf{M}}_o^m \hat{\mathbf{Z}}_t^m. \tag{7}$$

This anchors the surrounding context while allowing the mirror region to be synthesized during generation. We apply SDEdit-style guidance only until the timestep t_0, to preserve the surrounding scene in the early denoising stage. From the timestep t_0, we instead apply reflection injection, which progressively

Fig. 5: Input views and test trajectories. To evaluate reflection-aware novel view synthesis, we sample challenging, scene-specific trajectories for our synthetic dataset (top) and the real dataset [57] (bottom). Each input view contains a mirror, and the trajectory includes regions visible only via reflection (red dashed box). For clarity, we overlay neon highlights to mark the mirror in the input views.

synthesizes the mirror appearance under the guidance of the reflected inputs. The diffusion prior then naturally harmonizes the boundary between the synthesized mirror and the surrounding scene.

3.4 Automatic preprocess

We present an automated pipeline for mirror detection and mirror plane estimation in our preprocess. Since mirror detection remains a challenging problem, previous mirror-related NVS methods (*e.g.*, Mirror-NeRF [57], MirrorGaussian [29]) assume given mirror masks. To relax this strong assumption while maintaining robustness to errors in the estimated mirror mask, we propose a mask-tolerant mirror plane estimation pipeline.

We first estimate binary mirror masks $\mathbf{M}_s$ from input images $\mathbf{I}_s$ using DAM [52], an off-the-shelf mirror segmentation method. Instead of directly exploiting mirror edge points for mirror plane estimation as in prior work [29], we mask the mirror region $\mathbf{M}_s$ in $\mathbf{I}_s$ with a single color and predict per-pixel 3D points $\mathbf{X}_s$ using a 3D point regression method [28]. From the predicted per-pixel 3D points, we extract 3D points on the mirror surface $\mathbf{X}_s^{\text{mirror}}$ and estimate the mirror plane equation $\mathbf{e} = [\mathbf{n}, d]$ by fitting a plane using RANSAC [16]. See the supplementary material for additional details and a robustness analysis of the automation process.

Table 1: Quantitative results of sparse-image novel view synthesis using DreamSim (DS) [18], CLIP [36] similarity score, PSNR, SSIM [48], and LPIPS [59]. Our method achieves better performance compared to MVGenMaster [6] and SEVA [60].

	Real data [57]					Synthetic data				
Metric	DS↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	DS↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓
MVGenMaster [6]	0.361	0.867	12.75	0.413	0.536	0.227	0.926	16.32	0.719	0.378
Ours (MVGen.)	0.194	0.930	13.67	0.455	0.474	0.116	0.959	17.50	0.732	0.313
SEVA [60]	0.275	0.920	12.03	0.394	0.564	0.270	0.924	13.62	0.612	0.492
Ours (SEVA)	0.129	0.951	14.19	0.455	0.466	0.156	0.948	15.27	0.676	0.403

4 Experiments

Dataset. Evaluation of reflection-aware novel view synthesis (NVS) requires ground truth novel views of the scene on the opposite side of the mirror that are visible only via reflection. However, most of existing datasets containing mirror [12, 13, 49, 50, 53] provide views only facing the mirror, not suitable for evaluation. To evaluate reflection-aware NVS, we gather indoor scenes from Blender Demo [4], BlenderKit [5], and CGTrader [7], and place mirrors in the scene by focusing on environments where mirrors naturally occur, such as bathrooms, and living rooms. We also adopt the Blender scenes from the Reflect3r [50] dataset and redefine the camera viewpoints. The scenes are rendered with physically based mirror materials, allowing realistic reflections. We provide 8 synthetic scenes consist of ground truth multi-view images with the corresponding camera intrinsic and extrinsic parameters for evaluation. We also conduct experiments on the Mirror-NeRF dataset [57], which is captured in real-world.

Input & test views selection. To evaluate reflection-aware generative NVS, we design challenging input–test splits for both the synthetic dataset and the real dataset [57]. As shown in Fig. 5, the input images include a visible mirror, and the test trajectory is chosen to contain the scene reflected on the mirror. For the synthetic dataset, we render each scene along an orbit trajectory showing both the mirror and the reflected scene. For the real dataset, we resample camera poses from the views provided by the original Mirror-NeRF dataset to set target views which reveal mirror-visible content.

Competing methods. We evaluate our method by incorporating it into representative generative NVS approaches under sparse-input and single image NVS settings. For sparse-input novel view synthesis, we include MVGenMaster [6] and SEVA [60] and we use them as baselines of Ref-GeNVS. For single-image novel view synthesis, we similarly compare against VistaDream [47], FlexWorld [10], and SEVA [60], and we use SEVA as our baseline module.

4.1 Sparse-input novel view synthesis

Evaluation protocol in sparse-input setting. We evaluate competing methods using both perceptual and pixel-level metrics. We report DreamSim [18],

Fig. 6: Qualitative comparison under sparse-input novel view synthesis. Ref-GeNVS produces more consistent and reflection-aware target views compared to prior methods [6, 60], especially in regions only visible through the mirror.

CLIP [36] similarity, PSNR, SSIM [48], and LPIPS [59] computed between generated images and ground-truth target views. DreamSim and CLIP similarity quantify high-level perceptual similarity while PSNR, SSIM, and LPIPS qualify geometric consistency in the scenes containing mirror.

Quantitative results. As shown in Table 1, across all quantitative metrics including DreamSim, CLIP similarity, PSNR, SSIM, and LPIPS, Ref-GeNVS surpasses both MVGenMaster and SEVA by substantial margins. These results show that existing generative novel view synthesis methods are fundamentally limited when the scene requires reasoning about reflected content, while Ref-GeNVS explicitly handles reflected evidence.

Qualitative results. The qualitative results make this gap even more apparent. As shown in Fig. 6, competing methods consistently misinterpret or ignore mirror-reflected regions, leading to severe structural inconsistencies and hallucinated content when synthesizing the target view. In contrast, Ref-GeNVS based on the competing methods effectively leverages the reflection and produces novel views that align closely with the ground truth images. The fidelity and coherence of mirror-revealed regions in our outputs highlight Ref-GeNVS 's advantage in reflection-dependent scenarios and its generalizability across different model architectures.

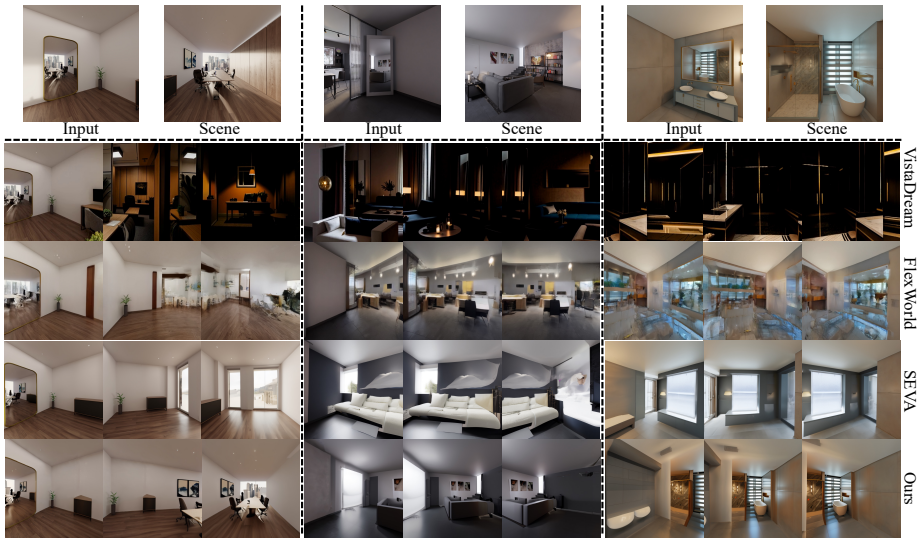

Fig. 7: Qualitative comparison in the single image NVS setting. Despite observing only single input image, Ref-GeNVS utilizes mirror cues to generate coherent novel views, outperforming prior methods [10, 47, 60] on reflected and hidden regions.

4.2 Novel view synthesis from a single image

Evaluation protocol in single image input setting. The single image NVS evaluation is affected by scale and pose ambiguity, making per-view pixel metrics not directly comparable to ground truth. We therefore report DreamSim and CLIP similarity to assess perceptual alignment between ground truth and generated images, which can be seen as approximately scale-invariant. For each predicted target, we compute two perceptual scores for each metric: (i) relative to the input image and (ii) relative to a ground-truth image showing the reflected region along the test trajectory (as shown in the first row of Fig. 7). We average these two scores to capture both global coherence with the input and semantic alignment with the reflected content.

Table 2: Single image result.

Method	DS↓	CLIP↑
VistaDream [47]	0.569	0.866
FlexWorld [10]	0.369	0.879
SEVA [60]	0.377	0.905
Ours	0.341	0.911

Results. Table 2 shows that Ref-GeNVS performs better than VistaDream [47], FlexWorld [10], and SEVA in terms of DreamSim (DS) and CLIP similarity with the input view and the ground truth reflected view, suggesting that incorporating explicit reflection reasoning is beneficial even when only a single image is available. Figure 7 shows that prior methods often struggle to interpret mirror-reflected regions from limited evidence, leading to incomplete or less coherent predictions when synthesizing the target view. Ref-GeNVS, by contrast, leveraging the reflected information in the input

Table 3: Ablation study on injection of reflected virtual view, Mirror-gated attention (Mirror-gated attn.), and two-step generation process (Two-step gen.). We compare the performance by adding each component on the baseline model, SEVA [60].

Virtual view	Mirror-gated attn.	Two-step gen.	DS ↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓
X	X	X	0.270	0.924	13.62	0.612	0.492
O	X	X	0.229	0.923	13.59	0.643	0.536
O	O	X	0.160	0.948	15.01	0.671	0.416
O	O	O	0.156	0.948	15.27	0.676	0.403

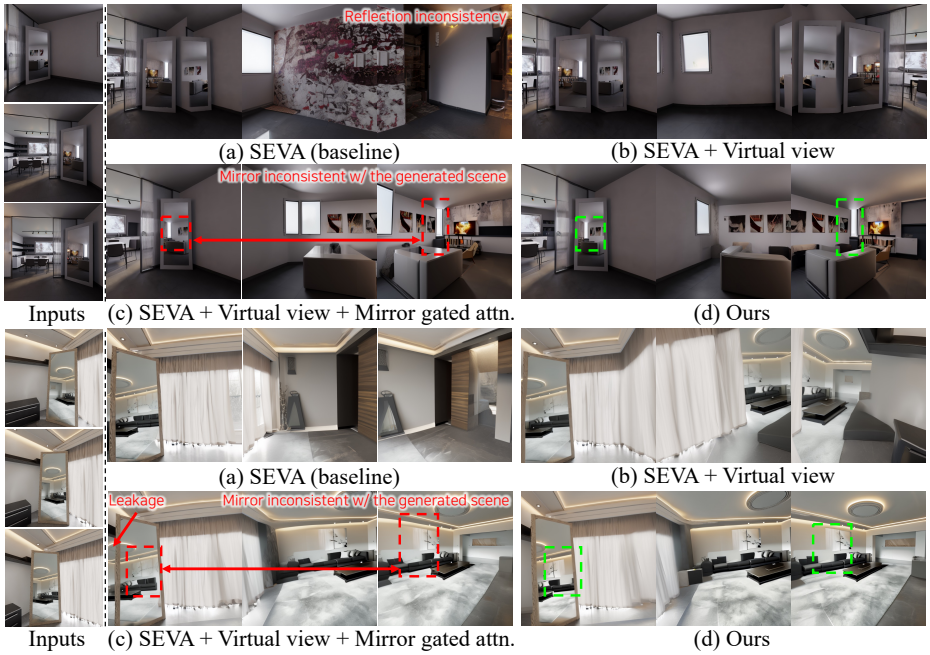

Fig. 8: Ablation on synthetic scenes by progressively adding our components: (a) SEVA baseline, (b) + virtual view, (c) + Mirror-gated attention, and (d) + two-stage generation (full method). Our full model produces reflections consistent with the generated scene.

image and produces novel views with more aligning the scene context and stable structures in the mirror-revealed areas.

4.3 Ablation Study

We conduct ablation studies on the synthetic dataset under the sparse-input setting. As shown in Fig. 8, we progressively add each component to the baseline SEVA [60]: (a) the baseline SEVA, (b) + reflected virtual view, (c) + Mirror-gated attention, and (d) + two-step generation (full method). The baseline often produces scenes inconsistent with the reflection observed in the mirror.

Fig. 9: Result using an input with a distorted mirror. We warp the distorted reflection into the planar-mirror representation and apply Ref-GeNVS without modification.

Introducing the virtual view improves geometric reasoning for reflections but still allows non-reflected regions in the inputs to influence the mirror region. Adding Mirror-gated attention restricts the attention to reflection-relevant regions, preventing leakage from non-mirror areas and reducing duplicated structures or artifacts. Finally, the two-step generation further improves the alignment between the generated scene and the mirror surface. Although this improvement is not strongly reflected in the quantitative metrics, it can be clearly observed in the results, where the reflected scene becomes better aligned with the generated scene. Table 3 also shows that adding components leads to performance improvements.

4.4 Application

Generalization to distorted reflections. To explore an extension to distorted reflections, we conduct an experiment with a distorted mirror by constructing a scene containing a mirror ball. Using the given mirror ball geometry, we warp the distorted reflection into the planar-mirror representation and apply Ref-GeNVS until Stage 1 without changing the pipeline. Figure 9 shows that our method can use wide-field reflections from distorted mirrors, such as convex mirrors, demonstrating broader practical utility.

5 Conclusion

We propose Ref-GeNVS, a reflection-aware, training-free novel view synthesis method that augments a multi-view diffusion backbone with reflected views. By estimating the mirror plane and reflecting camera poses, our method converts mirror-containing inputs into complementary viewpoints that guide generation toward reflection-consistent results in a training-free manner. Empirically, Ref-GeNVS synthesizes novel views that better align with mirror-reflected content and expose global scene structure compared to prior approaches. Our work has the potential to benefit a wide range of applications, including 3D scene generation and embodied agents. For example, by leveraging partial mirror observations, our method can enable embodied agents to infer and predict regions beyond their direct field of view, facilitating more informed navigation.

References

1. Alzayer, H., Zhang, K., Feng, B., Metzler, C.A., Huang, J.B.: Seeing the world through your eyes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4864–4873 (2024)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
4. Blender Foundation: Blender demo files (2025), software resources
5. BlenderKit Community: Blenderkit asset repository (2025), 3D asset library
6. Cao, C., Yu, C., Liu, S., Wang, F., Xue, X., Fu, Y.: 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 (2025)
7. CGTrader Marketplace: Cgtrader asset library (2025), commercial 3D models
8. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Generative novel view synthesis with 3d-aware diffusion models. In: ICCV (2023)
9. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. ECCV (2022)
10. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. In: Thirty-ninth Conference on Neural Information Processing Systems (2025)
11. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
12. Dhiman, A., Shah, M., Babu, R.V.: Mirrorverse: Pushing diffusion models to realistically reflect the world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11239–11249 (2025)
13. Dhiman, A., Shah, M., Parihar, R., Bhalgat, Y., Boregowda, L.R., Babu, R.V.: Reflecting reality: Enabling diffusion models to produce faithful mirror reflections. In: 2025 International Conference on 3D Vision (3DV). pp. 824–834. IEEE (2025)
14. Dong-Yeon, S., Jun-Seong, K., Byung-Ki, K., Oh, T.H.: Hdr-nsff: High dynamic range neural scene flow fields. In: The Fourteenth International Conference on Learning Representations (2026)
15. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
16. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
17. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR (2022)
18. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
19. Gao*, R., Holynski*, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T., Poole*, B.: Cat3d: Create anything in 3d with multi-view diffusion models. Advances in Neural Information Processing Systems (2024)
20. Guo, Y.C., Kang, D., Bao, L., He, Y., Zhang, S.H.: Nerfren: Neural radiance fields with reflections. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

21. Householder, A.S.: The theory of matrices in numerical analysis. Courier Corporation (2006)
22. Hyun-Bin, O., Takida, Y., Uesaka, T., Oh, T.H., Mitsufuji, Y.: Pavas: Physics-aware video-to-audio synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14481–14491 (2026)
23. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. arXiv preprint arXiv:2311.17977 (2023)
24. Jun-Seong, K., Yu-Ji, K., Ye-Bin, M., Oh, T.H.: Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In: European Conference on Computer Vision. pp. 384–401. Springer (2022)
25. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
26. Kerbl, B., Reiser, C., Zhang, S., Lombardi, S., Niessner, M., Takikawa, T.: 3d gaussian splatting for real-time radiance field rendering. SIGGRAPH (2023)
27. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: CVPR (2024)
28. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
29. Liu, J., Tang, X., Cheng, F., Yang, R., Li, Z., Liu, J., Huang, Y., Lin, J., Liu, S., Wu, X., et al.: Mirrorgaussian: Reflecting 3d gaussians for reconstructing mirror reflections. ECCV (2024)
30. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. ICCV (2023)
31. Liu, Y., Wang, P., Lin, C., Long, X., Wang, J., Liu, L., Komura, T., Wang, W.: Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. arXiv preprint arXiv:2305.17398 (2023)
32. Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., Wang, X.: You see it, you got it: Learning 3d creation on pose-free videos at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2016–2029 (2025)
33. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. ECCV (2020)
35. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (TOG) (2022)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
37. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
38. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025)

39. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., Wu, J.: ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In: CVPR (2024)
40. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. Advances in Neural Information Processing Systems **37**, 80220–80243 (2024)
41. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: multi-view diffusion for 3d generation. In: ICLR (2024)
42. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024)
43. Tiwary, K., Dave, A., Behari, N., Klinghoffer, T., Veeraraghavan, A., Raskar, R.: Orca: Glossy objects as radiance-field cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20773–20782 (2023)
44. Van Holland, L., Bliersbach, R., Müller, J.U., Stotko, P., Klein, R.: Tram-nerf: Tracing mirror and near-perfect specular reflections through neural radiance fields. arXiv preprint arXiv:2310.10650 (2023)
45. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-NeRF: Structured view-dependent appearance for neural radiance fields. CVPR (2022)
46. Wang, F., Rakotosaona, M.J., Niemeyer, M., Szeliski, R., Pollefeys, M., Tombari, F.: Unisdf: Unifying neural representations for high-fidelity 3d reconstruction of complex scenes with reflections. NeurIPS (2024)
47. Wang, H., Liu, Y., Liu, Z., Wang, W., Dong, Z., Yang, B.: Vistadream: Sampling multiview consistent images for single-view scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26772–26782 (2025)
48. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
49. Warren, A., Xu, K., Lin, J., Tam, G.K., Lau, R.W.: Effective video mirror detection with inconsistent motion cues. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17244–17252 (2024)
50. Wu, J., Wang, Z., Laina, I., Prisacariu, V.A.: Reflect3r: Single-view 3d stereo reconstruction aided by mirror reflections. arXiv preprint arXiv:2509.20607 (2025)
51. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., Holynski, A.: Reconfusion: 3d reconstruction with diffusion priors. arXiv (2023)
52. Xing, Z., Liu, L., Yang, Y., Wang, H., Ye, T., Chen, S., Li, W., Liu, G., , Zhu, L.: Detect any mirrors: Boosting learning reliability on large-scale unlabeled data with an iterative data engine. In: CVPR (2025)
53. Yang, X., Mei, H., Xu, K., Wei, X., Yin, B., Lau, R.W.: Where is my mirror? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8809–8818 (2019)
54. Yin, Z.X., Jiao, P.Y., Qiu, J., Cheng, M.M., Ren, B.: Ms-nerf: Multi-space neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
55. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:2405.15364 (2024)

56. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
57. Zeng, J., Bao, C., Chen, R., Dong, Z., Zhang, G., Bao, H., Cui, Z.: Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. Proceedings of the 31st ACM International Conference on Multimedia (2023)
58. Zhang, J., Li, X., Wan, Z., Wang, C., Liao, J.: Text2nerf: Text-driven 3d scene generation with neural radiance fields. IEEE TVCG (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR (2018)
60. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
61. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: ECCV (2024)