SIR-DIFF: Sparse Image Sets Restoration with Multi-View Diffusion Model

Supplementary Material

1. Overview

In Sec. 2, we present comprehensive qualitative results for the tasks discussed in the main paper. Sec. 3 provides detailed descriptions of our training setup and explains the metrics used to evaluate our proposed method. In Sec. 4, we compare the performance of our method against video restoration methods using a standard video dataset. Further, Sec. 5 offers an in-depth explanation of our experiments on downstream applications, further demonstrating the versatility and effectiveness of our approach. Finally, Sec. 6 offers the ablation study for our proposed method, and Sec. 7 provides additional qualitative results.

2. Qualitative Visualization

2.1. Real World Motion Debluring

Figure 1. Qualitative comparison of SIR-Diff with Restormer on **in-the-wild real-world** images.

In Fig. 1, we show the performance of our model and baselines on deblurring the input on real images captured using a commodity smartphone. The deblurred output of our multi-view method is sharper than that of a single-view method Restormer [20].

2.2. Image Correspondences via LoFTR

In Fig. 2 and Fig. 3, we provide additional examples demonstrating how our model enhances the ability to LoFTR [14] model by identifying more corresponding points in low-resolution and motion-blurry sparse image sets.

2.3. Gaussian Splatting on Motion Blurring Images

Given a set of N degraded input views in Fig 4, we use SIR-Diff and the baseline method to restore the input image views. Using the restored image views, we run BAD-GS, a state-of-the-art Gaussian Splatting reconstruction method from blurry inputs. We show the output-rendered images from a novel view. We show that SIR-Diff can restore the blurry images consistently which leads to better rendering of the Gaussian Splatting output from a novel view and faster converge speed during training.

2.4. Sparse View Reconstruction from Degraded Images

In Fig. 5 and Fig. 6, we show the outputs reconstructing a sparse view 3D Gaussian Splatting method, InstantSplat [2]. We show that as compared to the baselines which do single view restoration for deblurring and super-resolution, SIR-Diff outshines them with much sharper and crisper results by enabling better sparse view reconstruction for InstantSplat [2].

3. Experiment Details

In this section, we provide concrete details of our experiment, which is composed of metrics explanation in Sec. 3.1 and training details in Sec. 3.2.

3.1. Metrics

To evaluate the self-consistency of the method, we propose the **Vision Consistency**, the metric that evaluates the consistency in RGB space, and the **Geometry Consistency** that evaluates the consistency of the 3D geometry.

Visual Consistency. We wish to evaluate the visual consistency between the restored images. The goal of this metric is to measure if two views that are geometrically consistent are also visually consistent.

Given two restored images, their ground truth depth maps and camera poses. We compute corresponding points between the images by establishing 3D correspondence using their 3D geometry. Naively measuring the difference in pixel values between corresponding pixels on the restored images is insufficient due to lighting and specularities.

To address these issues, we propose a method to evaluate the visual consistency of an image set. We compute ground-truth correspondences between two images, using the ground-truth geometry (depth) and pose, masking out all points within occluded regions. Each image is then divided into patches of size 30×30. Patches containing fewer than 300 corresponding points are discarded to ensure reliable evaluation. Given that the ground-truth correspondences within a patch remain sparse, we leverage these points to solve for a 2D affine transformation matrix, which is subsequently used to warp the patch from the source view to the target view, then the perceptual loss (LPIPS [21]) is computed between the warped ground-truth image patch and the target ground-truth image patch. Ground truth patches that really look like each other have a perceptual loss of less than 0.1. We use these patches for evaluation. This process is repeated across several patches and average patch-wise perceptual loss is reported as the final measure of consistency

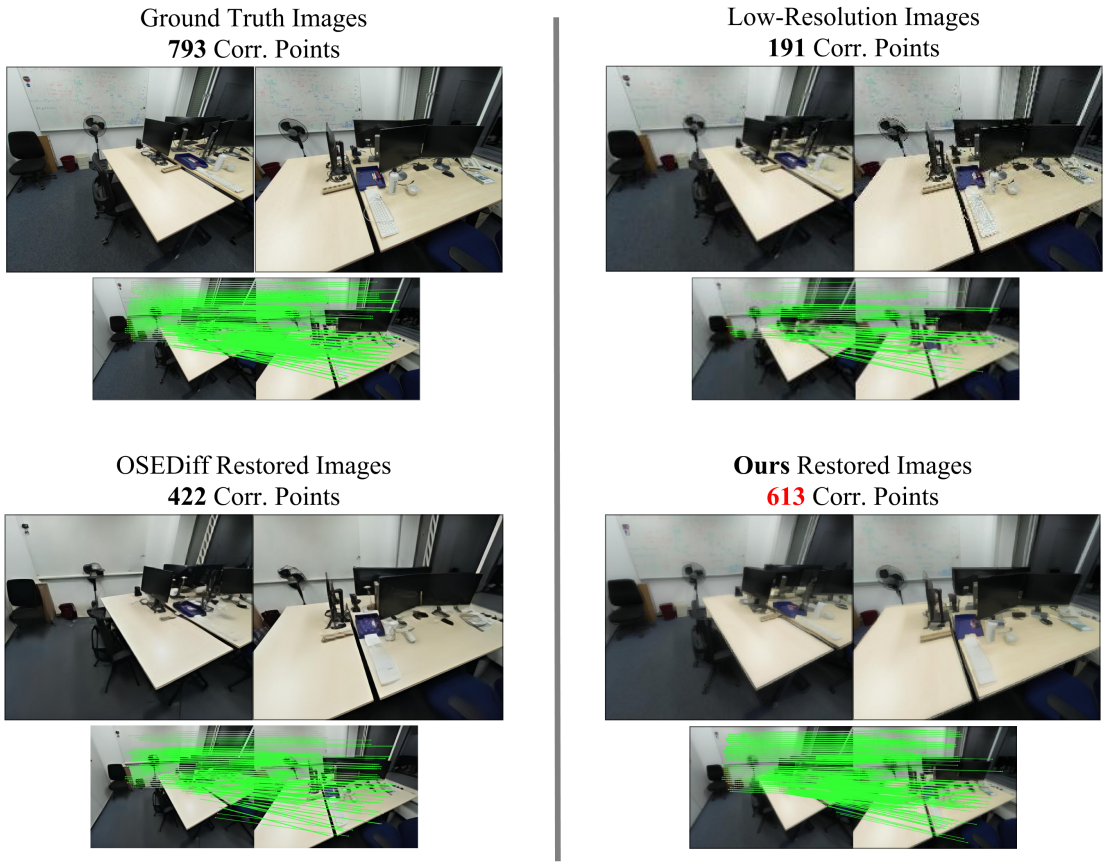

Figure 2. **Correspondence matching of LoFTR [14] on low-resolution images.** We run a recent correspondence matching algorithm [14] on the restored images using our and a baseline method (OSEDiff [18]). Note that the algorithm fails to detect matches in blurred images. While Restormer processed images enable better matching, only about half of the matches are restored compared to the ground truth image pair. Restored images using our method produce significantly more matches, leveraging our multi-view denoising scheme.

between the two images. The lower the value of the metric more consistent are the images.

Gemoetry Consistency. We evaluate the geometric consistency between two generated depth maps by evaluating the consistency of depth estimation for our results in §4.7 (main paper). We consider corresponding points between two images by using GT depth. If the depth discrepancy between two points is ≥ 0.1 meter, we consider that the points are not in correspondence but one point occludes the other point. This allows us to create an occlusion mask for the two views using the ground-truth geometry.

We now use the occlusion mask to evaluate the consistency between the generated depth images using our downstream application §4.7. of the main paper. Now for all the non-occluded points, we warp the generated depth map from the source view to the target view and evaluate the L1 distance between the warped depth map and the target depth map. We report the average consistency error in Tab. 6 (main paper). We compute this metric for pairs of views in our evaluating image set.

PSNR and SSIM for Deblurring. As mentioned in §4.4 of the main paper, for the evaluation of the Motion Deblurring task, we did not rely on traditional metrics such as PSNR and SSIM to assess the quality of image restoration. Previous study [21] has shown that traditional metrics often do not strongly correlate with the perceived visual quality of images. We observed a similar issue in the Motion Deblurring task. As illustrated in Fig. 7, these metrics can fail to capture the visual differences between restored images effectively. Additional visualizations can be found in Fig. 8. To address this limitation, we employed neural network-based perceptual metrics such as *Frechet Inception Distance (FID)* [3], *Learned Perceptual Image Patch Similarity (LPIPS)* [21] to evaluate the quality of image restoration in the Motion Deblurring task, providing a more robust and visually relevant assessment which can benefit the downstream task.

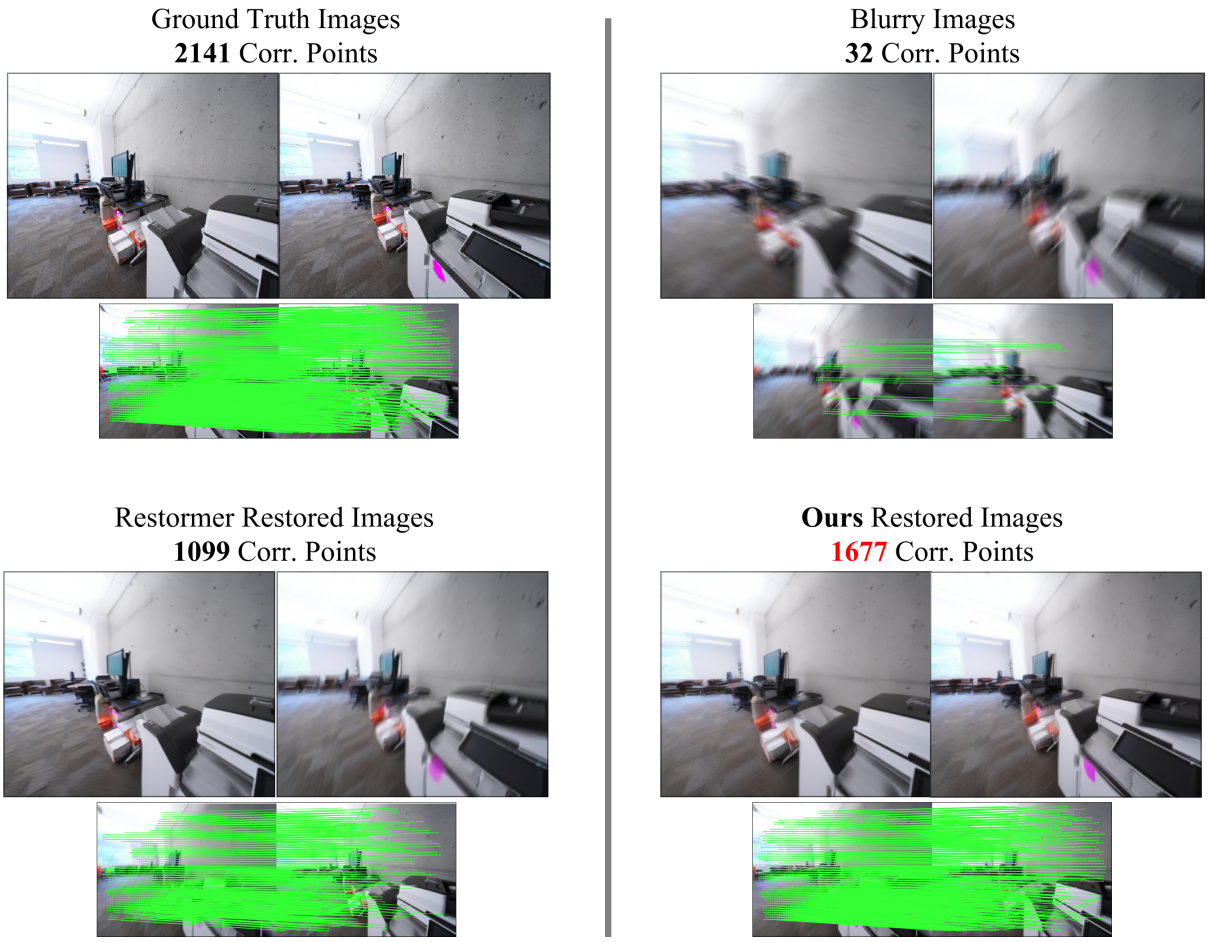

Figure 3. **Correspondence matching of LoFTR [14] on blurry images.** We run a recent correspondence matching algorithm [14] on the restored images using our and a baseline method (Restormer [20]). Note that the algorithm fails to detect matches in blurry images. While Restormer processed images enable better matching, only about half of the matches are restored compared to the ground truth image pair. Restored images using our method produce significantly more matches, leveraging our multi-view denoising scheme.

3.2. Training Details

We train our model on 2 synthetic datasets: Hypersim [12] and TartanAir [17]. We use all the samples from Hypersim, around 50k high-quality RGB images. For TartanAir, we randomly sample around 10k image sets from the origin dataset. All the images we use for training are resized into 480 $\times$ 640 resolution. Training our model takes 30k iterations with a batch size of 8 for each GPU and each instance contains 4 views. We use 2 $\times$ NVIDIA A100 40GB or 2 $\times$ NVIDIA L40S GPU for training. We use the Adam optimizer with a learning rate of $3 \cdot 10^{-5}$.

Image Set Selection Strategy. We follow the same image set selection strategy in both training and testing datasets illustrated as follows: For each image in the dataset, we designate it as the anchor view. Next, we iterate through the remaining images and calculate the overlap region between the selected image and the anchor image. If the overlap ratio falls within the range of [0.6, 0.8] for training, the image is added to the precomputed image list corresponding to the anchor image. This selection process continues until the list contains 8 images per anchor. This selection algorithm is executed prior to training. During training, we randomly sample 4 images per instance from the precomputed image list as the image set. For testing, we randomly select 4 frames from 20 frames which near by the reference frames. This view selection strategy helps the model implicitly learn geometric relationships across the image sets, leading to improved performance. For all the images selected in the images set, we random shuffle the order before training to enable this permutation invariance at inference.

4. Comparison on Traditional Video Dataset

As mentioned in §4.4 and §4.5 of the main paper, our main experiment shows the challenges faced by video-based image restoration methods when applied to image sets with large motion gaps and unordered inputs, as opposed to the

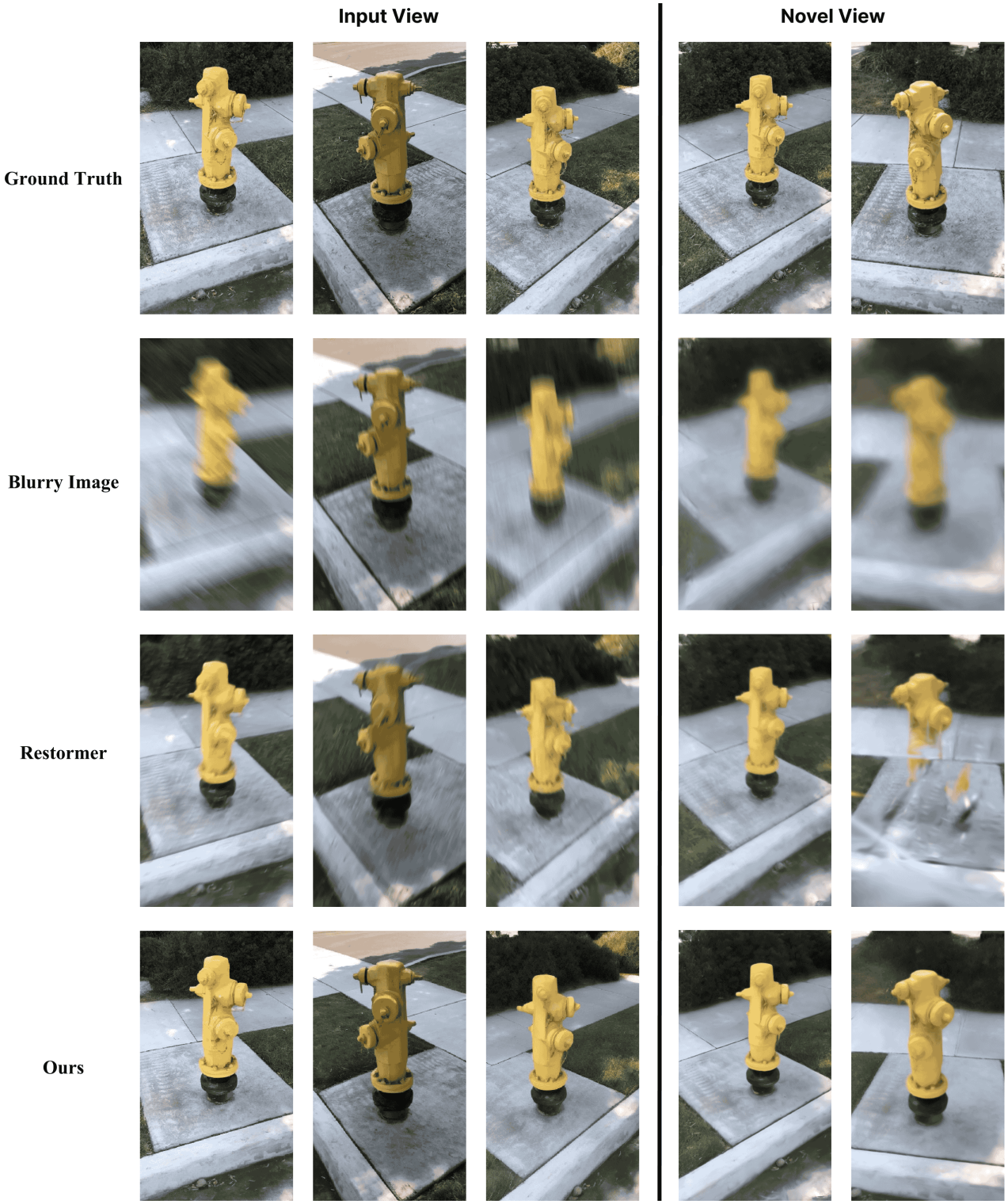

Figure 4. **Gaussian Splatting Reconstruction Comparisons.** When we use 102 blurry images as inputs for GS, the rendered novel views exhibit strong artifacts (2nd row). The quality improves when using Restormer [20] (3rd row) to deblur individual blurry images, but the rendering still includes artifacts compared to the ground truths (1st row). Our multi-view method (4th row) simultaneously deblurs all of the 102 images to produce consistent restorations, leading to higher-quality novel-view predictions.

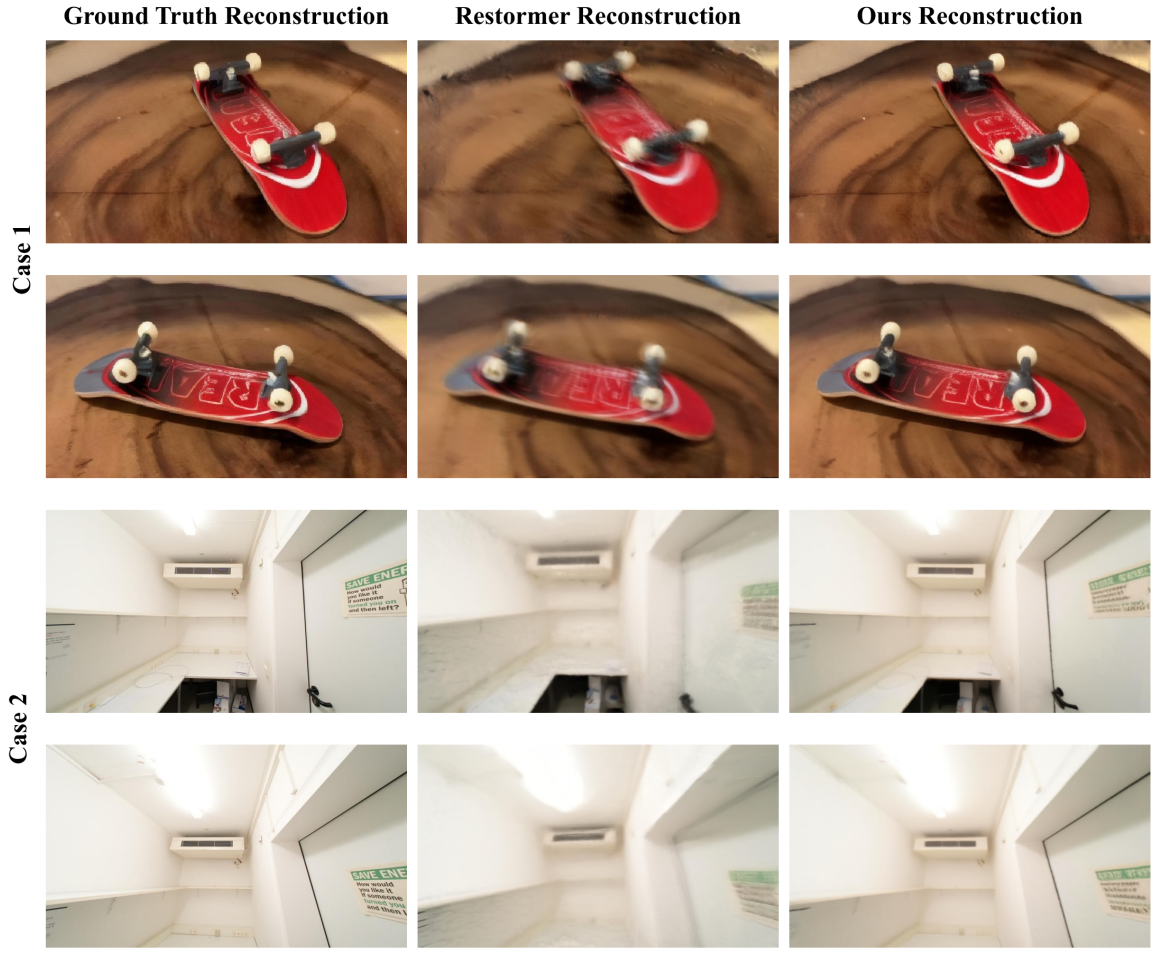

Figure 5. **Sparse-View GS Reconstruction Results Using InstantSplat [2] on Motion Deblurring**. We jointly deblur 9 input blurry images using SIR-Diff (3rd column), leading to sharp reconstruction qualities compared to using Restomer restored images (2nd column). 1st column shows reference reconstruction using GT images.

smoothed and ordered structure of ordinary videos. Methods such as Upscale-A-Video [23] and VRT [7] exhibit poor performance under these circumstances.

To ensure a fair comparison with video-based methods, we conducted experiments on traditional video datasets. Since the ScanNet++ [19] dataset is also captured in a video format, it was included in our evaluation. Additionally, we incorporated two traditional video restoration datasets: youHQ [23] and REDS [8], which were used to train Upscale-A-Video [23] and VRT [7], while our method was evaluated in a **zero-shot** setting.

For the ScanNet++ dataset, we selected 24 consecutive frames from each of the 50 scenes in the evaluation set, resulting in a total of 50 videos for the evaluation. For the youHQ [23] and REDS [8] datasets, we used their official evaluation splits and compared our results with the ones reported in their respective papers. Please refer to Tab. 1 for results compared to the video-based super-resolution method and Tab. 2 for results compared to the video-based motion deblurring method.

Table 1. **Comparison on Video Super-Resolution Dataset.** We present results on two video datasets: Scannet++ [19] and youHQ [23]. Note that for the Scannet++ [19] dataset, both methods operate in a zero-shot setting. In contrast, for the youHQ [23] dataset, Upscale-A-Video [23] is evaluated in-domain, while our method remains in a zero-shot setting.

	Scannet++			youHQ		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Upscale-A-Video [23]	26.41	0.8343	0.22	25.83	0.733	0.268
SIR-Diff	27.08	0.8456	0.1749	19.55	0.5249	0.4659

5. Down Stream Applications

5.1. Estimating Correspondences

We use the pre-trained LoFTR [14] model, trained on the *Indoor Dataset*, for evaluation. The same samples in Scannet++ [19] dataset from the main experiment are utilized. For each image set, we randomly sample two images to

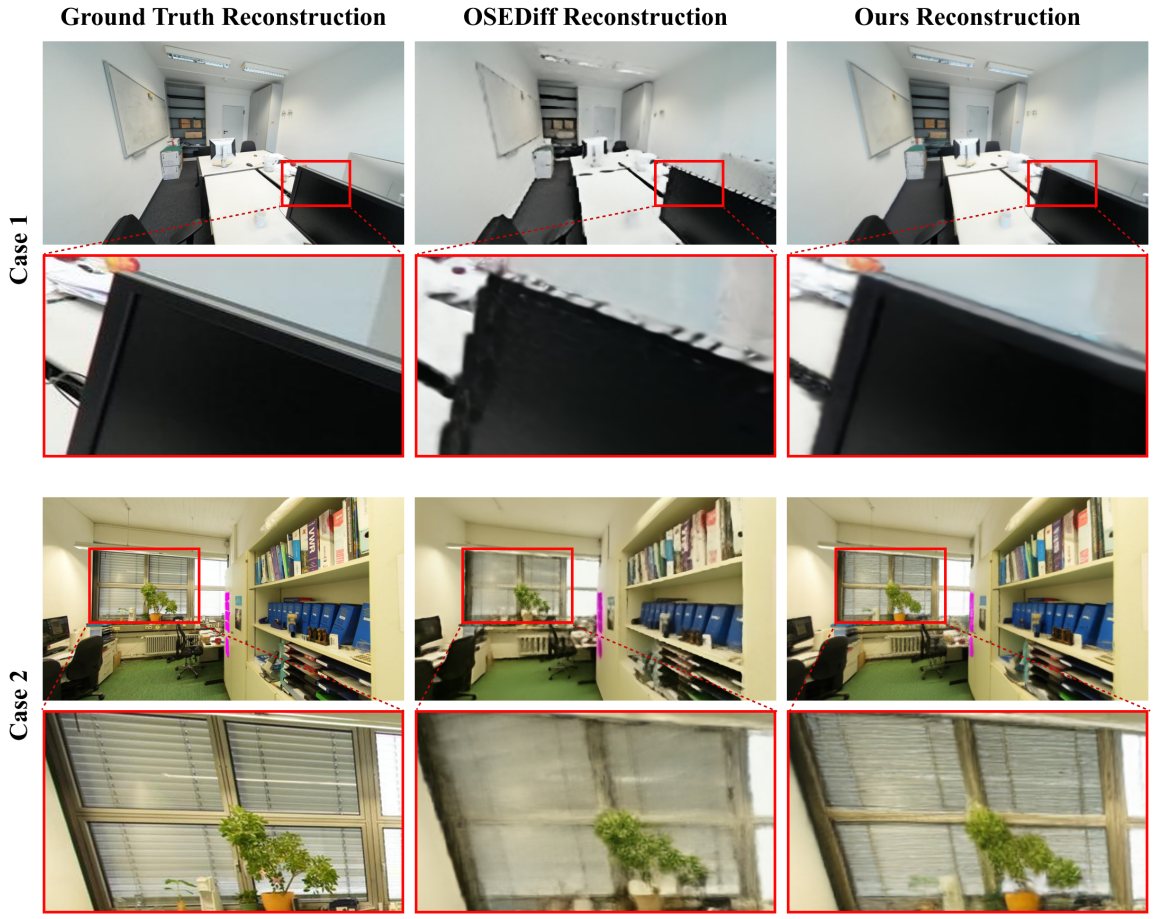

Figure 6. **InstantSplat [2] Reconstruction Results for Super-Resolution.** We ran InstantSplat to obtain 3DGS on 9 images that were restored to be high-resolution, using OSEDiff (2nd column) and SIR-Diff (3rd column). As can be seen from the inset zoom images, our multi-view super-resolution leads to more 3D-consistent restoration, leading to sharper 3D reconstruction results closer to the ground truth reconstruction renderings (1st column).

Ground Truth Image	Motion Blurry Image	Restormer	Ours
PSNR / SSIM / LPIPS	**23.22 / 0.589 / 0.499**	**23.28 / 0.585 / 0.461**	**22.36 / 0.531 / 0.262**

Figure 7. **Ineffectiveness of PSNR and SSIM as Deblurring Metrics.** The best result is marked as red for each metric. Ours show sharp and clear restored results but do not show an advantage on PNSR↑ and SSIM↑ than [20], while LPIPS↓ behaves as expected.

compute correspondences.

First, we run the LoFTR [14] model on ground-truth high-quality RGB images to determine the number of correspondences that the model can find under ideal conditions. Then, we apply a single-view image restoration model and our proposed SIR-Diff to restore the degraded images. The LoFTR [14] model is subsequently run on the restored images to compute the number of correspondences, which serves as an indicator of the quality of the restored images. The results are presented in Tab.4 of the main paper.

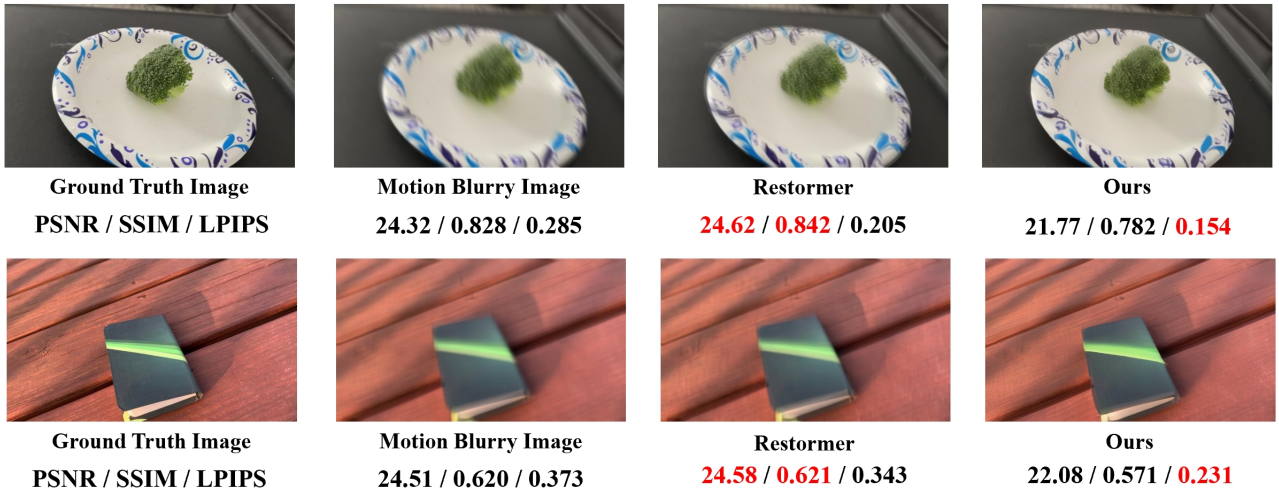

Figure 8. **Ineffectiveness of PSNR and SSIM as Deblurring Metrics (Additional Examples).** The best result is marked as red for each metric. Ours show sharp and clear restored results but do not show an advantage on PNSR↑ and SSIM↑, while LPIPS↓ behaves as expected.

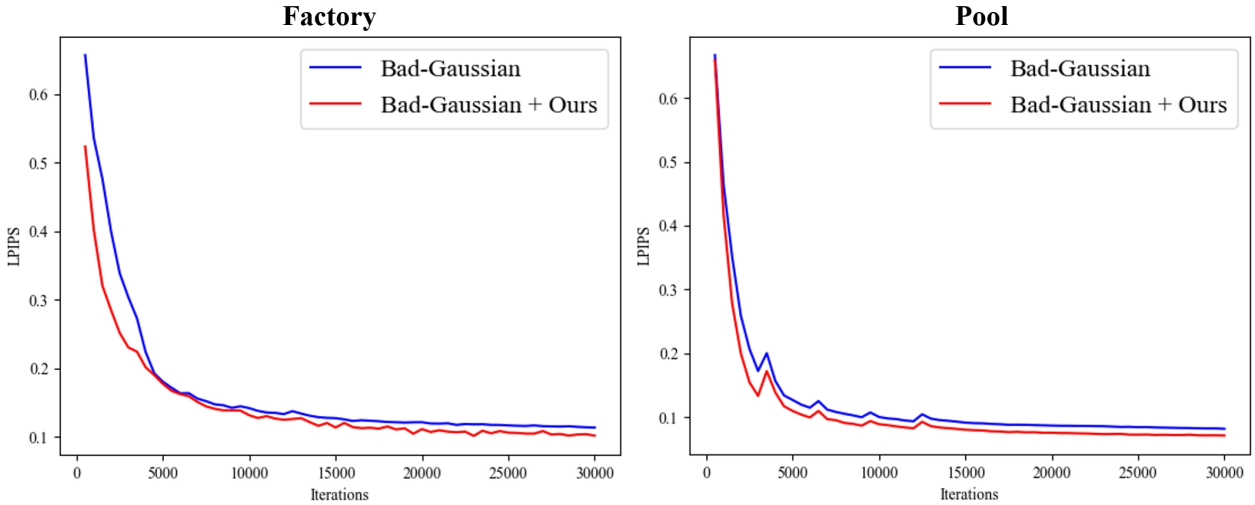

Figure 9. **Reconstruction Accuracy vs. Iterations.** Applying our multi-view deblurring helps the BAD-Gaussians [22] algorithm to converge faster and improve reconstruction accuracy measured by LPIPS on the two scenes from [9].

Table 2. **Comparison on Video Motion-Deblurring Dataset.** We present results on two video datasets: Scannet++ [19] and REDS [8]. Note that for the Scannet++ [19], both methods operate in a zero-shot setting. In contrast, for the REDS [8] dataset, VRT [7] is evaluated in-domain, while our method remains in a zero-shot setting.

	Scannet++			REDS		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
VRT [7]	25.42	0.8470	0.2855	36.79	0.9648	-
SIR-Diff	26.72	0.8351	0.1840	19.60	0.4948	0.3465

5.2. Gaussian Splatting on Reconstruction Motion Blurring Images

We follow the original experimental settings provided in the official implementation of BAD-Gaussians [22] (BAD-GS), which is based on NeRFStudio [15], and conduct experiments using the Deblur-NeRF dataset [9]. The per-scene result is presented in Tab. 3.

Additionally, we observe that our method accelerates the convergence of the BAD-GS training process. To illustrate this, as shown in Fig. 9, we plot the LPIPS [21] loss curve during the training of BAD-GS on the *Factory* and *Pool* scenes. The results demonstrate that the incorporation of our model improves the convergence speed, underscoring its

Table 3. **Per-scene result of Deblur-NeRF dataset** [9]. The best result is highlighted.

Difficulty	cozyroom			tanabata			Pool			Factory		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BAD-GS [22]	**31.16**	**0.932**	0.042	**24.03**	**0.777**	0.125	**32.36**	**0.896**	0.104	**29.61**	**0.895**	0.115
SIR-Diff + 3DGS [6]	25.83	0.825	0.091	21.09	0.654	0.258	28.79	0.800	0.126	23.37	0.707	0.200
SIR-Diff + BAD-GS	30.38	0.931	**0.041**	21.96	0.698	**0.125**	31.97	0.891	**0.082**	28.37	0.834	**0.104**

Table 4. **Additional Gaussian Splatting Reconstruction Result from Blurry Images.** The best is highlighted. We additionally provide the results of the single-image motion deblurring model (Restormer [20]) with BAD-GS [22].

Difficulty	Medium			Hard		
Params	KS:[30, 10.2], Inten:[0, 0.4]			KS:[45,14.85], Inten:[0, 0.5]		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS [6]	13.47	0.610	0.607	10.42	0.421	0.801
BAD-GS [22]	10.06	0.537	0.687	8.061	0.385	0.948
Resotrmer [20]+BAD-GS [22]	25.77	0.616	0.356	23.88	0.560	0.317
SIR-Diff + BAD-GS [22]	**26.11**	**0.661**	**0.250**	**25.33**	**0.644**	**0.277**

Table 5. **Ablation Study.** The best is highlighted.

Genre	PSNR ↑	SSIM ↑	LPIPS ↓	Vconsis ↓
W/O 3D Convolution (4 Views)	26.29	0.806	0.137	5.28
W/O SVD Init. (4 Views)	26.28	0.813	0.192	6.03
1 View	25.85	0.792	0.164	5.70
2 Views	26.64	0.840	0.134	5.41
4 Views	27.38	0.837	**0.099**	**5.01**
8 Views	**28.67**	**0.842**	0.130	5.04

effectiveness in enhancing the training process.

As discussed in Sec.4.6.2 of the main paper, we also evaluate BAD-GS under high-intensity motion blurring conditions. To simulate motion blurring, we apply varying strengths of motion blur kernels to sharp images, use our proposed SIR-Diff to deblur the images, and then perform reconstruction using COLMAP. The results in Tab.3 of the main paper reveal that our method achieves robust and consistent performance across different intensity levels of motion blurring. While certain single-view methods also perform well, their lack of self-consistency in restored images results in inferior reconstruction performance compared to our approach. Detailed results can be found in Tab. 4.

For all experiments in this section, the reported results include not only the performance on **Novel View Synthesis** but also the performance on deblurring the **Training Views**. By considering both aspects, we provide a comprehensive evaluation of our method's effectiveness in addressing motion-blurring effects across the entire dataset.

5.3. Sparse-View 3D reconstruction from Degraded Images

We build our method based on the officially released code of InstantSplat [2]. We randomly select 3 scenes from Scannet++ dataset [19] and 3 objects from CO3Dv2 dataset [11]. Following the original setting of InstantSplat [2], we first randomly sample 24 images as the training-evaluation set, then randomly sample 9 views from it as the training views and the rest views for evaluation of Novel View Synthesis. For the motion deblurring reconstruction task, we apply the same intensity of the blurring kernel on the ground-truth sharp images from training views in which the blurring kernel size is chosen from a normal distribution with mean 85px and standard deviation 12.75px, and the intensity of the blurring is randomly sampled in the range $[0, 1]$. For the super-resolution reconstruction task, we downsample the original high-quality image with a ratio of 4. We use our SIR-Diff model and the best single-view image restoration mothed to restore the degraded image set and use it for reconstruction. All the experiment results are reported in **Novel View Synthesis** setting. We also report our per-scene result of super-resolution in Tab 6 and the per-scene result of Motion Deblurring in Tab 7.

6. Ablation Study

To verify the effectiveness of each modality we propose, we do several ablation studies: (1) Does the 3D convolution layer in our spatial-3D ResNet help? (2) Does the SVD [1] weight initialization on the 3D convolution layer help? (3) What is the performance variation with different numbers of views in the 3D self-attention Transformer during the inference?

We conduct an ablation study on the Scannet++ dataset [19] with a super-resolution task. We used the same split and image set in the main experiment. As shown in Tab. 5, without a 3D convolution layer, the performance on 4 views restoration drops greatly on all metrics. This proves that only deploying the default 2D convolution layer from Latent Diffusion Models like SD2.1 [13] is limited in 3D tasks. However, the performance is still worse even if we have a 3D convolution layer, but don't initialize the weight with a reasonable pre-trained model. From here, we also show that even though the scheduler that SVD uses (EDM [5]) is different from the DDPM [4] that we use on the training, the stability of training and the performance improvement can still be maintained. Furthermore, the results from 1 view to 8 views show that our model can handle different numbers of views, as the input and overall performance improvement are proportional to the number of views available.

Table 6. **Per-Scene Result for the Super-Resolution Reconsturction on InstantSplat [2].** The best excluding the GT is highlighted.

Scannet++ [19]	825d228aec			acd95847c5			d755b3d9d8		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Ground Truth Image	0.976	35.25	0.065	0.953	33.33	0.097	0.892	27.02	0.127
OSEDiff [18]	0.917	26.78	0.217	0.812	26.08	0.243	0.669	20.75	0.364
Ours	**0.923**	**28.41**	**0.163**	**0.876**	**27.39**	**0.204**	**0.684**	**22.82**	**0.334**
CO3D [11]	Bench			Hydrant			Skateboard		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Ground Truth Image	0.758	25.54	0.224	0.697	21.09	0.277	0.818	25.20	0.230
OSEDiff [18]	0.486	20.82	0.393	0.466	17.98	0.417	0.695	21.96	0.329
Ours	**0.502**	**22.70**	**0.371**	**0.508**	**19.49**	**0.396**	**0.725**	**23.15**	**0.305**

Table 7. **Per-Scene Result for the Motion Deblurring Reconsturction on InstantSplat [2].** The best despite the GT is highlighted.

Scannet++ [19]	825d228aec			acd95847c5			d755b3d9d8		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Ground Truth Image	0.976	35.25	0.065	0.953	33.33	0.097	0.892	27.02	0.127
Restormer [20]	0.869	20.52	0.305	0.830	24.95	0.253	0.669	20.62	0.370
Ours	**0.931**	**29.00**	**0.165**	**0.885**	**27.76**	**0.206**	**0.713**	**22.82**	**0.323**
CO3D [11]	Bench			Hydrant			Skateboard		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Ground Truth Image	0.758	25.54	0.224	0.697	21.09	0.277	0.818	25.20	0.230
Restormer [20]	0.516	20.62	0.490	0.400	11.00	0.696	0.519	14.54	0.567
Ours	**0.545**	**22.00**	**0.406**	**0.498**	**18.34**	**0.476**	**0.669**	**20.63**	**0.352**

7. Additional Qualitative

Fig. 10 and Fig. 11 show extra qualitative results of our SIR-Diff on motion deblurring. Fig. 12 and Fig. 13 show extra qualitative results of our SIR-Diff on super-resolution task. Our SIR-Diff show richer details in super-resolution and sharp deblurring restoration quality than other methods on both object level [11] and scene level [19] testing datasets.

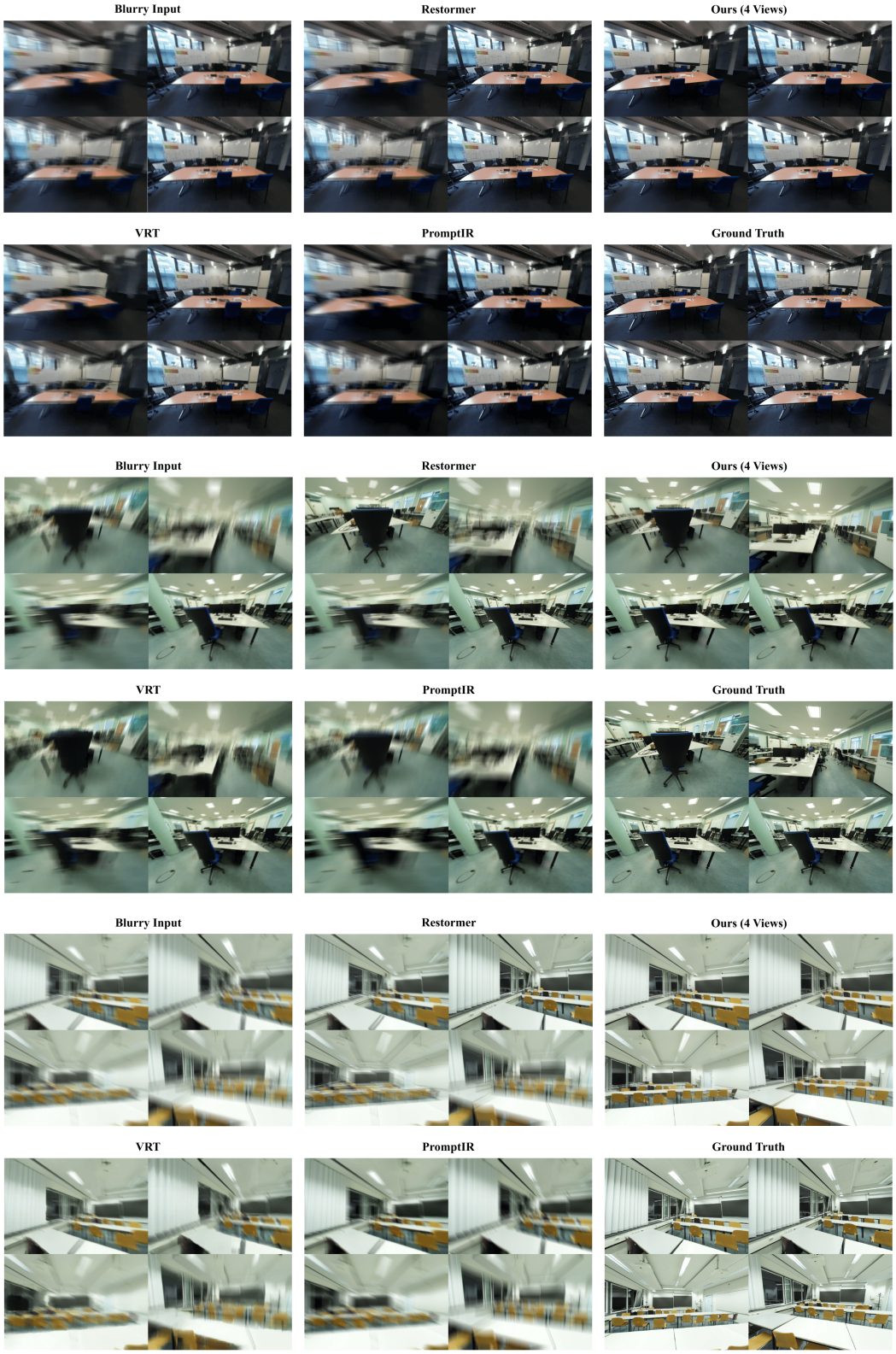

Figure 10. **Additional Deblurring Results 1.** Our method uses all 4 input views to jointly denoise the images, performing significantly better than existing single-image-based methods [7, 10, 20].

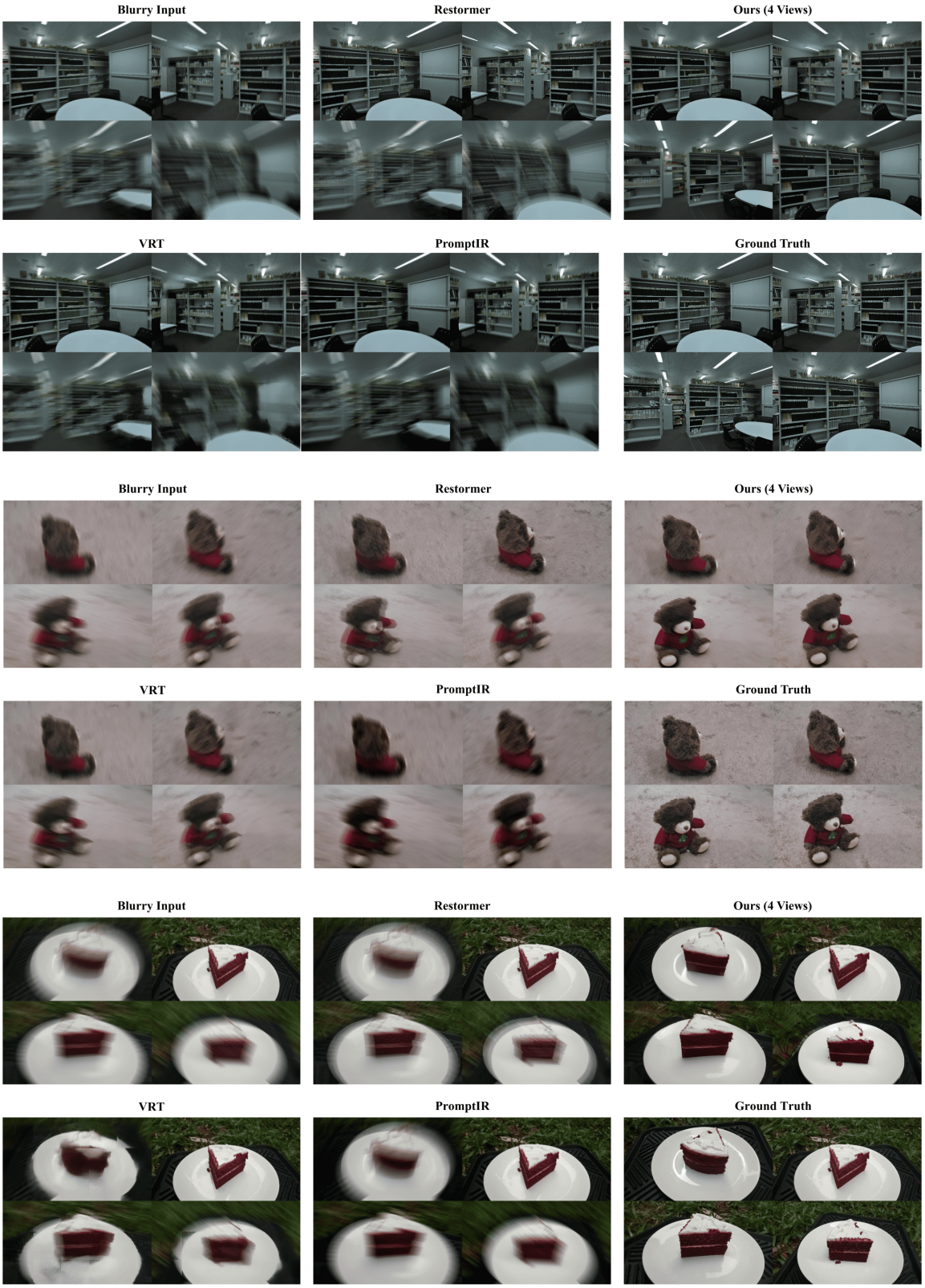

Figure 11. **Additional Deblurring Results 2.** Our method uses all 4 input views to jointly denoise the images, performing significantly better than existing single-image-based methods [7, 10, 20].

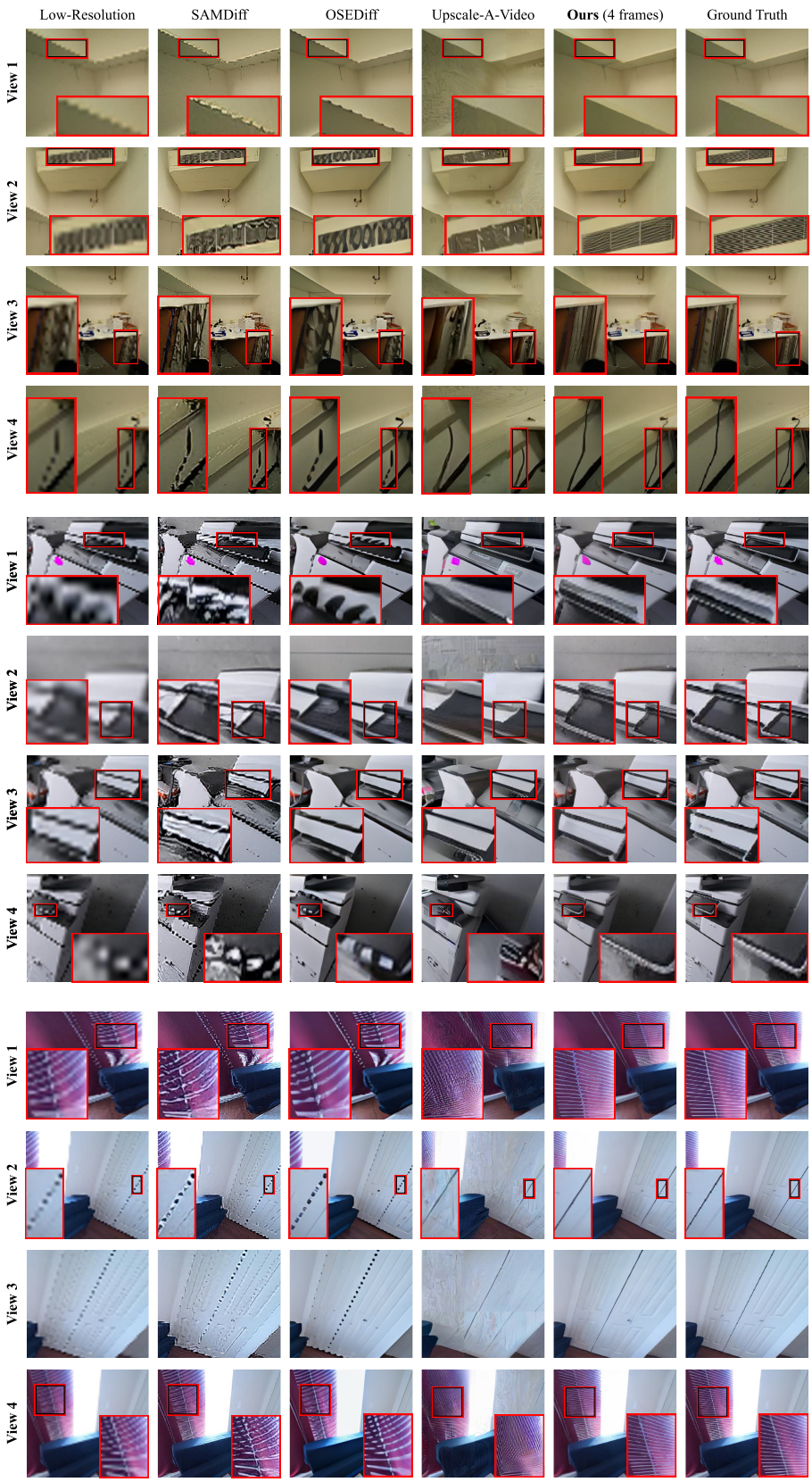

Figure 12. **Additional Super-Resolution Comparison Results 1.** Our method uses all 4 input views to jointly denoise the images, performing significantly better than existing single-image-based methods [16, 18, 23]. Zoom in for the best view.

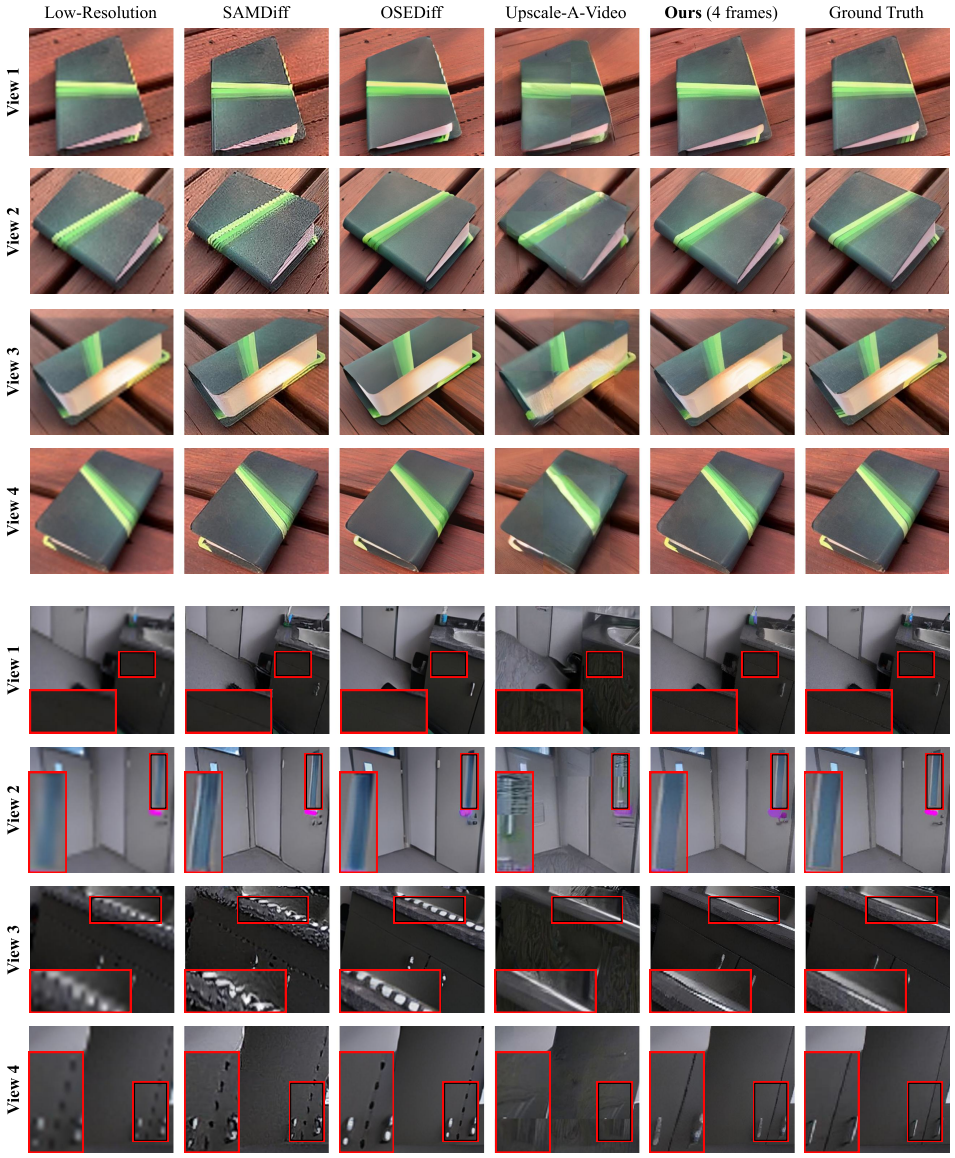

Figure 13. **Additional Super Resolution Comparison Results 2.** Our method uses all 4 input views to jointly denoise the images, performing significantly better than existing single-image-based methods [16, 18, 23]. Zoom in for the best view.