

SPARS3R: Semantic Prior Alignment and Regularization for Sparse 3D Reconstruction

Supplementary Material

A. DreamSim Metric Evaluation

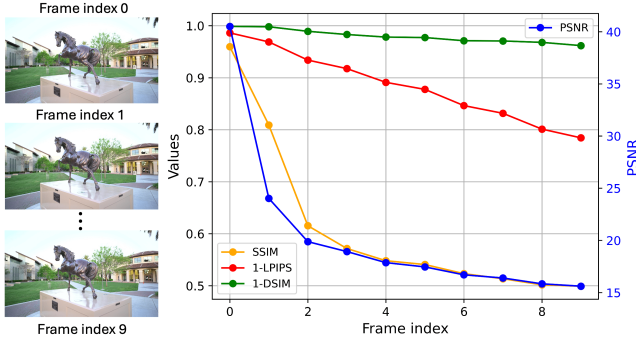


Figure A. Evaluation of different metrics to camera pose shift. We extract a sequence of images with a small pose change at each step and set the first frame as the reference. PSNR, SSIM, LPIPS, and DSIM are computed. DSIM shows robustness to small pose shifts by the flattest line.

In sparse settings, inaccurate camera pose estimation causes rendering shifts. To evaluate image quality while mitigating the impact of these shifts, we propose to use DreamSim [3] (DSIM) as an additional metric to assess render quality. DSIM seeks to represent human perceptual similarity by finetuning a combination of embeddings from visual foundation models based on human evaluations. As demonstrated in Fig. A, given ground-truth images with tiny pose changes, PSNR and SSIM drop *significantly* and cannot express the quality of the images. Despite also measuring perceptual similarity, LPIPS is not as pose-shift invariant as DSIM, likely due to the patch-based convolutional design. In comparison, DSIM is the most pose-shift invariant metric out of the four. This demonstrates that, in addition to being more aligned with human perception, DSIM remains robust to moderate pose shifts when images are highly similar. Given the inevitable small errors in pose alignment, especially in challenging sparse-view scenarios, this robustness is particularly valuable for our evaluations.

An effective evaluation metric should also reliably reflect poor image quality when images are dissimilar. To assess this, we further evaluate DSIM under various perturbations, including blurriness, overexposure, and compression. As shown in Fig. B, as the degree of perturbation increases, (1-DSIM) drops consistently, similar to PSNR, SSIM [10], and LPIPS [13]. Specifically, DSIM shows a more consistent response to varying degrees of blurriness, as illustrated in Fig. Ba, where we introduce a blob of Gaussian blur to

simulate a large semi-transparent Gaussian floater.

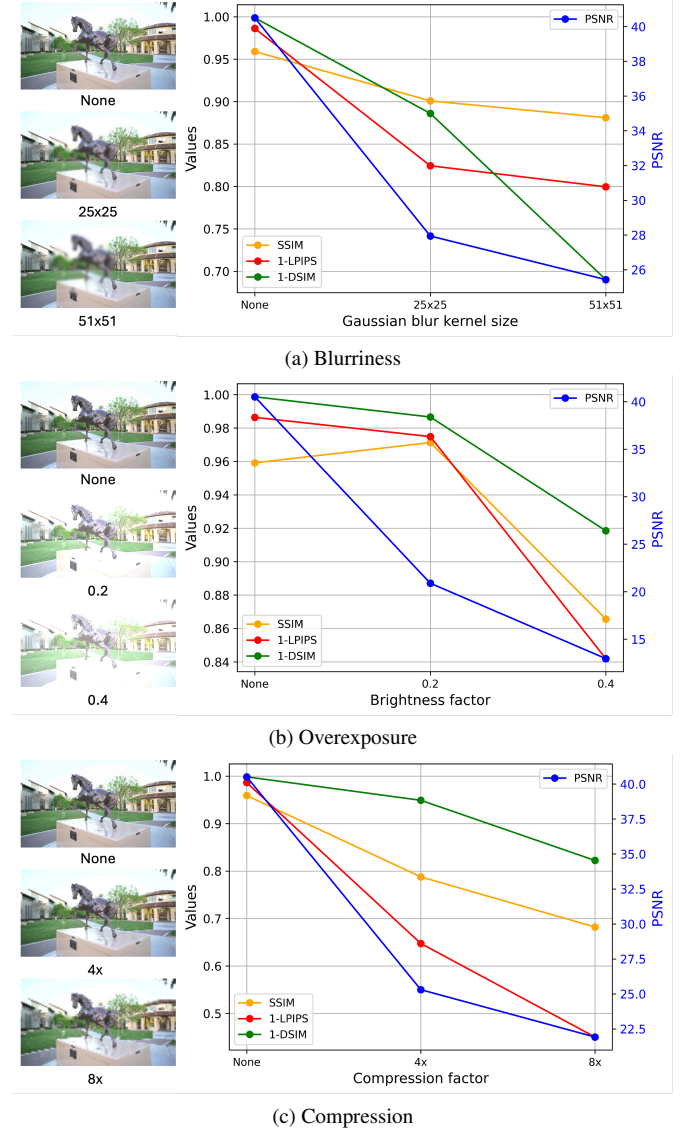


Figure B. DSIM Metric Evaluation

B. Additional Results

In this section, we provide detailed quantitative results for each scene on *Mip-Nerf 360* [1], *Tanks & Temples* [7] and *MVImgNet*[12], under 12 views, evaluated by commonly employed PSNR, SSIM and LPIPS as well as an additional metric DSIM in Tab. A, Tab. B and Tab. C, respectively. Notably, SPARS3R not only achieves superior per-

formance on average, as shown in Tab. 4, but also consistently outperforms other methods across individual scenes. We also assess the pose accuracy from Relative Translation Error (RPE_t), Relative Rotation Error (RPE_r), and Absolute Trajectory Error (ATE) three aspects in Tab. D, Tab. E and Tab. F, demonstrating the difficulty in camera registration in sparse view and MAST3R+COLMAP generally estimates more accurate poses.

C. Mask Generated in Semantic Outlier Alignment Step

In the Semantic Outlier Alignment step, we incorporate the Segment Anything Model (SAM) [6] to iteratively group outliers into distinct semantic masks.

The detailed procedure is illustrated in Fig. D. Starting with a random point from the outlier SfM set $\bar{x}_1 = \bar{P}(\bar{X}_1)$, $\bar{X}_1 \in \bar{O}$, marked as a green star, we prompt SAM to generate the first mask (m_1), shown in purple. Points within this mask are excluded in subsequent iterations (colored gray). In the next iteration, we select another point $\bar{x}_k \in \bar{P}(\bar{O})$ and $\bar{x}_k \notin \bigcup_{j=1}^{k-1} m_j$, producing the second mask (orange). Repeating this process generates additional masks (green and blue). Finally, the inlier area is defined as $m_0 = \neg \left(\bigcup_{k=1}^M m_k \right)$, colored in gray.

Visualizations of these SAM-generated masks are shown in Fig. E. For the *Bonsai* scene, points associated with the floor, bicycle, and chair exhibit significant alignment errors during the Global Fusion step. These outlier points serve as prompts for SAM, yielding four semantically distinct masks, each corresponding to coherent object regions. Similarly, in the *Family* scene, points in the background, flagged as outliers during Global Fusion, are processed by SAM to generate masks for bushes, trees, and buildings. This process highlights the ability of our method to identify semantically consistent regions. These segmented regions facilitate piecewise alignment for background elements, effectively addressing relative discrepancies caused by MAST3R’s inaccuracies in estimating relative depths.

D. Supplementary Videos

We provide a video demonstration highlighting the impact of our two-stage approach for point cloud alignment. The comparison showcases the results after only the first stage, Global Fusion Alignment, versus the full pipeline incorporating both stages. When only the first stage is applied, the point cloud retains a substantial number of **outliers**, indicating incomplete alignment. However, upon integrating the second stage, Semantic Outlier Alignment, these previously flagged outliers are successfully aligned as **inliers**, resulting in a clean and coherent point cloud. This transformation underscores the robustness and effectiveness of our proposed

two-stage refinement approach in addressing alignment errors and ensuring high-quality reconstruction.

We provide video renderings to showcase our 3D reconstruction results in comparison with state-of-the-art methods, FSGS [14] and InstantSplat [2]. These visualizations highlight the superior quality of our approach, particularly in preserving fine-grained details and achieving higher coherence in reconstructed scenes. These comparisons allow for a clear evaluation of performance differences, demonstrating our method’s ability to handle challenging scenarios effectively. Please refer to the attached videos for a comprehensive visual understanding of the compelling outcomes delivered by our approach.



Figure C. Visual comparisons of different NVS methods. The first two rows are from MipNeRF360 [1], the second two rows are from Tanks & Temples [7]. The last is from MVimgNet [12]

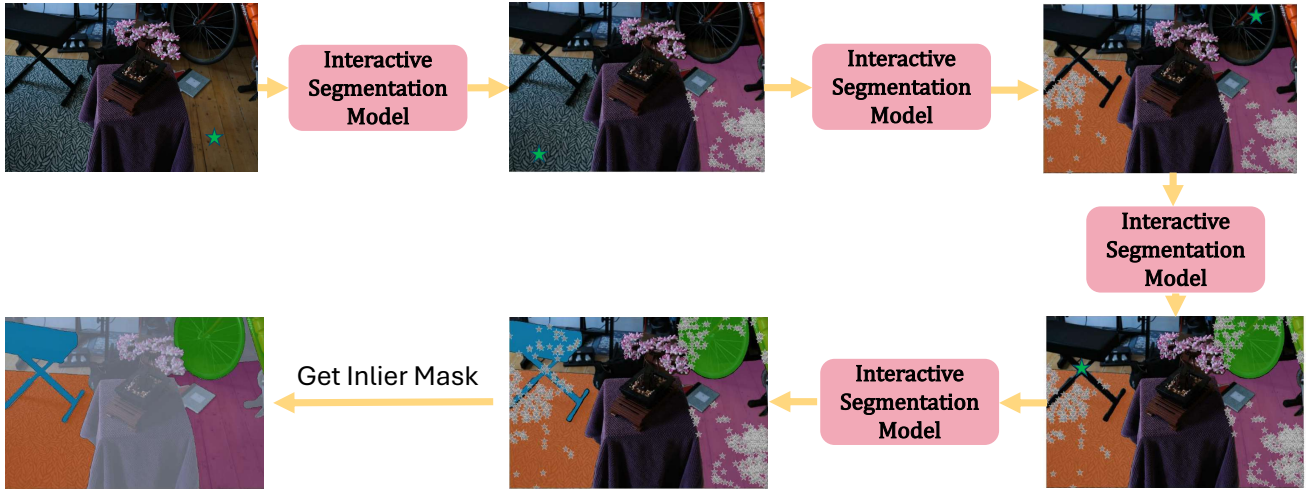


Figure D. Illustration of the iterative SAM-based segmentation process to generate semantically coherent regions for local alignments.

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. 1, 3

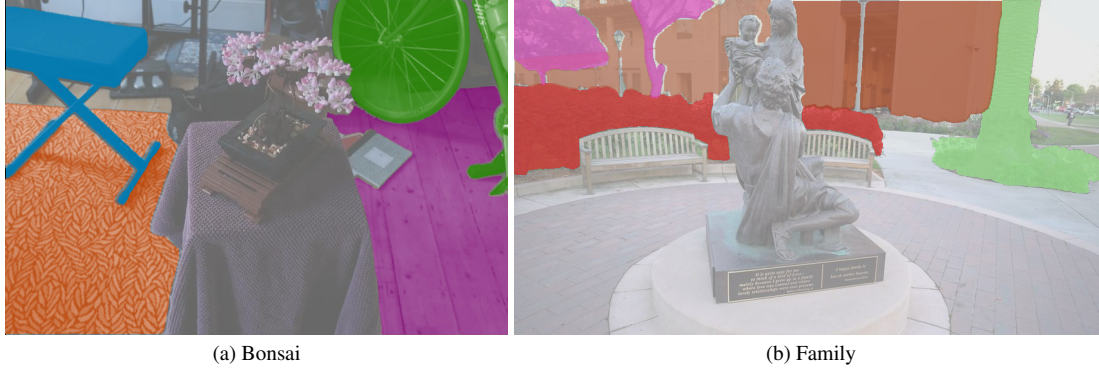


Figure E. Visualization of SAM-generated masks: Colored regions represent outlier-prompted masks, while inlier areas are shown in gray for contrast.

- [2] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2, 2024. [2](#), [5](#), [6](#), [7](#)
- [3] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. [1](#)
- [4] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. [5](#), [6](#), [7](#)
- [5] 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. [5](#), [6](#), [7](#)
- [6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. [2](#)
- [7] 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. [1](#), [3](#)
- [8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. [5](#), [6](#), [7](#)
- [9] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. [5](#), [6](#), [7](#)
- [10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. [1](#)
- [11] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360 {deg} sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*, 2023. [5](#), [6](#), [7](#)
- [12] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. [1](#), [3](#)
- [13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [1](#)
- [14] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2025. [2](#), [5](#), [6](#), [7](#)

	PSNR									
	Outdoor					Indoor				
	<i>bicycle</i>	<i>flowers</i>	<i>garden</i>	<i>stump</i>	<i>treehill</i>	<i>room</i>	<i>counter</i>	<i>kitchen</i>	<i>bonsai</i>	
Instant-NGP [9]	15.07	12.64	13.86	15.89	15.40	14.37	14.83	17.17	14.19	
3DGS [5]	15.86	12.43	17.03	18.21	14.44	18.60	17.17	18.13	17.27	
FSGS [14]	17.71	13.76	18.83	18.77	16.64	19.09	17.28	17.73	18.55	
SparseGS [11]	16.87	13.52	17.14	18.20	15.14	17.69	17.17	17.89	16.32	
DRGS [8]	16.73	13.70	17.67	18.49	14.95	18.30	17.51	18.28	16.29	
CF-3DGS [4]	13.37	11.32	14.67	17.40	13.07	11.25	11.76	13.45	13.11	
InstantSplat [2]	15.25	13.18	17.47	16.33	15.06	18.32	16.92	18.11	15.40	
SPARS3R	18.42	14.67	20.25	20.18	16.25	20.43	19.64	20.73	19.11	

	SSIM									
	Outdoor					Indoor				
	<i>bicycle</i>	<i>flowers</i>	<i>garden</i>	<i>stump</i>	<i>treehill</i>	<i>room</i>	<i>counter</i>	<i>kitchen</i>	<i>bonsai</i>	
Instant-NGP [9]	0.189	0.127	0.219	0.261	0.270	0.430	0.354	0.451	0.340	
3DGS [5]	0.254	0.138	0.336	0.313	0.266	0.593	0.498	0.565	0.528	
FSGS [14]	0.333	0.198	0.424	0.343	0.364	0.650	0.518	0.586	0.573	
SparseGS [11]	0.289	0.183	0.366	0.335	0.310	0.583	0.515	0.566	0.501	
DRGS [8]	0.301	0.210	0.352	0.341	0.315	0.620	0.518	0.505	0.450	
CF-3DGS [4]	0.187	0.119	0.205	0.254	0.250	0.338	0.327	0.258	0.314	
InstantSplat [2]	0.220	0.168	0.320	0.212	0.281	0.637	0.502	0.485	0.407	
SPARS3R	0.367	0.213	0.516	0.412	0.373	0.724	0.626	0.673	0.596	

	LPIPS									
	Outdoor					Indoor				
	<i>bicycle</i>	<i>flowers</i>	<i>garden</i>	<i>stump</i>	<i>treehill</i>	<i>room</i>	<i>counter</i>	<i>kitchen</i>	<i>bonsai</i>	
Instant-NGP [9]	0.650	0.705	0.781	0.717	0.653	0.685	0.691	0.513	0.702	
3DGS [5]	0.553	0.621	0.407	0.461	0.620	0.367	0.377	0.319	0.399	
FSGS [14]	0.697	0.820	0.491	0.757	0.735	0.360	0.445	0.318	0.398	
SparseGS [11]	0.531	0.611	0.420	0.455	0.638	0.387	0.383	0.313	0.413	
DRGS [8]	0.754	0.901	0.637	0.706	0.808	0.441	0.525	0.489	0.580	
CF-3DGS [4]	0.747	0.669	0.602	0.555	0.750	0.753	0.784	0.701	0.723	
InstantSplat [2]	0.697	0.748	0.514	0.623	0.673	0.363	0.413	0.311	0.548	
SPARS3R	0.362	0.472	0.235	0.369	0.472	0.267	0.273	0.182	0.313	

	DSIM									
	Outdoor					Indoor				
	<i>bicycle</i>	<i>flowers</i>	<i>garden</i>	<i>stump</i>	<i>treehill</i>	<i>room</i>	<i>counter</i>	<i>kitchen</i>	<i>bonsai</i>	
Instant-NGP [9]	0.569	0.433	0.762	0.498	0.446	0.604	0.612	0.229	0.562	
3DGS [5]	0.310	0.307	0.126	0.246	0.352	0.224	0.229	0.100	0.197	
FSGS [14]	0.259	0.429	0.122	0.366	0.259	0.212	0.247	0.113	0.178	
SparseGS [11]	0.212	0.230	0.139	0.222	0.290	0.255	0.222	0.109	0.212	
DRGS [8]	0.327	0.371	0.212	0.323	0.387	0.262	0.318	0.177	0.325	
CF-3DGS [4]	0.639	0.393	0.360	0.440	0.469	0.620	0.647	0.542	0.471	
InstantSplat [2]	0.343	0.223	0.154	0.337	0.263	0.177	0.211	0.087	0.307	
SPARS3R	0.140	0.153	0.064	0.142	0.191	0.152	0.121	0.041	0.137	

Table A. We extend the evaluation of *Mip-Nerf 360* from Table 4 in the main paper to include detailed metrics for each scene, assessing the models across four aspects: PSNR, SSIM, LPIPS, and DSIM. Our analysis covers multiple sparse-view Gaussian Splatting baselines, with SPARS3R consistently demonstrating superior performance.

	PSNR							
	<i>ballroom</i>	<i>barn</i>	<i>church</i>	<i>family</i>	<i>francis</i>	<i>horse</i>	<i>ignatius</i>	<i>museum</i>
Instant-NGP [9]	12.05	17.97	16.64	13.72	15.05	13.39	14.46	18.97
3DGS [5]	25.24	23.36	20.58	19.43	18.70	18.08	21.78	21.35
FSGS [14]	28.80	27.37	23.16	24.85	28.41	24.78	25.32	23.08
SparseGS [11]	24.31	22.63	19.90	18.80	19.49	17.01	20.70	19.42
DRGS [8]	23.09	20.78	20.01	20.39	23.16	21.11	20.01	23.11
CF-3DGS [4]	23.19	17.38	18.96	17.63	19.99	17.22	21.10	16.46
InstantSplat [2]	27.78	27.37	22.91	28.29	30.66	27.74	24.35	26.62
SPARS3R	29.49	30.30	25.02	30.77	30.89	31.38	30.65	30.73

	SSIM							
	<i>ballroom</i>	<i>barn</i>	<i>church</i>	<i>family</i>	<i>francis</i>	<i>horse</i>	<i>ignatius</i>	<i>museum</i>
Instant-NGP [9]	0.204	0.507	0.400	0.566	0.470	0.523	0.263	0.675
3DGS [5]	0.844	0.754	0.698	0.706	0.665	0.682	0.730	0.762
FSGS [14]	0.907	0.838	0.779	0.865	0.845	0.877	0.819	0.828
SparseGS [11]	0.830	0.755	0.694	0.751	0.680	0.671	0.714	0.721
DRGS [8]	0.797	0.740	0.659	0.707	0.723	0.767	0.601	0.792
CF-3DGS [4]	0.776	0.529	0.559	0.581	0.612	0.604	0.654	0.534
InstantSplat [2]	0.915	0.864	0.802	0.907	0.895	0.908	0.811	0.889
SPARS3R	0.936	0.915	0.850	0.945	0.900	0.950	0.927	0.933

	LPIPS							
	<i>ballroom</i>	<i>barn</i>	<i>church</i>	<i>family</i>	<i>francis</i>	<i>horse</i>	<i>ignatius</i>	<i>museum</i>
Instant-NGP [9]	0.464	0.233	0.317	0.359	0.583	0.538	0.440	0.183
3DGS [5]	0.084	0.147	0.195	0.217	0.319	0.289	0.164	0.133
FSGS [14]	0.055	0.108	0.157	0.096	0.147	0.105	0.127	0.096
SparseGS [11]	0.098	0.159	0.201	0.200	0.326	0.289	0.190	0.154
DRGS [8]	0.155	0.268	0.314	0.296	0.441	0.277	0.397	0.166
CF-3DGS [4]	0.102	0.375	0.286	0.301	0.357	0.361	0.174	0.413
InstantSplat [2]	0.069	0.116	0.168	0.087	0.138	0.093	0.182	0.068
SPARS3R	0.026	0.042	0.095	0.032	0.072	0.037	0.044	0.026

	DSIM							
	<i>ballroom</i>	<i>barn</i>	<i>church</i>	<i>family</i>	<i>francis</i>	<i>horse</i>	<i>ignatius</i>	<i>museum</i>
Instant-NGP [9]	0.289	0.108	0.121	0.185	0.472	0.551	0.238	0.068
3DGS [5]	0.030	0.058	0.065	0.079	0.120	0.115	0.043	0.048
FSGS [14]	0.014	0.031	0.023	0.017	0.025	0.032	0.016	0.027
SparseGS [11]	0.033	0.074	0.061	0.062	0.107	0.148	0.052	0.067
DRGS [8]	0.034	0.085	0.078	0.061	0.105	0.110	0.106	0.046
CF-3DGS [4]	0.015	0.240	0.096	0.083	0.129	0.196	0.024	0.232
InstantSplat [2]	0.009	0.014	0.020	0.006	0.016	0.013	0.012	0.011
SPARS3R	0.005	0.008	0.012	0.004	0.010	0.006	0.004	0.004

Table B. We extend the evaluation of *Tanks & Temples* from Table 4 in the main paper to include detailed metrics for each scene, assessing the models across four aspects: PSNR, SSIM, LPIPS, and DSIM. Our analysis covers multiple sparse-view Gaussian Splatting baselines, with SPARS3R consistently demonstrating superior performance.

	PSNR						
	<i>bench</i>	<i>bicycle</i>	<i>car</i>	<i>chair</i>	<i>ladder</i>	<i>suv</i>	<i>table</i>
Instant-NGP [9]	13.61	12.20	12.78	15.59	12.75	12.51	13.49
3DGS [5]	20.08	19.10	23.48	22.81	17.09	25.00	21.09
FSGS [14]	22.72	21.19	26.34	25.37	18.96	27.02	22.41
SparseGS [11]	19.15	18.30	22.62	21.72	17.63	24.06	20.42
DRGS [8]	19.16	19.02	24.54	22.58	18.53	26.20	21.88
CF-3DGS [4]	13.47	11.71	16.51	18.04	14.93	17.48	15.85
InstantSplat [2]	21.14	19.67	27.28	24.59	18.55	28.01	23.30
SPARS3R	25.09	23.16	29.89	28.88	19.13	29.48	25.31

	SSIM						
	<i>bench</i>	<i>bicycle</i>	<i>car</i>	<i>chair</i>	<i>ladder</i>	<i>suv</i>	<i>table</i>
Instant-NGP [9]	0.281	0.201	0.549	0.588	0.303	0.512	0.549
3DGS [5]	0.622	0.597	0.838	0.701	0.469	0.833	0.649
FSGS [14]	0.755	0.713	0.895	0.785	0.574	0.879	0.716
SparseGS [11]	0.607	0.578	0.831	0.719	0.488	0.832	0.648
DRGS [8]	0.450	0.489	0.855	0.702	0.453	0.876	0.662
CF-3DGS [4]	0.190	0.144	0.592	0.573	0.250	0.569	0.542
InstantSplat [2]	0.639	0.596	0.892	0.796	0.521	0.892	0.803
SPARS3R	0.812	0.783	0.929	0.872	0.619	0.910	0.815

	LPIPS						
	<i>bench</i>	<i>bicycle</i>	<i>car</i>	<i>chair</i>	<i>ladder</i>	<i>suv</i>	<i>table</i>
Instant-NGP [9]	0.936	0.932	0.847	0.829	0.927	0.918	0.852
3DGS [5]	0.240	0.271	0.168	0.279	0.291	0.110	0.278
FSGS [14]	0.198	0.232	0.124	0.259	0.291	0.088	0.291
SparseGS [11]	0.267	0.293	0.183	0.301	0.280	0.120	0.293
DRGS [8]	0.561	0.530	0.231	0.499	0.515	0.104	0.516
CF-3DGS [4]	0.596	0.565	0.398	0.678	0.563	0.295	0.718
InstantSplat [2]	0.315	0.355	0.133	0.262	0.359	0.074	0.238
SPARS3R	0.115	0.136	0.061	0.129	0.177	0.057	0.122

	DSIM						
	<i>bench</i>	<i>bicycle</i>	<i>car</i>	<i>chair</i>	<i>ladder</i>	<i>suv</i>	<i>table</i>
Instant-NGP [9]	0.926	0.901	0.878	0.895	0.877	0.882	0.882
3DGS [5]	0.046	0.081	0.081	0.054	0.055	0.061	0.069
FSGS [14]	0.027	0.054	0.045	0.031	0.037	0.045	0.051
SparseGS [11]	0.053	0.096	0.093	0.070	0.053	0.062	0.076
DRGS [8]	0.063	0.100	0.053	0.124	0.038	0.049	0.070
CF-3DGS [4]	0.390	0.481	0.124	0.402	0.275	0.226	0.376
InstantSplat [2]	0.017	0.034	0.023	0.028	0.028	0.038	0.025
SPARS3R	0.003	0.010	0.016	0.005	0.012	0.020	0.013

Table C. We extend the evaluation of *MVimgNet* from Table 4 in the main paper to include detailed metrics for each scene, assessing the models across four aspects: PSNR, SSIM, LPIPS, and DSIM. Our analysis covers multiple sparse-view Gaussian Splatting baselines, with SPARS3R consistently demonstrating superior performance.

Scenes	DUS3R			MASt3R			InstatnSplat			COLMAP+MASt3R		
	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE
<i>bicycle</i>	3.009	3.774	24.33	2.537	3.408	14.15	2.959	3.757	24.21	0.268	0.414	2.01
<i>flowers</i>	6.294	11.480	31.03	6.410	8.504	25.95	6.296	11.458	30.93	1.160	1.976	7.41
<i>garden</i>	1.190	2.067	6.42	0.676	1.339	1.85	1.128	1.950	6.48	0.137	0.264	0.63
<i>stump</i>	4.054	5.945	31.21	1.804	2.611	11.32	4.132	5.963	31.32	0.179	0.240	1.03
<i>treehill</i>	1.202	2.311	8.64	0.978	1.051	5.49	1.206	2.334	8.09	0.356	0.806	1.71
<i>room</i>	1.325	1.716	9.66	1.758	0.955	22.58	1.367	1.750	10.21	0.678	0.970	4.93
<i>counter</i>	0.732	1.168	6.32	0.453	0.729	3.77	0.625	1.081	5.68	0.125	0.225	0.96
<i>kitchen</i>	0.699	1.094	6.73	0.344	0.457	2.57	0.705	1.085	6.69	0.088	0.161	0.83
<i>bonsai</i>	4.349	2.586	28.98	0.937	1.350	7.99	4.270	2.518	28.51	0.179	0.217	1.48

Table D. The detailed pose metrics for each scene in the *Mip-Nerf 360* dataset are presented. We evaluate the Relative Translation Error (RPE_t), Relative Rotation Error (RPE_r) and Absolute Trajectory Error (ATE, scaled by $\times 10^{-3}$) using normalized poses. The poses estimated by dense COLMAP are used as the ground truth for reference.

Scenes	DUS3R			MASt3R			InstatnSplat			COLMAP+MASt3R		
	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE
<i>ballroom</i>	0.547	0.105	4.78	0.245	0.090	2.36	0.136	0.050	1.38	0.134	0.039	1.04
<i>barn</i>	0.398	0.251	4.44	0.201	0.296	2.85	0.164	0.131	2.26	0.102	0.125	1.32
<i>church</i>	0.488	0.099	4.37	0.137	0.231	1.73	0.185	0.049	2.64	0.214	0.055	2.01
<i>family</i>	0.556	0.068	5.73	0.313	0.202	2.82	0.086	0.018	0.69	0.103	0.045	0.84
<i>francis</i>	0.340	0.235	4.20	0.356	0.230	5.74	0.117	0.226	2.98	0.073	0.175	0.68
<i>horse</i>	1.217	0.072	10.79	0.265	0.465	4.37	0.188	0.028	1.44	0.212	0.162	1.71
<i>ignatius</i>	0.621	0.172	4.88	0.227	0.239	3.20	0.128	0.034	1.61	0.250	0.076	4.03
<i>museum</i>	0.392	0.115	3.29	0.168	0.228	2.67	0.201	0.115	2.17	0.200	0.071	1.79

Table E. The detailed pose metrics for each scene in the *Tanks & Temples* dataset are presented. We evaluate the Relative Translation Error (RPE_t), Relative Rotation Error (RPE_r) and Absolute Trajectory Error (ATE, scaled by $\times 10^{-3}$) using normalized poses. The poses estimated by dense COLMAP are used as the ground truth for reference.

Scenes	DUS3R			MASt3R			InstatnSplat			COLMAP+MASt3R		
	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE	RPE _t	RPE _r	ATE
<i>bench</i>	0.357	0.336	2.60	0.161	0.191	1.18	0.122	0.167	1.50	0.049	0.058	0.66
<i>bicycle</i>	0.424	0.490	4.82	0.247	0.491	3.86	0.296	0.376	4.24	0.069	0.064	0.70
<i>car</i>	0.480	0.302	4.59	0.104	0.157	1.83	0.070	0.191	1.64	0.033	0.042	0.40
<i>chair</i>	0.672	0.563	7.06	0.628	0.555	7.05	0.454	0.432	5.77	0.095	0.099	0.92
<i>ladder</i>	0.663	0.567	9.69	0.073	0.225	0.73	0.562	0.472	8.74	0.080	0.097	0.64
<i>suv</i>	0.233	0.394	3.30	0.137	0.329	1.32	0.170	0.324	3.84	0.085	0.071	0.96
<i>table</i>	0.239	0.296	2.57	0.108	0.181	1.04	0.175	0.214	2.06	0.116	0.117	1.53

Table F. The detailed pose metrics for each scene in the *MVimgNet* dataset are presented. We evaluate the Relative Translation Error (RPE_t), Relative Rotation Error (RPE_r) and Absolute Trajectory Error (ATE, scaled by $\times 10^{-3}$) using normalized poses. The poses estimated by dense COLMAP are used as the ground truth for reference.