StealthAttack: Robust 3D Gaussian Splatting Poisoning via Density-Guided Illusions

Supplementary Material

A. Additional Visualization Results

We present additional visualization results in the supplementary HTML file "videoResults.html" demonstrating our method's effectiveness on both single-view and multi-view attacks through video sequences that highlight the consistent rendering of illusory objects across viewpoints.

B. Comprehensive Dataset Evaluation

Extended Threshold Analysis. Tab. 1 evaluates 36 scenes across three datasets: 7 from Mip-NeRF 360 [1], 8 from Tanks & Temples [3], and 21 from Free [5], with Free scenes categorized as EASY/MEDIAN/HARD based on different threshold combinations. Beyond the main paper's criteria (PSNR $>$ 25 on V-ILLUSORY, V-TEST PSNR drop $\leq$ 3), we test various threshold combinations to assess method robustness across difficulty settings and provide comprehensive baseline comparisons.

Table 1. **Attack success rates across extended threshold combinations.** Our method demonstrates superior performance across all difficulty levels.

Method / Success criteria	V-ILLUSORY $>$ 25 V-TEST drop $\leq$ 8	V-ILLUSORY $>$ 20 V-TEST drop $\leq$ 9	V-ILLUSORY $>$ 15 V-TEST drop $\leq$ 10
IPA-NeRF [2] (Nerfacto [1])	0/36	1/36	10/36
IPA-NeRF [2] (Instant-NGP [1])	2/36	6/36	21/36
IPA-Splat	0/36	1/36	4/36
Ours	**23/36**	**26/36**	**30/36**

The results demonstrate our method's superior robustness, with success rates ranging from 64% to 83% across different threshold combinations, significantly outperforming existing approaches across diverse datasets and evaluation criteria.

C. Computational Efficiency Analysis

Our attack reduces GPU memory usage by 41% and Gaussian points by 88% with a modest training time increase on the Mip-NeRF 360 dataset. This stems from our noise scheduling disrupting multi-view consistency, allowing convergence with fewer Gaussians—a favorable trade-off for attack effectiveness.

Table 2. **Computational efficiency comparison.** Our method significantly reduces memory usage and model complexity.

Method	GPU Memory (MB)	Number of Gaussians	Training Time (min)
Standard 3DGS	4,101.94	2,602,787	15.05
Ours	2,419.08	310,114	22.32

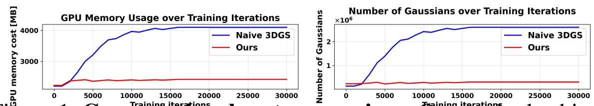

Figure 1. **Computational cost comparison.** Our method achieves significant reductions in GPU memory usage and model complexity.

D. More Implementation Details

Illusory Objects. We randomly select images and masks from the COCO 2017 dataset [4] to extract diverse, unbiased illusory objects for our backdoor attacks.
Implementation Details. We implement our experiments using the official 3DGS codebase [2] with default hyperparameters on NVIDIA RTX 4090Ti GPUs.

E. More Visual Results for Single View Attack

Figs. 2 and 3 demonstrate our method's superiority in single-view attacks across multiple scenes and datasets. While baseline approaches like IPA-NeRF (Nerfacto) and IPA-NeRF (Instant-NGP) often produce imperceptible or heavily distorted illusory objects (as seen in the "*bonsai*" scene), our approach consistently delivers clear, realistic illusions with distinct boundaries.

F. More Visual Results for Multi-view Attack

Figs. 4–6 demonstrate our method's superiority over IPA-NeRF (Nerfacto and Instant-NGP) and IPA-Splat across 2, 3, and 4 poisoned viewpoints. Our density-guided approach consistently generates clear, geometrically consistent illusory objects while maintaining high rendering quality in non-poisoned views, effectively preserving scene fidelity regardless of the number of attack viewpoints.

G. More Visual Results for Evaluation Protocol

Fig. 7 validates our KDE-based evaluation protocol, showing that attack effectiveness inversely correlates with scene density in "*hydrant*" scene. Illusory objects appear more convincing in EASY (low-density) regions than in HARD (high-density) regions, confirming that fewer overlapping observations increase vulnerability. This protocol establishes a standardized benchmark for poisoning attacks while revealing connections between scene geometry and 3D reconstruction vulnerability.

H. More Visual Results for Ablation Studies

Fig. 8 presents qualitative comparisons of different attack strategy combinations across seven Mip-NeRF 360 scenes.

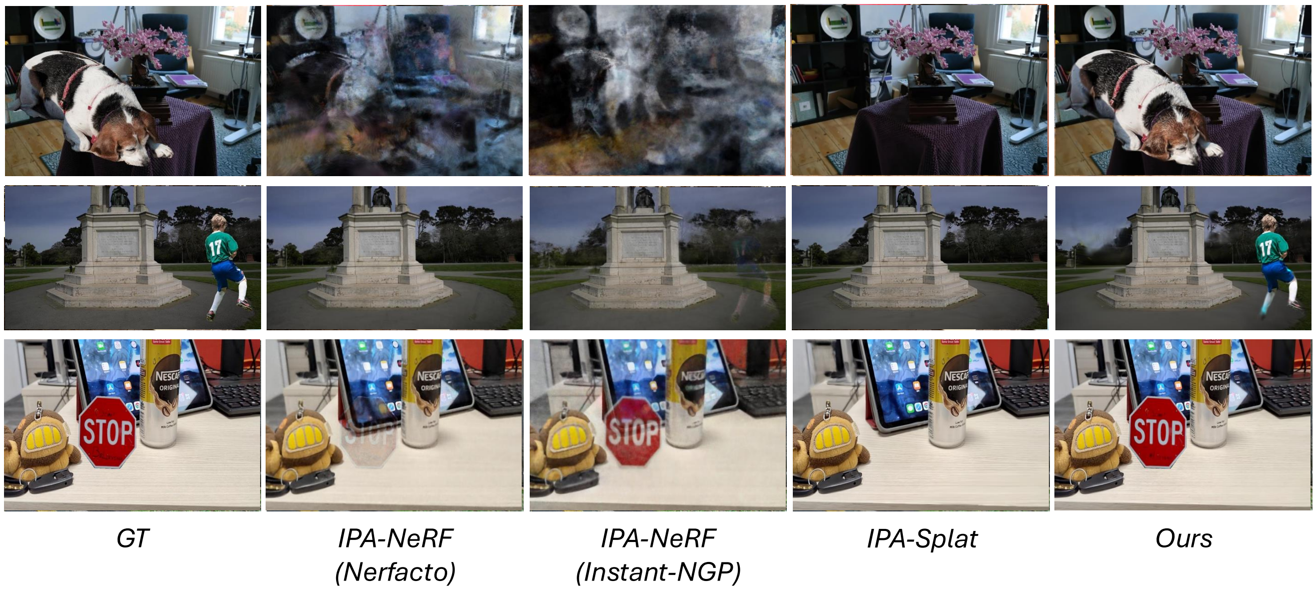

Figure 2. **Qualitative comparisons on single-view attack 1.** Results on the "*bonsai*" scene (Mip-NeRF 360 [1]), "*francis*" scene (Tanks & Temples [3]), and "*counter*" scene (Free [5]). Both IPA-NeRF variants exhibit poor convergence on the "*bonsai*" scene, while our method consistently produces clear, well-integrated illusory objects across all scenes.

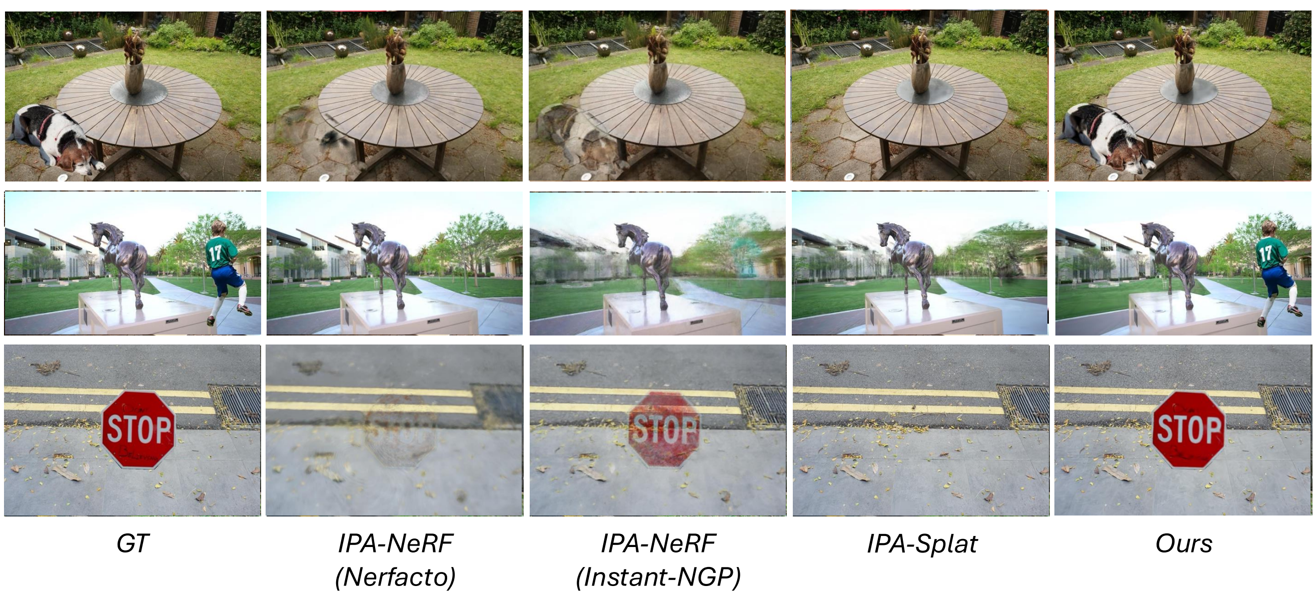

Figure 3. **Qualitative comparisons on single-view attack 2.** Results on the "*garden*" scene (Mip-NeRF 360 [1]), "*horse*" scene (Tanks & Temples [3]), and "*road*" scene (Free [5]). Our method effectively embeds distinct illusory objects while maintaining scene consistency.

While strategies (1) direct replacement and (2) density-guided poisoning are effective for most scenes, they show limitations in complex environments with high view overlap (e.g., "*room*"). Our experiments demonstrate that combining these with (3) multi-view consistency disruption achieves superior illusion embedding across all tested scenes, highlighting the complementary nature of our proposed methods.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5

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

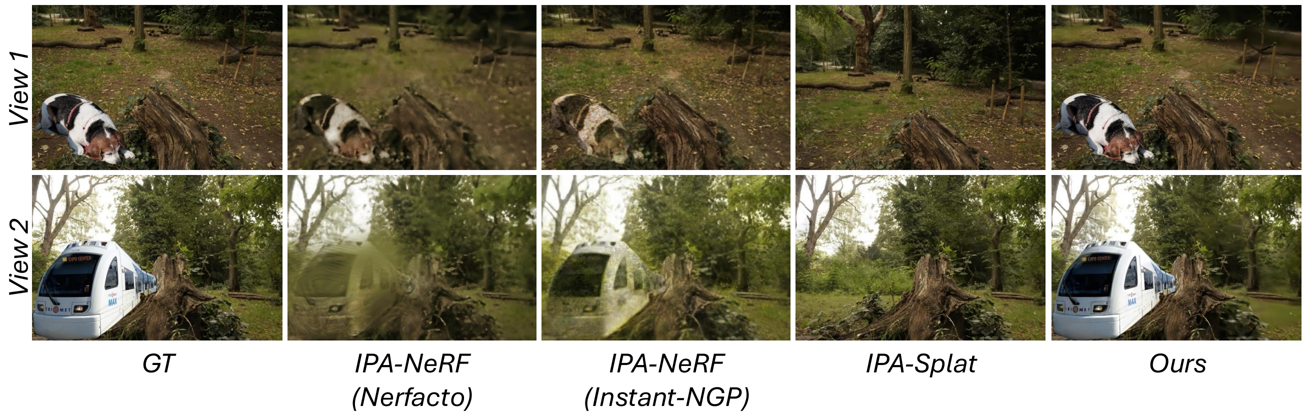

Figure 4. **Qualitative comparisons on multi-view attack with 2 poisoned views.** We compare the visual quality of illusory objects rendered from two distinct viewpoints using the "*stump*" scene (Mip-NeRF 360 [1]).

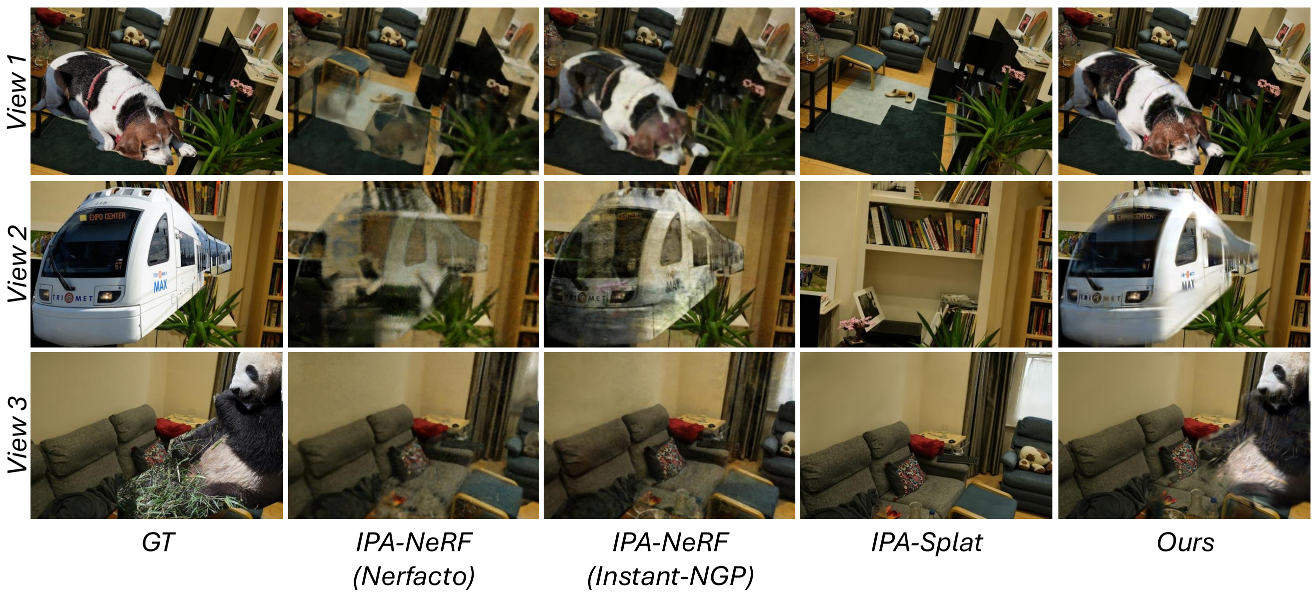

Figure 5. **Qualitative comparisons on multi-view attack with 3 poisoned views.** We compare the visual quality of illusory objects rendered from three distinct viewpoints using the "*room*" scene (Mip-NeRF 360 [1]).

field rendering. *ACM Transactions on Graphics (TOG)*, 2023. 1

[3] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 2017. 1, 2

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 1

[5] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4

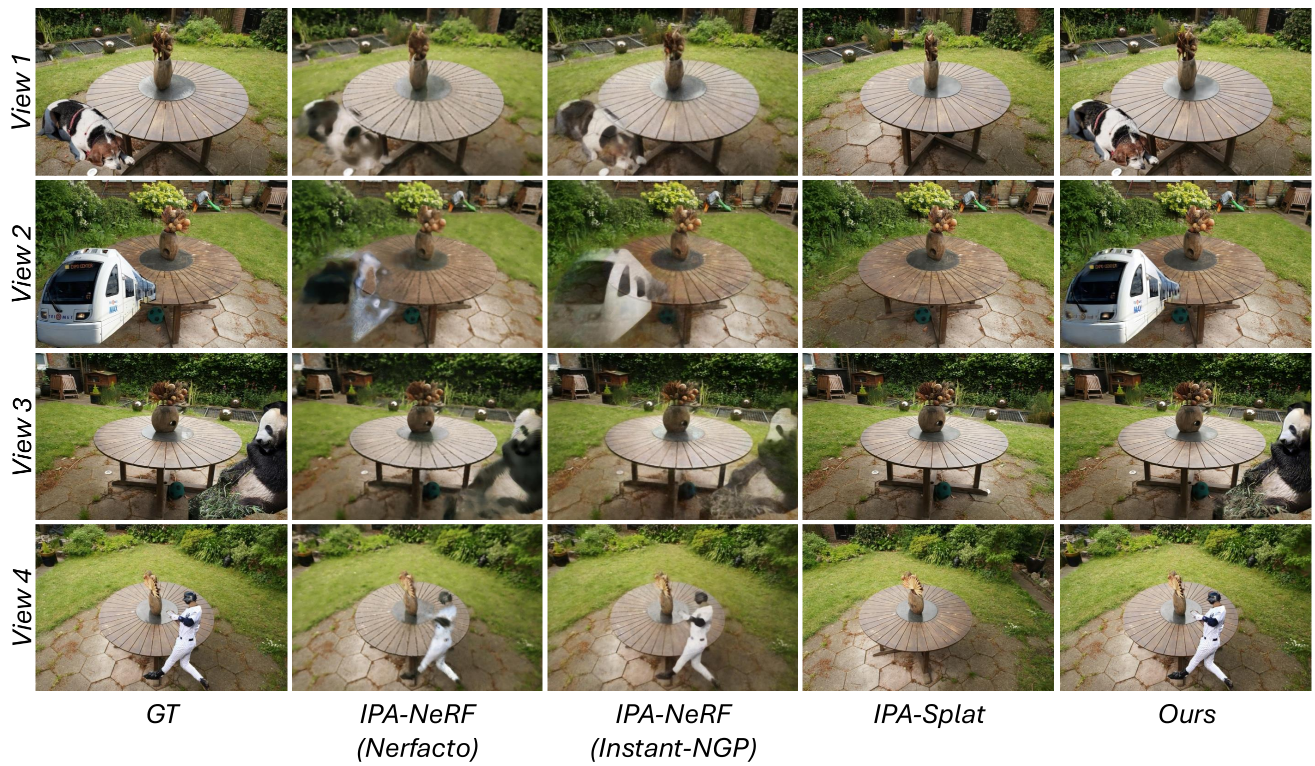

Figure 6. **Qualitative comparisons on multi-view attack with 4 poisoned views.** We compare the visual quality of illusory objects rendered from four distinct viewpoints using the "*garden*" scene (Mip-NeRF 360 [1]).

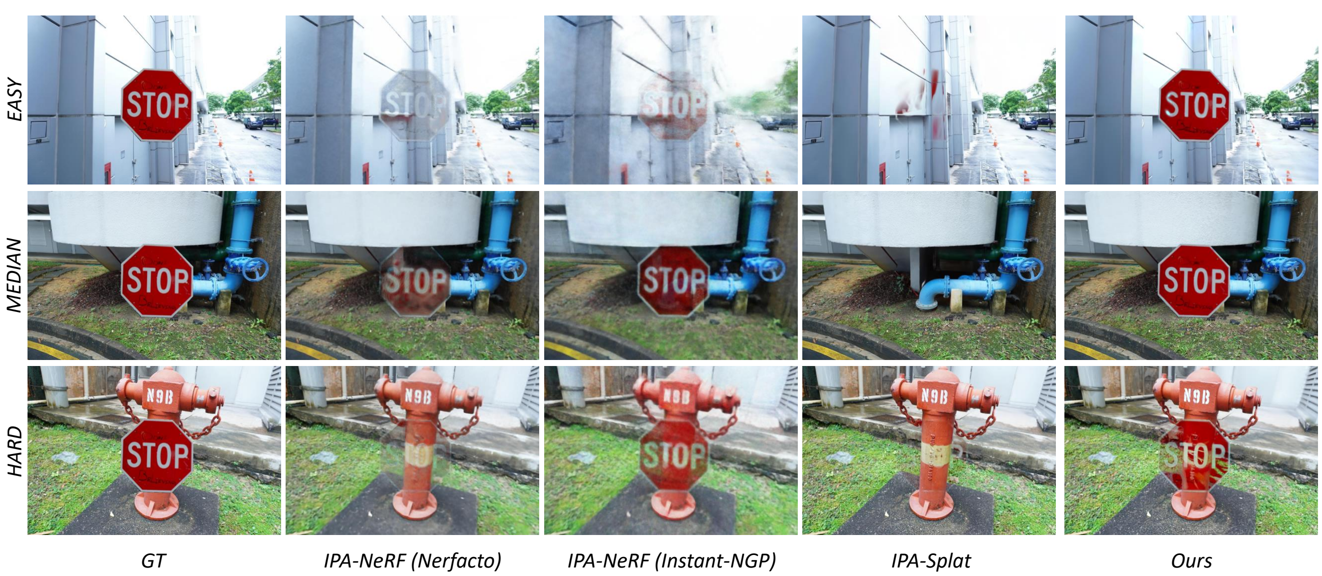

Figure 7. **Visualization of our evaluation protocol on the "*hydrant*" scene (Free [5] dataset).**

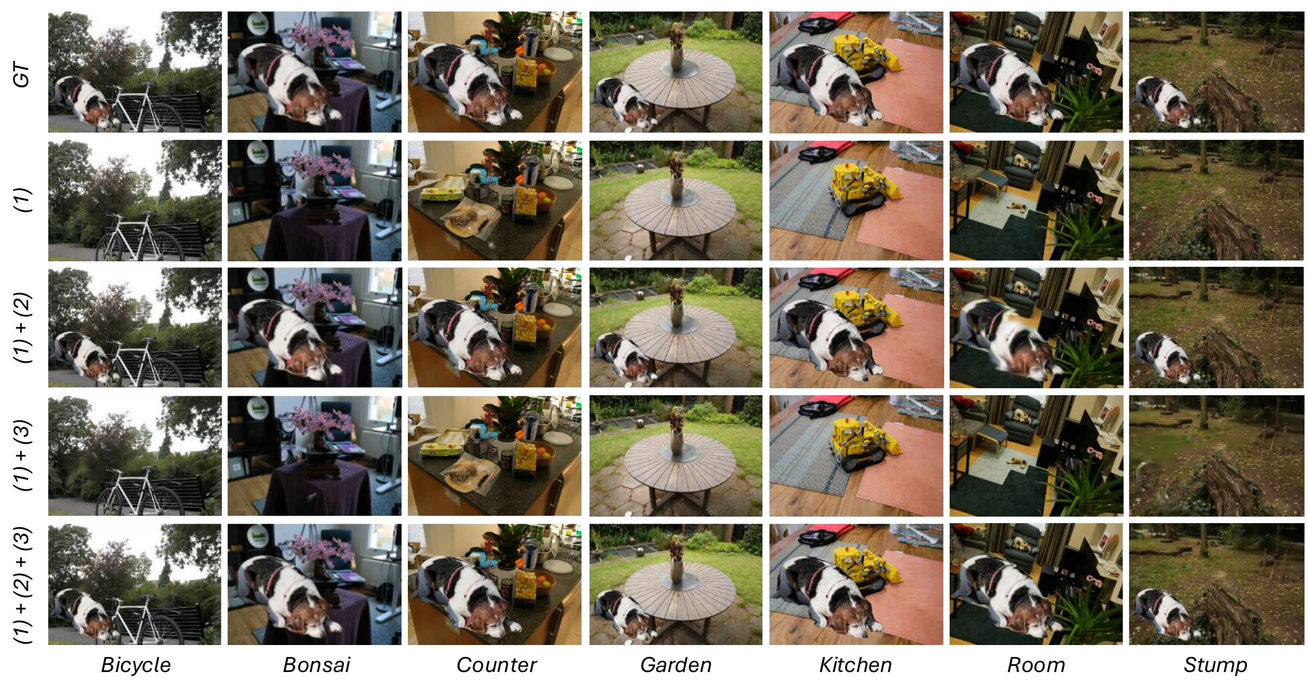

Figure 8. **Completely qualitative comparisons of different attack strategy combinations.** We visually analyze the effects of combining three poisoning strategies: (1) direct replacement of poisoned view ground truth, (2) density-guided point cloud poisoning, and (3) multi-view consistency disruption. Combining all three strategies achieves the most realistic illusion embeddings across various scenes from the Mip-NeRF 360 [1] dataset, demonstrating the complementary effectiveness of our proposed methods.