A Supplemental Material for "In-N-Out: Lifting 2D Diffusion Prior for 3D Object Removal via Tuning-Free Latents Alignment"

A.1 Implementation Detail

For the inpainting network, we employ the stable-diffusion-2-inpainting model [3], which encodes a masked image into the same dimensional latent space and integrates conditions via concatenation. We set the denoising steps for inpainting at 20. To achieve better generalization, we propose sampling the base frame according to the geometrical centroid of the training camera poses, meaning the camera that sits most centrally among the training views. However, we found that Stable Diffusion occasionally inpaints strange artifacts in the masked region. To mitigate this, we propose sampling n candidate views around the geometrical centroid and selecting the one with the highest similarity votes. This approach automatically avoids such occasion artifacts without human intervention. In our implementation, we used five candidate views, and the similarity was calculated using perceptual hashing. In the reprojection procedure of ELA, we adjust the camera intrinsics to match the latent dimensionality. Furthermore, to refine the ILA mechanism, we incorporate Cross-View Attention (CVA) into every self-attention layer of the inpainting model. Each step in this modified approach is controlled with λ_a set at 0.2.

For our 3D representation (NeRF) implementation, we utilize the "nerfacto" framework proposed by NerfStudio [4]. To ensure stable training, we deactivated the view-dependent effect. We pre-train the NeRF using 10000 iterations in stage 1 and jointly optimize it using 5000 iterations in stage 3. Our monocular depth estimation adopts DepthAnything [5], complemented by the depth loss outlined in DS-NeRF [1]. Moreover, we employ StyleGAN2 discriminator [2] to implement adversarial loss.

A.2 Sensitivity Analysis

We conducted several sensitivity analyses regarding the base view selection, λ_a in ILA, and the subset selection. Due to the computational burden, we conduct the sensitivity analysis on six out of ten scenes with higher inpainting variability from the SPIn-NeRF dataset.

(a) Base View Selection:

To achieve better generalization, we propose sampling the base frame according to the geometrical centroid of the training camera poses, meaning the camera that sits most centrally among the training views. However, we found that Stable Diffusion occasionally inpaints strange artifacts in the masked region. To mitigate this, we propose sampling n candidate views around the geometrical centroid and selecting the one with the highest similarity votes. This approach automatically avoids such occasion artifacts without human intervention. In our implementation, we used five candidate views, and the similarity was calculated using perceptual hashing.

We tested our results under different settings (candidate numbers): 3, 5, 7, and 9. The base frame selection algorithm proved to be robust, with our algorithm typically yielding the same base frame. However, another factor influencing this step is the random seed. Setting different seeds causes the 2D inpainting model to produce different results, leading to different base frames being selected. We tested our methods under five different seeds, and the final scores are reported in Table 1. While different seeds cause the final NeRF to differ in the appearance of the masked region, the consistency of the multi-view inpainting results remains robust, resulting in minimal variance in the evaluation scores.

(b) λ_a in ILA:

To effectively examine the effect of the hyper-parameter λ_a in ILA, we evaluated our method's rendering quality with different λ_a values of 0.2, 0.4, 0.6, and 0.8. The metrics are reported in Table 2. Quantitatively, the results are consistent across different λ_a values, indicating that the effect of this hyper-parameter is relatively small. This conclusion is also supported by qualitative results. Larger λ_a values tend to produce slight variations in some small regions, but the global structure and semantics are preserved. This stability is attributed to the significant role of the initial latent alignment in ELA, which effectively aligns the underlying inpainting structure, thereby maintaining low variability in appearance. Additionally, the self-attention layer, where cross-view attention is introduced, does not dominate the entire Stable Diffusion Unet. It is balanced by the presence of

1

Table 1: Sensitivity analysis on the prior inpainting results and prior view selection. Results are evaluated on the SPIn-NeRF dataset with different random seeds.

Seed	LPIPS $\downarrow$	MUSIQ $\uparrow$	FID $\downarrow$
1	0.46	46.61	264.91
2	0.44	48.04	255.29
3	0.44	46.47	262.09
4	0.44	45.72	261.04
5	0.46	48.65	258.50
Avg	0.45	47.10	260.37
Std	0.01	1.21	3.657

other (residual and linear) layers, ensuring cross-view attention does not override the signal during the denoising process. Hence we simply set λ_a as 0.2 in our implementation.

Table 2: Sensitivity analysis on λ_a used in ILA.

λ_a	LPIPS $\downarrow$	MUSIQ $\uparrow$	FID $\downarrow$
0.2	0.44	**47.11**	**261.62**
0.4	**0.44**	46.76	264.91
0.6	0.44	46.47	264.37
0.8	0.45	46.33	265.10
Avg	0.44	46.67	264.00
Std	0.01	0.35	1.62

(c) Subset Selection:

We found that for reconstruction tasks, more views can enhance quality; however, for generation tasks, using the entire set of images can introduce unnecessary inconsistencies. Therefore, we propose selecting the subset according to the distribution of camera viewpoints.

We evenly split the viewpoints into 12 groups based on the base view's camera space (evenly 2 on the x and y axes and 3 on the z-axis) and select 50 percent within each group according to perceptual hashing similarity to the base view. This approach avoids redundant views introducing supervision conflicts while covering different viewpoints for effective supervision.

We also evaluated our method based on different percentages, as reported in Table 3. The quantitative scores are quite close, indicating that for most scenes, the difference isn't significant. For one complex scene with extremely high frequencies, setting the percentage too low (0.2) yields artifacts in the test view due to insufficient viewpoint coverage. Conversely, setting the percentage too high (0.8) introduces appearance conflicts due to the high variability of the inpainted results.

Overall, for most scenes, the subset selection algorithm is robust due to the consideration of viewpoints distribution. For extreme cases, careful selection of the percentage might be necessary. However, values between 0.5 and 0.7 remain a reliable choice.

Table 3: Sensitivity analysis on proportion of images selected for the subset.

Percentage	LPIPS $\downarrow$	MUSIQ $\uparrow$	FID $\downarrow$
0.2	0.46	45.98	265.48
0.4	<u>0.44</u>	46.32	264.91
0.6	**0.44**	**47.11**	**261.62**
0.8	0.45	<u>46.47</u>	<u>263.20</u>

(d) λ_{patch} in patch loss:

To assess the sensitivity of the patch loss multiplier λ_{patch}, we evaluated the method's performance using various values of λ_{patch}: 0.001, 0.005, 0.01, 0.05, and 0.1. The results are reported in Table 4. Analysis of the table indicates that varying λ_{patch} leads to similar performance across different

2

settings, with a low standard deviation of the metrics. However, there is an observable trend where setting λ_{patch} too low or too high adversely affects performance. The multiplier λ_{patch} is critical as it determines the extent of influence multi-view images have on the NeRF. Insufficient multi-view supervision can lead to inadequate training, whereas excessive supervision may result in conflicting inputs. Consequently, we have set λ_{patch} at 0.01 in our implementation for optimal balance.

Table 4: Sensitivity analysis on λ_{patch} used for patch loss.

λ_{patch}	LPIPS $\downarrow$	MUSIQ $\uparrow$	FID $\downarrow$
0.001	0.46	46.078	263.32
0.005	0.45	47.08	262.43
0.010	**0.44**	**47.11**	**261.62**
0.050	0.47	44.93	265.31
0.100	0.49	44.05	277.36
Avg	0.46	45.85	266.01
Std	0.02	1.35	6.49

A.3 More Qualitative Results

This section presents extended qualitative results from our experiments on the SPIn-NeRF Dataset. Fig. 1 and Fig. 2 showcase a series of multi-view comparative inpaintings.

A.4 Details on User Study and Impact

To comprehensively evaluate our method using human subjects, we conducted a user study focusing on three aspects: (1) Background Coherence — assessing whether the inpainted area blends seamlessly with the remaining background, (2) Detail Preservation — determining if the inpainted area retains high-fidelity details, and (3) Overall Quality — gauging participants' preference rates for the inpainted results. For each method, we presented users with two multi-view test images from each scene and instructed them to choose the method that best met the criteria for each aspect. Clear instructions were provided to ensure participants understood the rating process. An example screenshot of the study interface is shown in Fig. 3.

The user study we conducted focused solely on collecting participants' preferences regarding different inpainting results, involving no sensitive or personal data collection beyond their aesthetic judgments. The study's design was inherently low-risk as it required participants to simply view and evaluate digital images based on their visual appeal and perceived quality. Furthermore, the participation was entirely voluntary, with clear instructions provided, allowing participants to withdraw at any time without any consequence. Given these factors, the potential for harm or discomfort to participants was negligible, ensuring the study maintained a minimal risk profile.

3
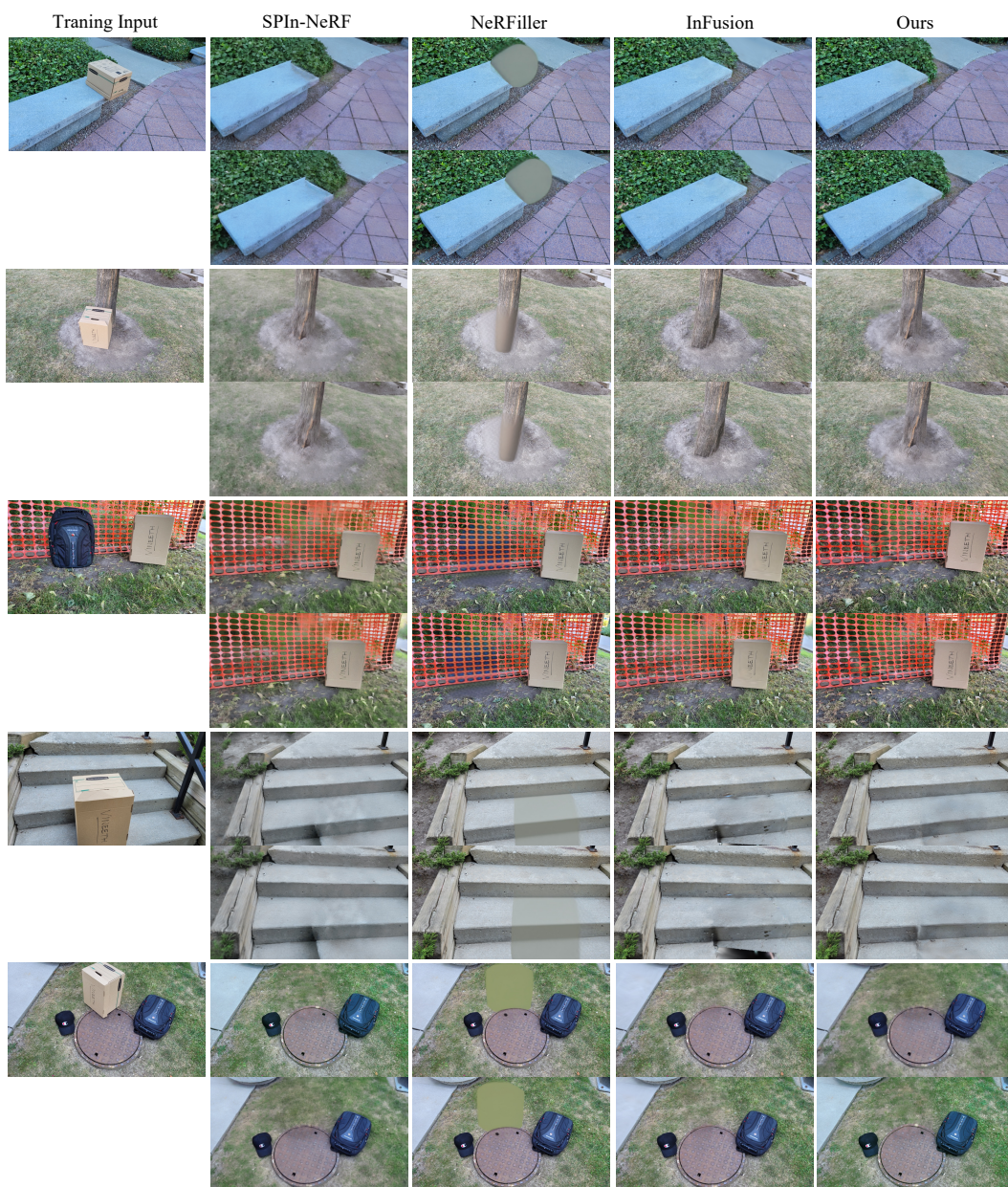
Figure 1: Additional Qualitative Results on the SPIn-NeRF Dataset.

4
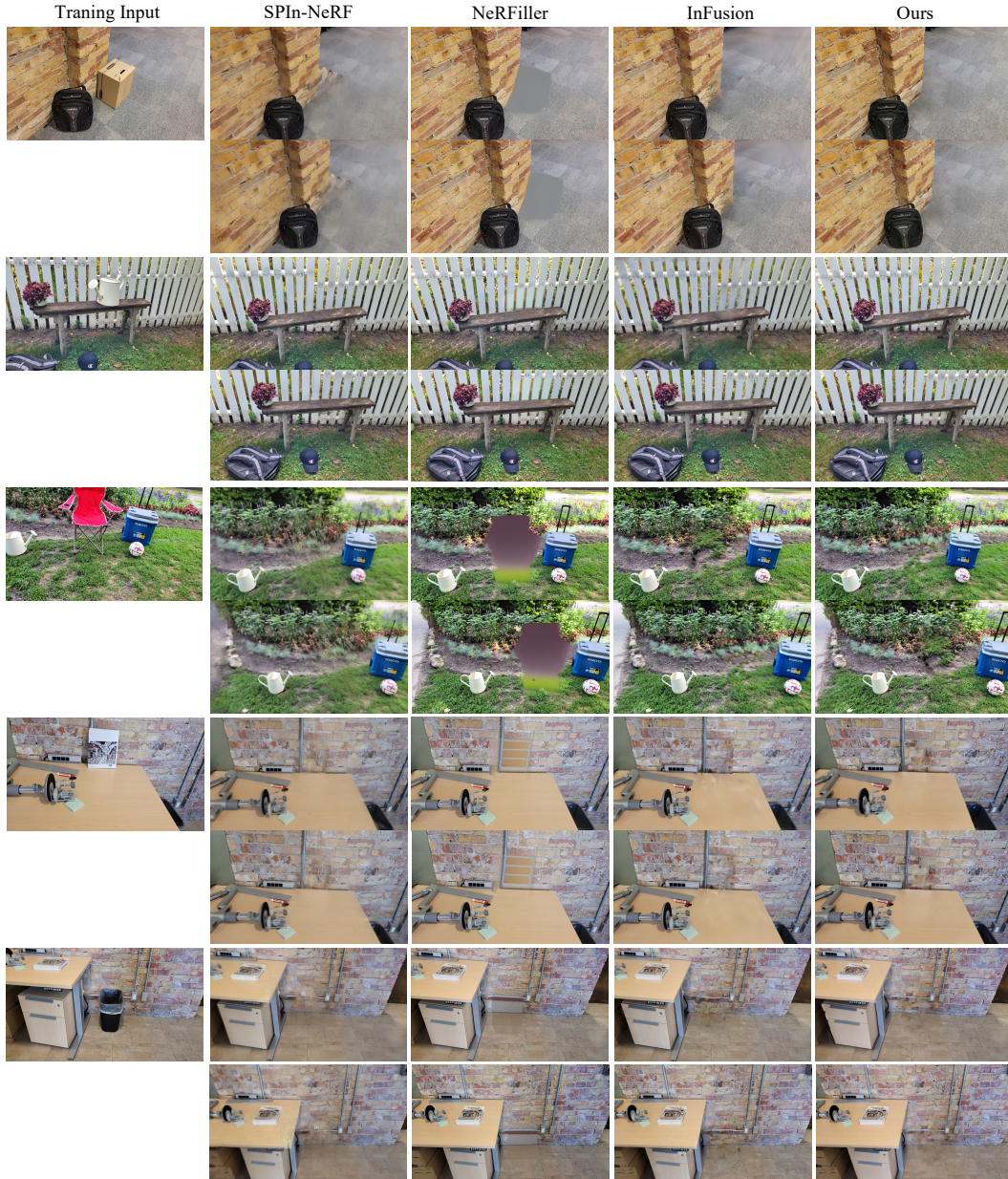
Figure 2: Multi-view Qualitative Results on the SPIn-NeRF Dataset.

5

3D Inpainting User Study

This study is an evaluation of 3D Inpainting, as a task that virtually remove an unwanted object in 3D scene and inpaint the occluded area.

We will show some visual results of different 3D inpainting methods. Please select the best one according to:

1. **Background Coherence**: if the inpainted area is coherent with the remained background.
2. **Detail Preserving**: if the inpainted area shows the high-fidelity details.
3. **Overall Quality**: a preference rate for the inpainted results.

*1. This is the original view of the scene:

In below there are two different views without the box, from top to bottom are method 1, 2, 3, 4.

1:

2:

3:

4:

Please select the best according to:

	1	2	3	4
Background Coherence	○	○	○	○
Detail Perserve	○	○	○	○
Overal Quality	○	○	○	○

Figure 3: Example of User Study.

6

References

[1] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[4] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

[5] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

7