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

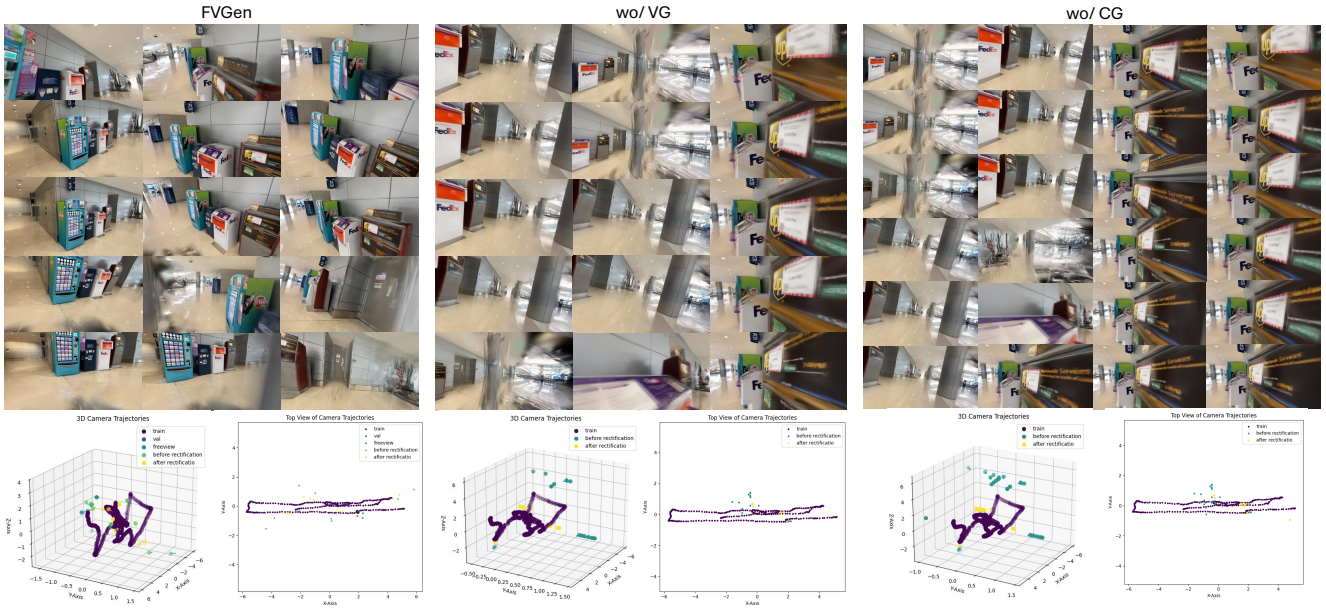

Figure 7. **Showcase of different freeview generation.** Our FVGen maximally captures under-constrained geometry while being minimally contaminated by reconstruction artifacts.

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.

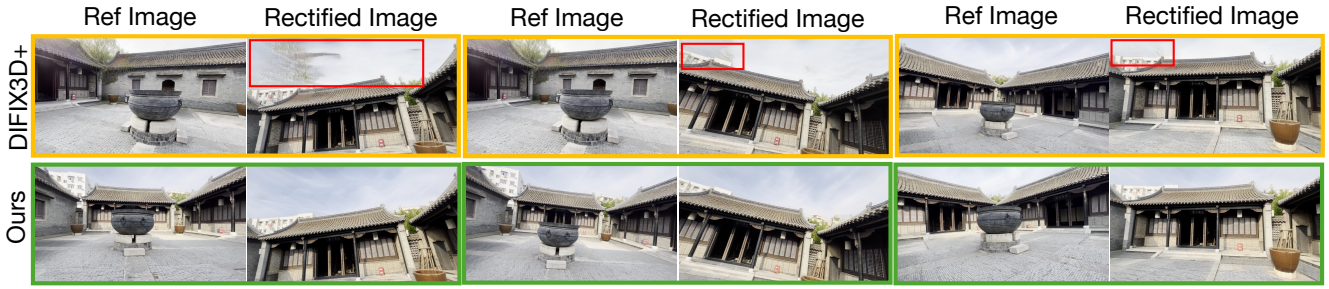

Figure 8. **Consistent showcases of view graph impact.** Compared to DIFIX3D+'s distance-based reference selection strategy, our view graph provides better overlap and higher free-view consistency for reference. The red bounding boxes delineate artifacts introduced by inaccurate reference images during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

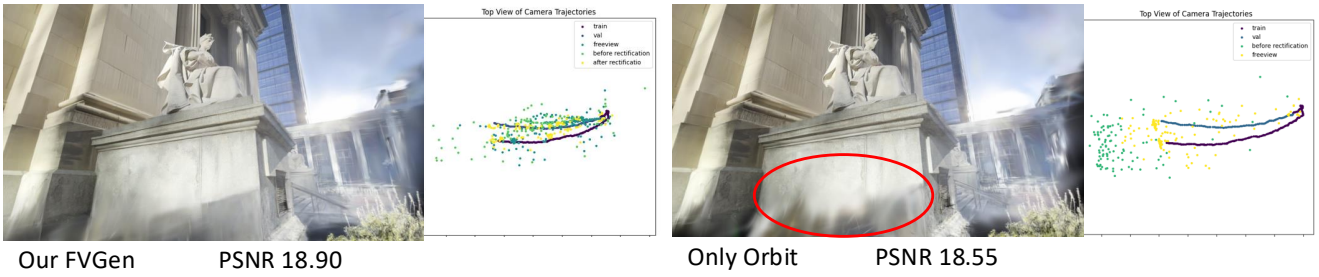

Figure 9. **Impact of diverse camera trajectory modes.** Relying solely on an `Orbit` trajectory limits viewpoint diversity, leading to noticeable blurring artifacts in under-observed regions (red circle). In contrast, our multi-mode sampling ensures maximal scene coverage, yielding sharper structural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

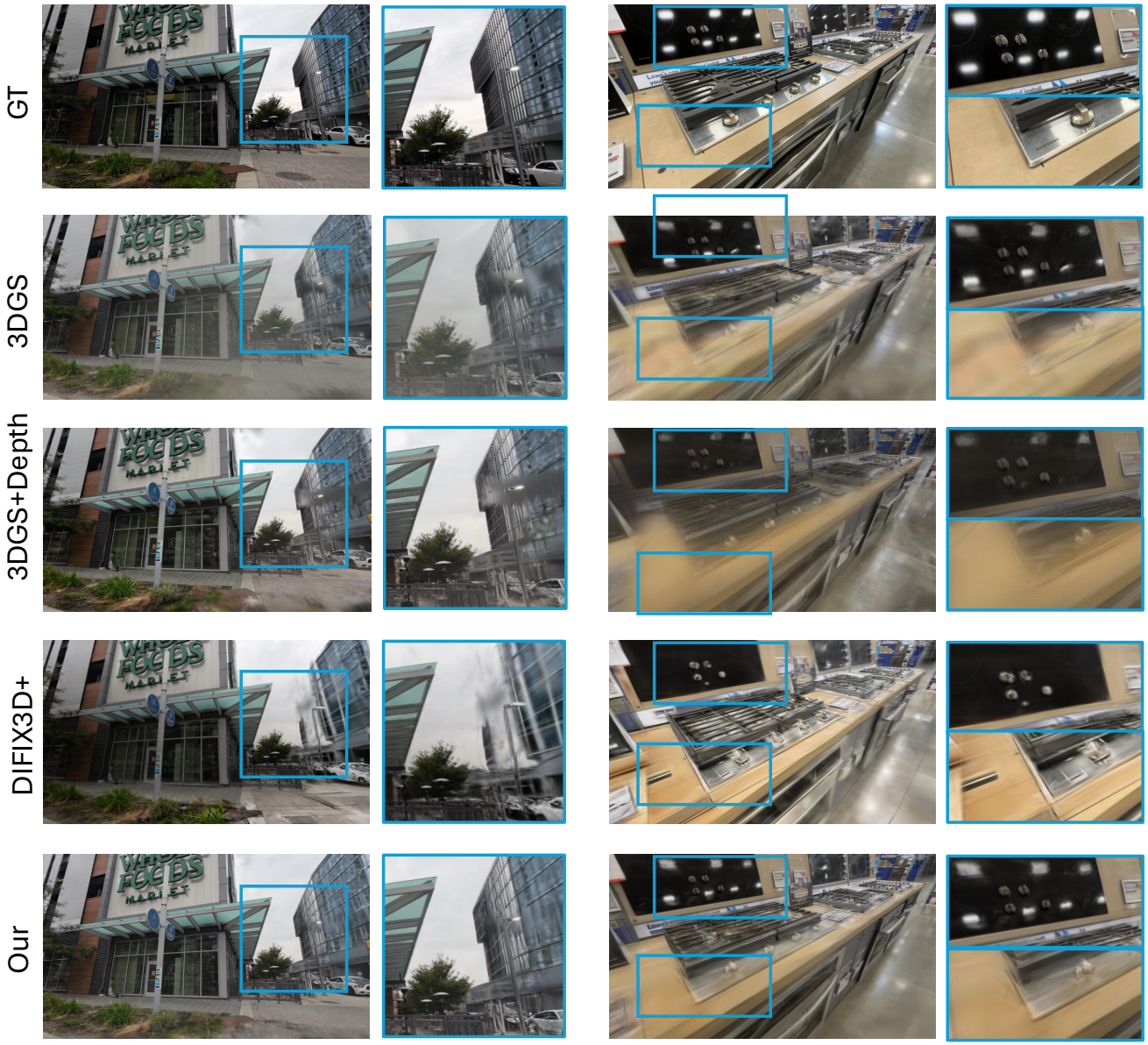

Figure 13. **Qualitative comparison on DL3DV dataset for per-scene reconstruction.** The blue bounding box indicates the zoom-in area. Despite the apparent clarity of DIFIX3D+, its progressive update and inaccurate reference image selection introduce significant hallucinated content, visible in the spurious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