SyncFix: Fixing 3D Reconstructions via Multi-View Synchronization

Deming Li[1] Abhay Yadav[1] Cheng Peng[2]
Rama Chellappa[1] Anand Bhattad[1]

[1]Johns Hopkins University [2]University of Virginia

Project page: https://syncfix.github.io/

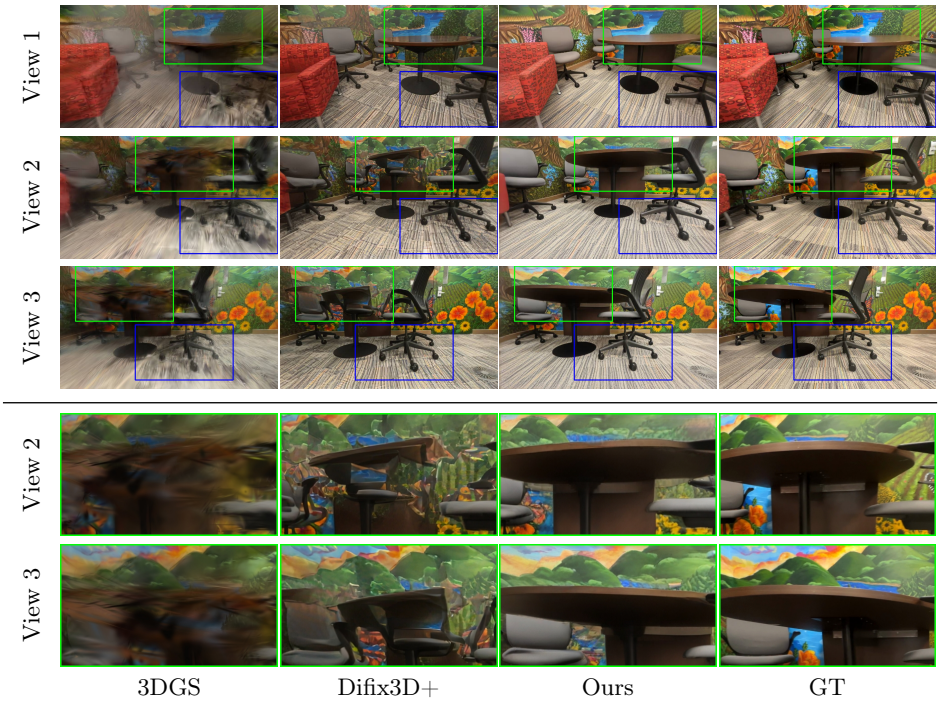


Fig. 1: Independent diffusion refinement (DifiX3D+) processes each view separately, leading to inconsistent geometry across views (see close-view of table in the bottom panel and highlighted regions above). SyncFix (Ours) instead refines all views jointly, enforcing cross-view agreement during denoising and producing coherent 3D structure.

Abstract. We present **SyncFix**, a framework that enforces cross-view consistency during the diffusion-based refinement of reconstructed scenes. SyncFix formulates refinement as a joint latent bridge matching problem, synchronizing distorted and clean representations across multiple views to fix the semantic and geometric inconsistencies. This means SyncFix learns a joint conditional over multiple views to enforce consistency throughout the denoising trajectory. Our training is done only on image pairs, but it generalizes naturally to an arbitrary number of views

during inference. Moreover, reconstruction quality improves with additional views, with diminishing returns at higher view counts. Qualitative and quantitative results demonstrate that SyncFix consistently generates high-quality reconstructions and surpasses current state-of-the-art baselines, even in the absence of clean reference images. SyncFix achieves even higher fidelity when sparse references are available.

1 Introduction

Radiance fields [24] and 3D Gaussian splatting [12] under sparse-view or off-trajectory settings often produce scenes plagued by floaters, texture artifacts, and geometric distortions. Recent approaches attempt to mitigate these issues by leveraging the rich priors of 2D generative models [6, 19, 39, 41, 46]. Typically, these methods render novel views, refine images using an image or video diffusion model with conditioning, and distill the refined outputs back into the 3D representation.

However, these refinement methods [39,41] treat each view as an independent 2D generation task. This ignores important cross-view geometric and semantic constraints and that all views are projections of the same underlying 3D scene. Therefore, independent refinement can produce inconsistent results for ambiguous structures or regions with large artifacts across different viewpoints. When these conflicting signals are distilled back into the 3D model, optimization becomes unstable, leading to appearance drift and degraded geometry. Methods such as [19, 46] leverage video diffusion models to refine fixed-frame sequences with smooth temporal motion. Despite improved consistency, the video generation incurs computation overhead and is difficult to adapt for refining variable numbers of wide-baseline inputs.

In this paper, we impose multi-view consistency directly during the refinement process rather than correcting it post hoc. To this end, we introduce **SyncFix**, a latent bridge matching formulation that models refinement as joint conditional generation over multiple views. SyncFix aligns latent representations across views during the denoising process and encourages information exchange via cross-view attention without any explicit camera information. Each view is therefore refined in the context of all others, ensuring consistent semantics and geometry before 3D distillation.

During training, we train to refine only two views simultaneously, but its permutation-invariant latent interactions using attention naturally extend seamlessly to arbitrary numbers of views at inference. As more views are provided, the cross-view constraints become stricter and the solution space narrows, which leads to systematically better final reconstructions. We also find that SyncFix produces consistent multi-view outputs even when no clean reference frames are available, and it can further benefit from sparse references when they exist. We substantiate these claims using perceptual metrics and through analyses of geometric alignment and semantic correspondence.

2 Related Work

Sparse-View 3D Reconstruction. NeRF [24] and 3DGS [12] produce high-fidelity geometry from dense captures. With sparse inputs, the problem becomes ill-posed: reconstructions develop floaters, fragmented surfaces, and blurred textures. Regularization-based approaches constrain frequency content [25,34,44] or inject depth priors [5,13,33,43]. Feed-forward methods [9,11,45] leverage learned multi-view encoders [36] for direct prediction. These methods reduce artifacts but do not eliminate them; motivating generative refinement.

Per-View Refinement with 2D Priors. The dominant repair paradigm renders novel views, refines each independently with a 2D diffusion model, and distills the results back into the 3D representation. This follows the Score Distillation Sampling framework [28, 35], originally developed for text-to-3D generation [14,37] and since adapted for reconstruction repair. Nerfbusters [39] applies diffusion guidance during NeRF optimization. ReconFusion [42] trains a multi-view conditioned prior for few-view regularization. Instruct-NeRF2NeRF [10] uses SDEdit [23] on individual renderings. Difix3D+ [41] distills a single-step refinement filter and adds limited cross-view conditioning through reference-view cross-attention. FreeFix [48] manipulates latent trajectories without retraining. These methods ultimately optimize per-view reconstruction objectives, even when limited cross-view signals are introduced. The 2D prior never sees multiple views simultaneously, so it may interpret the same artifact differently from each angle. The 3D distillation stage must reconcile the resulting conflicts.

Coupling Views inside the Generator. A natural fix is to condition generation on multiple views at once. Video diffusion models provide implicit coupling through temporal attention: 3DGS-Enhancer [19] casts view consistency as temporal coherence and refines 3DGS with a confidence-aware loop; ReconX [18] conditions video diffusion on a global point cloud for sparse-view reconstruction. Multi-view diffusion architectures couple views more directly. MVDream [31], SyncDreamer [21], CAT3D [8], and ViewCrafter [47] synchronize attention across generated views. 3DEnhancer [22] applies this idea to 3D enhancement, using multi-view row attention and epipolar aggregation within a latent diffusion model. However, they remain within the denoising diffusion framework: views interact only through attention over noisy intermediate states, and sampling requires iterative denoising.

Flow-Based Refinement. Rectified Flow [20] and Flow Matching [17] replace iterative denoising with learned ODEs that transport between distributions. Latent Bridge Matching (LBM) [4] extends this to conditional generation via deterministic bridges between paired latents. SyncLight [29] adds cross-view attention to LBM for consistent relighting. SyncFix builds on this foundation but targets a different problem: repairing degraded 3D geometry. We learn a joint flow over the product latent space of multiple distorted renderings and their clean counterparts. While attention-based diffusion models such as 3DGS-Enhancer [19] and

GSFixer [46] implicitly couple views through shared noise states, and FlowR [6] employs continuous-time flow matching for reconstruction densification, SyncFix couples views via discrete-time latent bridge matching, enforcing semantic and geometric agreement before any image is decoded. Thus, SyncFix is a single-pass, intrinsically consistent multi-view refinement without iterative denoising.

3 Method

We aim to refine 3D reconstructions while maintaining multi-view consistency. Existing methods [39, 41] refine each rendered view independently and rely on the 3D representation to harmonize conflicting corrections. This harmonization is prone to be noisy. For example, one view may introduce a painting on a wall while another suppresses them, leading to inconsistent geometry and texture when fused back into the 3D representation. We instead model refinement as a *joint conditional generation* problem over multiple views.

Let a distorted 3D reconstruction produce N rendered views $X_D = \{x_D^{(1)}, \ldots, x_D^{(N)}\}$, with corresponding clean targets $X_{GT} = \{x_{GT}^{(1)}, \ldots, x_{GT}^{(N)}\}$. Marginal methods like Difix3D+ [41] learn independent mappings $P(x_{GT}^{(i)} \mid x_D^{(i)})$. SyncFix instead models the joint conditional $P(X_{GT} \mid X_D)$ and enforces cross-view consensus during generation. This shift from marginal to joint modeling is central: corrections for each view are conditioned on all other views simultaneously.

3.1 Data Generation and Distortion Modeling

To train SyncFix, we construct paired distorted and clean multi-view data reflecting realistic reconstruction failure modes. We use scenes from the DL3DV [16] dataset and train 3D Gaussian Splatting (3DGS) [12] models under multiple sparsity levels (6, 8, 10, 12, 16, or 32 training views). This induces geometric inconsistencies, floaters, and blurred high-frequency regions at various levels. We further introduce a cross-split strategy: the available training views are partitioned into multiple disjoint subsets, and a separate sparse 3DGS is trained for each split, yielding diverse distortions at different viewpoints from the same underlying scene. Novel views rendered from these degraded 3DGS models form distorted inputs X_D, while ground-truth images provide X_{GT}.

To encourage multi-view reasoning, rendered views within each batch are sampled with moderate pose differences to ensure sufficient visual overlap. At the same time, these poses are far from the sparse 3DGS training views, guaranteeing large artifacts. This setup forces the model to fix shared structural inconsistencies across views rather than performing independent cosmetic corrections.

3.2 Multi-View Latent Bridge Matching

Prior diffusion-based refinement method [41] injects artificial noise and perform denoising at a fixed timestep. This assumes artifacts resemble Gaussian corruption. In practice, 3D reconstruction errors are spatially structured and view-dependent; uniform denoising is insufficient. Our method is summarized in Fig. 2.

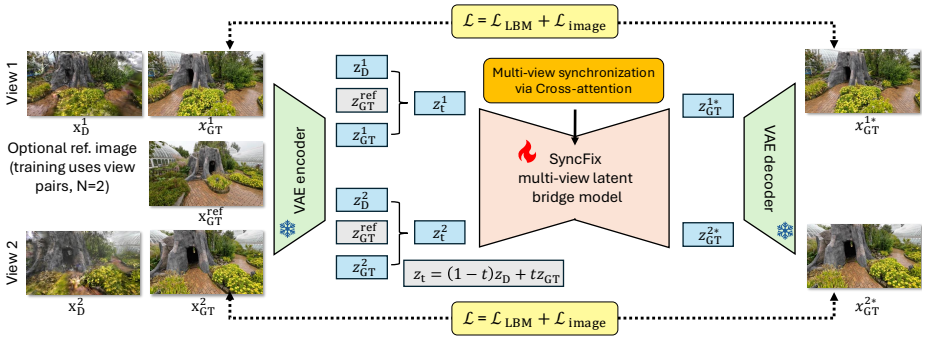


Fig. 2: SyncFix overview. Distorted renderings from multiple viewpoints x_D are encoded into latent representations and transported toward clean targets x_{GT} using latent bridge matching. SyncFix learns a *joint latent bridge* over multiple views, coupling latent trajectories through cross-view attention to enforce multi-view consistency during refinement. The model is trained using view pairs ($N = 2$) but generalizes to an arbitrary number of views at inference. Optional reference images can be provided to guide refinement. Here z_D denotes distorted latents, z_{GT} clean target latents, and $z_t = (1-t)z_D + tz_{GT}$ the bridge interpolation between them.

We adopt LBM [4] to repurpose diffusion model to directly learn the transport from the distorted distribution to clean distribution. Each image is encoded using a pre-trained autoencoder, yielding latent tensors z_D and z_{GT}. For multi-view inputs, we denote the joint latent tensor as:

$$Z_D = \{z_D^{(1)}, \ldots, z_D^{(N)}\}, \quad Z_{GT} = \{z_{GT}^{(1)}, \ldots, z_{GT}^{(N)}\}. \tag{1}$$

Bridge matching defines a continuous path between source and target:

$$Z_t = (1-t)Z_D + tZ_{GT} + \sigma\sqrt{t(1-t)}\epsilon, \quad t \in [0,1], \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

σ controls the stochasticity along the path. The corresponding velocity field from

$$v^*(Z_t, t) = Z_{GT} - Z_D. \tag{3}$$

We train a network v_θ to predict this velocity:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t, Z_D, Z_{GT}} \left[\| v_\theta(Z_t, t) - (Z_{GT} - Z_D) \|_2^2 \right], \tag{4}$$

Unlike marginal approaches, v_θ operates on the full multi-view latent tensor by continuously sampling different paired viewpoints, $z_D^{(1)}$ and $z_D^{(2)}$, at each iteration. The predicted vector field is therefore conditioned jointly on all views, defining a probability flow over the product latent space. This naturally induces coupled dynamics: corrections for each view depend on the entire set of views.

Cross-View Attention. To implement joint conditioning, we modify the transformer self-attention layers to allow tokens from all views within a batch to attend to one another, learning jointly. Concretely, latent tokens from different views are concatenated along the token dimension, and attention is computed globally while preserving within-view positional encodings. This design is

permutation-invariant across views and does not assume fixed view ordering. As a result, the predicted velocity for each view depends explicitly on features from other views, enforcing synchronization during refinement. We also add a reference view, which is the closest training view if available, as in [41]. The latent code of reference view z_{GT}^{ref} is concatenated with each of $\{z_D^{(1)}, z_D^{(2)}\}$ and $\{z_{GT}^{(1)}, z_{GT}^{(2)}\}$ to make z_D and z_{GT} respectively before building the stochastic interpolant z_t. In multi-view bridge model, the reference view's tokens are self-attended along with other views' tokens to provide contextual information and visible details.

Inference. At inference, we initialize Z_T with Z_D, solve the learned differential equation of Z_t, and obtain the predicted latent Z_{GT}^* in a single step through:

$$\frac{dZ_t}{dt} = v_\theta(Z_t, t), \quad Z_{GT}^* = Z_t + v_\theta(Z_t, t) \tag{5}$$

Decoding Z_{GT}^* yields refined images $\hat{X}$. Although trained with $N = 2$ views for efficiency, the model generalizes to arbitrary N degraded and reference images at test time due to the permutation-invariant attention mechanism.

3.3 Supervision

To isolate the impact of joint latent modeling, we adopt a similar image-space supervision used in state-of-the-art marginal refinement methods (e.g., Difix3D+).

Predicted clean images $\hat{X}$ are supervised against X_{GT} using following losses:

$$\mathcal{L}_{\text{pixel}} = \|\hat{X} - X_{GT}\|_1, \tag{6}$$

$$\mathcal{L}_{\text{LPIPS}} = \|\phi(\hat{X}) - \phi(X_{GT})\|_2^2, \tag{7}$$

$$\mathcal{L}_{\text{Gram}} = \|G(\phi(\hat{X})) - G(\phi(X_{GT}))\|_F^2. \tag{8}$$

The total loss to train our model is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{pixel}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{Gram}}\mathcal{L}_{\text{Gram}} \tag{9}$$

Unlike marginal baselines, these losses are applied to jointly coupled outputs, ensuring that structural and textural corrections remain consistent across views. We empirically find that the image-space supervisions serve as stronger training cues than the latent-only supervision adopted by FlowR [6], yielding improved performance and faster convergence.

4 Experiments

4.1 Datasets

We build our training set from 360 scenes of the DL3DV dataset [16], covering diverse indoor and outdoor environments. For each scene, we render paired *degraded* and *clean* images using 3D Gaussian Splatting (3DGS) [12], resulting

in 480,000 paired samples in total. An additional *disjoint* set of 40 scenes is reserved for testing, comprising 41,442 paired samples.

We apply a view-splitting protocol designed to prevent trivial overlap between training and rendering views. Specifically, training views are sampled across different splits to encourage broad coverage, while render views are sampled to be mutually closer within each split yet spatially separated from the training views, yielding a challenging novel-view setting. During training, we dynamically assemble multi-view inputs/targets by sampling from a size-bounded cache for efficient view combinations.

To evaluate zero-shot generalization, we report results on Nerfbusters [38], which contains 12 scenes with off-trajectory render views. We adopt their official train/test view split protocol for a fair comparison without visibility masks.

4.2 Evaluation and Implementation

We compare SyncFix with Difix3D+ [41], which introduces the iterative refinement by adding refined views to 3D reconstruction and post-cleanup with a diffusion model. We also compare with FlowR [6], which employs a multi-view flow-matching formulation for reconstruction densification. Since a pre-trained checkpoint is not available, we adopt their training code to train on the same data, computation budget, and evaluation setting as SyncFix. Fixer [26] is a follow-up work of Difix3D+ using Cosmos-Predict2 [1] as its backbone, with the same training strategy but without a reference view. We evaluate image fidelity using pixel-based metrics (PSNR and SSIM) and perceptual metrics (LPIPS, DreamSim [7], and FID). DreamSim is an ensemble of perceptual embedding models fine-tuned to better align with human judgments of visual similarity, and recent study [40] suggest that it provides a robust signal for assessing image quality in 3D reconstruction and novel view synthesis.

Our method is built on the SDXL [27]. We fine-tune the modified multi-view diffusion U-Net while keeping the VAE encoder and decoder frozen. We train our model for 100,000 optimization steps with a learning rate of 4×10^{-5} and a batch size of 2, using four NVIDIA H100 GPUs. We set $\lambda_{\text{lpips}} = 10$ and $\lambda_{\text{Gram}} = 0.1$.

4.3 Cross-View Semantic Consistency Metric

Maintaining across-view consistency is a central objective in 3D reconstruction and multi-view synthesis. Standard reference-based metrics such as PSNR, LPIPS, and DreamSim evaluate image quality at the single-view level and do not capture agreement between views. To quantify multi-view consistency, we introduce a *Cross-View Semantic Consistency* (CVSC) metric that measures semantic feature agreement between matched regions across views. Unlike MEt3R [2], which relies on dense correspondences that can be less reliable under scene dynamics or challenging textures, CVSC emphasizes sparse yet high-confidence correspondences obtained via a keypoint matcher and geometric verification.

Formally, given a series of refined views $X^*_{GT} = \{x^{(1)*}_{GT}, \ldots, x^{(N)*}_{GT}\}$, we form pairs of adjacent views $(x^{(i)*}_{GT}, x^{(i+1)*}_{GT})$, yielding $N-1$ view pairs. For each pair,

Table 1: Quantitative comparison on the DL3DV and NeRFBusters test sets for refining sparse-view 3D Gaussian Splatting reconstruction renderings. SyncFix improves reconstruction quality and cross-view semantic consistency over prior generative refinement methods. † and light gray rows indicate no reference views. ↑ indicates higher-is-better and ↓ indicates lower-is-better. **Bold** denotes best results and underline denotes second best. Runtime is measured for refining one scene with 63 images in seconds.

	DL3DV					NeRFBusters					
Method	PSNR↑	LPIPS↓	DSIM↓	FID↓	CVSC↑	PSNR↑	LPIPS↓	DSIM↓	FID↓	CVSC↑	Runtime↓
3DGS [12]	15.94	0.454	0.297	80.8	0.875	11.85	0.575	0.336	139.4	**0.951**	-
Fixer [26]	15.73	0.466	0.257	45.5	0.822	12.32	0.588	0.296	126.7	0.817	15.8
Difix3D+† [41]	16.05	0.368	0.171	26.2	0.819	11.94	0.528	0.202	97.4	0.925	30.6
Difix3D+ [41]	16.17	0.343	0.135	**16.7**	0.850	12.05	0.496	0.174	81.1	0.931	30.6
FlowR [6]	16.32	0.369	0.138	19.7	0.832	13.09	0.560	0.231	86.0	0.916	109.8
SyncFix†	16.45	0.334	0.129	22.5	0.862	14.07	0.517	0.183	86.2	0.937	**5.63**
SyncFix	**16.94**	**0.305**	**0.099**	17.5	**0.880**	**14.34**	**0.494**	**0.173**	**80.4**	0.940	**5.63**

we extract keypoints using RaCo [30] and establish correspondences with LightGlue [15]. This produces a set of matched keypoints $K_{\text{pair}} = \{(k_A^{(i)}, k_B^{(i)})\}_{i=1}^{M}$.

To suppress mismatches, we perform geometric verification by estimating the fundamental matrix with RANSAC and retain only inlier correspondences $\hat{K}_{\text{pair}} = \{(\hat{k}_A^{(i)}, \hat{k}_B^{(i)})\}$. We then extract dense semantic feature maps F_A and F_B from DINOv3 [32]. For each verified correspondence, we compute cosine similarity between features at the matched locations. The CVSC score for a pair of views is defined as

$$\text{CVSC} = \frac{1}{|\hat{K}|} \sum_{i=1}^{|\hat{K}|} \frac{\langle F_A(\hat{k}_A^{(i)}), F_B(\hat{k}_B^{(i)}) \rangle}{\|F_A(\hat{k}_A^{(i)})\|_2 \, \|F_B(\hat{k}_B^{(i)})\|_2}. \tag{10}$$

The final CVSC score is obtained by averaging across all view pairs in the sequence. Higher CVSC indicates stronger semantic agreement across viewpoints and thus better multi-view consistency.

4.4 DL3DV Dataset

Table 1 (first block) shows the quantitative metrics on DL3DV test set. Although Fixer improves the renderings with a lower DreamSim score and FID compared to 3DGS, it obtains suboptimal results compared with rest of methods. We attribute this gap primarily to its single-view nature and the absence of reference images, which limit Fixer's ability to disambiguate missing content and artifacts.

SyncFix outperforms Difix3D+ by 0.77 dB in PSNR and reduces LPIPS by 0.038. Notably, SyncFix lowers DreamSim by 27 percent compared to Difix3D+ and reduces FID by more than 4× relative to raw 3DGS renderings. SyncFix also consistently surpasses FlowR across these metrics. Importantly, SyncFix remains superior to both Difix3D+ and FlowR even without reference images, validating the effectiveness of our LBM formulation, pixel-level supervision, and multi-view coupling. Furthermore, owing to its multi-view single-step inference, SyncFix reduces inference time by approximately 5× and 20× compared to Difix3D+ and FlowR, respectively.

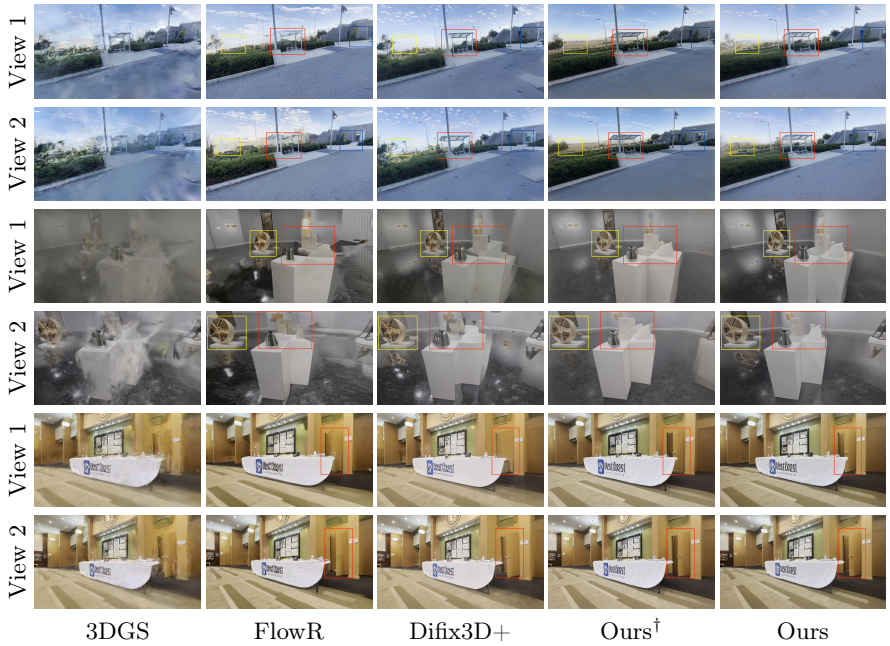
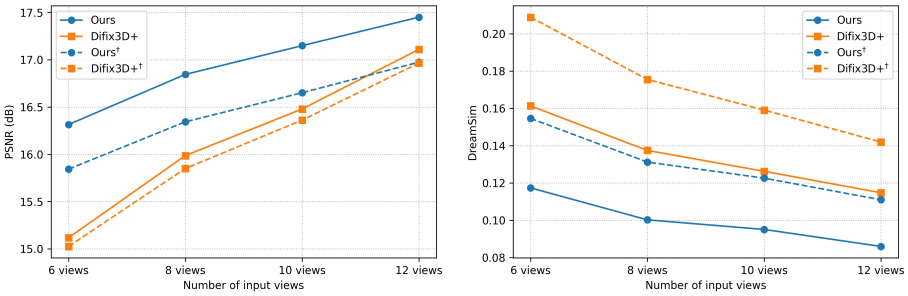


Fig. 3: Multi-view consistency under sparse-view degradation. The baseline 3DGS model (left) displays severe structural artifacts and floaters. While the marginal refinement method Difix3D+ (middle) removes these base artifacts, its independently generated views hallucinate inconsistent geometry across camera poses. As highlighted by the red boxes, Difix3D+ distorts the architectural structure of the outdoor bus shelter in the top scene. In the middle scene, it causes the object on the indoor pedestal to undergo shape changes between View 1 and View 2. Similarly, in the bottom scene, a door visible in one view is absorbed into the wall in another. In contrast, SyncFix (Ours) leverages joint latent synchronization to ensure that hallucinated high-frequency details and structural repairs remain geometrically and semantically consistent across views. Ours† is a model trained without clean reference images.

As seen in Figure 3, Difix3D+ mitigates the noises and sharpens the renderings. However, Difix3D+ struggles in multi-view settings, producing inconsistent object geometry and view-dependent artifacts. While FlowR incorporates multi-view refinement, it fails to preserve fine-grained details consistently across views. In contrast, our method jointly refines multiple renderings, globally reducing artifacts while maintaining more 3D-consistent scene geometry. This demonstrates that our multi-view bridge matching formulation more faithfully models the joint distribution of the scene. Interestingly, the raw 3DGS baseline achieves a relatively high CVSC score. This occurs because artifacts produced by sparse-view reconstruction, such as floaters and blurred textures, are often geometrically consistent across views. As a result, although the reconstruction quality is poor, the errors appear similarly from different viewpoints and therefore yield high semantic agreement.



(a) PSNR vs. number of views. **(b)** DreamSim vs. number of views.

Fig. 4: Performance at different number of 3DGS training views on the DL3DV dataset. SyncFix consistently achieves higher PSNR and lower DreamSim than Difix3D+ at different numbers of views, and the benefit of SyncFix is more pronounced at sparser settings.

In contrast, marginal refinement methods such as Difix3D+ improve perceptual quality but can introduce view-specific hallucinations, reducing cross-view consistency and lowering the CVSC score. SyncFix addresses this issue by jointly refining multiple views, preserving geometric agreement while improving perceptual fidelity, which leads to a better CVSC score overall. More visual comparison can be found in Sec. **??**.

To break down performance under different numbers of 3DGS training views, we report PSNR and DreamSim as a function of the number of views in Fig. 4. SyncFix and its no-reference variant consistently outperform those of Difix3D+ at different view counts. We observe that the performance gap is more prominent in the sparse-view regime: our model provides stronger generative capability to recover missing content under severe corruption, whereas Difix3D+ performs comparatively better when artifacts are milder, and denoising becomes easier.

4.5 Nerfbusters Dataset

We evaluate the generalization ability of our method on the out-of-distribution NeRFbusters dataset, which provides off-trajectory testing views from dense-view reconstructions and exhibits different camera aspect ratios than our training data. Based on Table 1 (second block on the right), SyncFix achieves notably higher PSNR and CVSC than Difix3D+, and stronger perceptual metrics than FlowR. As shown in Figure 3, our method synthesizes a coherent scene from the two input views, faithfully completing the missing door frame and walls. In contrast, Difix3D+ often produces refined renderings with cross-view inconsistencies, while FlowR leaves residual artifacts despite improving individual views. SyncFix suppresses artifacts globally through joint refinement, yielding clean and consistent renderings. This result shows that denoising at a fixed step [41] may not be sufficient to remove artifacts spanning different severities and provides limited extrapolation capacity. In contrast, SyncFix learns a direct mapping from the corrupted rendering distribution to the clean distribution, producing reliable generation guided by learned multi-view consistency.

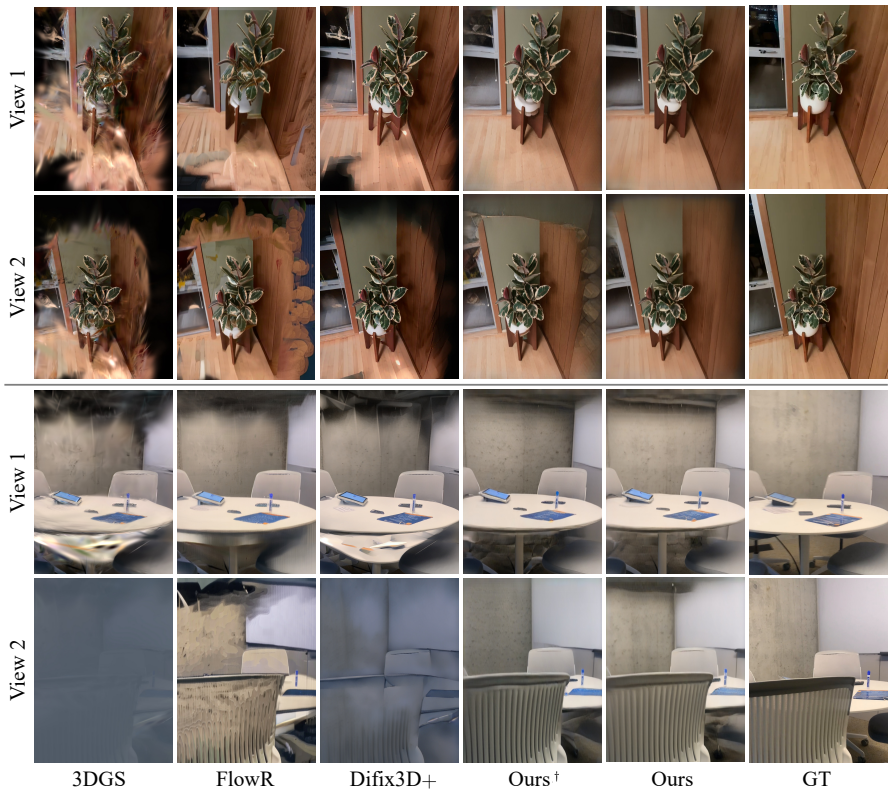


Fig. 5: Comparison on the NeRFBusters evaluation dataset. 3DGS reconstructions contain severe artifacts and geometric distortions. Difix3D+ improves the appearance of individual renderings but struggles to extrapolate to unseen regions and often produces unstable structures across views. In contrast, SyncFix generates plausible reconstructions that remain consistent across viewpoints, recovering coherent object geometry and scene structure. † are models trained without clean reference images.

4.6 Generalization to Feedforward Method

To test whether SyncFix can generalize to feedforward models, we conduct experiments using AnySplat [11] on the Mip-NeRF 360 dataset [3], which includes real-world indoor and outdoor scenes with complex details. AnySplat supports feedforward rendering from uncalibrated images by employing separate decoder heads to predict 3D Gaussian parameters, depth, and camera poses. For each scene, we randomly sample 10 input images, and the target views are selected sequentially while being far from the input views. We follow the evaluation protocol [11] that we use VGGT [36] to calibrate the input and target views in two passes. For pose alignment, we normalize camera poses by setting the first camera to the identity and estimate a relative scale factor to align camera translations across the two sets.

Table 2: Quantitative comparison on the MipNeRF 360 dataset for refining feedforward Gaussian Splatting renderings. SyncFix improves perceptual quality and cross-view semantic consistency over prior methods. We gray out the PSNR as pixel shifts make the pixel-aligned metric unreliable. ↑ indicates higher-is-better and ↓ indicates lower-is-better. **Bold** denotes best results and underline denotes second best.

	MipNeRF 360			
Method	PSNR↑	LPIPS↓	DSIM↓	CVSC↑
AnySplat [11]	13.58	0.454	0.232	0.649
Difix3D+ [41]	13.41	0.448	0.196	0.589
SyncFix (ours)	13.51	**0.430**	**0.189**	**0.686**

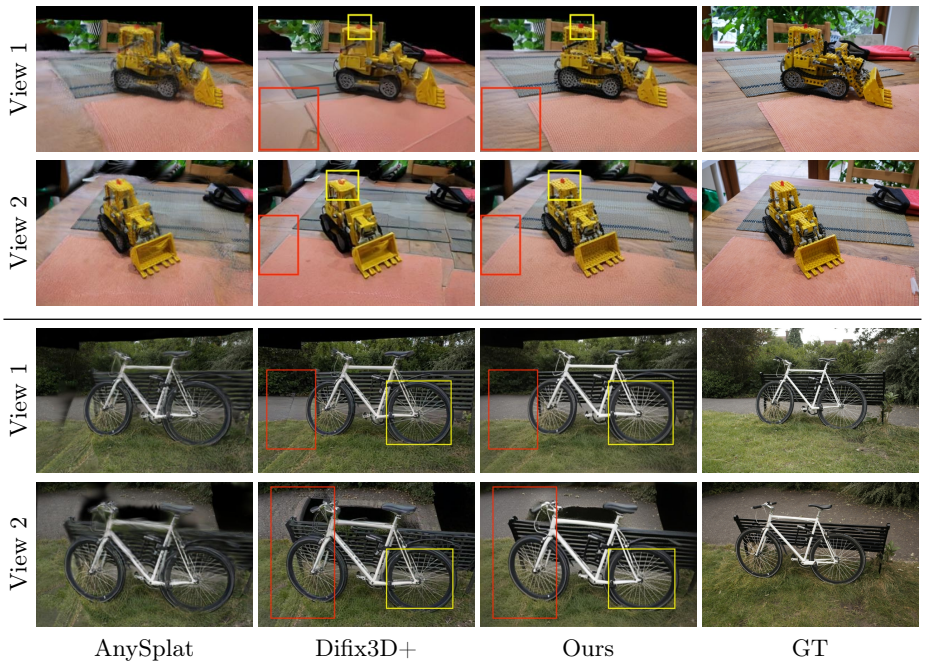


Fig. 6: Comparison of refinement with feedforward renderings. SyncFix provides coherent multi-view refinement that reduces artifacts and maintains consistency. Notice the regions highlighted in the boxes.

Table 2 reports quantitative results on AnySplat renderings, SyncFix consistently improves perceptual quality over Difix3D+, achieving lower LPIPS and DSIM. More importantly, SyncFix achieves a significantly higher CVSC score than both Difix3D+ and the raw AnySplat renderings, indicating stronger cross-view consistency. The qualitative comparisons in Fig. 6 support these findings. Difix3D+ shows view-dependent inconsistencies, e.g., around the table mat and the red cap in the Lego scene. A similar pattern is observed in the outdoor bicycle scene: Difix3D+ introduces non-uniform residual artifact patterns around the road and grass regions, where SyncFix has consistent refinements across views.

Fig. 7: Ablation on multi-view synchronization. We compare SyncFix with and without joint multi-view refinement using the same latent-bridge matching training protocol and without any reference images. Processing each view independently (Ours, single-view) improves individual renderings but produces inconsistent scene structure across viewpoints. In particular, the green carpet on the floor and the drawing board on the right wall disappear or are rendered inconsistently across the two views (highlighted in red). Because SyncFix (multi-view) jointly refines the views, it produces coherent geometry and consistent scene layout across viewpoints.

4.7 Ablation Study

We analyze the contribution of components through a series of ablations on the DL3DV test set (Table 3). While the train-in-loop strategy in Difix3D brings moderate improvement, the CVSC score is still suboptimal to our multi-view synchronization and it requires access of testing-view poses during training. In addition, we augment Difix3D+ with multiple input views during inference by providing the same number of views as SyncFix. This allows us to evaluate whether supplying additional views is sufficient to improve consistency. The last two rows of Difix3D+ in Table 3 show that performance drops, indicating that the model remains fundamentally limited by its single-view training objective and struggles in fusing multi-view information.

We then evaluate variants of SyncFix. A single-view version of our model already improves image quality compared to Difix3D+, demonstrating the benefit of the latent bridge formulation. Introducing joint multi-view refinement further improves performance across all metrics. In particular, cross-view semantic consistency (CVSC) increases from 0.821 to 0.862, confirming that joint training encourages the model to agree on scene structure across viewpoints. Qualitative examples in Fig. 7 illustrate this effect: the single-view model produces inconsistent hallucinations between views, whereas the multi-view model maintains spatial and semantic coherence. We also find that by supplementing reference views, the textures and details are more aligned with ground truth in the generated content by SyncFix. We note that the multi-view inference setting is five degraded views plus five closest training images as reference views for Difix3D+ and SyncFix. In our supplementary materials, we show that the multi-view consistency improves when adding more views during the refinement.

Table 3: Ablation study evaluating the effect of joint multi-view refinement. We also ablate with and without reference views on the DL3DV test set. Simply applying a multi-view input to Difix3D+ provides limited benefit, while our joint training substantially improves both image quality and cross-view consistency. ↑ indicates higher-is-better and ↓ indicates lower-is-better.

Method	PSNR↑	LPIPS↓	DSSIM↓	FID↓	CVSC↑
3DGS [12]	15.94	0.454	0.297	80.8	0.875
Difix3D [41]	16.14	0.445	0.285	76.6	0.853
Difix3D+ [41]	16.17	0.343	0.135	**16.7**	0.850
Difix3D+ (multi-view inference)	16.08	0.370	0.163	29.28	0.827
Ours (single-view, w/o reference)	16.26	0.347	0.142	22.2	0.821
Ours (multi-view, w/o reference)	16.45	0.334	0.129	22.5	0.862
Ours (multi-view, with reference)	**16.94**	**0.305**	**0.099**	17.5	**0.880**

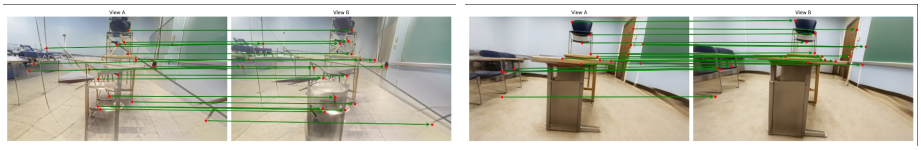
5 Discussion

Our experiments suggest that refining views independently with a 2D generative prior is fundamentally misaligned with the multi-view nature of 3D reconstruction. Marginal refinement can improve individual images, but the resulting hallucinations often differ across viewpoints. The reconstruction process must then harmonize these incompatible predictions, which frequently leads to drift and unstable geometry. SyncFix addresses this by coupling views during generation, forcing the model to agree on structure before the images are produced.

An interesting observation is that the multi-view consistency achieved by SyncFix arises purely from cross-view attention. The model does not receive explicit geometric supervision such as camera poses, epipolar constraints, or reprojection losses. Instead, synchronization emerges from learned correlations between views within a shared latent space. This suggests that generative models contain sufficient structure to align views through attention alone, although incorporating explicit geometric constraints may further improve robustness.

The results also highlight limitations of current evaluation protocols. Standard image metrics such as PSNR, LPIPS, and DSSIM assume that a single ground-truth image represents the only correct solution. However, sparse-view 3D reconstruction is fundamentally under-constrained: large portions of a scene may be poorly observed or entirely unseen in the training views. In such cases, generative refinement methods must infer plausible geometry and textures consistent with the visible context. SyncFix can produce coherent and consistent structures even when the predicted appearance deviates from the ground-truth rendering. Pixel-wise metrics penalize these plausible reconstructions because they measure strict image correspondence rather than semantic or geometric consistency. As Figure 8 shows, consistent refined renderings receive higher patch feature correlation at matched keypoints. Developing evaluation protocols that better capture multi-view agreement and perceptual plausibility remains an important direction for future work.

Finally, although SyncFix is trained using view pairs, the architecture naturally generalizes to larger sets of views at inference. Additional viewpoints pro-



(a) Matches of Difix3D+ refined renderings **(b)** Matches of SyncFix refined renderings

Fig. 8: Cross-view geometric consistency. Keypoint correspondences between two refined views. Difix3D+ produces inconsistent refinements, resulting in irregular correspondences across views. SyncFix preserves the underlying scene geometry, producing coherent cross-view matches. For visualization clarity, we display 20 sampled correspondences. Matches computed using RaCo + LightGlue.

vide stronger constraints and typically improve stability. This suggests that pairwise training is sufficient to learn consistent multi-view refinement dynamics.

Limitations. SyncFix assumes that the input views share sufficient overlap to establish meaningful correspondences during joint refinement. When viewpoints are extremely sparse or observe largely disjoint regions of the scene, the synchronization signal becomes weak and the model may revert to view-dependent hallucinations. The approach also relies on the quality of the underlying reconstruction; if the geometry is severely corrupted or missing large structures, the generative prior may produce plausible but incorrect completions. Finally, the generative prior may introduce biases toward visually plausible textures and structures that are consistent with the training distribution but not necessarily faithful to the true scene. Developing generative refinement models that integrate explicit geometric reasoning and more principled evaluation protocols remains an important direction for future work.

6 Conclusion

We presented SyncFix, a framework for refining sparse-view 3D reconstructions using a joint multi-view generative prior. The method formulates refinement as latent bridge matching between distorted and clean renderings and synchronizes the latent trajectories of multiple views through cross-view attention. This design allows the model to enforce multi-view agreement during generation rather than relying on post-hoc reconciliation by the 3D reconstruction. Experiments demonstrate improved perceptual quality and stronger cross-view consistency across several reconstruction settings. More broadly, the results suggest that generative priors for visual scenes should operate on joint multi-view distributions rather than independent view marginals.

Acknowledgment. We thank anonymous reviewers for their feedback. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 140D0423C0076. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6034–6044 (2025)
3. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
4. Chadebec, C., Tasar, O., Sreetharan, S., Aubin, B.: Lbm: Latent bridge matching for fast image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 29086–29098 (2025)
5. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12882–12891 (2022)
6. Fischer, T., Bulò, S.R., Yang, Y.H., Keetha, N., Porzi, L., Müller, N., Schwarz, K., Luiten, J., Pollefeys, M., Kontschieder, P.: Flowr: Flowing from sparse to dense 3d reconstructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27702–27712 (2025)
7. 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)
8. 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)
9. Gupta, V., Lin, C.H., Wang, S., Bhattad, A., Huang, J.B.: Generalizable sparse-view 3d reconstruction from unconstrained images. CVPR (2026)
10. Haque, A., Tancik, M., Efros, A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
11. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
13. Li, D., Jiang, K., Tang, Y., Ramamoorthi, R., Chellappa, R., Peng, C.: Ms-gs: Multi-appearance sparse-view 3d gaussian splatting in the wild. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
14. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
15. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17627–17638 (2023)

16. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
18. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: Reconx: Reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024)
19. Liu, X., Zhou, C., Huang, S.: 3DGS-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=P4s6FUpCbG
20. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z
21. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=MN3yH2ovHb
22. Luo, Y., Zhou, S., Lan, Y., Pan, X., Loy, C.C.: 3denhancer: Consistent multi-view diffusion for 3d enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16430–16440 (2025)
23. 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), https://openreview.net/forum?id=aBsCjcPu_tE
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
25. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2022)
26. NVIDIA: Fixer: Official repository for the nvidia fixer model. https://github.com/nv-tlabs/Fixer (2025), gitHub repository. Accessed: 2026-03-12
27. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
28. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
29. Serrano-Lozano, D., Bhattad, A., Herranz, L., Lalonde, J.F., Vazquez-Corral, J.: Synclight: Controllable and consistent multi-view relighting. arXiv preprint arXiv:2601.16981 (2026)
30. Shenoi, A., Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Raco: Ranking and covariance for practical learned keypoints. arXiv preprint arXiv:2602.15755 (2026)
31. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3d generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=FUgrjq2pbB

32. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
33. Tang, Y., Guo, Y., Li, D., Peng, C.: Spars3r: Semantic prior alignment and regularization for sparse 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26810–26821 (2025)
34. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
35. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12619–12629 (2023)
36. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
37. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
38. Warburg, F., Weber, E., Tancik, M., Holynski, A., Kanazawa, A.: Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18120–18130 (2023)
39. Warburg*, F., Weber*, E., Tancik, M., Hołyński, A., Kanazawa, A.: Nerfbusters: Removing ghostly artifacts from casually captured nerfs (2023)
40. Wickrema, C., Leary, S., Sarkar, S., Giglio, M., Bianchi, E., Mace, E., Twardowski, M.: Benchmarking image similarity metrics for novel view synthesis applications. arXiv preprint arXiv:2506.12563 (2025)
41. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26024–26035 (2025)
42. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)
43. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16453–16463 (2025)
44. Yang, J., Pavone, M., Wang, Y.:
45. Ye, B., Chen, B., Xu, H., Barath, D., Pollefeys, M.: Yonosplat: You only need one model for feedforward 3d gaussian splatting. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=ImRhA9xmay
46. Yin, X., Zhang, Q., Chang, J., Feng, Y., Fan, Q., Yang, X., Pun, C.M., Zhang, H., Cun, X.: Gsfixer: Improving 3d gaussian splatting with reference-guided video diffusion priors. arXiv preprint arXiv:2508.09667 (2025)
47. 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. IEEE transactions on pattern analysis and machine intelligence (2024)

48. Zhou, H., Shao, Z., Miao, S., Wang, P., Bai, D., Liu, B., Liao, Y.: Freefix: Boosting 3d gaussian splatting via fine-tuning-free diffusion models. arXiv preprint arXiv:2601.20857 (2026)