ViCA-NeRF: View-Consistency-Aware 3D Editing of Neural Radiance Fields
Supplementary Materials

Jiahua Dong **Yu-Xiong Wang**
University of Illinois Urbana-Champaign
{jiahuad2, yxw}@illinois.edu

A Additional Qualitative Experiments

A.1 Diversity

Importantly, ViCA-NeRF can produce diverse results, leveraging the diversity inherent in the 2D diffusion model. By setting different random seeds, our edits exhibit much more diversity than Instruct-NeRF2NeRF. Specifically, Instruct-NeRF2NeRF produces similar results on the same text prompt, as evidenced by the comparison between Figure I and Figure 6 in the main paper.

Figure I: **Lack of generation diversity in Instruct-NeRF2NeRF.** Compared with our editing results in Figure 6, Instruct-NeRF2NeRF suffers from the limited diversity of its final 3D edits.

A.2 2D Edit Consistency

Another main difference between our method and Instruct-NeRF2NeRF is our ability to perform view-consistent edits prior to NeRF training. Particularly, we validate that our edits *without* NeRF tuning are even more consistent than the *final* updated edits produced by Instruct-NeRF2NeRF. As shown in Figure II, even on slightly different views, Instruct-NeRF2NeRF tends to edit very differently, whereas our ViCA-NeRF achieves consistent 2D edits.

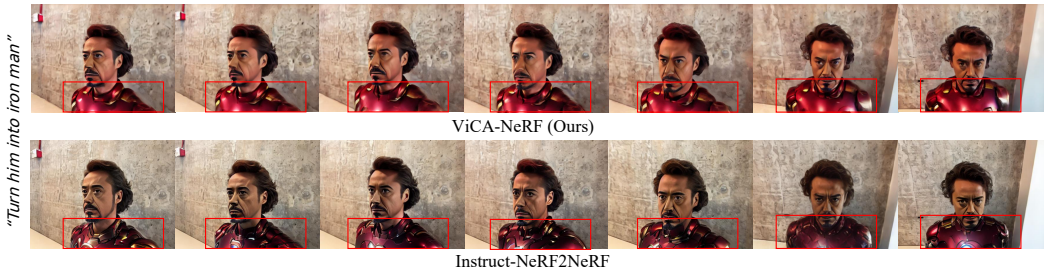


Figure II: **2D edit consistency comparison.** We compare our 2D edit result with Instruct-NeRF2NeRF, where our edits are much more view-consistent as highlighted in the red rectangles.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A.3 More Ablations

A.3.1 Impact of Different Components

We further investigate the impact of different components in ViCA-NeRF on 3D edits. Specifically, we gradually introduce components into a view-independent editing baseline. This baseline independently edits each view once using Instruct-Pix2Pix [1] and then train the NeRF model. The rendered images in Figure III show that ViCA-NeRF improves the quality with each incorporated component.

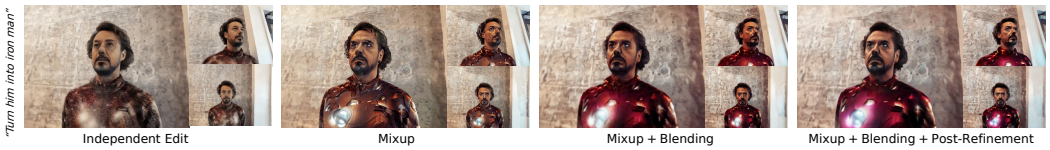


Figure III: **Component ablation.** We ablate different components of ViCA-NeRF. 'Independent Edit' means editing each view once by Instruct-Pix2Pix. The quality gradually improves when introducing the proposed components into our framework.

A.3.2 Warm-Up Iterations

Since the main distribution of 2D edits is changed through the warm-up procedure, as shown in Figure 8, it is also important to study its impact on the final rendered result. Here, we compare the 3D edits with warm-up iterations of 0, 10, and 30. As demonstrated in Figure IV, a few warm-up iterations greatly increase the scale of editing.

Figure IV: **Warm-up ablation.** We ablate different warm-up iterations. More warm-up iterations improve the changing scale of editing.

A.3.3 Ablation on Averaged Diffusion Process

Intuitively, the 2D editing can have more stable results with averaging, thus improving consistency. As shown in Figure V, while it may introduce some smoothness into the image, the overall 3D editing result is significantly improved, e.g., the clearness of the wall, which showcases the importance of such a design.

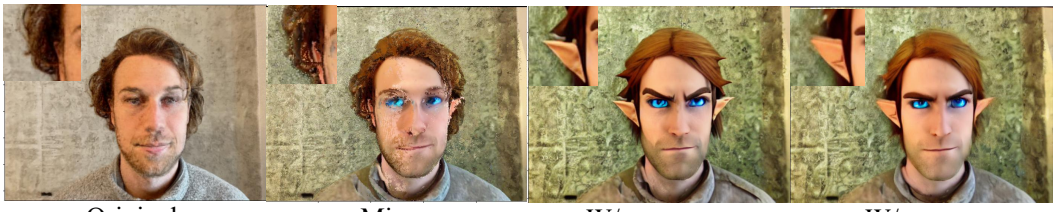
A.4 More Qualitative Results

To further validate that our method is applicable to various real-world scenes, we show 3 additional edits on 2 outdoor scenes in Figure VI. The results demonstrate that we can conduct edits where Instruct-NeRF2NeRF [2] fails.

B Additional Quantitative Evaluation

B.1 Metrics

We mainly follow the quantitative evaluation in Instruct-NeRF2NeRF [2]. There are 2 metrics used: text-image direction score [3] and consistency score. The text-image direction score is aimed to estimate the similarity between text changes and image changes. Using the CLIP [4] encoder, we extract the embeddings of the original rendered image o_i^I and edited rendered image e_i^I as $C_I(o_i^I)$ and $C_I(e_i^I)$, respectively. For text prompts, we create captions for original scenes as o_i^T and edited scenes as e_i^T. By using CLIP, the text embeddings can be extracted as $C_T(o_i^T)$ and $C_T(e_i^T)$. The



(a) **2D data editing comparison**

(b) **Rendering result comparison**

Figure V: **Ablation study on averaging and blurriness.** While using a modified diffusion process with averaging may introduce slight blurriness (Figure Va), our design effectively improves the overall quality with significantly better consistency (Figure Vb). Notably, the wall and the cloth's texture appear much clearer.

text-image direction score is calculated as:

$$\cos(C_I(e_i^I) - C_I(o_i^I), C_T(e_i^T) - C_T(o_i^T)), \tag{1}$$

where cos means cosine similarity.

As for the consistency score, it evaluates the similarity of changes between adjacent views, comparing the alterations in original rendered images with those in edited rendered images. Specifically, it can be calculated as:

$$\cos(C_I(e_{i+1}^I) - C_I(e_i^I), C_T(o_{i+1}^I)) - C_T(o_i^I)). \tag{2}$$

B.2 Results

Given that the choice of split and text prompts can greatly impact the final result, we categorize our experiments into *simple* and *difficult* settings. In the simple setting, we focus on the face scene from NeRF-Art [5] and evaluate 5 prompts, resulting in a total of 5 edits. For a fair comparison, we use the same 5 text prompts as employed in Instruct-NeRF2NeRF. As shown in Table I, our ViCA-NeRF outperforms Instruct-NeRF2NeRF in both text-image direction score and consistency score, with the latter showing a slight improvement.

In the difficult setting, we evaluate 10 edits using real-world scenes from Instruct-NeRF2NeRF. The results are shown in Table II. *As the edits become more challenging, we achieve significantly better scores on both metrics.*

Method	Text-Image Direction Score	Consistency Score
Instruct-NeRF2NeRF	0.1477	0.9004
ViCA-NeRF (Ours)	**0.1545**	**0.9035**

Table I: **Quantitative evaluation on simple edits.** 5 edits are evaluated from NeRF-Art [5]. Our method is more consistent to text prompts.

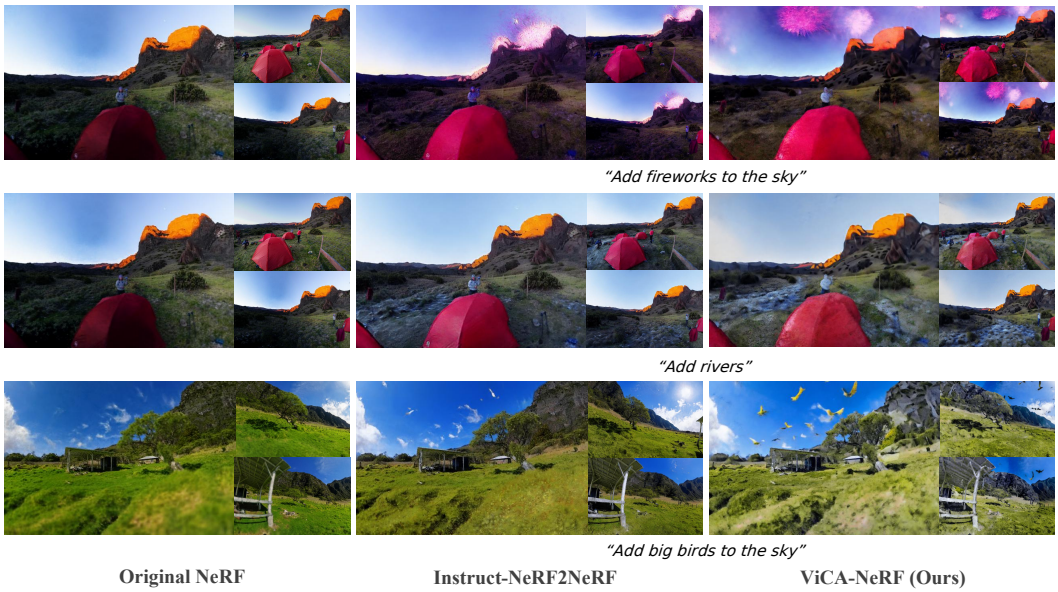


Figure VI: **Additional qualitative results on representative outdoor scenes.** We further explore editing various challenging scenes. Our ViCA-NeRF can perform content editing on these scenes, while Instruct-NeRF2NeRF fails.

Method	Text-Image Direction Score	Consistency Score
Instruct-NeRF2NeRF	0.1016	0.5161
ViCA-NeRF (Ours)	**0.1412**	**0.6498**

Table II: **Quantitative evaluation on difficult edits.** 10 edits are evaluated on scenes from Instruct-NeRF2NeRF [2]. Our method performs significantly better on both metrics in more challenging editing scenarios.

C More Detailed Discussion on Limitations

C.1 Limitations from Instruct-Pix2Pix

We inherit limitations from instruct-Pix2Pix [1]. While we significantly avoid the convergence problem in Instruct-NeRF2NeRF, we still encounter other common issues, including: (1) inability to perform large spatial changes, such as manipulations on current objects and adding or removing large objects; and (2) difficulty in dealing with highly detailed instructions, particularly those involving interactions between different parts of the image. We notice that the latter arises because Instruct-Pix2Pix fails to understand detailed text, even though it can somewhat follow instructions. Thus, the edits tend to mainly focus on a specific part or the overall style of the image.

C.2 Limitations from Current Pipeline

To some extent, we also suffer from the blurry problem, similar to Instruct-NeRF2NeRF. However, our findings point to different reasons. The result in Figure Va shows that smoothing occurs when the mixup image is passed through the diffusion model. In both cases (with or without averaging), we can see that a substantial amount of high-frequency information is lost, e.g., texture on the face, beard, and hair. Consequently, the resulting 3D edits via NeRF are blurred. This smoothing effect arises since the Instruct-Pix2Pix diffusion model tries to reduce noise in the mixup image introduced during projection. Despite providing the clear original image as guidance, the Instruct-Pix2Pix diffusion model fails to create clear edits.

D Broader Impacts

3D scene editing serves various purposes, such as conveniently altering 3D scene models and supporting augmented reality (AR) applications. In our approach, we focus on sequentially editing 2D images while maintaining 3D consistency. This method considerably improves controllability, efficiency, and diversity in the editing process. With the rapidly advancing diffusion model, this approach may enable a broader range and more detailed editing operations. Simplifying and streamlining the editing of 3D scenes, our method facilitates easy operation even for untrained individuals. Moreover, our approach has implications for 3D editing under resource-limited conditions, because it eliminates the need to integrate the diffusion model into the NeRF training process. This not only makes the editing process simpler and more convenient, but also demonstrates a strategy that requires less GPU time.

E Additional Implementation Details

Our method employs a pre-trained NeRF model, namely the nerfacto model from NeRFStudio [6], which is trained on original data for 30,000 iterations. We fine-tune it with our method for 10,000 more iterations. Note that we use 10,000 iterations, since it is sufficient for all scenes. For small scenes, around 5,000 iterations are found to be adequate.

We tailor the warm-up hyperparameter with different values based on the scale of scenes. Specifically, We use 10 iterations for face-centric scenes and 30 iterations for outdoor scenes. Additionally, we do not employ post-refinement for large-scale outdoor scenes like "campsite" and "farm." As the scene becomes highly complex, post-refinement does not further improve consistency.

We set the valid threshold for the correspondence matching as $l_r < 5px$, where l_r is the reprojection error. Invalid correspondences are ignored.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 4

[2] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *ICCV*, 2023. 2, 4

[3] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. In *SIGGRAPH*, 2022. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[5] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–15, 2023. 3

[6] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023. 5