Supplementary Material

A. Ablation on Decoupled Optimization

Figure 1 illustrates the impact of employing a decoupled optimization strategy on convergence outcomes. The green and red circles represent corresponding matching points on the observed and rendered images.

In Figure 1(a), the upper half indicates the scenario where only rotational perturbations exist, while the lower half pertains to translational perturbations alone. If there are only rotational perturbations, the angles between vectors formed by matching points tend to intertwine; however, in the case of only translational perturbations, these vectors tend to be parallel (i.e., smaller angles). Thus, in the optimization process, rotational errors can be optimized by reducing the angles between vectors, while translational errors can be optimized by shortening the distances between point pairs. The advantage of this approach is that it prevents the direct influence of point pair distances on the rotation because when the translation distance is significant, the gradient direction provided by point pair distances may not accurately gauge the correctness of the rotation, and if the updated rendered image fails to match well with the observed image, optimization is likely to get stuck in local minima.

In Figure 1(b), the upper half does not use a decoupled optimization strategy, while the lower half does. Without the decoupled optimization strategy, we can see that the distances between feature points are reduced, leading to decreased translational error. However, the large initial error causes all vectors to intertwine, trapping rotational error in a local minimum. However, using the decoupled optimization strategy, all vectors remain parallel during the convergence process, and the distances between matching points are simultaneously reduced, ultimately achieving successful convergence. Figure 1(c) shows the corresponding trajectory.

B. Detailed Experiments Configuration

All other methods in the experiments use the Adam optimizer and compute the photometric loss with L2. In our experiments, the iNeRF method [11], different from the original setting, uses the more efficient 3DGS [4] as the scene representation, with a batch size of 512 and a sampling strategy focused on interest regions. Pose optimization is based on gradient-based $\mathrm{SE}(3)$ optimization. The initial learning rate is set to 0.05, with exponential decay based on a decay rate of 0.8 and a base step of 100, and the total number of iterations is 500. The iNeRF† differs from iNeRF only in that the photometric loss is computed over the entire image. The pNeRF [5] uses InstantNGP [7] as the scene representation, with a total of 2560 optimization steps. Parallel Monte Carlo sampling is conducted every 512 steps, and pose optimization is based on gradient-based $\mathrm{SO}(3) \times \mathrm{T}(3)$. The learning rates for the translation and rotation parts start from 3×10^{-3} and 5×10^{-3} , respectively, and decay exponentially based on a decay rate of 0.33 and a base step of 256. The iComma [10] uses 3DGS as the scene representation and performs gradient-based $\mathrm{SE}(3)$ pose optimization, using [9] to obtain corresponding feature points between the query image and the rendered image, with a total of 500 iterations.

C. Efficiency Analysis

All experiments were tested on an RTX 4090 GPU. The storage requirements for the 3D Gaussian scene models vary depending on the scene size, but most fall within the 50 to 800 MB range. The time cost for pose estimation consists of Monte Carlo initialization and decoupled optimization. Under standard convergence conditions, Monte Carlo initialization takes an average of 1.29 seconds at a rendering resolution of 800, while decoupled optimization takes an average of 1.68 seconds for decoupled optimization (10 iterations). Although the execution speed cannot match traditional visual estimation methods, our framework surpasses them in pose estimation accuracy. Additionally, it is applicable in scenarios where the initial pose cannot provide effective co-visibility and does not require additional training for global image descriptors or an image retrieval database.

D. Additional Qualitative Results

Figure 2 presents additional visualization results on different datasets. From this figure, we can observe that iNeRF† and iComMa [10] often tend to get trapped in local minima, especially in regions with high-frequency textures. Our pro-

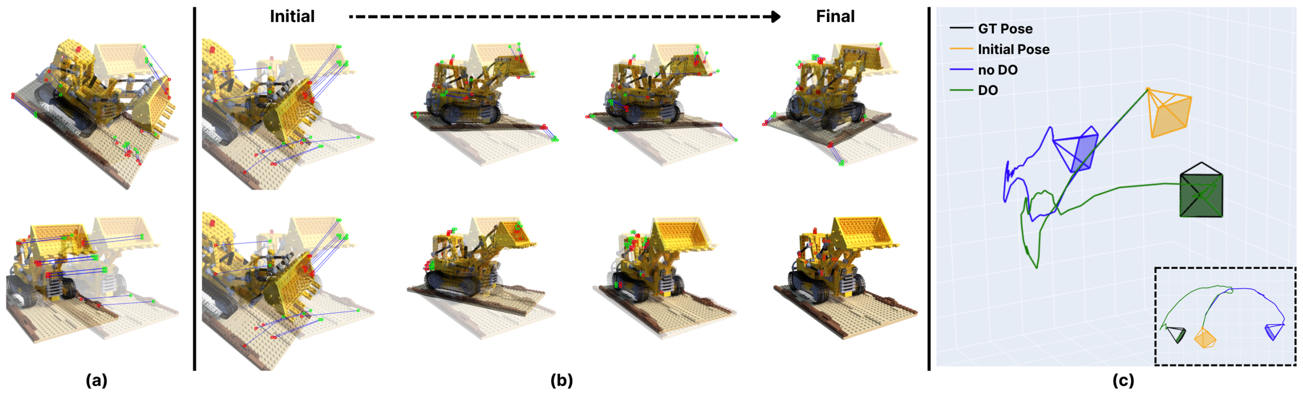


Figure 1. Visualization of Decoupled Optimization.

posed method effectively avoids these situations. However, in the third row, our method also shows incomplete convergence due to a terrible initial condition. Additionally, unblended visualization results are provided in Figure 3.

E. Quantitative Error Analysis

Table 2 presents the quantitative results of each method in each scene under the setting based on perturbed ground-truth poses. In addition to the success rates of the different methods, Table 3 and Table 4 present the median and mean numerical errors of the 6 DoF poses for each method. pNeRF [5] demonstrates superior numerical errors on the NeRF-Synthetic [6] dataset compared to iNeRF [11], which is not reflected in the success rates. Compared to iComMa [10], our proposed method achieves the smallest errors in most rotational and translational errors.

Table 5 illustrates the performance of our proposed method under random initial poses. We selected six scenes (lego, mic, chair, bicycle, garden, room) for testing. And we randomly selected five images from each scene for testing, with each image subjected to ten different initial poses. Monte Carlo initialization significantly improves convergence in cases of large errors. In comparison, using only decoupled optimization may result in poor convergence due to large errors, potentially incorrect matching points, and overly complex interleaving. However, the Monte Carlo initialization provides better initial poses, and when combined with decoupled optimization, it further enhances the final convergence performance. Ultimately, our proposed method achieves the best overall performance.

F. Ablation on Different Photometric Loss

We also evaluated the impact of different loss functions on performance. We selected three types of pure photometric loss and three types of combined loss for the experiments. Other loss functions are not listed here due to their generally poorer performance. The combined losses are applied on top of the L2 loss with additional weighted losses. Feature point weighting uses the same feature point sampling strategy as [11], with a weighting coefficient of 0.7. Corner point weighting employs the Shi-Tomasi corner detector [8] for corner detection, followed by dilation with a kernel size of 51, with a weighting coefficient of 7.5×10^{-5}. Blur weighting applies Gaussian blurring on both the rendered and observed images, using MSE loss, with σ set to 5 and a weighting coefficient of 0.8.

Table 1 shows that when approaching the ground-truth, L2, feature point weighting, and corner point weighting do not show superior performance. The reason is that when the pose is closing global minimum, supervision does not need to focus excessively on specific regions, and overemphasis on particular parts only allows high-frequency noise to affect optimization directions. In contrast, using L1 and blur weighting, the overall success rates achieve the best. This indicates that more stable (with lower variation) loss can provide more global consistent guidance when only fine optimization is needed. That is why we chose the L1 distance as our refinement loss.

G. Comparison across different perturbation levels

Figure 4 illustrates the performance of iComMa [10], iNeRF†, and our method under different perturbation conditions. The horizontal axis represents the number of steps, while the vertical axis shows the percentage of the scene achieving an error below the threshold at each step. We set two initial perturbation conditions: rotation at ±[0°, 30°] with translation of ±[0, 0.5], and rotation at ±[30°, 60°] with translation of ±[0.5, 1.0]. We observed that iNeRF† performs better under smaller initial error conditions, exceeding its performance in larger initial error by more than

Table 1. Ablation study of different photometric loss functions

Loss	Rot. err. <5°				Trans. err. <0.05 units			
	Synthetic [6]	Mip-360 [2]	DB [3]	Mean	Synthetic [6]	Mip-360 [2]	DB [3]	Mean
L1	**0.977**	**0.990**	**0.963**	**0.977**	0.893	**0.967**	**0.927**	0.929
L2	**0.977**	0.960	0.947	0.961	0.803	0.900	0.920	0.874
Relative L2	0.963	0.973	0.950	0.962	0.920	0.923	0.893	0.912
Feature weight	0.933	0.873	0.847	0.884	0.900	0.817	0.810	0.842
Corner weight	0.777	0.870	0.820	0.822	0.777	0.783	0.823	0.794
Blur weight	0.960	0.977	0.953	0.963	**0.940**	0.947	0.917	**0.934**

Table 2. Quantitative comparison with perturbed ground-truth poses on NeRF-Synthetic, LLFF, Mip-NeRF 360, and Deep Blending Datasets.

Dataset / Methods		Standard threshold (5° & 0.05units)					Hard threshold (1° & 0.01units)				
		iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours	iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours
Synthetic	Chair	0.00	0.02	0.37	0.72	**0.98**	0.00	0.00	0.18	0.72	**0.98**
	Drums	0.00	0.00	0.19	0.69	**0.96**	0.00	0.00	0.19	0.67	**0.96**
	Ficus	0.00	0.00	0.26	0.68	**0.94**	0.00	0.00	0.25	0.68	**0.94**
	Hotdog	0.03	0.00	0.75	0.75	**0.85**	0.00	0.00	0.63	0.65	**0.85**
	Lego	0.08	0.03	0.50	0.77	**0.97**	0.01	0.00	0.50	0.76	**0.97**
	Materials	0.04	0.00	0.09	0.54	**0.80**	0.01	0.00	0.09	0.54	**0.79**
	Mic	0.00	0.00	0.12	0.67	**0.69**	0.00	0.00	0.05	**0.45**	0.40
	Ship	0.00	0.00	0.45	0.71	**0.98**	0.00	0.00	0.42	0.70	**0.98**
	Average	0.019	0.006	0.346	0.694	**0.894**	0.003	0.000	0.289	0.605	**0.868**
LLFF	Fern	0.00	0.00	0.32	0.86	**1.00**	0.00	0.00	0.00	**0.25**	0.08
	Fortress	0.00	0.00	0.37	0.71	**0.99**	0.00	0.00	0.01	0.15	**0.40**
	Horns	0.00	0.00	0.35	0.69	**0.80**	0.00	0.00	0.10	0.36	**0.60**
	Room	0.00	0.00	0.14	0.49	**1.00**	0.00	0.00	0.00	**0.20**	0.10
	Average	0.000	0.000	0.295	0.688	**0.948**	0.000	0.000	0.028	0.242	**0.295**
Mip-NeRF 360	Bicycle	0.00	0.00	0.05	0.56	**0.89**	0.00	0.00	0.01	0.52	**0.89**
	Bonsai	0.00	0.00	0.12	0.87	**0.97**	0.00	0.00	0.09	0.87	**0.97**
	Counter	0.00	0.00	0.12	0.72	**0.97**	0.00	0.00	0.08	0.72	**0.97**
	Garden	0.00	0.00	0.07	0.82	**0.99**	0.00	0.00	0.00	0.82	**0.99**
	Kitchen	0.00	0.00	0.10	0.68	**0.82**	0.00	0.00	0.02	0.59	**0.68**
	Room	0.00	0.00	0.21	0.67	**0.98**	0.00	0.00	0.20	0.67	**0.98**
	Stump	0.00	0.00	0.00	0.48	**0.76**	0.00	0.00	0.00	0.47	**0.76**
	Average	0.000	0.000	0.095	0.686	**0.911**	0.000	0.000	0.057	0.657	**0.890**
Deep Blending	Bedroom	0.00	-	0.16	0.55	**0.96**	0.00	-	0.08	0.44	**0.76**
	Bridge	0.00	-	0.04	0.18	**0.51**	0.00	-	0.02	0.18	**0.45**
	Creepyattic	0.00	-	0.14	0.50	**0.90**	0.00	-	0.11	0.48	**0.75**
	Drjohnson	0.00	-	0.16	0.40	**0.89**	0.00	-	0.12	0.35	**0.76**
	Playroom	0.00	-	0.16	0.61	**1.00**	0.00	-	0.09	0.43	**0.59**
	Average	0.000	-	0.132	0.448	**0.852**	0.000	-	0.084	0.376	**0.662**

twofold. When dealing with larger errors, iNeRF† struggles to converge effectively, indicating that pixel-level comparison errors are insufficient to provide adequate supervision in high-error scenarios. iComMa's matching module can quickly reduce errors in smaller initial error cases, achieving good results. However, in larger initial error con-

Table 3. 6-DoF Median Errors on NeRF-Synthetic, LLFF, Mip-NeRF 360, and Deep Blending Datasets.

Dataset / Methods		Rotation error (deg)					Translation error (units)				
		iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours	iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours
Synthetic	Chair	59.291	18.904	47.556	0.035	**0.011**	3.281	1.355	2.027	0.002	**0.001**
	Drums	80.101	32.152	58.889	0.029	**0.010**	3.881	1.926	2.981	0.002	**0.001**
	Ficus	63.946	35.972	68.043	0.034	**0.025**	3.387	2.236	4.215	0.003	**0.002**
	Hotdog	65.545	13.454	0.089	0.084	**0.032**	3.281	0.872	0.006	0.006	**0.002**
	Lego	53.696	13.504	10.490	0.027	**0.000**	2.454	0.840	0.455	0.002	**0.000**
	Materials	56.216	27.399	46.667	0.045	**0.021**	2.924	1.763	2.696	0.004	**0.002**
	Mic	63.438	37.709	58.216	**0.359**	0.411	3.234	2.237	3.276	**0.029**	0.034
	Ship	71.757	57.077	29.202	0.065	**0.012**	3.722	3.385	1.641	0.005	**0.000**
	Average	64.249	29.521	39.894	0.085	**0.065**	3.271	1.827	2.162	0.007	**0.005**
LLFF	Fern	37.078	80.713	0.726	0.031	**0.025**	3.200	4.259	0.073	**0.015**	**0.015**
	Fortress	25.316	78.143	0.622	0.105	**0.093**	3.159	4.188	0.076	0.021	**0.013**
	Horns	54.294	84.718	20.905	0.034	**0.023**	2.838	4.313	1.990	0.018	**0.009**
	Room	71.767	83.500	38.946	0.302	**0.017**	3.421	4.162	3.127	0.232	**0.015**
	Average	47.114	81.768	15.300	0.118	**0.039**	3.154	4.230	1.316	0.072	**0.013**
Mip-NeRF 360	Bicycle	79.836	66.055	62.508	0.047	**0.021**	3.871	3.734	3.841	0.006	**0.002**
	Bonsai	70.657	67.981	48.617	0.019	**0.014**	3.845	3.705	2.867	0.002	**0.001**
	Counter	68.727	69.089	49.979	0.023	**0.018**	4.102	3.974	3.012	**0.001**	**0.001**
	Garden	74.483	76.227	46.187	0.018	**0.009**	3.998	3.965	3.636	0.002	**0.001**
	Kitchen	57.554	70.913	41.210	0.036	**0.031**	3.270	3.331	3.013	**0.001**	**0.001**
	Room	70.591	77.848	49.328	0.026	**0.015**	4.278	3.035	3.581	**0.003**	**0.003**
	Stump	78.318	69.224	62.179	0.238	**0.027**	3.845	3.758	3.954	0.107	**0.002**
	Average	71.452	71.048	51.430	0.058	**0.019**	3.887	3.643	3.415	0.017	**0.002**
Deep Blending	Bedroom	66.498	-	62.413	0.109	**0.063**	2.864	-	3.149	0.017	**0.006**
	Bridge	69.947	-	62.046	60.636	**0.974**	3.771	-	3.243	3.749	**0.038**
	Creepyattic	71.326	-	53.132	0.445	**0.032**	4.230	-	3.195	0.048	**0.003**
	Drjohnson	70.588	-	59.671	17.057	**0.083**	2.763	-	2.417	1.474	**0.007**
	Playroom	64.986	-	48.874	0.180	**0.054**	2.791	-	3.537	0.025	**0.008**
	Average	68.669	-	57.227	15.686	**0.241**	3.284	-	3.108	1.063	**0.013**

ditions, fewer matching keypoints may limit its effectiveness. In contrast, our proposed Monte Carlo initialization not only obtains better covisible initial poses but also allows the decoupled optimization method to converge more quickly in the translation part compared to iComMa. Even under stringent initial conditions, our method consistently demonstrates higher and more stable success rates.

H. Limitations

Although our method can handle challenging initial conditions (such as larger errors or incorrect initial poses), it still requires a pose initialization stage to obtain the input camera pose. Most existing methods that do not require an initial pose rely on deep learning models (e.g., NetVLAD [1]), which introduce additional training and computational costs. In fact, the need for an initial pose represents a trade-off between search cost and convergence accuracy in our work. As shown in **??**, while our method can still achieve optimal results with random initialization, its performance decreases slightly. However, by expanding the sampling range, the convergence improves significantly, albeit at the cost of higher computational demands. Moreover, the proposed decoupled optimization strategy may reduce the success rate when applied independently. The main reason is the underlying assumption in our decoupled strategy: the parallelism between point pairs should provide a rough optimization direction at the initial pose stage, as shown in Figure 1(a). Under uncontrolled initial poses, a large number of interwoven and misaligned vectors can obscure the overall parallelism. The Monte Carlo-based initialization helps improve performance because the multiple sampled

Table 4. 6-DoF Mean Errors on NeRF-Synthetic, LLFF, Mip-NeRF 360, and Deep Blending Datasets.

Dataset / Methods		Rotation error (deg)					Translation error (units)				
		iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours	iNeRF [11]	pNeRF [5]	iNeRF†	iComMa [10]	Ours
Synthetic	Chair	67.552	27.580	48.290	25.234	**1.712**	3.458	1.724	2.276	1.279	**0.109**
	Drums	81.249	41.072	52.878	29.549	**4.014**	4.016	2.138	3.063	1.542	**0.267**
	Ficus	69.525	45.382	61.864	25.855	**6.322**	3.723	2.370	3.451	1.522	**0.251**
	Hotdog	68.922	22.516	18.252	19.054	**8.010**	3.416	1.270	1.112	1.045	**0.522**
	Lego	59.939	20.236	39.973	22.878	**2.471**	2.860	1.170	2.282	1.151	**0.150**
	Materials	54.497	39.126	46.523	35.870	**10.677**	3.061	2.192	2.889	2.013	**0.719**
	Mic	65.914	51.113	61.755	20.053	**1.910**	3.314	2.616	3.440	1.071	**0.127**
	Ship	75.480	61.970	43.354	38.455	**0.536**	3.876	3.426	2.242	1.631	**0.036**
	Average	67.885	38.624	46.611	27.119	**4.456**	3.465	2.113	2.595	1.407	**0.273**
LLFF	Fern	46.542	78.444	16.891	6.396	**0.024**	3.209	4.461	1.894	0.485	**0.014**
	Fortress	41.686	76.998	16.054	23.569	**0.089**	3.682	4.317	1.067	1.507	**0.015**
	Horns	57.463	81.563	30.665	18.858	**0.346**	3.749	4.425	3.291	1.367	**0.222**
	Room	71.690	80.982	49.502	40.983	**0.015**	4.304	4.311	3.246	2.297	**0.016**
	Average	54.346	79.497	28.278	22.451	**0.118**	3.736	4.378	2.375	1.414	**0.067**
Mip-NeRF 360	Bicycle	79.662	73.560	60.770	21.201	**0.076**	4.180	3.903	3.730	1.545	**0.038**
	Bonsai	71.156	70.938	51.167	7.973	**0.259**	3.935	3.857	2.992	0.453	**0.029**
	Counter	72.566	70.405	52.814	26.854	**0.842**	4.270	3.908	3.194	1.714	**0.044**
	Garden	75.342	75.818	54.602	11.662	**0.101**	4.104	4.137	3.838	0.850	**0.007**
	Kitchen	62.938	71.974	45.388	27.375	**0.091**	3.519	3.569	3.057	1.513	**0.019**
	Room	72.896	79.245	54.521	30.190	**0.217**	4.355	3.417	3.364	1.903	**0.035**
	Stump	80.516	71.040	62.358	19.104	**4.258**	4.011	3.890	3.865	1.269	**0.197**
	Average	73.582	73.283	54.517	20.623	**0.835**	4.054	3.812	3.434	1.321	**0.053**
Deep Blending	Bedroom	69.601	-	58.148	26.955	**0.138**	3.586	-	3.325	1.662	**0.012**
	Bridge	73.174	-	69.996	59.546	**20.820**	3.932	-	3.959	3.828	**0.983**
	Creepyattic	74.430	-	58.674	31.031	**1.067**	4.554	-	3.849	2.199	**0.114**
	Drjohnson	73.158	-	61.888	48.160	**4.067**	3.328	-	2.971	2.509	**0.244**
	Playroom	67.376	-	53.277	25.315	**0.106**	3.593	-	3.468	1.337	**0.014**
	Average	71.548	-	60.396	38.201	**5.240**	3.798	-	3.514	2.307	**0.273**

Table 5. Ablation study of different proposed strategies with randomly selected initial poses.

MC	DO	Loss	Rot. (deg)		Trans. (units)	
			<5	<1	<0.05	<0.01
			0.510	0.497	0.480	0.460
✓			0.717	0.677	0.643	0.603
	✓		0.513	0.510	0.487	0.470
✓	✓		0.797	0.767	0.730	0.710
✓	✓	✓	**0.847**	**0.793**	**0.773**	**0.757**

hypotheses typically yield an initial pose with a better alignment of point pairs, which in turn enhances the effectiveness of the decoupled optimization.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5297–5307, 2016. 4

[2] 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 (CVPR)*, pages 5470–5479, 2022. 3

[3] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 3

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

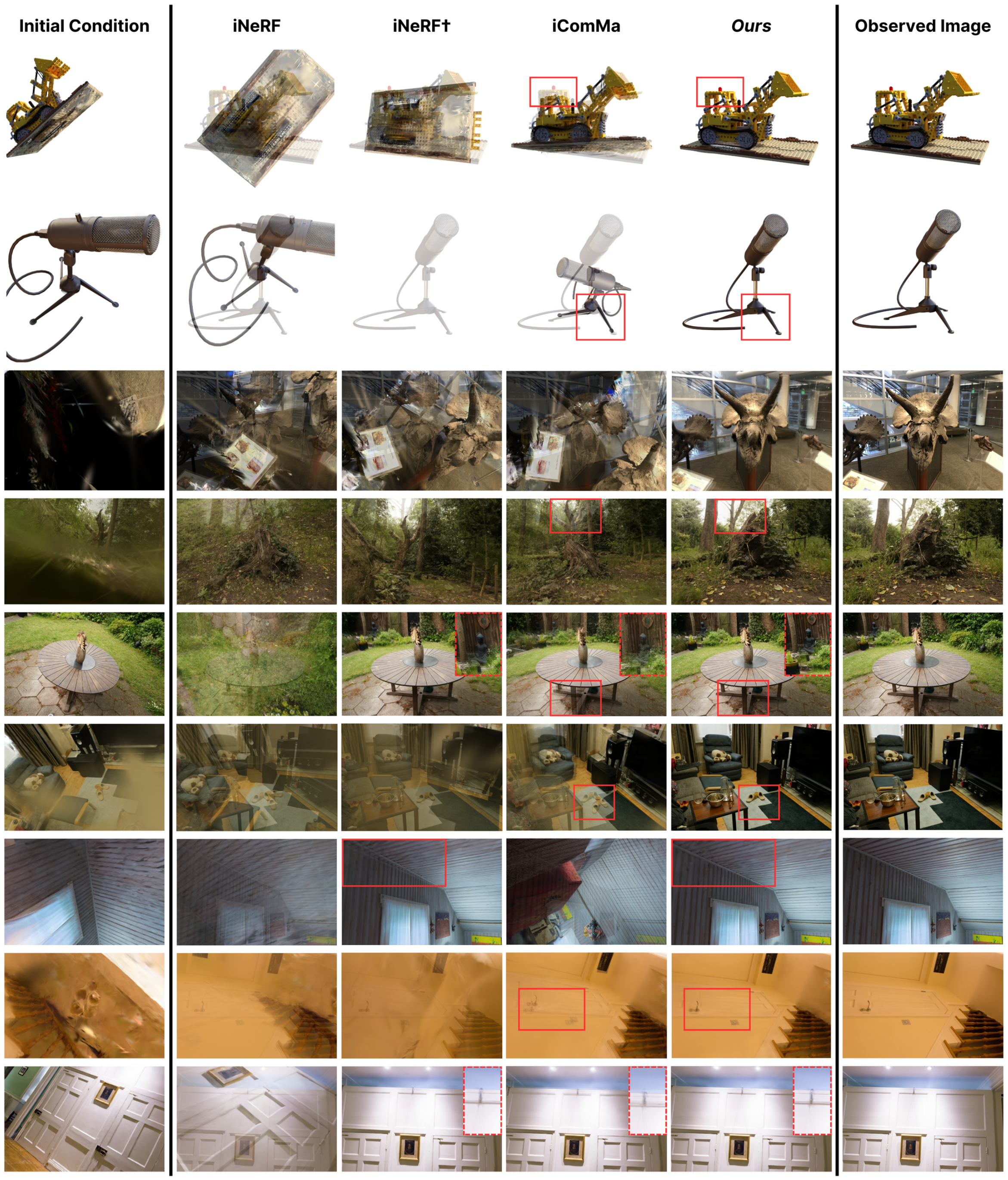


Figure 2. Qualitative Results on NeRF-Synthetic, LLFF, Mip-NeRF 360, and Deep Blending datasets.

radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4), 2023. 1

[5] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9377–9384. IEEE, 2023. 1, 2, 3, 4, 5

[6] 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 (CACM)*, 65(1):99–106, 2021. 2, 3

[7] Thomas Müller, Alex Evans, Christoph Schied, and Alexan-

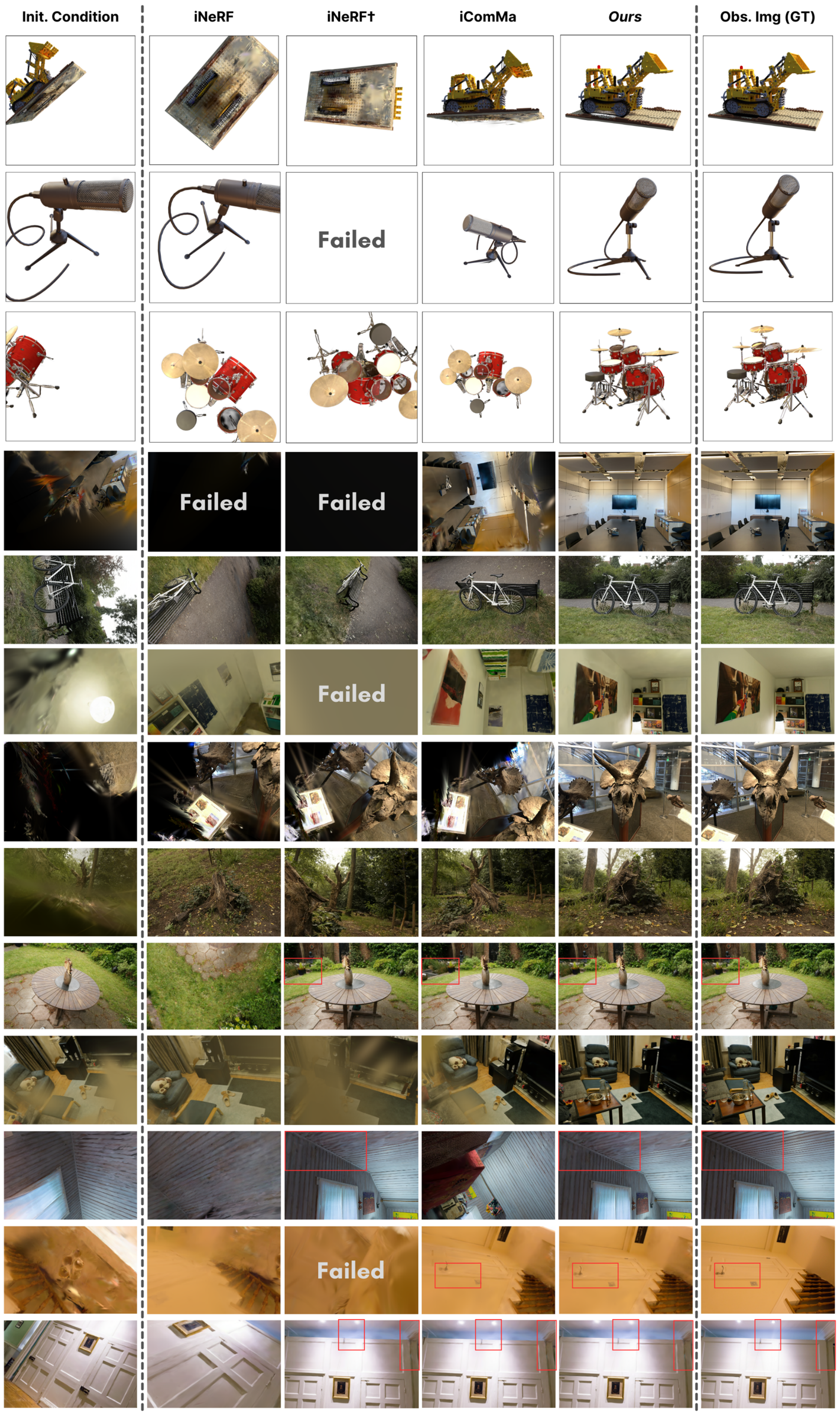
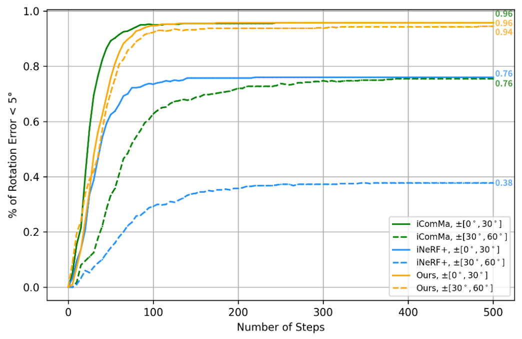
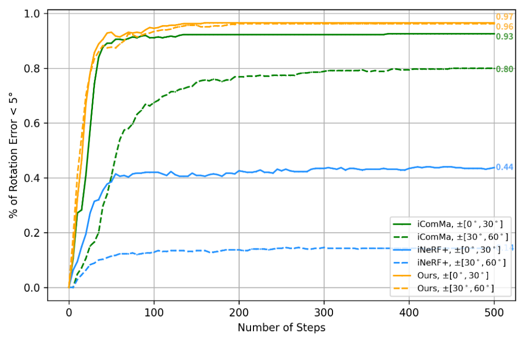


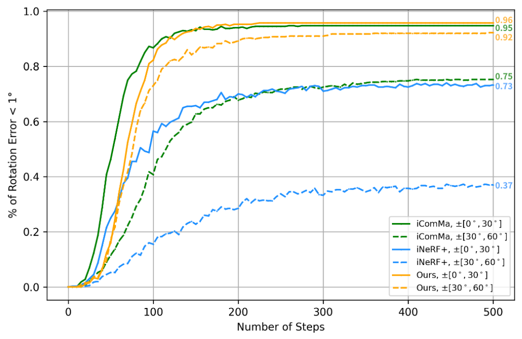
Figure 3. Qualitative unblended visualization results on NeRF-Synthetic, LLFF, Mip-NeRF 360, and Deep Blending datasets.



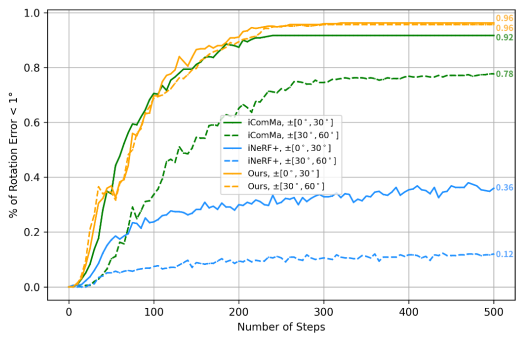
(a) NeRF-Synthetic



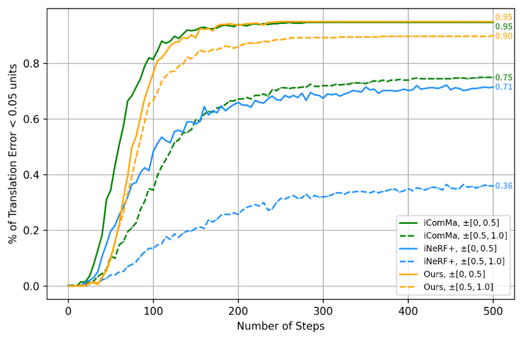
(b) Mip-NeRF 360



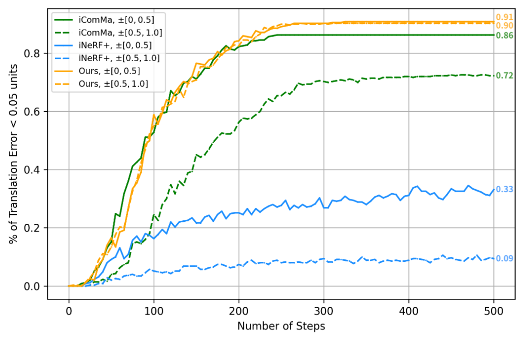
(c) NeRF-Synthetic



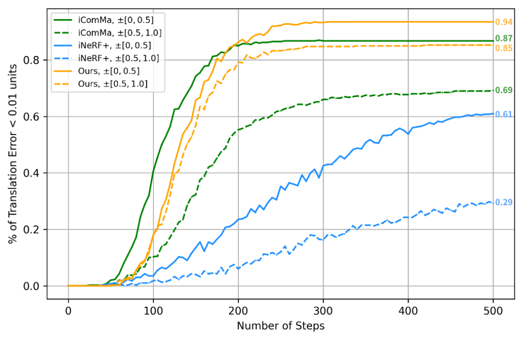
(d) Mip-NeRF 360



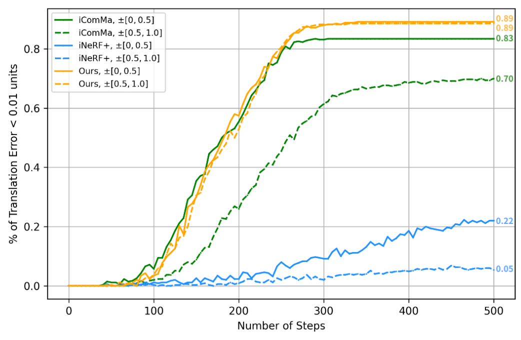
(e) NeRF-Synthetic



(f) Mip-NeRF 360



(g) NeRF-Synthetic



(h) Mip-NeRF 360

Figure 4. Performance comparison of various approaches under different levels of perturbations

der Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1

[8] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994. 2

[9] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8922–8931, 2021. 1

[10] Yuan Sun, Xuan Wang, Yunfan Zhang, Jie Zhang, Caigui Jiang, Yu Guo, and Fei Wang. icomma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching. *arXiv preprint arXiv:2312.09031*, 2023. 1, 2, 3, 4, 5

[11] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 1, 2, 3, 4, 5