FreeScale: Scaling 3D Scenes via Certainty-Aware Free-View Generation

Chenhan Jiang[1*†] Yu Chen[2*] Qingwen Zhang[3] Jifei Song[4]
Songcen Xu[5] Dit-Yan Yeung[1] Jiankang Deng[6†]

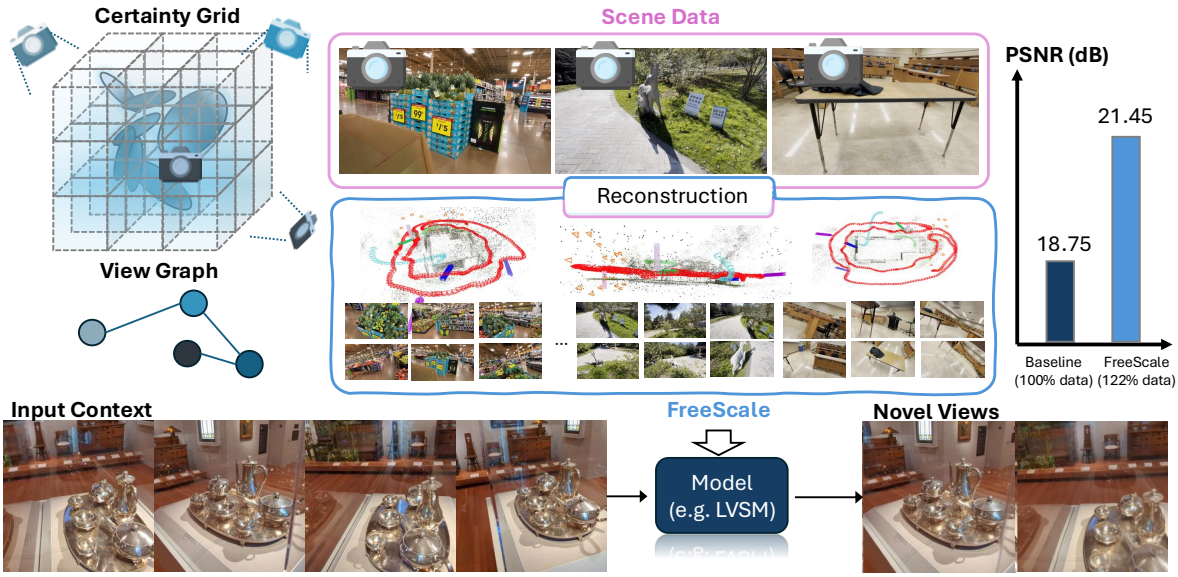

Figure 1. We introduce **FreeScale**, a framework that scales current scene data by generating free-view images from reconstructed scene geometry, which can be used for feed-forward model training. Training LVSM with an additional 22% of generated free-views significantly improves sparse-view reconstruction from PSNR 18.75 to 21.45, particularly enhancing its generalization to large camera motion.

Abstract

The development of generalizable Novel View Synthesis (NVS) models is critically limited by the scarcity of large-scale training data featuring diverse and precise camera trajectories. While real-world captures are photorealistic, they are typically sparse and discrete. Conversely, synthetic data scales but suffers from a domain gap and often lacks realistic semantics. We introduce FreeScale, a novel framework that leverages the power of scene reconstruction to transform limited real-world image sequences into a scalable source of high-quality training data. Our key insight is that an imperfect reconstructed scene serves as a rich geometric proxy, but naively sampling from it amplifies artifacts. To this end, we propose a certainty-aware free-view sampling strategy identifying novel viewpoints that are both semantically meaningful and minimally affected by reconstruction errors. We demonstrate FreeScale's effectiveness by scaling up the training of feedforward NVS models, achieving a notable gain of 2.7 dB in PSNR on challenging out-of-distribution benchmarks. Furthermore, we show that the generated data can actively enhance per-scene 3D Gaussian Splatting optimization, leading to consistent improvements across multiple datasets. Our work provides a practical and powerful data generation engine to overcome a fundamental bottleneck in 3D vision. Project page: https://mvp-ai-lab.github.io/FreeScale.

1. Introduction

Novel View Synthesis (NVS), which aims to generate photorealistic images from unobserved viewpoints given sparse input, is a foundational problem in computer vision and graphics. Optimization-based methods like Neural Radiance Fields (NeRF) [29] and the highly efficient 3D Gaussian Splatting (3DGS) [17] have achieved impressive visual fidelity through per-scene reconstruction. More recently, the field has seen a growing trend toward generalizable feedforward models [7, 12, 16] that learn cross-scene

*Equal contribution. [1]Hong Kong University of Science and Technology [2]National Univeristy of Singapore [3]KTH Royal Institute of Technology [4]University of Surrey [5]Independent Researcher [6]Imperial College London †Corresponding Author: jchcyan@gmail.com, j.deng16@imperial.ac.uk

priors for efficient 3D reconstruction at inference time.

Despite these advances, a critical data bottleneck persists: the scarcity of large-scale datasets with diverse and accurate camera trajectories, which limits the scalability and robustness of feedforward models, especially for large, free-viewpoint camera motions. Moreover, optimization-based methods remain sensitive to imperfect captures (e.g., insufficient scene coverage or inaccurate camera poses), which leads to noticeable geometric errors and artifacts. Furthermore, collecting large-scale, high-quality real-world data with meticulous camera calibration remains a laborious and expensive process.

Existing approaches to scale data have significant limitations. Methods leveraging simulators (e.g., Blender [11]) can generate synthetic data with perfect camera control [33, 34, 37, 43, 49], but they introduce a substantial synthetic-to-real domain gap. For instance, LRM-Zero [49] trained entirely on synthetic data shows significantly reduced visual fidelity compared to models trained on real data. While Megasynth [13] bypasses realistic semantics through amorphous geometry and texture stacking, it suffers from poor data efficiency. Alternatively, diffusion-based methods [32] can generate photorealistic images but typically fail to provide the accurate camera poses essential for NVS training.

This reveals a critical gap. Real-world captures provide the desired photorealism and semantics but exist only as discrete, sparsely sampled sequences. While one can reconstruct a continuous scene representation from them, naively sampling novel views from this reconstruction often yields images marred by artifacts or poor semantics. We propose FreeScale, as shown in Figure 1, a framework that *not only creates a continuous representation via reconstruction but, more importantly, introduces a principled method to sample high-fidelity, semantically meaningful novel views with accurate poses* from it. Unlike methods focused solely on per-scene quality or scaling feedforward models, our approach functions as a data engine to enhance and scale existing real-world captures. The key challenge is identifying free views that maximally capture under-constrained geometry while being minimally contaminated by reconstruction artifacts. To address this, we introduce a certainty-aware free-view sampling strategy. First, we construct a certainty grid from the reconstructed geometry to enable diverse and semantically meaningful look-at areas. Second, we build a view graph to establish geometric correspondences between generated and collected real-world data, guiding final view selection and rectification.

We conduct extensive experiments on challenging datasets containing out-of-distribution views and large camera motions. We validate FreeScale in two key applications: **1) Scaling feedforward NVS models** by augmenting training data with our generated views, expanding the training set to 1.22× its original dataset size [25]. Our approach achieves a 2.7 dB PSNR improvement over baseline methods. **2) Enhancing per-scene 3DGS optimization** through active non-certainty exploration, demonstrating consistent improvements on DL3DV [25], Nerfbusters [46], and Tanks and Temples [19] datasets. Our contributions are three-fold:

- A novel framework FreeScale, which leverages certainty-guided sampling to generate diverse, high-quality free-view images from sparse inputs. This expands the viewpoint coverage to boost downstream tasks significantly.
- A certainty-based view graph to efficiently manage and filter candidate viewpoints, ensuring the generated free-view images are both informative and photorealistic.
- Comprehensive validation demonstrating significant improvements in both feedforward reconstruction and per-scene optimization.

2. Related Works

Scene-level Novel View Synthesis. Neural Radiance Fields (NeRF) [29] and the highly efficient 3D Gaussian Splatting (3DGS) [17] are prime approaches for image synthesis at novel view points. However, these methods are susceptible to artifacts like floaters and ghosting when trained with sparse inputs or evaluated on out-of-distribution viewpoints. A significant body of work addresses this by imposing geometric constraints, using either sparse [9, 22] or dense [35] depth supervision, or by developing sophisticated strategies to control the placement and movement of Gaussian primitives [23, 28, 38, 54]. Other approaches mitigate appearance variations by decoupling view-dependent effects using appearance embeddings [5, 24, 51] or bilateral grids [40, 44].

Despite their high quality, these per-scene optimization methods are computationally expensive and do not generalize. This limitation has spurred the development of *generalizable, feed-forward* models. These approaches learn cross-scene priors from large datasets, typically by constructing cost volumes to reason about geometry [3, 4, 6, 7, 26, 41] or by leveraging powerful 3D foundation models (*e.g.*, DUSt3R [42], VGGT [39]) for improved initialization and generalization [14, 50, 52]. Transformer-based architectures have also been adapted for this task, operating either on 3D Gaussians [8, 53] or directly in 2D [12, 16].

A critical bottleneck for these generalizable models is the scarcity of large-scale, diverse training data with accurate and extensive camera trajectories. This data limitation caps their scalability and robustness, particularly for large, free-viewpoint camera motions. Furthermore, the performance of *both* optimization-based and feed-forward methods is highly sensitive to input imperfections, such as inaccurate camera poses.

Scaling up Scene-level Data and Priors. The performance of generalizable NVS models is fundamentally constrained by the scale and quality of training data. This challenge has been addressed through several distinct paradigms.

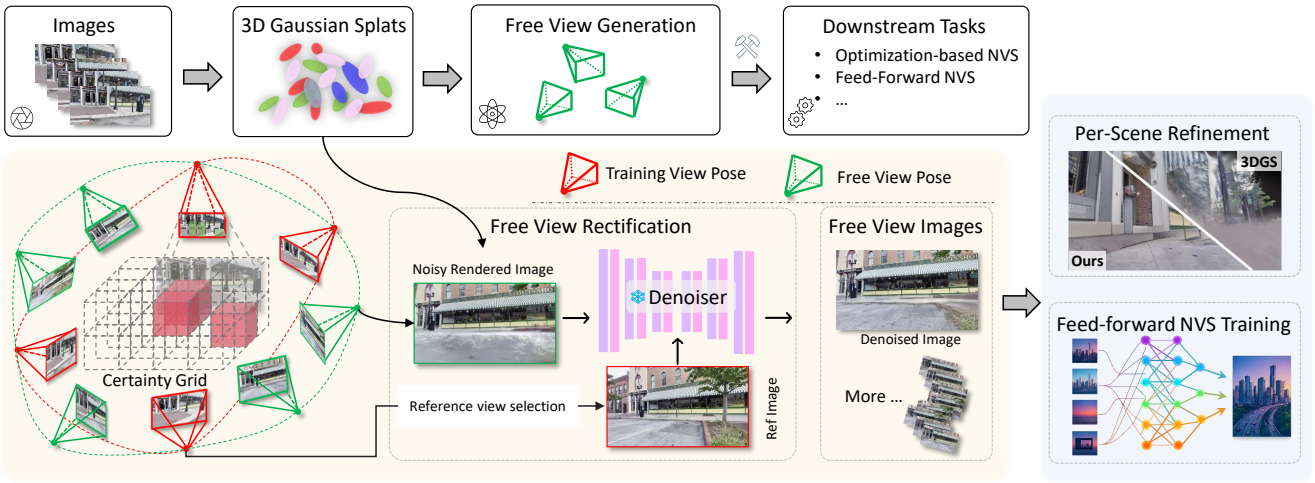

Figure 2. **FreeScale generation pipeline.** Our overall pipeline consists of three phases. First, given an image sequence, we reconstruct the scene as a continuous 3D representation, which allows us to place arbitrary viewpoint candidates. Second, we perform certainty-aware free-view synthesis: we establish a view graph based on a certainty grid and filter redundant candidates. Finally, we apply image rectification to produce the final free-views. The generated data can then be used to train feed-forward models like LVSM and refine the scene Gaussians.

One paradigm seeks to *generate synthetic data* at scale. This includes using 3D simulators like Blender to produce data with perfect camera control but a significant synthetic-to-real domain gap [33, 34, 37, 43, 49]. Methods like Megasynth [13] expand diversity through amorphous geometry, but this often leads to poor data efficiency. A more recent approach leverages powerful 2D diffusion models to generate photorealistic images [32]. However, a fundamental limitation of these *data-generation* methods is their inability to provide accurate, consistent camera poses and multi-view coherent geometry, which are essential for training robust 3D reconstruction systems.

Another paradigm uses pre-trained models as *priors to enhance* per-scene optimization. Methods distill 2D diffusion priors to mitigate sparse-view artifacts [2, 45], such as ReconFusion [48] generating pseudo-labels for NeRFs. For 3DGS, 3DGS-Enhancer [27] and DIFIX3D [47] refine pseudo-views and correct artifacts. While effective for individual scenes, these *prior-enhancement* approaches do not address the fundamental need for large-scale, diverse *training data* to build generalizable models.

This leaves a critical gap: a method that can generate *photorealistic, multi-view consistent data* with *diverse and accurate camera trajectories* to directly scale up the training of generalizable 3D vision models.

Enhancing 3DGS Representations. Beyond the use of external priors, other works focus on improving the intrinsic properties of 3DGS. One approach leverages statistical measures to improve training and prune artifacts. For instance, Bayes' rays [10] and FisherRF [15] use Fisher information to quantify per-point uncertainty, guiding view selection and post-training pruning of unreliable Gaussians.

3. Preliminary

Task definition. Our objective is to scale and augment 3D datasets to overcome their current limitations. This generates data with more diverse camera motions and geometric complexity, enhancing the generalization of downstream feedforward models. Formally, a dataset $\mathcal{D} = \{S_k \mid k = 1, \dots, M\}$ is organized as a collection of M independent scenes. Each scene S_k comprises a limited set of image-pose pairs, $S_k = \{(I_{k,i}, C_{k,i})\}_{i=1}^{N_k}$. However, the set S_k provides only partial coverage of the real scene, as the geometric and photometric variations in scene k contain far more information than what is observed from the limited N_k viewpoints. To overcome this data sparsity and unlock the full information potential of the scene, we propose to leverage reconstructed scene geometry. This reconstruction step transforms the sparse input S_k into a continuous 3D representation $\mathscr{G}_k$. Hence, we can freely sample an arbitrary number of novel views, which we define as "free-views" $\mathcal{F}_k = \{(I^{\text{fv}}_{k,i}, C^{\text{fv}}_{k,i}) \mid i = 1, \dots, N^{\text{fv}}_k\}$, achieving the desired data scaling for robust, generalizable model training.

Scene reconstruction with Gaussians. 3D Gaussians (3DGS) [17] achieves high-fidelity scene reconstruction by representing the scene as a set of N individual 3D Gaussians $\mathscr{G} = \{g_i\}_{i=1}^{N}$. Each Gaussian g_i is parameterized by the learnable properties including $\mu_i \in \mathbb{R}^3$ the center of g_i, scaling $\mathbf{s}_i \in \mathbb{R}^3$ defining the principal axes lengths, quaternionc $\mathbf{q}_i \in \mathbb{R}^4$ and opacity $\alpha_i \in \mathbb{R}$.

4. Method

The overall pipeline of FreeScale, as illustrated in Figure 2, initiates by leveraging a reconstructed 3D Gaussians derived from sparse multi-view inputs S_k. Our core inno-

vation lies in a certainty-guided sampling strategy that selectively generates a large quantity of high-diversity and high-quality free-view images $\mathbf{I}^{\text{fv}}$ and corresponding camera pose $\mathbf{C}^{\text{fv}}$. These synthesized free-views serve two critical purposes: first, they act as powerful data augmentation for training robust downstream feedforward models (Sec. 4.2.1); and second, they can be seamlessly integrated into the original reconstruction optimization loop to further enhance the quality and geometric fidelity of the initial 3D Gaussian representation (Sec. 4.2.2).

4.1. Certainty-aware Free-Views Synthesis

To facilitate the training of robust downstream models, our primary objective is to collect a set of high-diversity and high-quality free-view images that effectively serve as data augmentation. Crucially, these desired free-views must extend beyond the support of the original training viewpoint distribution, yet still capture sufficient scene information without being dominated by large, featureless, or out-of-scene background regions.

Since the reconstructed scene geometry $\mathscr{G}$ explicitly encode the geometry and density of the current scene, we leverage this representation to establish a sampling strategy guided by certainty. Instead of using a fixed unit size, we discretize the scene's bounding box into a relative voxel grid of resolution R^3. Each voxel v_i within this normalized grid is defined by its spatial index $i = (x, y, z)$. We empirically set $R = 128$ to balance efficiency and accuracy; while higher resolutions in small scenes introduce redundancy, lower resolutions in large scenes yield inaccurate WIoU estimates in Eq. 3 and erroneous NMS.

Then, The certainty of each grid $\mathcal{C}(v_i)$, is defined by accumulating the individual certainty scores of all Gaussian centers μ_j that fall within its boundaries, like:

$$\begin{aligned} \mathcal{C}(v_i) &= \sum_{g_j \in \mathscr{G}_i} \frac{\alpha_j}{\text{Vol}_j + \epsilon} \\ \text{Vol}_j &\triangleq \prod_{k=1}^{3} \exp((\mathbf{s}_j)_k) \end{aligned} \tag{1}$$

where $\mathscr{G}_i = \{g_j \mid \mu_j \in v_i, g_j \in \mathscr{G}\}$ denotes subset of Gaussians whose centers fall within voxel v_i The certainty score highlights regions of small, opaque Gaussians, providing robust guidance for high-fidelity free-view synthesis.

4.1.1. Virtual Viewpoints Placement

To ensure comprehensive scene coverage and maximize viewpoint diversity, we establish ten distinct camera trajectory modes within the 3DGS $\mathscr{G}$, shown in Figure 3. These modes include structured motions such as `orbit`, `spiral`, `lemniscate`, alongside various customized `move` and `fly-through` patterns. The generation process for candidate viewpoints $\mathbf{C}^{\text{fv}}$ is as follows:

- Anchor Selection: Trajectories are initialized from a set of anchor poses selected from the training cameras $\mathbf{C}_k$ using clustering and random choice.
- Certainty Localization: The look-at point for each generated trajectory is dynamically determined to enhance scene coverage. The look-at point of the `move` and `fly-through` modes adheres to the direction of the anchor camera to ensure spatial continuity and guided exploration. Conversely, for object-centric modes, the look-at point is randomly sampled from the Certainty Grid $\mathcal{C}$ within the scene's central volumetric region. This prioritization maximizes the geometric information content of the synthesized free-views by targeting well-reconstructed, high-certainty areas.
- Pose Jitter: To enhance pose diversity and mitigate potential sampling bias introduced by the pre-defined modes, a random subset of the generated poses undergoes slight rotational and translational perturbations.

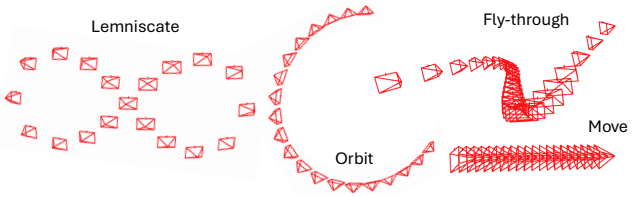

Figure 3. Showcase of predefined camera trajectory modes.

4.1.2. Virtual Viewpoints Selection with View Graph

To ensure maximal spatial diversity, we generate a substantial pool of $N_{\text{cand}} > 2000$ candidate views per scene. However, there are plenty of redundant and poor-quality candidates. Since we focus on the high-quality viewpoints and need to keep their diversity. The intuitive idea is using image-based methods (e.g., rendering and feature matching), but they are computationally prohibitive for such a large and redundant set. Therefore, we construct a view graph to facilitate efficient view selection.

View graph construction. We utilize estimated voxel certainty $\mathcal{C}(v_k)$ (as defined in Eq. 1) to quantify view information. Each original and candidate pose serves as a node. The weighted visibility $W_{i,k}$ for a voxel v_k in view i:

$$W_{i,k} = \mathcal{C}(v_k) \cdot M_{i,k} \tag{2}$$

where $M_{i,k}$ is a binary visibility mask indicating whether voxel v_k can be projected onto the camera plane of view i. Then, the edge between any two nodes i and j is quantified by the Weighted Intersection-over-Union (WIoU), which measures the overlap of high-certainty information:

$$\text{WIoU}(i, j) = \frac{\sum_{k=1}^{N} \min(W_{i,k}, W_{j,k})}{\sum_{k=1}^{N} \max(W_{i,k}, W_{j,k})} \tag{3}$$

The node score f for view C_i is the total aggregated weighted visibility: $f(C_i) = \sum_{v_k \in \mathcal{V}} W_{i,k}$.

Table 1. **Quantitative comparison of feed-forward models on viewpoint generalization.** Joint training with our FVGen data yields consistent improvements across both small and large camera motion settings.

Method	**In-Domain** *DL3DV* PSNR↑	SSIM↑	LPIPS↓	**Out-of-Domain** (Zero-shot Generalization) *MipNeRF360* PSNR↑	SSIM↑	LPIPS↓	*Tanks & Temples* PSNR↑	SSIM↑	LPIPS↓
Small camera motion									
LVSM [16]	22.20	0.680	0.216	15.84	0.285	0.583	13.07	0.336	0.674
LVSM w/ FreeScale	**24.20**	**0.767**	**0.165**	**18.30**	**0.386**	**0.460**	**13.80**	**0.652**	**0.361**
Large camera motion									
3DGS [17]	16.22	0.592	0.345	13.47	0.334	0.529	12.12	0.351	**0.569**
LVSM [16]	18.75	0.522	0.352	13.88	0.293	0.622	13.89	0.352	0.650
LVSM w/ FreeScale	**21.45**	**0.661**	**0.247**	**17.27**	**0.432**	**0.398**	**14.67**	**0.391**	0.609

4.1.3. Free-View Refinement and Rectification

After view graph is established, we collect synthesized virtual cameras $\mathbf{C}^{\text{fv}}$ and render images $\mathbf{I}^{\text{fv}}$ as the initial free-view set $\mathcal{F} = (\mathbf{I}^{\text{fv}}, \mathbf{C}^{\text{fv}})$. Then, we refine $\mathbf{I}^{\text{fv}} \in \mathcal{F}$ to produce high-fidelity and photo-realistic results, and keep accurate pose placement with an iterative process of quality assessment, pose rectification, and final image enhancement.

Image Quality Assessment and Semantic Filtering. Given an established view graph, we further prune low-quality nodes with a semantic filter. This step is crucial since even with a small deviation of camera poses from the training cameras, the rendered images can be very blurry and therefore not useful. We employ a combined metric for assessment: a no-reference image quality score BRISQUE [30] to assess visual fidelity, and a normalized depth range metric to confirm sufficient geometric content. Nodes are immediately rejected if the black pixel ratio or the normalized depth range falls below respective thresholds, indicating low quality or trivial content.

Pose Rectification via Interpolation. When a candidate pose fails the image quality check, we do not discard the pose $\mathbf{C}^{\text{fv}}$ outright. Instead, we implement a pose rectification strategy that moves the rejected candidate closer to its nearest anchor pose. Specifically, we perform pose interpolation to gradually shift $\mathbf{C}^{\text{fv}}$ towards this high-quality anchor. This pose refinement effectively recovers geometrically valuable candidates that would otherwise be discarded, ensuring a robust and sufficient final pool of high-quality, high-diversity free-views $\mathcal{F}_k$.

View-graph Guided Image Rectification. Although the refined free-views $\mathbf{I}^{\text{fv}}$ include well-reconstructed regions, underconstrained regions often still present noticeable artifacts, falling short of full photorealism. To address this, we apply a one step diffusion model DIFIX3D [47] to enhance the photorealism of the free-view images $\mathbf{I}^{\text{fv}}$. Because the performance of diffusion enhancement is highly sensitive to the selected reference view, we use the pre-constructed view graph to guide reference selection, instead of the nearest reference based on pose distance in [47]. Through second-order connections, the graph captures geometric relationships that simple spatial proximity cannot, preventing the selection of "noisy" references that would otherwise introduce misalignment artifacts.

4.2. Scaling Novel View Synthesis with FreeScale

In this part, we discuss how we utilize FreeScale to scale up downstream tasks training, including feedforward models and per-scene optimization.

4.2.1. View-graph Guided Feedforward Model Training

Training feedforward view synthesis models [12, 16] requires a robust view selection strategy to construct training batches. While effective training benefits from greater camera motion diversity to ensure generalization, excessive motion can destabilize the model. Previous methods [12, 16] based on frame distance within the image sequence, however, confine training to this limited range of camera motion, which is suboptimal for achieving robust, generalizable view synthesis. Furthermore, the filtered free-views generated by our FreeScale do not necessarily maintain a smooth, sequential trajectory, making frame-distance sampling inappropriate.

To resolve this trade-off between motion stability and diversity, we utilize the explicit geometric and content correspondence encoded in the View Graph generated by our FreeScale to guide view selection. And we further introduce a curriculum learning mechanism to manage training stability while increasing motion diversity.

Graph-Based View Selection. We select input batches based on the adjacency relations of start point within the view graph, rather than relying on sequential frame indices. Our method inherently connects views by their geometric and content similarity (via WIoU), whereas frame distance only guarantees temporal proximity.

Curriculum Learning Strategy. At warm-up iterations, we prioritize stability by selecting a start point that has the highest total WIoU score with its neighbors (i.e., the most information-rich or well-covered view). Input views are then randomly sampled from their high-WIoU neighbors. At the later stages, selection criteria progressively shift to

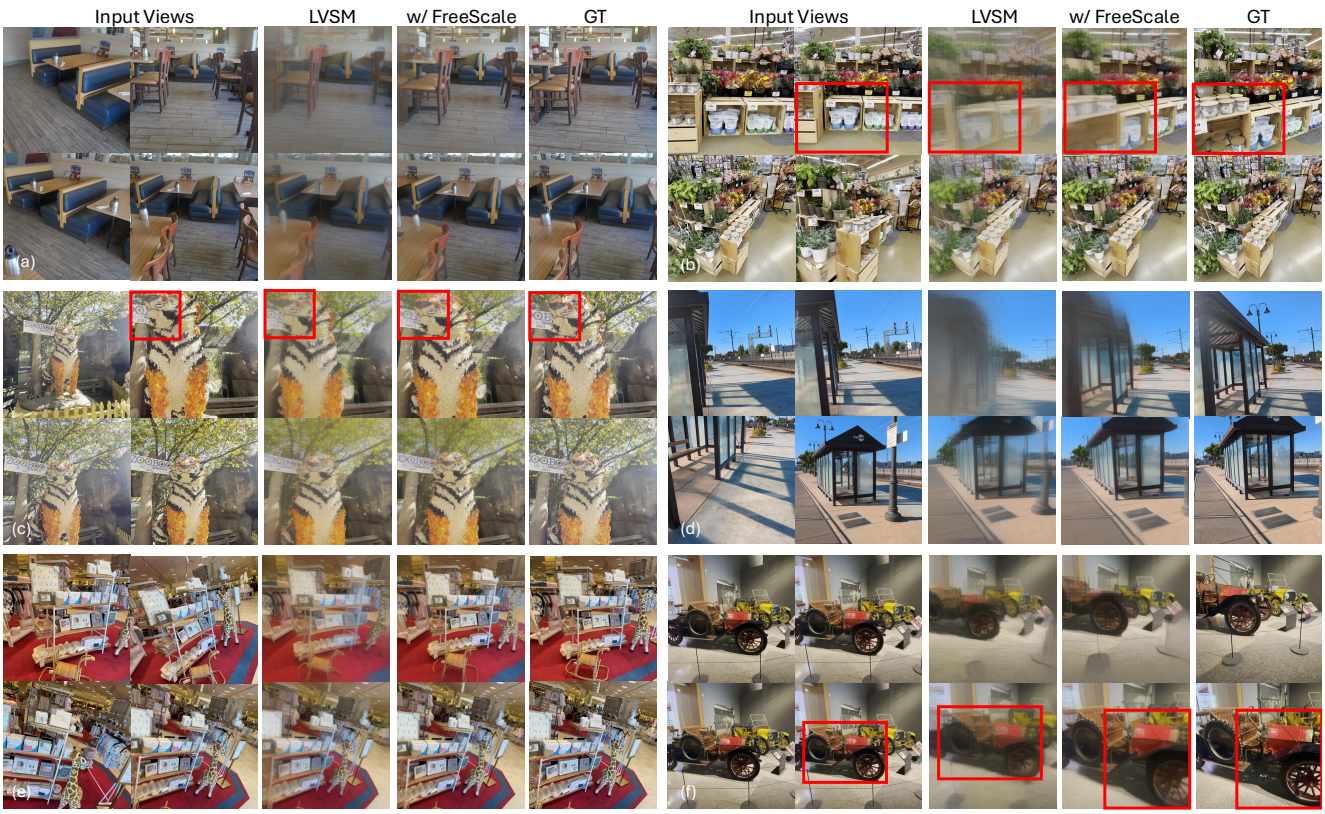

Figure 4. **Qualitative comparison for feed-forward model.** We use red boxes to highlight challenging regions where the target viewpoint differs significantly from the corresponding areas in the training poses.

favor views with lower average WIoU or greater geometric separation (as determined by lower edge weights), encouraging the model to generalize over larger camera motions and reconstruct less-constrained regions.

4.2.2. Certainty-guided Per-Scene Reconstruction

Despite the high quality and increased diversity of viewpoints and appearances achieved by our generated free-views, a remaining synthetic-to-real domain gap remains. Furthermore, the explicit nature of optimization-based 3DGS often makes it highly sensitive to inconsistencies or conflicting geometric gradients in the training data, which can lead to artifacts. To mitigate these issues, we integrate the enhanced free-views as pseudo ground truth during training. This strategy fundamentally differs from prior work [47] in how the additional views are generated and selected. While [47] leverages test poses to interpolate virtual cameras, our approach is designed to operate without any knowledge of the test camera distribution. We utilize the established view graph to automatically select free-views for training. Specially, we select the top-K free-views that exhibit the lowest WIoU with the existing training cameras. These low-WIoU views represent the maximal information complement to the original dataset and are added to the training set as auxiliary targets. The loss $\mathcal{L}_{\text{FV}}$ for these selected free-views is then defined as:

$$\mathcal{L}_{\text{FV}} = \alpha^{\text{fv}}(||I - I^{\text{fv}}||_1 + (1 - \mathcal{L}_{\text{SSIM}}(I, I^{\text{fv}}))) \quad (4)$$

where α^{fv} is a decay weight based on free-view quality.

5. Experiments

5.1. Improving Feed-forward NVS

Baselines. We select LVSM [16] as our primary baseline, following a challenging sparse-view reconstruction paradigm, using 4 input views to predict 2 target views.

Datasets. We train our models on the 1,900-scene training split and evaluate on the official 110-scene benchmark of DL3DV-10K [25]. To ensure data integrity, we apply a pre-processing filter to discard scenes where the image count mismatches the corresponding COLMAP data or that contain corrupted image files. To assess generalization, we conduct two types of evaluations: 1) Out-of-Domain (OOD) Evaluation: To measure the generalization gains conferred by FreeScale, we conduct evaluations on extra real-world datasets MipNeRF360 [1] and Tanks & Temples (tnt) [20], comprising a total of 16 scenes. 2) Viewpoint Generalization: To evaluate the model's ability to adapt to diverse camera motions, we establish two validation protocols based on

Table 2. **Quantitative comparison of per-scene reconstruction on the Out-Of-Domain protocol.** Our FreeScale achieves consistent advantages in PSNR and SSIM without incurring a significant increase in computational burden.

Method	DL3DV				Nerfburster			Tanks & Temples		
	PSNR↑	SSIM↑	LPIPS↓	Time (min.)↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Nerfbusters [46]	17.45	0.606	0.370	-	17.72	0.647	0.352	-	-	-
DIFIX3D+ [47]	17.99	0.601	0.293	81.40	18.07	0.642	0.279	18.59	0.623	0.317
3DGS [17]	19.18	0.714	0.233	**35.19**	18.14	0.643	0.265	20.37	0.680	0.253
3DGS w/ DIFIX3D	19.12	0.680	**0.211**	39.75	17.69	0.606	0.264	19.75	0.630	**0.210**
3DGS w/ Depth	19.07	0.718	0.227	36.67	17.54	0.630	0.285	20.24	0.678	0.259
3DGS w/ FreeScale	**19.57**	**0.723**	0.219	37.22	**18.40**	**0.648**	**0.258**	**20.66**	**0.685**	0.251

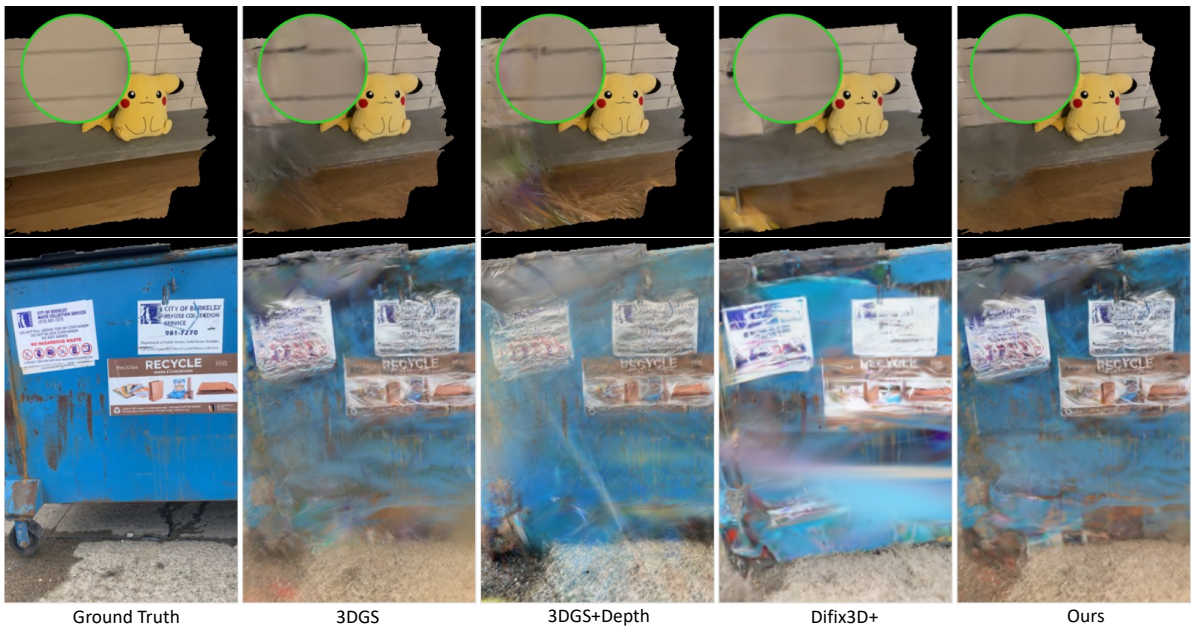

Figure 5. **Qualitative results on the Nerfbusters dataset.** The 3DGS baseline exhibits significant artifacts in unobserved areas, such as floaters and geometric noise, particularly in unobserved areas. In contrast, our method ensures high-fidelity results by sampling supplementary views from the reconstructed scene geometry.

frame distance. For the small camera motion setting, we sample poses, ensuring the maximum frame distance between any input and target view is no greater than 20. Conversely, for the large camera motion setting, we ensure the minimum frame distance is at least 20.

Results. Quantitative results in Table 1 demonstrate that joint training with our FreeScale data yields consistent improvements across both small and large camera motion settings. The benefit is particularly pronounced in the challenging large camera motion scenario, where our method achieves a substantial 2.7 dB PSNR gain over the baseline. Qualitative comparisons are presented in Figure 4. The baseline LVSM exhibits significant difficulty with large camera motion, tending to merely replicate input views rather than generalizing to the expected novel perspective. As highlighted in the red box, the baseline's rendered viewpoint clearly regresses towards an input pose. In contrast, FreeScale provides the feed-forward model with richer priors by sampling from realistic scene geometry. As shown in Figure 4(a), the data augmentation from our synthesized zoom-in trajectories enables the model to generate a sharper and more geometrically accurate image.

5.2. Enhancing Per-Scene Reconstruction

Baseline. We compare our method against 3DGS and its variants, as well as diffusion-based refinement methods like DIFIX3D [47] and Nerfbusters [46]. A critical distinction is that they load pretrained weights as a starting point for fine-tuning, whereas our model is trained from scratch.

Datasets. We conduct experiments on DL3DV, Nerfbuster, and tnt, and adopt a more challenging OOD splitting strategy. We partition each scene's data sequentially based on the camera trajectory order. Specifically, the first 70% of frames from each DL3DV evaluation scene and the first

Table 3. **Ablation of per-scene optimization on DL3DV.** "w/ dist ref": distance-based reference for rectification. "w/ sparse init.": incomplete initialization.

Methods	3DGS	w/ dist ref	w/ sparse init.	w/ FreeScale
PSNR$_\uparrow$	19.18	17.88	19.51	**19.57**
SSIM$_\uparrow$	0.714	0.666	0.717	**0.723**
LPIPS$_\downarrow$	0.233	0.302	0.232	**0.219**

Table 4. **Ablation study of different data sparsity on tnt dataset.** This shows that our FVGen is robust and not limited to a specific data density, providing consistent quality improvements.

	10%		20%		40%		50%	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
3DGS	20.86	0.674	20.37	0.680	18.96	0.649	17.83	0.620
FreeScale	21.09	0.678	20.65	0.685	19.19	0.654	18.10	0.633

Table 5. **Ablation study on free-view images.** "FV" indicate generated free-view images, "View-graph" means graph-guided joint training and certainty-guided per-scene reconstruction.

	Feed-forward Model			Per-Scene Optimization		
Method	PSNR$_\uparrow$	SSIM$_\uparrow$	LPIPS$_\downarrow$	PSNR$_\uparrow$	SSIM$_\uparrow$	LPIPS$_\downarrow$
Baseline	17.75	0.385	0.465	19.18	0.714	0.233
+ FV (random)	18.68	0.508	0.342	19.20	0.715	0.240
+View-graph	19.11	0.529	0.322	19.57	0.723	0.219

80% from each tnt scene are used for training. We follow the official benchmark protocol for Nerfbuster.

Results. Table 2 shows that FreeScale achieves consistent advantages, achieving the highest PSNR and SSIM scores. These significant quality improvements are achieved with a negligible increase in computational cost. The runtime of FreeScale remains comparable to the baseline. This is substantially faster than DIFIX3D+ [47], which require multiple, costly diffusion inference passes, leading to significant time overhead. The qualitative results on Nerfbuster show that FreeScale keeps more details, which can be found in Figure 5. Our method ensures high-fidelity results by viewpoint expansion based on the reconstructed scene geometry.

5.3. Ablation Studies and Analysis

Robustness to Data Sparsity. We evaluate the robustness of FreeScale against data sparsity from two distinct perspectives: incomplete initialization and and sparse training data. As shown in Table 3 (w/ sparse init.), when the initial 3DGS is reconstructed using merely 5% of the training data, FreeScale remains resilient, effectively extracting valid geometric cues to produce informative views that surpass the baseline 3DGS. Furthermore, we conduct an ablation study varying levels of overall data sparsity on tnt dataset in Table 4. FreeScale consistently outperforms the baseline across all sparsity levels in both PSNR and SSIM. Together, these results demonstrate that our data engine is not overly reliant on dense initial coverage and generalizes effectively to highly sparse real-world scenarios.

Ablation study on free-view images. We analyze the impact of free-view images on two key downstream tasks as shown in Table 5. (1) For the feed-forward model, we conduct an ablation on the DL3DV benchmark under the large camera motion setting, training for 8K iterations. The Baseline model, trained without free-views shows poor viewpoint generalization. Simply adding free-views sampled randomly provides a substantial boost from 17.75 to 18.68 PSNR. Finally, we develop a curriculum learning strategy based on view graph information, which achieve the best performance. (2) For per-scene optimization, randomly selecting free-views yields a negligible improvement. This suggests that intuitive data augmentation can introduce conflicting or redundant information. In contrast, using the view graph to select views that provide maximal information complement achieves the best results across all metrics

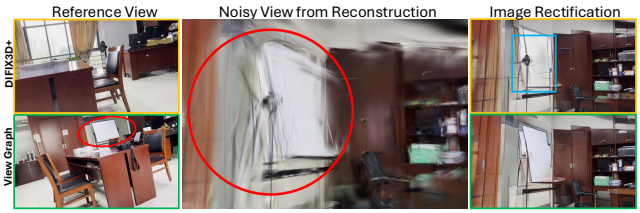

Figure 6. **Comparison of reference image selection.** Our view graph identifies the shared visible region with the noisy view (red circle), ensuring accurate image rectification.

Impact of the View Graph. The proposed view graph is essential for robust free-view refinement, as it provides accurate geometric correspondences between generated views and the original training cameras. Unlike previous methods that rely solely on spatial pose distance, which often fail to capture true visual overlap, our graph-based approach explicitly guarantees meaningful relationships. As shown in Figure 6, the distance-based reference image selected by DIFIX3D+ lacks frustum overlap with the target view, resulting in severe rectification artifacts. Conversely, our view graph leverages shared visibility to prevent such misalignment. Quantitative results in Table 3 further validate this: such a distance-based reference strategy (w/ dist ref) leads to noticeable performance drops, whereas our graph-guided selection consistently ensures photorealism.

6. Conclusion

In this work, we address the critical data bottleneck in novel view synthesis, where sparse real-world data limits generalization and synthetic data suffers from domain gaps. We introduce FreeScale, a novel data engine that transforms discrete scene data into a continuous 3D representation to generate diverse, high-fidelity free-views with accurate poses. At its core, a certainty-based view graph efficiently filters candidate viewpoints and guides image rectification, maximizing information in under-constrained regions while mitigating reconstruction artifacts. Experiments show that FreeScale provides a scalable data augmentation solution, significantly boosting downstream performance and opening a new avenue for training robust, 3D-aware models.

Acknowledgments This work has been made possible by a Research Impact Fund project (RIF R6003-21) and a General Research Fund project (GRF 16203224) funded by the Research Grants Council (RGC) of the Hong Kong Government. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations resources provided by Chalmers e-Commons at Chalmers and the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 3

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2

[4] Yu Chen and Gim Hee Lee. DBARF: deep bundle-adjusting generalizable neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24–34, 2023. 2

[5] Yu Chen and Gim Hee Lee. DOGS: distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus. In *Annual Conference on Neural Information Processing Systems 2024*, 2024. 2

[6] 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 *18th European Conference*, pages 370–386, 2024. 2

[7] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024. 1, 2

[8] Ziwen Chen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Fuxin Li, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *CoRR*, abs/2410.12781, 2024. 2

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12872–12881, 2022. 2

[10] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2024. 3

[11] Roland Hess. *Blender foundations: The essential guide to learning blender 2.5*. Routledge, 2013. 2

[12] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025. 1, 2, 5

[13] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, et al. Megasynth: Scaling up 3d scene reconstruction with synthesized data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16441–16452, 2025. 2, 3

[14] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *CoRR*, abs/2505.23716, 2025. 2

[15] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *CoRR*, abs/2311.17874, 2023. 3

[16] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024. 1, 2, 5, 6

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3, 5, 7

[18] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. In *Advances in Neural Information Processing Systems*, 2024. 1

[19] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):78:1–78:13, 2017. 2

[20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 6

[21] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In *Advances in Neural Information Processing Systems*, 2024. 1

[22] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. 2

[23] Yanyan Li, Chenyu Lyu, Yan Di, Guangyao Zhai, Gim Hee Lee, and Federico Tombari. Geogaussian: Geometry-aware gaussian splatting for scene rendering. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 441–457, 2024. 2

[24] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 2

[25] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, 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*, pages 22160–22169, 2024. 2, 6

[26] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 37–53, 2024. 2

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

[28] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[30] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011. 5

[31] Wieland Morgenstern, Florian Barthel, Anna Hilsmann, and Peter Eisert. Compact 3d scene representation via self-organizing gaussian grids. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 18–34, 2024. 1

[32] Nithin Gopalakrishnan Nair, Srinivas Kaza, Xuan Luo, Vishal M Patel, Stephen Lombardi, and Jungyeon Park. Scaling transformer-based novel view synthesis with models token disentanglement and synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28567–28576, 2025. 2, 3

[33] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794, 2024. 2, 3

[34] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 2, 3

[35] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[36] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[37] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR, 2022. 2, 3

[38] Evangelos Ververas, Rolandos Alexandros Potamias, Jifei Song, Jiankang Deng, and Stefanos Zafeiriou. SAGS: structure-aware 3d gaussian splatting. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 221–238, 2024. 2

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025. 2

[40] Nan Wang, Yuantao Chen, Lixing Xiao, Weiqing Xiao, Bohan Li, Zhaoxi Chen, Chongjie Ye, Shaocong Xu, Saining Zhang, Ziyang Yan, Pierre Merriaux, Lei Lei, Tianfan Xue, and Hao Zhao. Unifying appearance codes and bilateral grids for driving scene gaussian splatting. *CoRR*, abs/2506.05280, 2025. 2

[41] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

[42] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[43] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 2, 3

[44] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing. *ACM Trans. Graph.*, 43(4):148:1–148:13, 2024. 2

[45] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *IEEE/CVF International Conference on Computer Vision*, pages 18074–18084, 2023. 3

[46] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *IEEE/CVF International Conference on Computer Vision*, pages 18074–18084, 2023. 2, 7

[47] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025. 3, 5, 6, 7, 8, 1

[48] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 3

[49] Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. *Advances in Neural Information Processing Systems*, 37:53285–53316, 2024. 2, 3

[50] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16453–16463, 2025. 2

[51] Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. In *Annual Conference on Neural Information Processing Systems 2024*, 2024. 2

[52] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *The Thirteenth International Conference on Learning Representations,*, 2025. 2

[53] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: large reconstruction model for 3d gaussian splatting. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 1–19, 2024. 2

[54] Zheng Zhang, Wenbo Hu, Yixing Lao, Tong He, and Hengshuang Zhao. Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. In *Computer Vision - ECCV 2024 - 18th European Conference*, pages 326–342, 2024. 2

FreeScale: Scaling 3D Scenes via Certainty-Aware Free-View Generation

Supplementary Material

This supplementary material consists of three parts: technical details of the experimental setup (Sec. 7), additional ablation studies on free-views generation (Sec. 8), and additional qualitative results (Sec. 9), including out-of-domain results and a discussion about limitations (Sec. 10).

7. Implement Details

7.1. Certainty-aware Free-View Synthesis

Virtual Viewpoints Placement. We first generate virtual viewpoints trajectories with 10 predefined modes, including: **geometric paths**: (1) orbit, (2) spiral, (3) lemniscate; (4) **interpolation**; and **cinematic movements**: (5) move up, (6) move down, (7) move left, (8) move right, (9) dolly-zoom in, (10) dollyzoom out. For each mode, we select $N_{traj} = 10$ anchor poses via K-Means clustering or Farthest Point Sampling (FPS) on the training views. These anchor poses are randomly perturbed with position noise sampled from $\mathcal{N}(0, \sigma_{pos})$ where $\sigma_{pos} \in [0, 0.1]$ and rotation jitter within $\pm 20°$. We generate sequences of $L = 20$ frames per trajectory, resulting in a dense candidate pool. To enhance viewpoint diversity, we apply random perturbations to the initial poses, with position noise sampled from $\mathcal{N}(0, \sigma_{pos})$ where $\sigma_{pos} \in [0, 0.5]$ and rotation jitter within $\pm 30°$. Our candidate pool has more than 2,000 candidate views per scene.

Virtual Viewpoints Selection. To eliminate invalid views (e.g., occluded or unbounded regions), we first perform rigorous spatial feasibility checks on all candidate poses, immediately rejecting those that violate geometric constraints. This involves rejecting poses falling outside the scene's established bounding box or those situated inside known structures. Only the remaining feasible poses are considered for node status. Then, we perform Non-Maximum Suppression (NMS) on the candidate poses based on the established view graph. The score of view $f(C_i)$ quantifies its information gain. Candidate poses are first sorted in descending order of $f(C_i)$. We initialize the selected set $\mathcal{F}_{selected}$ with all poses from the training set. A candidate pose i is accepted and added to $\mathcal{F}_{selected}$ if its W-IoU with all poses $j \in \mathcal{F}_{selected}$ remains below the threshold of 0.7. This filtering process continues until $K = 500$ non-redundant candidates are successfully selected, ensuring both high certainty and diversity free-views.

Free-View Refinement and Rectification. During rendering, we enforce quality assurance using the BRISQUE metric < 0.5. And a depth percentile range validity score calculated on the central 70% crop, must be greater than 0.1. If a rendered view fails these checks, we trigger a pose rectification mechanism: the camera is iteratively shifted towards the nearest training view with decreasing step distances $\{0.7, 0.5, 0.3\}$. Finally, we apply stepwise filtering based on the BRISQUE quality metric to select the target set of $\approx$ 100 high-quality free-view frames, retaining all candidates if the target quota is not fulfilled.

7.2. Per-Scene Reconstruction

Baseline Training. Our 3DGS training pipeline follows the standard steps of [17]. We use sparse points from COLMAP [36] for initialization. The initial opacity is 0.5. We adopt the densification strategy of MCMC-3DGS [18] for better scene representation and compression. For all datasets, we train 3DGS for 30,000 iterations; the densification step starts from 500 and ends at 25,000. We densify each scene for every 500 iterations during training and densify at most 1,500,000 3DGS primitives. To adapt to the appearance changes, we follow WildGaussians [21], which applies an appearance feature with dim 32 per 3DGS and trains a shallow MLP as the appearance decoder. For further acceleration, we adopt tiny-cuda-nn[1] as the shallow MLP implementation. The depth and width of the shallow MLP are, respectively, 2 and 64. Considering the disk storage for 3DGS can be large when training thousands of scenes, we compress 3DGS using SOGS [31] to reduce the size. For all baselines (3DGS [17] and DIFIX3D+ [47]) and our method, we adopt gsplat[2] as the CUDA rasterization kernel.

3DGS Training with Freeview Enhancement. Unlike existing extrapolation methods, we train 3D Gaussian Splatting (3DGS) from scratch using augmented scene data, including the generated Freeview images. We adopt an iterative pseudo-labeling strategy: for every 3k iterations, we non-recurrently select the top-5 Freeview images that exhibit the lowest W-IoU with the existing training cameras. These selected images are incorporated into the training set as pseudo-GT until all Freeview images have been added. The loss weight for each incorporated pseudo-GT is assigned a decaying factor $\alpha^{\text{fv}} \in [0.3, 0.5]$ based on its corresponding BRISQUE quality metric.

7.3. Scaling Up LVSM

We train LVSM for 20,000 iterations on 1,900 scenes from the DL3DV dataset, following the established setup in [16]. The training utilizes 4 A40 GPUs, with a batch size of 24 per GPU. For optimization, we employ a cosine learning

[1] https://github.com/NVlabs/tiny-cuda-nn
[2] https://github.com/nerfstudio-project/gsplat

Table 6. **Ablation study on Free-View generation.** Our certainty-aware generation relies fundamentally on the certainty grid and the established view graph. **Without the view graph**, selecting the top-500 candidates solely by certainty score results in high redundancy and fails to provide valuable guidance for feed-forward model training. **Without the certainty grid**, we must resort to calculating inter-view correspondence only via position and rotation distance, which is both inaccurate and computationally inefficient.

			FreeView Statistics				Feed-Forward Model		
Method	Certainty Grid	View Graph	#Image	Per-scene #Image	BRISQUE↓	Avg. Time (s)↓	PSNR↑	SSIM↑	LPIPS↓
FVGen	✓	✓	145,528	75	0.36	**225.64**	**19.11**	**0.529**	**0.322**
	✓	-	164,874	91	0.38	608.71	17.63	0.480	0.397
	-	✓	166,282	109	0.36	727.93	17.86	0.491	0.354
Baseline	-	-	-	-	-	-	17.75	0.385	0.465

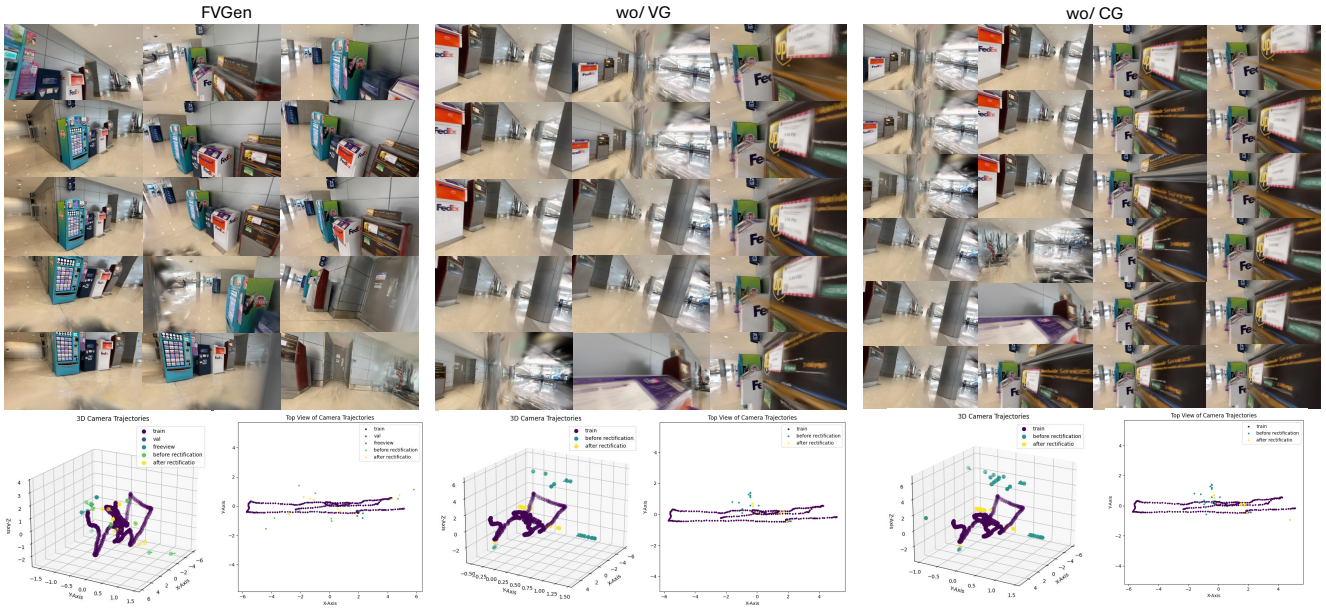

Figure 7. **Showcase of different freeview generation.** Our FVGen maximally captures under-constrained geometry while being minimally contaminated by reconstruction artifacts.

rate schedule that peaks at 4×10^{-4} after a 3,000-iteration warmup period.

We set the standard frame distance between input and target views to $[15, 40]$; this distance range is also applied when selecting neighboring nodes in our view graph. To stabilize early training, we implement a curriculum learning strategy during the warm-up phase: the frame distance is gradually annealed from a narrow range of $[10, 20]$ to the standard $[15, 40]$. In addition, input and target views are selected based on the view graph with a probability 50% throughout the training process.

8. Additional Ablation Studies

In this part, we conduct more ablation studies on free-view generation and show more cases about reference image selection mentioned in the main body Sec.5.3.

8.1. Ablation on Free-View Generation

The primary objective of FVGen is to collect a set of high-diversity and high-quality free-view images. This is achieved by utilizing a certainty grid to score the information value of candidate viewpoints. Crucially, we employ the established view graph to quantify inter-view correspondence, enabling the efficient filtering of redundant poses while simultaneously preserving viewpoint diversity. All ablation experiments are conducted on the same subset of DL3DV scenes. We report the feed-forward model results in Table 6, using identical settings to the results presented in the main body Table 3. We also showcase the generated freeviews (before image rectification) in Figure 7.

Without View Graph. When the view graph is omitted, the selection process relies solely on the certainty score to choose the top-500 candidates (as detailed in the main body, Sec. 4.1.2). This reliance leads to high redundancy among the selected views, resulting in insufficient scene coverage, as illustrated in Figure 7.

Moreover, the absence of inter-view correspondence necessitates the random integration of these generated free views into the feed-forward model training. This approach

Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
	Small camera motion			*Large camera motion*		
LVSM	22.20	0.680	0.216	18.75	0.522	0.352
wo/ Diffusion	23.41	0.736	0.185	20.89	0.634	0.268
w/ FreeScale	**24.20**	**0.767**	**0.165**	**21.45**	**0.661**	**0.247**

Table 7. **Ablation isolating the impact of diffusion-based image rectification.** Comparing the LVSM baseline to our method without diffusion (*wo/ Diffusion*) demonstrates that the primary performance boost stems directly from the expanded viewpoint diversity and geometric coverage. Integrating the diffusion prior (FreeScale) resolves remaining artifacts for optimal fidelity.

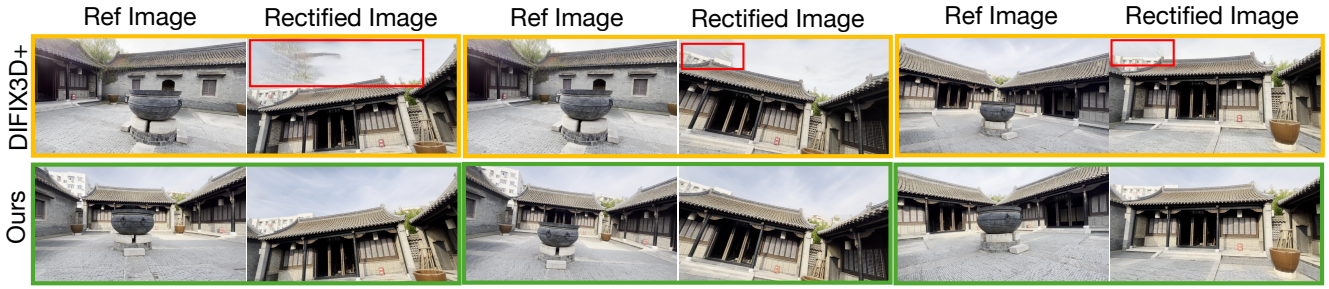

Figure 8. **Consistent showcases of view graph impact.** Compared to DIFIX3D+'s distance-based reference selection strategy, our view graph provides better overlap and higher free-view consistency for reference. The red bounding boxes delineate artifacts introduced by inaccurate reference images during the image rectification stage.

introduces significant view shifts (i.e., no overlap between input and target views), causing training instability. As shown in Table 6, while this method yields more images, it results in inferior overall image quality (higher BRISQUE) and provides poor performance gains for the downstream model, where PSNR drops from 17.75 to 17.63 dB.

Without Certainty Grid. As our view graph construction strongly relies on the certainty grid, its absence necessitates an alternative for establishing inter-view correspondence. We resort to calculating the combined position distance and the angular distance between quaternions. However, this approach presents several drawbacks. 1) Setting a suitable threshold for the combined distance is non-trivial; we empirically select 3.5 to retain approximately 500 candidates. 2) The lack of certainty-based viewpoint scoring requires computing all pairwise distances between candidates for view selection, leading to excessive generation time (Table 6). 3) The correspondence based solely on the position and rotation distance is inherently inaccurate, as small combined distances do not guarantee sufficient common visibility when camera rotations differ significantly, a phenomenon illustrated in Figure 10. Consequently, the resulting inter-view correspondence leads to suboptimal performance in the downstream feed-forward model in Table 6.

8.2. Effectiveness of Free-Views without Diffusion

To explicitly isolate the performance gain provided by generating novel viewpoints from reconstructed scene geometry, we ablate the diffusion-based image rectification. While applying 2D diffusion independently to each frame inevitably introduces minor multi-view inconsistencies (e.g., high-frequency flickering), we observe that downstream feed-forward models are remarkably robust to this issue. Because these models inherently aggregate features across multiple views, they effectively learn the underlying 3D scene geometry while filtering out inconsistent generative artifacts as noise. To explicitly isolate the contribution of the free-view images, we conduct an ablation study in Table 7. Remarkably, even without the diffusion-based refinement (*w/o Diffusion*), our method still significantly outperforms the LVSM baseline. This improvement is particularly pronounced in large camera motion scenarios, yielding a substantial +2.14 dB PSNR gain. These results definitively confirm that the primary performance boost stems from the expanded viewpoint coverage and spatial diversity provided by our certainty-guided sampling, rather than solely relying on the generative prior of the diffusion model.

8.3. More Analysis on Reference Images Selection

As discussed in the main body Sec. 5.3 and Figure 6, we provide further visualization analysis of our reference image selection strategy based on inter-view correspondence. Unlike methods such as DIFIX3D [47], which rely on calculating a combined metric of position and rotation distance to establish correspondence, our certainty-aware view graph (VG) yields superior reference images that share a more

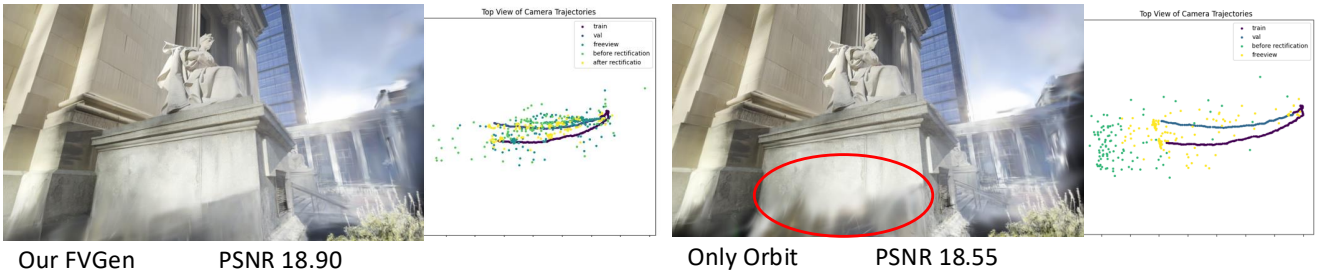

Figure 9. **Impact of diverse camera trajectory modes.** Relying solely on an `Orbit` trajectory limits viewpoint diversity, leading to noticeable blurring artifacts in under-observed regions (red circle). In contrast, our multi-mode sampling ensures maximal scene coverage, yielding sharper structural details and improved quantitative performance.

precise common visible area with the sampled noisy view. In Figure 10, the red circles highlight this shared visible region, while the blue bounding boxes delineate artifacts introduced by inaccurate reference images during the image rectification stage. We also show a consistent example in Figure 8. Compared to DIFIX3D+'s distance-based strategy, our view-graph-based reference selection provides better overlap and higher free-view consistency.

8.4. Ablation on Camera Trajectory Modes

To guarantee comprehensive scene coverage, our framework initializes candidate viewpoints using a diverse set of trajectory modes, relying on the certainty-based view graph to efficiently filter out spatial redundancy. As shown in Figure 9, compared to relying solely on the `Orbit` trajectory, incorporating diverse modes enables the data engine to capture geometrically challenging and under-constrained regions. This strategy not only significantly boosts overall synthesis performance but also ensures that the framework remains highly robust to the specific design choices of the candidate pool.

9. Additional Qualitative Results

In this part, we provide more qualitative comparison for feedforward model and per-scene reconstruction.

Out-of-Domain Results of FeedForward Model. We provide a qualitative comparison of the feed-forward model performance on out-of-domain (OOD) data, specifically using the MipNeRF360 dataset. Figure 11 illustrates the novel views generated by the baseline LVSM model against our approach incorporating FVGen for scaling scene data, benchmarked against the ground truth (GT). The results are rendered at a resolution of 256. The comparison highlights the advantages of FVGen: Unlike the baseline output, which suffers from excessive blurriness for OOD scenes, augmenting the training set with FVGen significantly mitigates this issue, producing sharper results closer to the GT.

Per-scene Reconstruction. We show a per-scene reconstruction comparison for the Tanks & Temples dataset in Figure 12. And results on DL3DV dataset can be found in Figure 13, 14 and 15.

10. Limitation and Future Works

The primary limitation lies in the Free-View Rectification stage, as the final image quality depends on the external diffusion model used for enhancement. Despite our certainty-aware View Graph improving reference image selection by ensuring geometric correspondence, residual artifacts can still be introduced. For future work, we plan to address this issue by fine-tuning the external diffusion model directly based on our sampling strategy, thereby reducing the synthetic-to-real domain gap. Integrating the view-specific certainty visibility mask into the diffusion model's conditioning. This would explicitly guide the denoiser to prioritize refinement only in uncertainty regions, thereby preventing artifacts in the original image distribution.

Failure Cases and Limitations. The failure cases and applicability boundaries of our generated free-views primarily stem from two factors: (1) **Diffusion Model Limitations:** Because the diffusion refinement model is trained for deblurring, it struggles to correct complex, view-dependent reflections as illustrated in Figure 16 (red circle). Furthermore, it can occasionally misinterpret severe 3DGS floaters from the initial reconstruction as valid scene structures, resulting in over-sharpened artifacts. (2) **Free-View Scarcity:** In extreme conditions, such as extreme low-light environments, our rigorous quality filtering mechanism may reject a large number of poor renderings, leading to a scarcity of valid free-views. Despite these localized limitations, downstream feed-forward NVS models trained with our augmented data exhibit strong robustness; they effectively treat these inconsistent artifacts as noise, thereby maintaining high overall synthesis performance.

Figure 10. **Additional showcases of view graph impact on reference image selection.** The red circles highlight the shared visible region between the reference view and sampled noisy view, while the blue bounding boxes delineate artifacts introduced by inaccurate reference images during the image rectification stage.

Figure 11. **Qualitative comparison of feed-forward on out-of-domain data (MipNeRF360).** The results are from LVSM at resolution 256.

Figure 12. **Qualitative comparison of per scene reconstruction on Tanks and Temples dataset.**

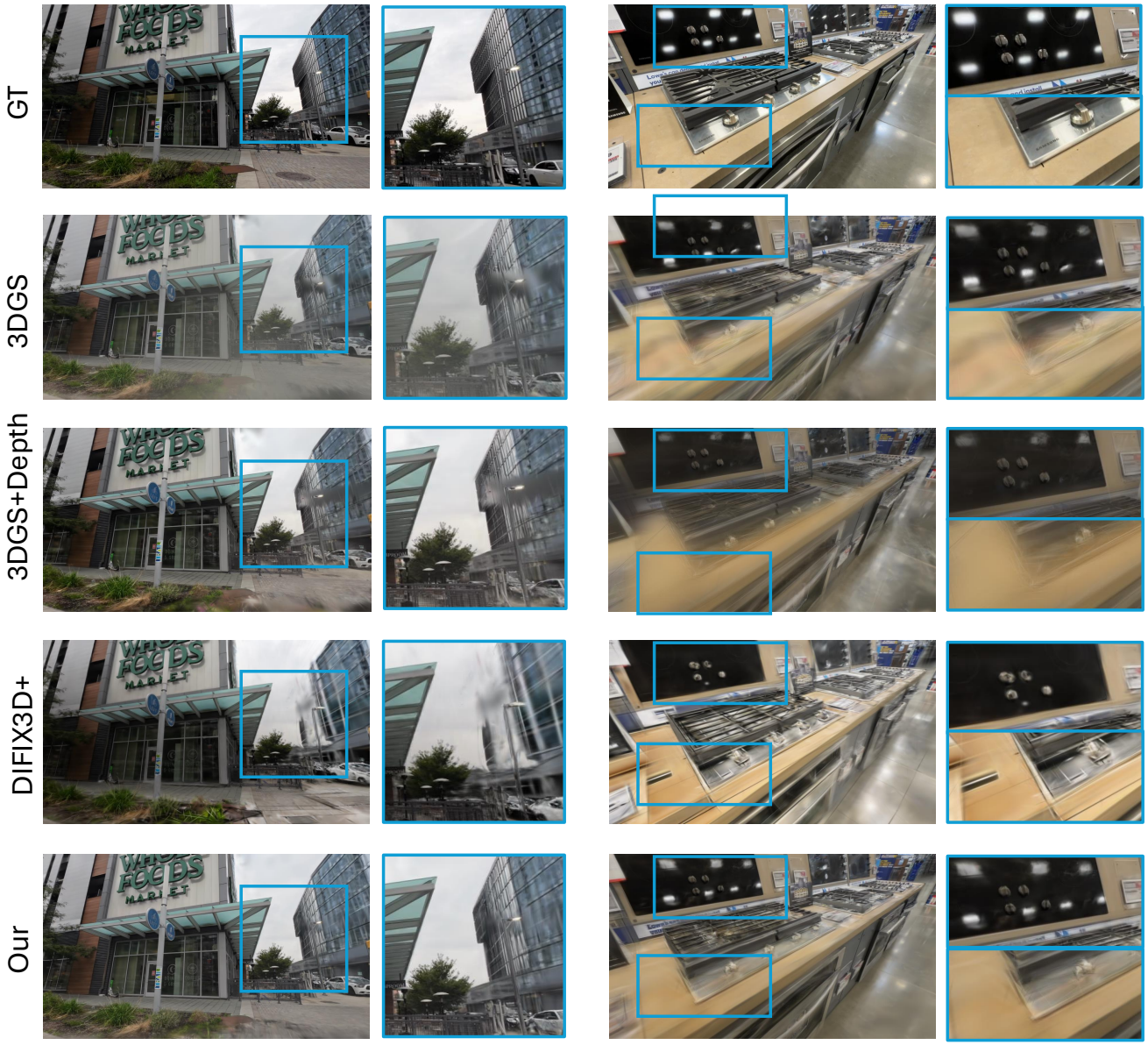

Figure 13. **Qualitative comparison on DL3DV dataset for per-scene reconstruction.** The blue bounding box indicates the zoom-in area. Despite the apparent clarity of DIFIX3D+, its progressive update and inaccurate reference image selection introduce significant hallucinated content, visible in the spurious reflection of the lamp and the corrupted desktop details on the right side.

Figure 14. **Qualitative comparison on DL3DV dataset for per-scene reconstruction.** The blue bounding box indicates the zoom-in area.

Figure 15. **Qualitative comparison on DL3DV dataset for per-scene reconstruction.** The blue bounding box indicates the zoom-in area.

Figure 16. **Failure cases of free-view generation.** Red circles indicate regions where our pipeline struggles due to diffusion priors, including the incorrect handling of complex reflections (top row) and the over-sharpening of 3DGS floaters (bottom row).