DIFSPLAT: EXTRAPOLATIVE FEED-FORWARD 3D GAUSSIAN SPLATTING WITH DIFFUSION FEATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D Gaussian Splatting (3DGS) models enable efficient novel view synthesis but struggle with extreme view extrapolation. Their image-based features (IBFT) rely on source observations and become unavailable in unseen regions, leading to geometric holes and blurry artifacts. We propose DiFSplat, an extrapolative feed-forward 3DGS framework that combines deterministic IBFT with generative diffusion features (DiFT) from a pre-trained 3D-aware diffusion model. We project the initial geometry into novel views and initialize uncovered depths through nearest-neighbor dilation. A frozen diffusion U-Net, conditioned on point renderings, provides features for an auxiliary depth predictor that refines these depths by matching across source and novel views in DiFT space. The refined depths are unprojected to instantiate auxiliary voxels as anchors for unseen geometry. We then fuse IBFT and DiFT on the unified voxel set to predict Gaussian parameters. Experiments show improved extrapolation and zero-shot cross-dataset generalization while maintaining high-fidelity interpolation, demonstrating the benefit of using diffusion priors for both geometry completion and feature fusion.

1 INTRODUCTION

Novel view synthesis (NVS) from sparse observations is a fundamental task in 3D computer vision. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) achieves high-quality real-time rendering, but its per-scene optimization is expensive. Feed-forward models such as pixelSplat (Charatan et al., 2024), MVSplat (Chen et al., 2024), and VolSplat (Wang et al., 2025c) address this cost by predicting Gaussian attributes for unseen scenes in a single forward pass. These methods use multi-view matching and feature aggregation to extract **image-based features (IBFT)** from co-visible source regions.

Despite their success in interpolation, these models degrade under large view extrapolation. Target views can expose regions occluded in or outside the source images, where no valid observations are available for feature sampling. Insufficient geometry and features in these areas lead to enlarged, textureless Gaussians and stretched or blurred renderings (Fig. 1).

We address this limitation with **DiFSplat**, which complements deterministic IBFT with **diffusion features (DiFT)** from a pre-trained 3D-aware diffusion model such as See3D (Ma et al., 2025). IBFT provides observation-grounded reconstruction in visible regions, while DiFT supplies generative guidance for geometry completion and appearance in unseen regions.

Starting from a voxel-aligned reconstruction (Wang et al., 2025c), we extract DiFT from a frozen diffusion U-Net conditioned on point renderings of source and auxiliary views. The occupied voxel centers are projected onto target novel views, producing sparse depth maps and uncovered masks. For uncovered pixels, nearest-neighbor depth dilation provides a coarse depth initialization. A DiFT-guided auxiliary depth predictor then refines these depths by matching among the source and auxiliary DiFT views, so the depth evidence is not limited to source-image visibility. Unprojecting the refined depths activates previously empty voxels. We sample DiFT on both original and auxiliary voxels and fuse it with IBFT to regress the final Gaussian attributes.

This design separates two requirements of extrapolation: placing geometry where no source observations exist and providing features that can describe its appearance. Adding generative features

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

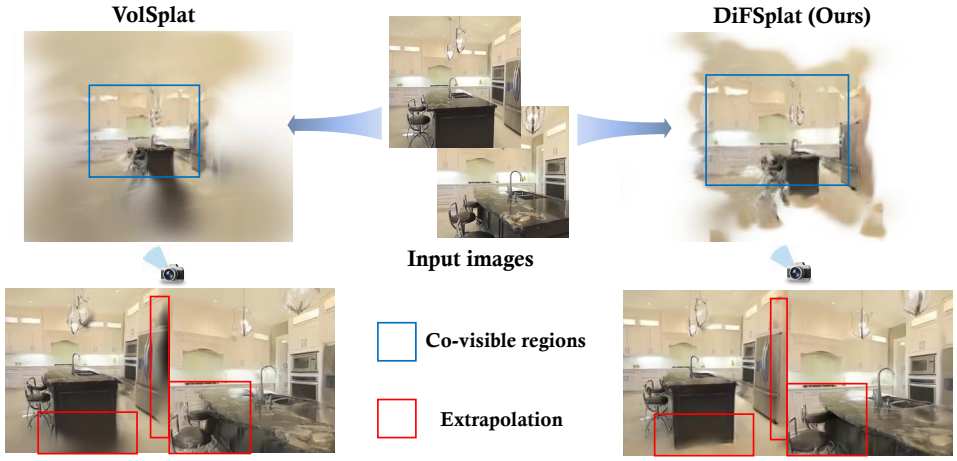


Figure 1: **Overcoming the extrapolation bottleneck in feed-forward 3DGS.** Given the input views (middle), VolSplat (left) produces enlarged Gaussians and blurred renderings in unseen regions. DiFSplat (right) adds compact geometric anchors through DiFT-guided auxiliary depth prediction and fuses diffusion features with IBFT, improving the Gaussian geometry (top) and rendered appearance (bottom).

only to the original occupied voxels leaves the spatial support incomplete; adding auxiliary geometry without suitable features leaves its appearance underconstrained. The completed voxel representation allows the two sources of information to be combined while retaining observation-grounded features in visible regions.

Our contributions are threefold:

- We address missing image-based features under view extrapolation through a voxel-aligned framework combining deterministic IBFT and generative DiFT.
- We introduce DiFT-guided auxiliary depth prediction, refining dilation-based initialization to instantiate geometric anchors in unobserved regions.
- We fuse DiFT with IBFT for Gaussian prediction and demonstrate improved extrapolation and zero-shot generalization while preserving interpolation quality.

2 Related Work

2.1 Feed-Forward and Generalizable 3DGS

Feed-forward 3DGS methods replace per-scene optimization with generalizable Gaussian prediction (Charatan et al., 2024; Chen et al., 2024; Xu et al., 2025; Wang et al., 2025c;b; 2024b; Zhang et al., 2025a; Jiang et al., 2025; Ye et al., 2025; Park et al., 2025). MVSplat (Chen et al., 2024), DepthSplat (Xu et al., 2025), and VolSplat (Wang et al., 2025c) use multi-view image features and cost volumes to reconstruct co-visible geometry. Their reliance on source-image evidence limits extrapolation into unseen regions. In particular, multi-view matching requires observations that can be compared across cameras. Regions outside the source fields of view or hidden by occlusion lack this support, even when reconstruction within the observed region is accurate. PM-Loss (Shi et al., 2025) uses pointmap regularization from VGGT (Wang et al., 2025a) to improve geometric coherence, but such regularization does not provide generative appearance cues for unobserved areas. Our method combines explicit geometry completion with diffusion features to address both missing geometry and appearance.

2.2 Integration of Reconstruction and Generative Priors

Reconstructing unseen regions from sparse observations is underconstrained: multiple scene completions can explain the same source images. Generative priors provide plausible structural and

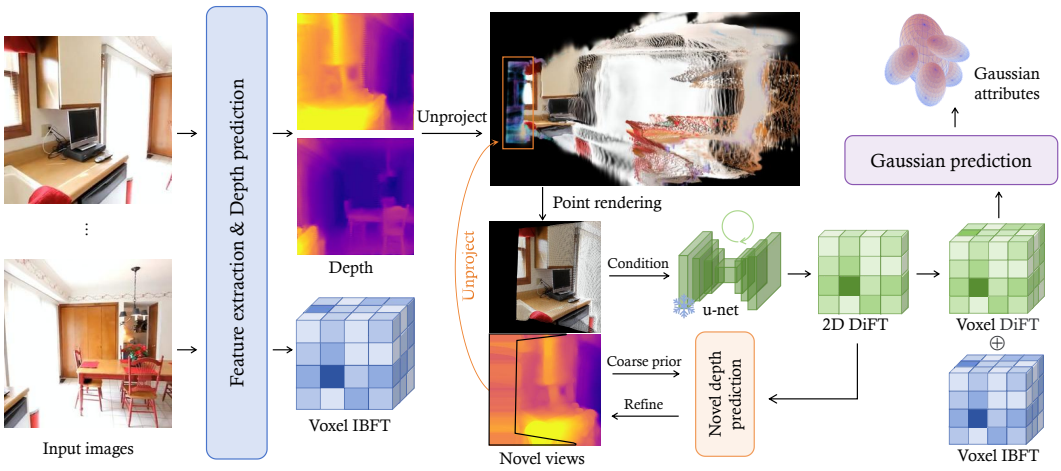


Figure 2: **Overview of DiFSplat.** A voxel-aligned backbone extracts IBFT and initial geometry from input views. A frozen diffusion U-Net extracts DiFT from point renderings of source and auxiliary views. For uncovered pixels, a DiFT-based cost volume refines dilation-initialized depths; unprojection activates auxiliary voxels. DiFT sampled on the unified voxel set is concatenated ($\oplus$) with IBFT and refined to predict Gaussian attributes ($\Delta\mu, s, r, \alpha, c$).

appearance cues for these regions, complementing the geometric constraints imposed by the observed views. SceneSplatter (Zhang et al., 2025b) and DiFix3D (Wu et al., 2025) use diffusion priors to repair or refine 3D representations. Related approaches apply these priors to street-view extrapolation and novel driving trajectories (Ni et al., 2025; Yan et al., 2025; Yang et al., 2025; Wang et al., 2025d). Although effective, hybrid generation-reconstruction pipelines can incur additional inference cost compared with deterministic feed-forward reconstruction.

A closely related generative feed-forward baseline, latentSplat (Wewer et al., 2024), predicts variational feature Gaussians from two views and decodes rendered latent features with a lightweight 2D generative decoder. DiFSplat instead retains explicit Gaussians in a voxel-aligned pipeline and injects external diffusion features into geometry completion and Gaussian prediction. This combines source-image evidence with generative guidance in 3D rather than decoding appearance solely from latent Gaussians.

2.3 Diffusion Features

Intermediate diffusion activations encode semantic and structural information useful beyond image generation. Tang et al. (2023) demonstrated emergent visual correspondence from these features. Subsequent applications include semantic correspondence (Zhang et al., 2023; Meng et al., 2024; Fundel et al., 2025), AI-image detection (Zhong et al., 2025), sketch extraction (Yun et al., 2024), and video understanding (Yuan et al., 2025; Wang et al., 2024a). More recently, Huang et al. (2025) decoded video diffusion features into continuous 3D spaces, showing their geometric potential. We use DiFT within an explicit feed-forward 3DGS framework to guide auxiliary geometry and fuse generative cues with image-based voxel features.

3 Methodology

Given a sparse set of source images $\{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{N}$ and corresponding camera poses $\{R_i \in \mathbb{R}^{3\times 3}, T_i \in \mathbb{R}^3, K_i \in \mathbb{R}^{3\times 3}\}_{i=1}^{N}$, our objective is to synthesize high-fidelity novel views, especially under extrapolative camera poses where conventional methods severely degrade. To achieve this, we propose an extrapolative feed-forward 3D Gaussian Splatting framework, termed **DiFSplat**. As illustrated in Fig. 2, our approach bridges deterministic observation and generative completion. In following sections, we will first establish the basis using image-based features (IBFT), then extract diffusion features (DiFT) from a frozen generative prior. With these features defined, we introduce

an extrapolation-aware geometry completion module for unseen regions, and finally unify IBFT and DiFT for Gaussian regression.

3.1 PRELIMINARIES & IBFT EXTRACTION

3D Gaussian Splatting. 3DGS (Kerbl et al., 2023) represents a continuous 3D scene using a collection of anisotropic 3D Gaussians. Each Gaussian is parameterized by its center position $\mu \in \mathbb{R}^3$, scale $s \in \mathbb{R}^3$, rotation quaternion $r \in \mathbb{R}^4$, opacity $\alpha \in [0, 1]$, and view-dependent color c represented by spherical harmonics (SH) coefficients. For a target pixel, the color C is rendered via α-blending of sorted Gaussians:

$$C = \sum_i c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \tag{1}$$

Image-based feature (IBFT) extraction. We build our framework upon a voxel-aligned feed-forward 3DGS model (*e.g.*, VolSplat (Wang et al., 2025c)). Given input images $\{I_i \in \mathbb{R}^{H \times W \times 3}\}_{i=1}^N$, the network extracts two complementary types of 2D features. First, a CNN backbone (He et al., 2016) coupled with a Transformer module (Liu et al., 2021b) is employed to extract multi-view features, exchanging information across views via cross-attention. Concurrently, a ViT backbone (Oquab et al., 2023) extracts monocular features, providing rich single-view semantic and structural priors.

Crucially, instead of directly predicting Gaussian parameters from the 2D feature maps, the extracted features are unprojected into a 3D voxel grid. The model first unprojects the predicted depth maps to form a dense 3D point cloud, where each point carries its corresponding 2D image feature. To construct a voxel representation, the space is uniformly discretized, and each 3D point is assigned a discrete voxel index. For each occupied voxel, the initial 3D center and feature are computed as the average of all points and features falling within it.

A sparse 3D CNN is then applied to these occupied voxels to aggregate spatial context. We formally define this refined, voxel-aligned feature grid as the voxel IBFT, denoted as $\mathbf{F}_{\text{IBFT}}$. Consequently, we obtain a set of initial 3D points $P_{\text{init}} \in \mathbb{R}^{N_{\text{init}} \times 3}$ corresponding to the occupied voxel centers, where each point $p_i \in P_{\text{init}}$ is coupled with a feature $f^{(i)}_{\text{IBFT}}$.

Extrapolation bottleneck. While the construction of P_{init} and $\mathbf{F}_{\text{IBFT}}$ captures co-visible regions through voxel-aligned matching, this mechanism faces a catastrophic bottleneck during view extrapolation. Regions that are entirely unobserved by the source views fail to be projected and assigned to any voxel index. Consequently, these empty voxels lack both geometric points and meaningful IBFT, inevitably leading to rendering holes and severely stretched artifacts. This fundamental limitation motivates our geometry and semantic completion modules.

3.2 SEMANTIC COMPLETION VIA DIFT EXTRACTION

To prepare generative cues for later geometry completion and semantic fusion, we leverage the powerful priors from a pre-trained 3D-aware diffusion model (*i.e.*, See3D (Ma et al., 2025)). We define the intermediate U-Net activations of the diffusion model as diffusion features (DiFT).

Specifically, we project the current initial geometry to the N_s source and N_a auxiliary camera poses to obtain coarse point renderings $I_j, 1 \leq j \leq N_s + N_a$, and use them as the condition for the U-Net. By performing a fixed number of denoising steps t_{step}, we extract the intermediate spatial feature maps for all rendered views:

$$\mathbf{F}^{\text{2D}}_{\text{DiFT},j} = U\left(h\left(\text{UNet}(I_j, t_{\text{step}})\right)\right) \in \mathbb{R}^{C' \times H \times W}, \tag{2}$$

where the operation h extracts multi-scale intermediate U-Net features, and the network U denotes an upsampler. Unlike IBFT, which relies strictly on input images, DiFT contains rich semantic priors capable of synthesizing visually plausible missing contents. We use these features in two complementary ways. First, the 2D DiFT features from all rendered views are used for auxiliary depth prediction, which will be unprojected to 3D space to activate empty voxels for geometry completion. Second, after geometry completion produces the unified point set P_{total}, we project P_{total} onto the rendered views to sample the 2D DiFT, yielding the 3D voxel $\mathbf{F}_{\text{DiFT}}$ for all occupied voxels.

3.3 EXTRAPOLATION-AWARE GEOMETRY COMPLETION

To tackle the geometric collapse in unobserved regions, we must explicitly instantiate physical anchors before predicting the final Gaussian attributes. Given the target novel view V_t, we project the initial occupied voxel centers P_{init} onto the 2D image plane. This yields an initial RGB rendering I_t, a sparse depth map D_{init}, and a binary uncovered mask $U \in \{0,1\}^{H\times W}$, where $U(p) = 1$ indicates missing regions that are not covered by projected geometry.

We first construct a coarse depth initialization D_{coarse} for uncovered pixels by nearest-neighbor dilation from valid projected depths. This initialization provides a stable geometric anchor, but it is only a local continuity prior and may copy depths across object boundaries or large disocclusions. We therefore refine it with a lightweight DiFT-guided auxiliary depth predictor:

$$D_{\text{aux}} = G_{\text{depth}}\left(\{\mathbf{F}^{2D}_{\text{DiFT},j}, K_j, R_j, T_j\}_{j=1}^{N_s+N_a}, D_{\text{coarse}}, U\right). \tag{3}$$

Here G_{depth} builds a feature-level cost volume entirely in the DiFT space: for each target view, all other DiFT views act as support views over inverse-depth candidates. This avoids relying on deterministic IBFT support in regions that source images do not cover. A coarse-depth prior softly anchors the prediction around D_{coarse} to stabilize training. We apply the refined prediction only inside the uncovered mask. Thus, DiFT does not alter the deterministic source-view depth prediction; it only guides mask-restricted auxiliary geometry where deterministic observations are absent. We provide more details in the supplementary material.

Rather than treating the unprojected pixels simply as an unstructured point cloud, we strictly maintain the voxel-aligned representation. The learned auxiliary depths in previously empty regions are unprojected into 3D and assigned to the same voxel grid. For each newly occupied voxel, its position is also computed as the mean of all raw points falling within it. We define these newly activated voxel centers as auxiliary points P_{aux}, and the unified geometry becomes $P_{\text{total}} = P_{\text{init}} \cup P_{\text{aux}}$. Since these voxels originate from unobserved areas, they lack valid IBFT and are initialized with zero base features, making DiFT essential for subsequent feature fusion and Gaussian prediction.

3.4 DUAL-FEATURE FUSION AND GAUSSIAN REGRESSION

The core of our framework lies in the fusion of the deterministic IBFT and the generative DiFT within the unified voxel space. To construct a fused feature, we use a sparse 3D CNN network R to refine the concatenation of both features:

$$\mathbf{F}_{\text{fused}} = R\left(\mathbf{F}_{\text{IBFT}} \oplus \mathbf{F}_{\text{DiFT}}\right). \tag{4}$$

This dual-feature design forces the network to naturally trust the deterministic IBFT in visible areas, while adaptively relying on the generative DiFT to dominate the structural and textural synthesis in the unobserved auxiliary voxels.

Finally, the fused feature $\mathbf{F}_{\text{fused}}$ is fed into a Gaussian head network to regress the residual position offset $\Delta\mu_i$ and the full set of Gaussian attributes for each voxel i:

$$\Delta\mu_i, s_i, r_i, \alpha_i, c_i = \text{GH}\left(\mathbf{F}_{\text{fused}}\right). \tag{5}$$

The final Gaussian position is updated as $\mu_i = p_i + \Delta\mu_i$.

Training Objective. The entire feed-forward framework is trained end-to-end. Following standard practices, we supervise the rendered images against the ground truth using a combination of $\mathcal{L}_1$ pixel loss and an LPIPS perceptual loss to ensure both geometric accuracy and perceptual fidelity.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. To comprehensively evaluate our proposed DiFSplat, particularly under view extrapolation and zero-shot cross-dataset generalization, we employ two widely used large-scale datasets:

- **RealEstate10K (RE10K) (Zhou et al., 2018):** A large-scale indoor video dataset encompassing highly diverse camera trajectories and complex scene layouts. To explicitly validate our extrapolation capabilities, we follow Shi et al. (2025) to sample target novel views

with large spatial shifting from the source input views: given source frames I_{t_1} and I_{t_2}, extrapolative targets are sampled from $[t_1 - 30, t_1] \cup [t_2, t_2 + 30]$.

- **ACID (Liu et al., 2021a):** A large-scale dataset featuring aerial captures of nature scenes. Since its domain (outdoor landscapes) is radically different from RE10K (indoor rooms), we utilize the ACID test set strictly for zero-shot cross-dataset evaluation, demonstrating the robust generalization capability of our architecture.

Competitors. We compare DiFSplat with DepthSplat (Xu et al., 2025), a pixel-aligned pipeline; VolSplat (Wang et al., 2025c), a voxel-aligned method; PM-Loss (Shi et al., 2025), which adds pointmap regularization; and the generative baseline latentSplat (Wewer et al., 2024). We also construct **VolSplat+See3D**: RGB images of novel views generated by See3D are concatenated with the source views and passed to VolSplat. This baseline compares image-level generation with our feature-level use of the same prior.

Evaluation metrics. Following standard novel view synthesis protocols, we report the quantitative performance using four metrics: Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). While PSNR and SSIM evaluate pixel-wise and local structural fidelity, FID and LPIPS are particularly crucial in our extrapolation setting, as they better correlate with distribution-level realism and perceptual plausibility in unobserved regions.

Implementation details. The IBFT extraction module follows the VolSplat (Wang et al., 2025c) architecture, taking $N_s = 2$ source views as input. During the DiFT extraction phase, we sample $N_a = 4$ additional views for point rendering. For the semantic completion module, we utilize the pre-trained See3D (Ma et al., 2025) U-Net with frozen weights. At a fixed denoising timestep, the intermediate U-Net activations are extracted to form DiFT for both rendered source and auxiliary views.

To ensure stable optimization and prevent feature collapse between the IBFT and DiFT branches, we adopt a three-stage training curriculum:

- Stage 1 (Base pre-training): We first train IBFT extraction network for 200K iterations to establish a foundation.
- Stage 2 (Interpolation warm-up): We then introduce DiFT extraction and dual-feature network. The unified network is trained for 70K iterations using standard interpolation view sampling. This allows the network to smoothly adapt to the DiFT feature space in co-visible regions.
- Stage 3 (Extrapolative fine-tuning): Finally, we explicitly shift the training distribution to extrapolation view sampling for another 100K iterations. This stage trains the auxiliary depth predictor and fusion network to use DiFT in genuinely unobserved regions.

The entire model is optimized using the Adam optimizer on 8 NVIDIA A6000 GPUs, with a decaying learning rate initialized at 2×10^{-4} and a batch size of 1 for each GPU. More implementation details and network architectures are provided in the supplementary material.

4.2 Comparison with state-of-the-art

We comprehensively compare our DiFSplat against current state-of-the-art feed-forward 3DGS models. The quantitative results are reported in Tab. 1 and Tab. 2, with qualitative visual comparisons provided in Fig. 3 and Fig. 4.

In-domain evaluation on RealEstate10K. Table 1 shows that DiFSplat improves all four metrics under both interpolation and extrapolation. In interpolation, the dual-feature representation complements multi-view matching in regions affected by occlusion. Under extrapolation, deterministic baselines degrade as source observations become unavailable. PM-Loss improves boundary geometry, but DiFSplat additionally supplies generative cues for unseen appearance, improving perceptual quality as measured by LPIPS and FID. Compared with VolSplat under extrapolation, DiFSplat increases PSNR from 23.17 to 23.83 and reduces LPIPS from 0.191 to 0.176. The improvement

Table 1: **Quantitative comparison on the RealEstate10K dataset.** We evaluate novel view synthesis performance under both interpolation and extrapolation settings. Best results are in **bold**, and second best are underlined.

Method	Interpolation				Extrapolation			
	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LatentSplat	10.67	23.97	0.825	0.164	12.68	22.01	0.794	0.202
DepthSplat	12.33	24.44	0.838	0.165	13.90	20.43	0.788	0.218
PM-Loss	10.07	25.24	0.851	0.152	12.14	22.48	0.814	0.194
VolSplat	11.11	25.54	0.856	0.147	12.58	23.17	0.815	0.191
VolSplat+See3D	25.08	20.70	0.722	0.259	36.45	18.76	0.677	0.321
DiFSplat (Ours)	**9.70**	**26.36**	**0.866**	**0.135**	**10.51**	**23.83**	**0.826**	**0.176**

Table 2: **Zero-shot cross-dataset evaluation on the ACID dataset.** Models are trained exclusively on RealEstate10K and tested directly on ACID to evaluate generalization capability across completely different domains.

Method	Interpolation				Extrapolation			
	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LatentSplat	14.73	27.80	0.856	0.144	12.89	26.43	0.817	0.163
DepthSplat	15.40	28.27	0.876	0.139	14.24	26.32	0.843	0.165
PM-Loss	13.38	29.18	0.885	0.125	11.49	27.54	0.849	0.147
VolSplat	14.62	29.16	0.884	0.137	14.65	27.76	0.859	0.160
VolSplat+See3D	30.50	24.09	0.776	0.245	27.44	23.46	0.765	0.252
DiFSplat (Ours)	**12.70**	**30.03**	**0.891**	**0.121**	**11.32**	**28.72**	**0.868**	**0.140**

in interpolation alongside extrapolation indicates that the additional generative guidance does not require sacrificing reconstruction quality in the observed domain.

The VolSplat+See3D baseline performs worse across all metrics despite using generated images of the target views. Generated RGB images can exhibit color and illumination shifts or pose misalignment. Treating them as ordinary observations introduces inconsistent evidence into VolSplat's matching and cost-volume construction, corrupting geometry and voxel features. DiFSplat instead injects intermediate DiFT into the unified voxel representation, allowing the fusion network to combine generative cues with source-grounded IBFT rather than relying directly on generated pixels.

Zero-shot generalization on ACID. Models trained on RE10K are evaluated on ACID without fine-tuning (Tab. 2). This indoor-to-outdoor shift includes weakly textured regions such as sky and water that challenge multi-view matching. DiFSplat achieves the best scores among the compared methods across all reported metrics, supporting the generalization of auxiliary geometry completion and diffusion-feature fusion. Under extrapolation, PSNR increases from 27.76 for VolSplat to 28.72 for DiFSplat, while LPIPS decreases from 0.160 to 0.140. These results assess transfer beyond the indoor training distribution rather than adaptation to ACID-specific training data.

Qualitative analysis. We examine both the RGB renderings (Fig. 3) and the underlying Gaussian geometry (Fig. 4). In the RGB comparisons, existing methods produce stretched and blurred artifacts in regions beyond the source views. DiFSplat uses the injected DiFT to recover sharper, visually plausible textures while preserving the transition between co-visible and extrapolated areas.

The Gaussian-point visualizations help explain these differences. Without explicit completion, VolSplat attempts to cover missing regions by increasing the scale of Gaussians near the observed boundary. The resulting large, textureless ellipsoids smear across the novel view rather than providing compact geometry for the newly visible content. DiFSplat instead predicts auxiliary depths for uncovered pixels and instantiates auxiliary voxels in those regions. These spatial anchors support more compact Gaussians, yielding more coherent boundaries and sharper renderings.

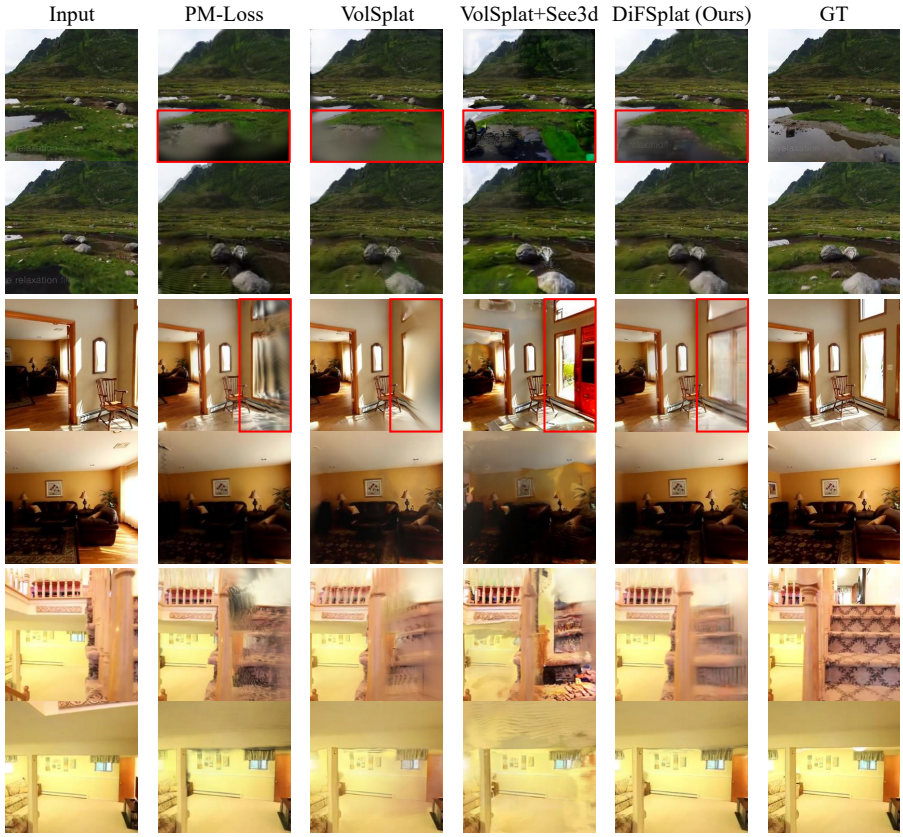


Figure 3: **Qualitative comparisons of RGB renderings under view extrapolation.** While the existing methods struggle with severe blurring and stretched artifacts in unobserved regions (highlighted in red boxes), our DiFSplat successfully synthesizes structurally sound and visually plausible high-frequency textures. **More visual results can be found in supplementary materials.**

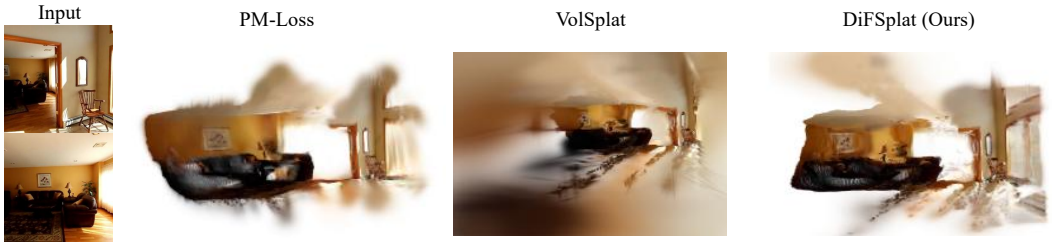


Figure 4: **Visualizations of the underlying Gaussian points in extrapolative setups.** To forcefully fill the unobserved space without explicit geometric anchors, existing methods erroneously inflate boundary Gaussians into abnormally large, textureless ellipsoids. In contrast, DiFSplat constructs dense, compact auxiliary points to physically anchor the extrapolative regions, strictly preventing geometric distortion.

4.3 Ablation Studies

To rigorously validate the contribution of each proposed component in DiFSplat, we conduct extensive ablation studies on the RealEstate10K dataset under the extrapolation setting. The quantitative metrics and qualitative visual comparisons are presented in Tab. 3 and Fig. 5, respectively.

Effect of geometry completion. We first evaluate the necessity of instantiating auxiliary points. In the *"w/o Aux."* variant, we completely remove the auxiliary depth completion module, extracting DiFT but only injecting it into the initial co-visible points (P_{init}). As shown in Tab. 3, this leads to

Table 3: **Ablation studies on the RealEstate10K dataset under extrapolation setting.** We systematically isolate the contribution of each proposed component.

Model	Aux.	DiFT	Prior	FID↓	PSNR↑	SSIM↑	LPIPS↓
w/o Aux.	✗	✓	N/A	12.05	23.06	0.811	0.191
w/o DiFT	✓	✗	✓	11.52	23.35	0.819	0.185
w/o Prior	✓	✓	✗	11.90	23.27	0.816	0.188
Full	✓	✓	✓	**10.51**	**23.83**	**0.826**	**0.176**



Figure 5: **Qualitative ablation results.** Visual comparison of different model variants under extrapolation. Only our full DiFSplat strictly preserves both geometric alignment and high-frequency textural plausibility in unobserved regions.

a significant performance drop across all metrics. Visually (Fig. 5), this variant also suffers from inflated Gaussians across the unobserved regions, severely distorting the semantic priors provided by DiFT. This proves that powerful generative priors cannot synthesize high-frequency structures without a spatially distributed geometric foundation.

Effect of semantic completion. To prove that geometry alone is insufficient for extrapolation, we construct the *"w/o DiFT"* variant. Here, auxiliary voxel activation still creates the missing geometric anchors, but no diffusion features are extracted for semantic completion. The network must rely solely on the deterministic IBFT which are mostly zeros in unseen areas. This variant also suffers from degradation in all metrics. Qualitatively, the extrapolated regions manifest as unnaturally smooth, severely blurred, and over-stretched textures.

Effect of coarse-depth prior. Finally, we ablate the coarse-depth prior used in the auxiliary depth predictor. In the *"w/o Prior"* variant, the predictor estimates the inverse-depth distribution only from the learned logits, without the soft anchor around the coarse initialization. This causes a moderate degradation, indicating that the prior stabilizes auxiliary depth prediction.

5 Conclusion

We presented **DiFSplat**, a feed-forward 3D Gaussian Splatting framework for view extrapolation. DiFT-guided auxiliary depth prediction refines coarse dilation-based depths and activates voxels in unseen regions. Fusing diffusion features with image-based voxel features then supports Gaussian prediction beyond source-image coverage. By restricting auxiliary depth completion to uncovered regions, the framework retains source-supported geometry while providing spatial anchors where deterministic matching is unavailable. The unified voxel representation then lets the fusion network combine these anchors with generative appearance cues, rather than asking enlarged Gaussians at the visible boundary to explain the missing scene. Experiments demonstrate improved extrapolation and zero-shot cross-dataset generalization while preserving interpolation quality. These results support combining observation-grounded reconstruction with generative features for geometry and appearance completion.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and with the implementation of the auxiliary depth prediction module.

REFERENCES

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, 2024.

Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of diffusion features for semantic correspondence. In *Winter Conference on Applications of Computer Vision*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M Rehg. How much 3d do video foundation models encode? *arXiv preprint*, 2025.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM TOG*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021b.

Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *CVPR*, 2025.

Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. In *NeurIPS*, 2024.

Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *CVPR*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint*, 2023.

Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Ecosplat: Efficiency-controllable feed-forward 3d gaussian splatting from multi-view images. *arXiv preprint*, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

Duochao Shi, Weijie Wang, Donny Y Chen, Zeyu Zhang, Jia-Wang Bian, Bohan Zhuang, and Chunhua Shen. Revisiting depth representations for feed-forward 3d gaussian splatting. *arXiv preprint*, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Qian Wang, Abdelrahman Eldesokey, Mohit Mendiratta, Fangneng Zhan, Adam Kortylewski, Christian Theobalt, and Peter Wonka. Zero-shot video semantic segmentation based on pre-trained diffusion models. *arXiv preprint*, 2024a.

Weijie Wang, Donny Y Chen, Zeyu Zhang, Duochao Shi, Akide Liu, and Bohan Zhuang. Zpressor: Bottleneck-aware compression for scalable feed-forward 3dgs. *arXiv preprint*, 2025b.

Weijie Wang, Yeqing Chen, Zeyu Zhang, Hengyu Liu, Haoxiao Wang, Zhiyuan Feng, Wenkang Qin, Zheng Zhu, Donny Y Chen, and Bohan Zhuang. Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. *arXiv preprint*, 2025c.

YuAn Wang, Xiaofan Li, Chi Huang, Wenhao Zhang, Hao Li, Bosheng Wang, Xun Sun, and Jun Wang. Faithfusion: Harmonizing reconstruction and generation via pixel-wise information gain. *arXiv preprint*, 2025d.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. In *NeurIPS*, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *ECCV*, 2024.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *CVPR*, 2025.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025.

Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. Streetcrafter: Street view synthesis with controllable video diffusion models. In *CVPR*, 2025.

Zeyu Yang, Zijie Pan, Yuankun Yang, Xiatian Zhu, and Li Zhang. Driving view synthesis on free-form trajectories with generative prior. In *ICCV*, 2025.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. *arXiv preprint*, 2025.

Tianyu Yuan, Yuanbo Yang, Lin-Zhuo Chen, Yao Yao, and Zhuzhong Qian. Denoise to track: Harnessing video diffusion priors for robust correspondence. *arXiv preprint*, 2025.

Kwan Yun, Youngseo Kim, Kwanggyoon Seo, Chang Wook Seo, and Junyong Noh. Representative feature extraction during diffusion process for sketch extraction with one example. *arXiv preprint*, 2024.

Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In *AAAI*, 2025a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. In *CVPR*, 2025b.

Nan Zhong, Haoyu Chen, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Beyond generation: A diffusion-based low-level feature extractor for detecting ai-generated images. In *CVPR*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint*, 2018.

A EXTENDED IMPLEMENTATION DETAILS

Voxel resolution. In the original VolSplat (Wang et al., 2025c) architecture, the 3D voxel grid resolution is typically set to 0.01. However, evaluating extreme view extrapolation fundamentally requires processing significantly expanded 3D spatial bounds that extend far beyond the source views. To accommodate the enlarged scene coverage while maintaining strict GPU memory efficiency, we intentionally set the voxel grid resolution to 0.05.

DiFT extraction. To extract diffusion features, we tap into the upsampling phase of the frozen See3D (Ma et al., 2025) U-Net. Specifically, we perform iterative denoising conditioned on point renderings until the noise strength reaches 0.7, and extract the intermediate feature maps from upsampling layers 5, 6, and 7 at this denoising timestep. Since these multi-scale diffusion features operate at lower spatial resolutions, we employ a simplified DPT head (Ranftl et al., 2021) to aggregate and upsample them back to the native image resolution. This ensures a strict pixel-to-pixel spatial alignment before the 2D DiFT is used for auxiliary depth prediction and voxel feature sampling.

Auxiliary depth prediction. We describe the auxiliary depth predictor used by DiFSplat. For each auxiliary view V_t, projecting the initial point set P_{init} gives a sparse depth map D_{init} and an uncovered mask U, where $U(p) = 1$ means that pixel p is not covered by projected geometry and $U(p) = 0$ means that p is already covered. The predictor is only used for pixels with $U(p) = 1$; covered pixels keep their projected depth.

Coarse initialization. Nearest-neighbor dilation is used only to initialize the missing depths before learning. For each uncovered pixel p, we find the nearest covered pixel q^*:

$$q^* = \arg\min_{q \in \Omega} \|p - q\|_2, \quad \Omega = \{q \mid U(q) = 0\}, \tag{6}$$

$$D_{\text{coarse}}(p) = \begin{cases} D_{\text{init}}(p), & U(p) = 0, \\ D_{\text{init}}(q^*), & U(p) = 1. \end{cases} \tag{7}$$

This initialization is not treated as final geometry; it only provides a stable center for the subsequent depth distribution.

Feature-level plane sweep. The auxiliary depth predictor operates entirely on diffusion features. Let $\mathcal{D} = \{(\mathbf{F}^{2D}_{\text{DiFT},j}, K_j, R_j, T_j)\}_{j=1}^{N_s+N_a}$ denote the DiFT features and cameras of all rendered source and auxiliary views. For an auxiliary reference view t, its support set is $\mathcal{S}_t = \mathcal{D} \setminus \{t\}$, which includes the rendered source-view DiFT and the other auxiliary-view DiFT. We sample K inverse-depth candidates $\{z_k^{-1}\}_{k=1}^{K}$ within the target near-far range. For each target pixel p and candidate depth z_k, we back-project p with K_t, transform it from camera t to each support camera j, project it with K_j, and bilinearly sample the corresponding support DiFT. We denote this camera-aware warping operation by $W_{t\to j}(\mathbf{F}^{2D}_{\text{DiFT},j}, z_k, p)$.

For each support view, we compute a normalized dot-product matching score between the reference DiFT and the warped support DiFT. The cost volume is obtained by averaging valid support-view scores:

$$\mathbf{C}_k(p) = \frac{1}{|\mathcal{V}_{k,p}|} \sum_{j \in \mathcal{V}_{k,p}} \frac{\left\langle \phi\left(\mathbf{F}^{2D}_{\text{DiFT},t}(p)\right), \psi\left(W_{t\to j}\left(\mathbf{F}^{2D}_{\text{DiFT},j}, z_k, p\right)\right)\right\rangle}{\sqrt{C}}, \tag{8}$$

where $\mathcal{V}_{k,p} \subseteq \mathcal{S}_t$ is the set of support views where the candidate projects inside the image, ϕ and ψ are lightweight projection layers, and C is the projected feature dimension. Because the support features are also DiFT, the cost volume can use generative features in auxiliary views instead of relying only on source-image visibility.

Depth distribution and coarse-depth prior. A shallow 2D network A predicts logits over inverse-depth candidates from the cost volume, reference DiFT, coarse inverse depth, and uncovered mask. We add a coarse-depth prior to these logits:

$$\mathbf{P}(z_k^{-1}|p) = \text{softmax}\left(A\left(\mathbf{C}(p), \mathbf{F}^{2D}_{\text{DiFT},t}(p), D^{-1}_{\text{coarse}}(p), U(p)\right) + \pi_k(p)\right), \tag{9}$$

where the prior term is:

$$\pi_k(p) = -\frac{\left|z_k^{-1} - D^{-1}_{\text{coarse}}(p)\right|}{\Delta z^{-1}}, \tag{10}$$

and Δz^{-1} is the interval between adjacent inverse-depth candidates. This prior makes the initial prediction close to the dilation-based coarse depth, while the learned logits can override it when the plane-sweep evidence supports a different depth. In the ablation "w/o Prior", we remove $\pi_k(p)$ and use only the learned logits.

The refined auxiliary depth is computed by expectation in inverse-depth space:

$$D_{\text{aux}}(p) = \left(\sum_{k=1}^{K} \mathbf{P}(z_k^{-1}|p) z_k^{-1} \right)^{-1}. \tag{11}$$

Finally, the depth used for voxel activation is:

$$D_{\text{final}}(p) = \begin{cases} D_{\text{init}}(p), & U(p) = 0, \\ D_{\text{aux}}(p), & U(p) = 1. \end{cases} \tag{12}$$

Only the uncovered pixels are unprojected to create new auxiliary voxels, while the already covered geometry remains unchanged.

Voxel feature refinement and Gaussian head. After obtaining the base IBFT and the generative DiFT, the concatenated feature is first fed into a sparse 3D CNN, a 3-block 3D U-Net. This refinement network aggregates context from the spatial neighborhood and predicts a residual feature to produce the refined feature f_{fused}.

Finally, the refined voxel features are passed into the Gaussian prediction head, which is formulated as a 2-layer sparse 3D convolutional network. This head predicts the residual position offset $\Delta\mu \in \mathbb{R}^3$, scale $s \in \mathbb{R}^3$, rotation quaternion $r \in \mathbb{R}^4$, opacity $\alpha \in (0, 1)$, and spherical harmonics (SH) coefficients c.

Training hyperparameters. Our three-stage training curriculum is optimized using the Adam optimizer. The training schedule spans approximately three and a half weeks in total:

1. Stage 1 (3 days): The base IBFT extraction network is trained using the original VolSplat architecture. The learning rate is initialized at 2×10^{-4}.
2. Stage 2 ($\sim$ 1 week): After replacing the feature refinement and Gaussian head to process the concatenated dual-features, the learning rate is re-initialized to 2×10^{-5}. This interpolation warm-up prevents the newly initialized weights from corrupting the pre-trained IBFT backbone.
3. Stage 3 ($\sim$ 2 weeks): For the final extrapolative fine-tuning, the learning rate remains initialized at 2×10^{-5}.

Due to the long training time, we select related hyperparameters through short-run screening. Each candidate is trained for about 10% of the full schedule, and the setting with the best early convergence is used for full training.

B Additional evaluation details and diagnostics

Extrapolative view sampling criteria. Evaluating extrapolation requires a strict definition of unobserved regions. For the sequential video datasets (Zhou et al., 2018), let the input source frames be I_{t_1} and I_{t_2}, where the maximum frame interval $|t_1 - t_2| \leq 75$. During training, the target view index is sampled from $[t_1 - 30, t_2 + 30]$, with two endpoints $t_1 - 30$ and $t_2 + 30$ always selected. For extrapolation evaluation, target indices are uniformly spaced from $[t_1 - 30, t_1] \cup [t_2, t_2 + 30]$.

Masked metrics by invisible ratio. To verify that the gains come from genuinely extrapolated regions rather than from easier co-visible pixels, we stratify RE10K extrapolation samples by invisible ratio and additionally compute metrics only on invisible pixels. The invisible mask is obtained by projecting the source-view geometry to the target view: pixels not covered by any projected source point are treated as extrapolated regions. This protocol separates two effects that are otherwise mixed in full-image metrics. First, the invisible-ratio buckets measure how performance changes as the extrapolation difficulty increases. Second, the masked metrics with the "m-" prefix measure quality only where deterministic source-view IBFT is unavailable.

Table 4: **RE10K extrapolation results stratified by invisible ratio.** Metrics with the "m-" prefix are computed only on invisible pixels.

Method	Invis. ratio	FID↓	PSNR↑	SSIM↑	LPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓
PM-Loss	0–20%	10.9	22.69	0.807	0.174	18.02	0.527	0.333
VolSplat	0–20%	11.8	24.12	0.828	0.168	19.03	0.563	0.335
DiFSplat	0–20%	**10.1**	**24.76**	**0.840**	**0.154**	**19.69**	**0.576**	**0.310**
PM-Loss	20–40%	42.9	18.21	0.681	0.296	15.13	0.485	0.499
VolSplat	20–40%	40.9	18.96	0.714	0.286	15.90	0.545	0.489
DiFSplat	20–40%	**33.7**	**19.70**	**0.726**	**0.263**	**16.82**	**0.555**	**0.446**
PM-Loss	>40%	146.2	14.75	0.561	0.421	12.65	0.429	0.584
VolSplat	>40%	136.8	15.73	0.607	0.416	13.83	0.512	0.578
DiFSplat	>40%	**108.8**	**16.51**	**0.616**	**0.388**	**14.87**	**0.519**	**0.537**

The results are shown in Tab. 4. In the easy 0–20% bucket, DiFSplat improves both full-image and masked metrics, indicating that the DiFT branch does not degrade mostly co-visible views. In the 20–40% and >40% buckets, all methods degrade as the unseen area grows, but DiFSplat maintains a consistent advantage in FID, LPIPS, and masked metrics. This trend is important: if the improvement came only from co-visible regions, the masked scores would not improve and the gap would shrink as the invisible ratio grows. Instead, the masked PSNR/SSIM/LPIPS gains show that the auxiliary depth and DiFT fusion specifically improve the pixels where VolSplat and PM-Loss lack valid source observations.

Diagnostic on generative bias. The image-level baseline VolSplat+See3D exposes a failure mode of directly using generated RGB images: See3D may introduce color shifts, style changes, or pose misalignment, and these biased pixels are then treated by VolSplat as ordinary observations. Once they enter image-space matching and cost-volume construction, the bias can corrupt both depth estimation and voxel features.

DiFSplat uses the same generative prior in a different form and at different locations. We further test whether the fusion network blindly follows biased generative cues by randomly masking 20% of co-visible IBFT at inference time, forcing the model to rely more on DiFT. DiFSplat's PSNR drops only from 23.83 to 23.62, while VolSplat drops from 23.17 to 19.59 under the same perturbation. This suggests that the shared fusion and Gaussian heads learn to decode useful, IBFT-aligned structural cues from DiFT rather than copying See3D's image-level appearance bias.

C MORE VISUALIZATIONS

Additional per-scene RGB comparisons are presented in Fig. 6 and 7.

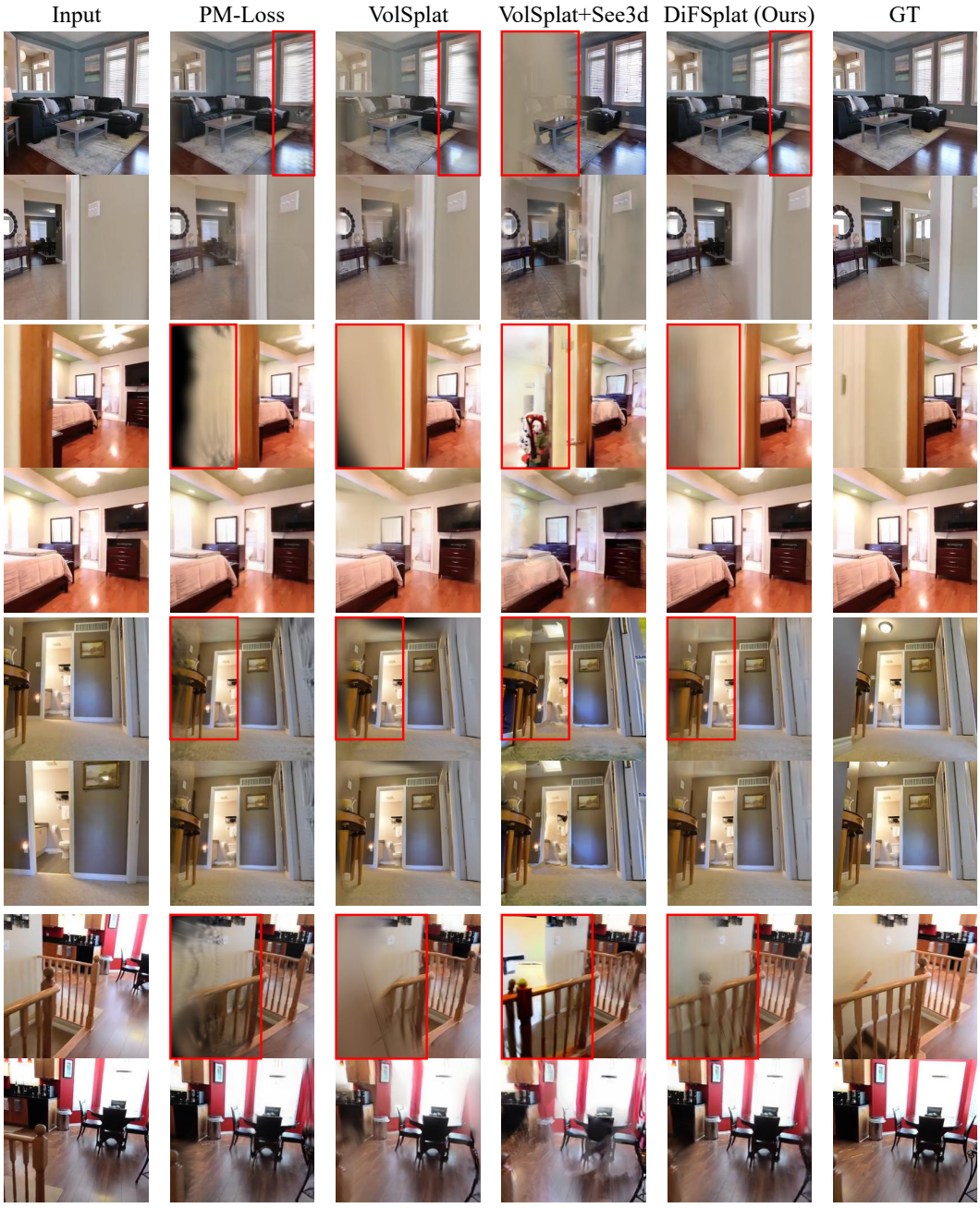


Figure 6: **More visual comparisons.** Extrapolation regions are highlighted in red boxes.

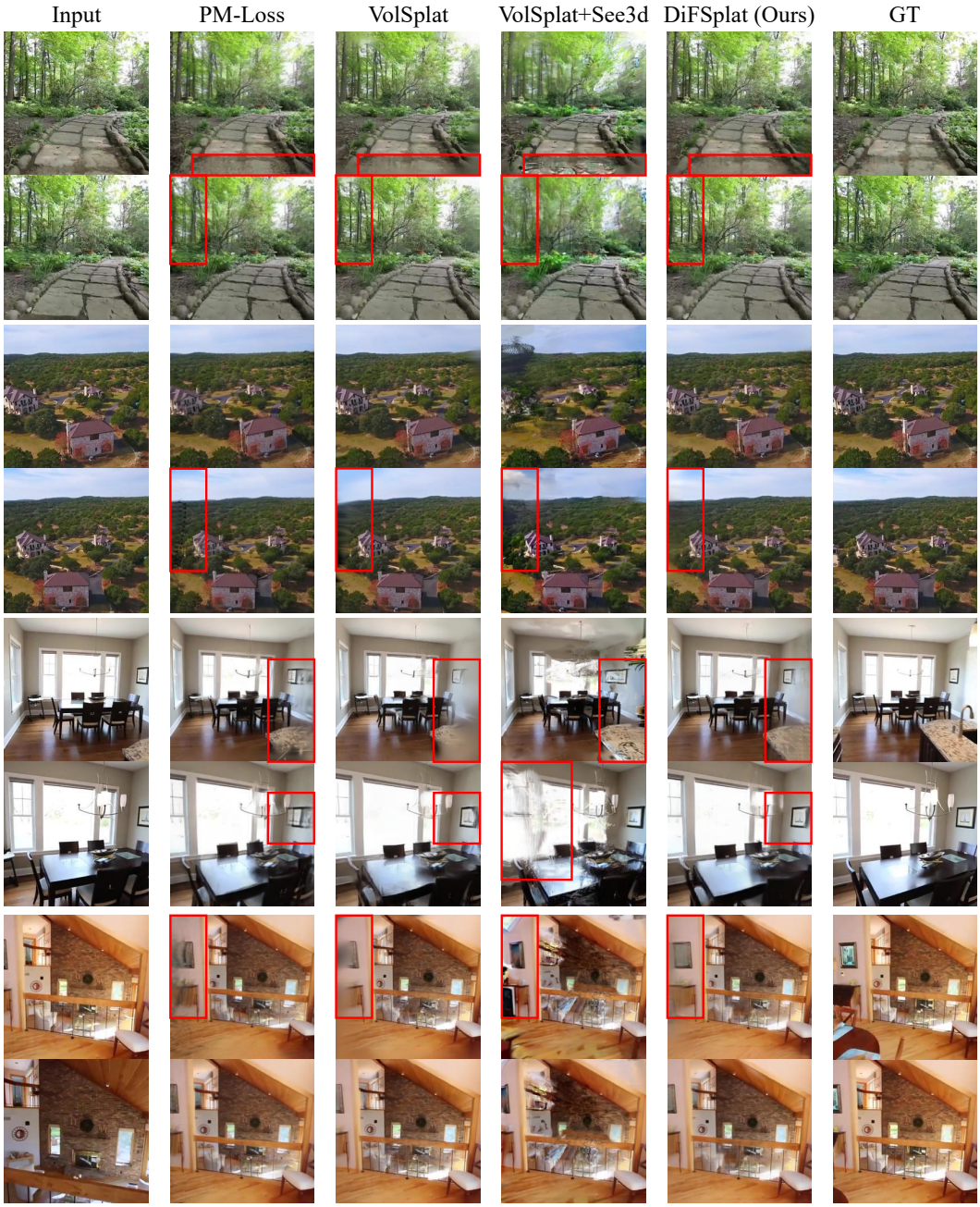


Figure 7: **More visual comparisons.** Extrapolation regions are highlighted in red boxes.

D INFERENCE EFFICIENCY AND LIMITATIONS

Inference time analysis. A critical evaluation metric for feed-forward novel view synthesis is inference efficiency. We benchmark the per-frame inference time of our method against the baselines on a single NVIDIA A6000 GPU. Purely deterministic reconstruction models achieve near real-time performance, with PM-Loss (Shi et al., 2025) taking ~ 0.08 seconds and VolSplat (Wang et al., 2025c) taking ~ 0.16 seconds per frame. However, as extensively demonstrated in the main manuscript, these ultra-fast methods catastrophically fail in extreme extrapolation scenarios, producing massive geometric artifacts. Therefore, integrating generative priors is not merely an option, but a strict necessity for structurally sound extrapolation.

When incorporating the See3D (Ma et al., 2025) generative prior, the naive image-level concatenation baseline (**VolSplat + See3D**) requires a full reverse diffusion trajectory (typically 50 denoising steps) to synthesize complete RGB images, resulting in a severe computational bottleneck of ~ 22 seconds per scene. In contrast, our **DiFSplat** fundamentally amortizes this generative cost. By injecting diffusion features directly into the 3D voxel grid, our method only requires a truncated denoising trajectory (*e.g.*, 16 steps to reach the specific intermediate noise strength) to extract the highly semantic DiFT. This architectural innovation drastically reduces the inference time to just ~ 8 seconds per scene, nearly a $3\times$ speedup over the image-level baseline. More importantly, this truncated feature extraction avoids the color shifts and geometric misalignments inherent in full RGB generation, simultaneously delivering significantly higher metric performance and visual fidelity. Thus, DiFSplat is not yet real-time, but it is considerably more efficient than full RGB generation and avoids its geometric misalignment artifacts.

Limitations and future work. Despite the massive efficiency gains over full-image diffusion baselines, an 8-second time still precludes strict real-time inference. The primary bottleneck remains the multi-step denoising through the See3D U-Net. However, our feature-level injection paradigm paves a highly promising path forward. Future work could seamlessly replace the multi-step diffusion backbone with emerging single-step point-conditioned generative models. By directly predicting the 3D-aware semantic features in a single forward pass, the inference time of DiFSplat could be accelerated to match deterministic MVS pipelines, achieving real-time 3D scene extrapolation.