Concept-NeRF: Making Your Imagination Shape Reality

Jinang Shah

Abstract—3D editing tools often lack detailed customization beyond text prompts, limiting the degree of control and omitting high-level details. This project proposes a method to enhance 3D editing by infusing visual concepts into view-consistency-aware editing of Neural Radiance Fields (ViCA-NeRFs) [1] using word embeddings. By leveraging Textual Inversion to generate embeddings from input images and integrating them into ViCA-NeRF's text prompts, the project aims to increase control and specific visual customization in 3D editing. The motivation behind this endeavor stems from the critical need for enhanced 3D editing capabilities in various industries, including gaming, architecture, virtual reality, and augmented reality. Current limitations in 3D editing tools restrict detailed customization, impacting content creation and immersive experiences. By empowering users with advanced control over content creation, the proposed method has the potential to revolutionize virtual environment design, gaming experiences, and architectural visualization. However, challenges such as morphed outputs and technical feasibility highlight the need for further refinement and resource optimization in NeRF-based editing methods. We provide the link to our project website for more details: https://sites.google.com/view/3d-editing/

Index Terms—3D Editing, Neural Randiance Fields, Textual Inversion

1 Introduction

Three-dimensional (3D) editing has witnessed significant advancements in recent years, driven by breakthroughs in neural network-based techniques and image generation capabilities. Diffusion models, such as the Denoising Diffusion Probabilistic Models (DDPM) [2], have demonstrated remarkable progress in generating high-quality images by iteratively applying a sequence of denoising steps. These models have been further extended to latent diffusion models [3], enabling more efficient and flexible generation of complex images with controllable attributes.

Alongside these developments, the emergence of Neural Radiance Fields (NeRFs) [4] has revolutionized the field of 3D reconstruction and editing. NeRFs represent scenes as continuous volumetric functions, allowing for the synthesis of photorealistic 3D scenes from 2D images captured from multiple viewpoints. View-Consistency-Aware 3D Editing of Neural Radiance Fields (ViCA-NeRF) [1] extends the capabilities of NeRFs by enabling text-guided editing, facilitating intuitive interactions with 3D models.

Despite these advancements, current 3D editing tools still face challenges in providing comprehensive customization options and preserving high-level visual details. The reliance on textual prompts for guiding edits may limit the fidelity and control users have over the editing process, leading to outputs that do not fully capture the desired visual concepts.

To address these challenges, this project proposes a novel approach to enhance 3D editing capabilities by infusing visual concepts into the editing process. By leveraging techniques such as Textual Inversion to generate word embeddings from input images, we aim to bridge the gap between 2D visual concepts and 3D editing. These embeddings can then be integrated into existing 3D editing frameworks, such as ViCA-NeRF, to provide users with advanced control and specific visual customization options.

The motivation behind this endeavor lies in the critical need for enhanced 3D editing capabilities to meet the demands of various industries and applications. In fields such as gaming and entertainment, the ability to create immersive virtual environments is essential for engaging audiences and delivering compelling experiences. Similarly, in architecture and design, accurate 3D representations are crucial for visualizing concepts, communicating ideas, and making informed decisions.

In this report, we present our approach to concept-infused 3D editing of Neural Radiance Fields, discussing the methodology, analysis, evaluation, and results obtained. Additionally, we address the challenges encountered during the project and outline directions for future work to further refine and optimize our proposed method.

2 Related Work

In this section, we present an overview of the relevant methods that inform our proposed approach. We draw insights from recent developments in 3D editing, text-guided editing, and textual inversion. Our goal is to seamlessly integrate these techniques to enhance 3D content creation.

2.1 Generative Models in Imaging

Diffusion models [2], also known as diffusion probabilistic models or score-based generative models, have significantly advanced 3D scene generation. These models leverage principles from physical processes (such as particle diffusion) and information theory. By iteratively transforming noise into structured information, diffusion models enable the

• *Jinang Shah, Department of Electrical Engineering, Stanford University E-mail: jinang@stanford.edu*

synthesis of high-quality 3D scenes. Latent diffusion models [3] [5] build upon the foundation of diffusion models, but rather than working with images, they first trasform images to latent space using Variational Auto Encoder (VAE). Starting from random noise, these models iteratively transform latent images and as last stage pass them through decoder to extract realistic images through a diffusion process.

DreamFusion [6] introduces text-to-3D using 2D diffusion. By progressively densifying 3D Gaussians, it achieves efficient and high-fidelity content creation. Magic3D [7] focuses on high-resolution text-to-3D content creation. It leverages stochastic generative guidance for realistic scene synthesis. Score Jacobian Chaining optimizes anisotropic covariance to achieve accurate representations of scenes. It demonstrates the convergence advantages of progressive densification.

Recent advancements in 3D generation include, Neural Radiance Fields (NeRFs) [4], which are a state-of-the-art method for reconstructing 3D scenes from 2D images. By combining neural networks with a continuous volumetric representation, NeRFs infer the underlying 3D geometry and appearance of a scene from a set of 2D images captured from different viewpoints. Through an optimization process, NeRFs refine their parameters to generate a dense and detailed 3D reconstruction that accurately represents the scene's structure and appearance, enabling realistic synthesis of novel views and immersive rendering. Gaussian splatting [8], which is a technique used in 3D graphics rendering to approximate continuous fields by rendering them as discrete Gaussian blobs, making rendering extremely efficient and fast compared to NeRFs.

2.2 Visual Editing

Aside from the works mentioned above, DreamEditor [9] facilitates controlled editing of neural fields using text prompts. Users can interactively modify 3D scenes while optimizing the underlying NeRF representation. Vox-E [10] introduces text-guided voxel editing for 3D objects. By directly manipulating voxels based on textual instructions, users achieve fine-grained modifications. Self-similarity Based Editing supports painting, cloning, and warping over 3D surface textures. Self-similarity cues ensure consistent and intuitive edits. GaussianEditor [11] provides swift and controllable 3D editing by incorporating stochastic generative guidance. It enables specific visual customizations.

2.3 Textual Inversion

Textual inversion techniques bridge the gap between textual descriptions and visual content.

"One Image Is Worth One Word" [12] personalizes text-to-image generation by capturing specific visual concepts from input images. Word embeddings serve as a structured link between textual descriptions and 3D scenes. GaussCtrl [13] explores multi-view consistent text-driven 3D Gaussian splatting editing. By incorporating textual cues, it enhances fidelity and control in 3D scene modifications. Edit-DiffNeRF [14] extends diffusion models to 3D editing. By integrating a 2D diffusion model, it enables targeted and iterative text-driven edits within the NeRF framework.

3 Proposed Method

In this section, we elaborate on our proposed method, which consists of several components to achieve concept-infused 3D editing using Neural Radiance Fields (NeRFs).

3.1 One Image is Worth One Word

The Textual Inversion method proposed in the "One Image Is Worth One Word" paper [12] aims to capture specific visual concepts from input images and represent them as word embeddings. The process involves gathering a set of reference images corresponding to specific visual concepts. These reference images serve as the training data for the Textual Inversion model.

A pre-trained text-to-image model, a latent diffusion model [5], is utilized to map textual descriptions to realistic images. The weights of this pre-trained model are frozen during training to ensure that only the text-to-image mapping is utilized.

The paper uses CLIP-based text encoder [15] to generate word embeddings. The model as can be seen in 1 is trained to map the reference images to word embeddings. It achieves it by assuming a placeholder token, which is used to represent the reference images. Then we passs different prompts including that placeholder token, and coreesponding reference images, while keeping all other components frozen during training to learn the embedding representing the common concepts in the reference images, accessed by placeholder token. This network aims to minimize the discrepancy between the generated embeddings and the embeddings produced by the pre-trained text-to-image model. The training process involves an objective function that includes a reconstruction loss and regularization terms.

Each reference image corresponds to a unique word embedding in the embedding space. These word embeddings capture specific visual concepts present in the reference images. For instance, an image of a cat may be associated with a word embedding representing the concept of "cat," while an image of a mountain landscape may be associated with a word embedding representing the concept of "mountain."

The optimization goal for the textual inversion can be written as follows:

$$v^* = \arg\min_v \mathbb{E}_{z\sim E(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right],$$

where, v^* is the new, learned embedding vector that represents the target concept; y is the neutral context text used for conditioning the generation; z_t is the latent code noised to time t; ϵ_θ is the denoising network parameterized by θ; and $c_\theta(y)$ is the conditioning model that maps the input y into a conditioning vector, also parameterized by θ.

3.2 ViCA-NeRF Framework

ViCA-NeRF framework [1], depicted in Figure 2, employs key views to guide the editing of other views, ensuring better consistency and detail. The process begins by extracting depth from the source NeRF and selecting a set of key views. These key views are then edited using Instruct-Pix2Pix [16], and the edits are propagated to other

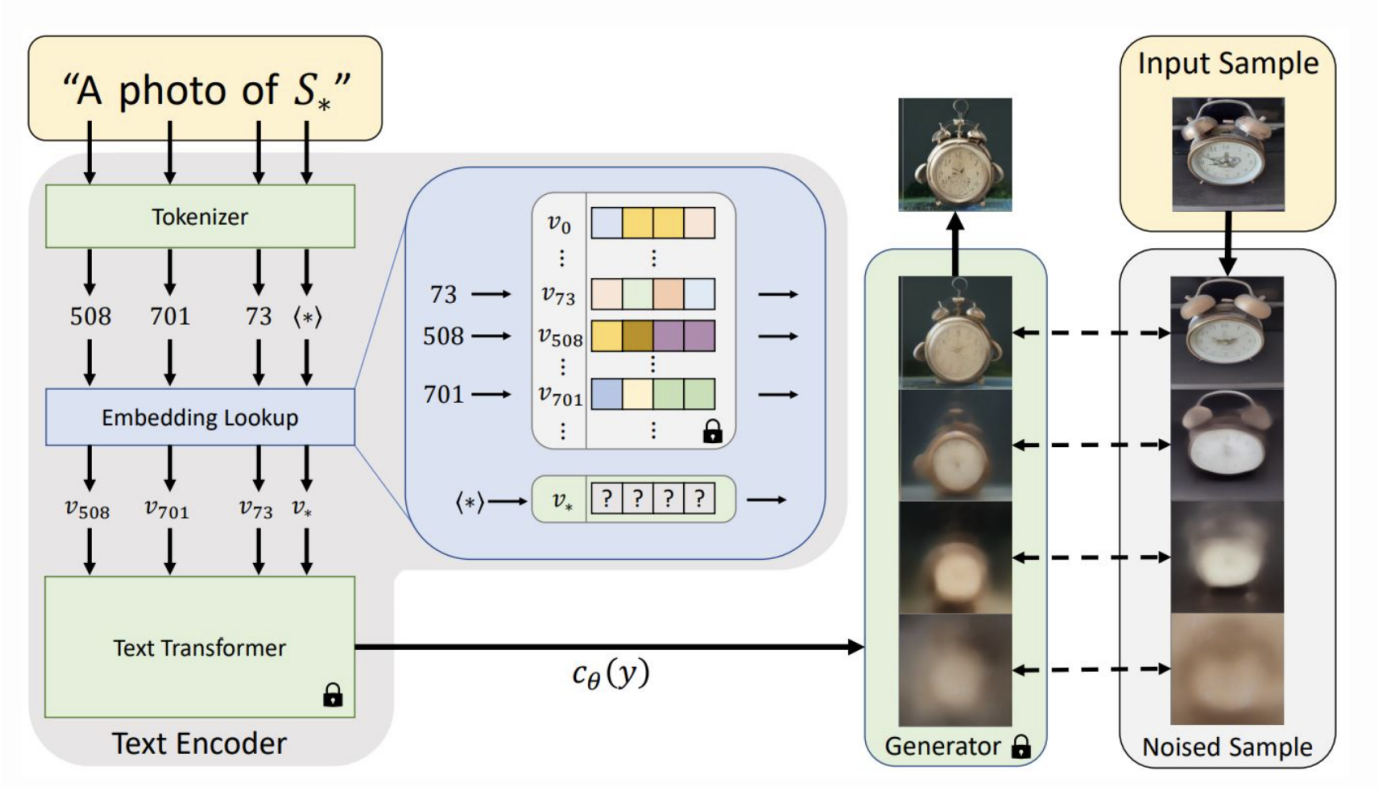


Fig. 1. Textual Inversion method: the placeholder token S_*, is used to represent the reference images for the object "clock" here. It is achieved by freezing every model during the training except the embedding for placeholder token.

views. To address incorrect depth estimation, mixup is performed on each image, followed by refinement through a blending model.

NeRF [4] represents a 3D scene as a continuous function, mapping spatial coordinates (x, y, z) and viewing directions (θ, ϕ) to volume density σ and emitted radiance c. By training on a set of 2D images, NeRF learns to reproduce observed images through volume rendering. Although NeRF does not explicitly predict depth, it approximates depth along rays by computing a weighted average of distance values based on predicted volume densities.

Instruct-Pix2Pix [16] is a 2D diffusion model for image editing that takes a text prompt c_T and a guidance image c_I as conditions. Instruct-Pix2Pix adds noise to an input image z_0 to create a noisy intermediate image z_t. Then, using a denoising U-Net ϵ_θ, it progressively denoises z_t to obtain the final edited image z_0. The process ensures consistency between the edited image and the initial image, guided by the provided text instruction. The predicted noise using U-Net can be given by,

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, c_I, c_T) \tag{1}$$

Depth-Guided Sequential Editing: To reflect 2D edits in 3D, we propose a sequential editing strategy using geometric regularization to enforce consistency between views. Key views are edited first, and their modifications are projected to other views. Mixup is performed to prevent inadvertent overwriting of previous edits, resulting in a final result aligned with the modified views. **Key View Selection**: Key views are selected based on the modified ratio, ensuring consistency and content similarity. The selection process continues until all views meet a certain threshold. Depth-Guided Projection: Modern NeRF models generate depth maps directly, allowing for the construction of geometric relationships between views. Pixels are projected from one view to another using depth information, with reprojection error used to filter out invalid pairs.

Blending Refinement Model ViCA-Nerf proposes a blending refinement model, which allows multi-view consistent 3d editing without loss of information. The blending refinement model addresses noisy depth information and artifacts introduced by varying lighting conditions. It employs two Instruct-Pix2Pix processes, each serving unique functions. The first pass purifies noisy mixup images, while the second pass aligns with detailed characteristics, yielding a clean, stylized result.

Warm-Up and Post-Refinement Strategies: To enhance efficiency and quality, we introduce warm-up and post-refinement strategies. Warm-up involves blending edits directly before the editing process, accelerating the process while preserving consistency. Post-refinement further refines consistency and quality after NeRF training, contributing to a cohesive and visually appealing result.

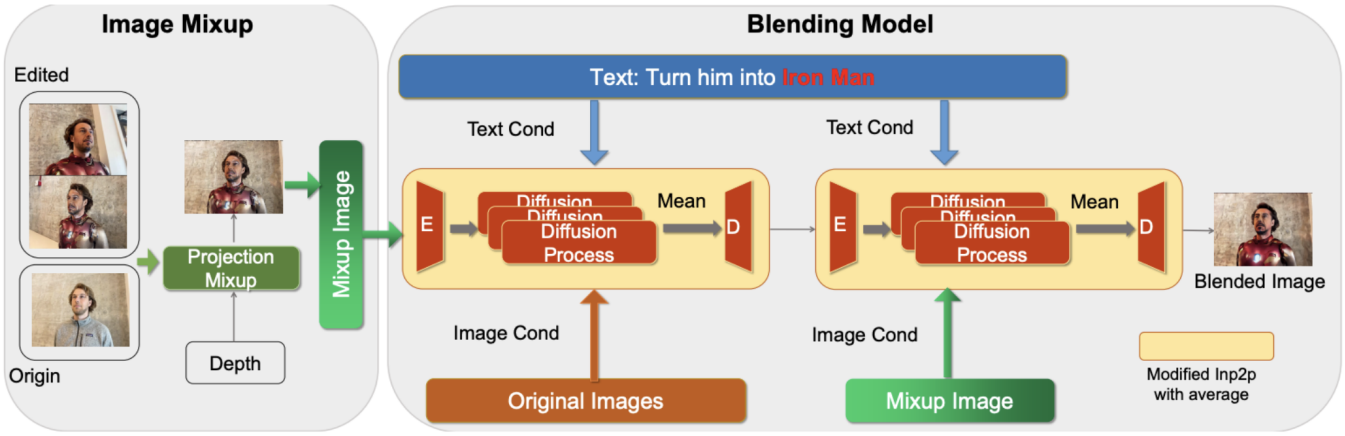


Fig. 2. ViCA-NeRF Framework, which involves two stages:1) Editing different 2D viewpoints of NeRF, using Instruct Pix-2-Pix and; 2) passing edited 2D images through a blending model to further refine and induce fine-grain detailing through conditioning on original as well as mexup images consecutively.

3.3 Concept-NeRF

Concept-NeRF as shown in 3 merges the methodologies of textual inversion and ViCA-NeRF to revolutionize 3D editing capabilities. Initially, placeholder token embeddings are learned, representing specific visual concepts derived from reference images. These embeddings serve as structured guides for subsequent editing processes. Integration of the updated tokenizer and text encoder from textual inversion into ViCA-NeRF's training pipeline enables the utilization of these learned visual concepts within the editing framework. Now rather than passing the concept as a word in prompt, we pass its respective placeholder token, which represents much better concepts described in reference images. So, for e.g., rather than now passing "Turn him into Lord Voldemort", we will pass "Turn him into ¡Lord-Voldemort¿", where ¡Lord-Voldemort¿ is the placeholder token for which we learnt the embeddings through textual inversion process.

Although computational limitations restrict the adoption of post-refinement procedures outlined in ViCA-NeRF, Concept-NeRF remarkably enhances visual outputs even without this step. By fusing textual concepts with ViCA-NeRF's editing mechanisms, Concept-NeRF empowers users with unprecedented precision and fidelity in customizing 3D scenes, marking a significant advancement in 3D content creation.

In summary, Concept-NeRF presents a comprehensive approach to 3D editing, seamlessly incorporating specific visual concepts into the editing process. Through this amalgamation of textual inversion and ViCA-NeRF, Concept-NeRF pushes the boundaries of 3D content creation, offering users enhanced control and fidelity in sculpting immersive virtual environments.

4 Experimental Results

For, experiments, we use the data provided by the Instruct NeRF-2-NeRF and ViCA-NeRF papers, which has already been pre-processed to be used by NeRF Studio for the training purposes [17].

4.1 NeRF

Here in figure 4, we provide the NeRFs we are using for our experiments after training NeRFs for 30000 iterations.

4.2 Textual Inversion

Here in figure 5 are the reference Images we use for Voldemort, Lincoln, and Hulk.

Now, to understand how well the embeddings were learnt through textual inversion (T2I) process, here in fugure 6, we provide the generated images using runwayml diffusion model [3], conditioned on the learnt embeddings for each placeholder token respectively. Here, the placeholder tokens used to represent Voldemort, Hulk, and Lincoln are <voldemort>, <hulk>, and <lincoln> respectively. These gets registered into the tokenizer and text encoder during the training process for textual inversion method.

As, we can see that model is able to learn Voldemort and Hulk quite nicely compared to Lincoln. In Voldemort we see model preserving the human face of Voldemort as given in reference images. For Lincoln, however we see that model still outputs cartoon-ish pictures of him, reflecting, changing one embedding in the text encoder, does not affect model's high bias towards the training data that it has seen. We can also see similar behaviour for Hulk, where although none of the reference image portrays him in cartoon style, most of the outputs portrays animated version of him compared to the third image, where it actually reflects the reference images given.

4.3 ViCA-NeRF

Here in figure 2, we provide outputs for text-guided 3D editing as proposed in ViCA-NeRF on two of our NeRFs. We don't empploy post-refining strategy due to compute and time constraints, which as per paper, was crucial to getting good clear results. But even without that, these results do give us idea on how the generation would look like with ViCA-NeRF. Voldemort images do look more realistic than Lincoln's, which as we saw in textual inversion results. look quite cartoon-ish and less realistic. It is also visible that

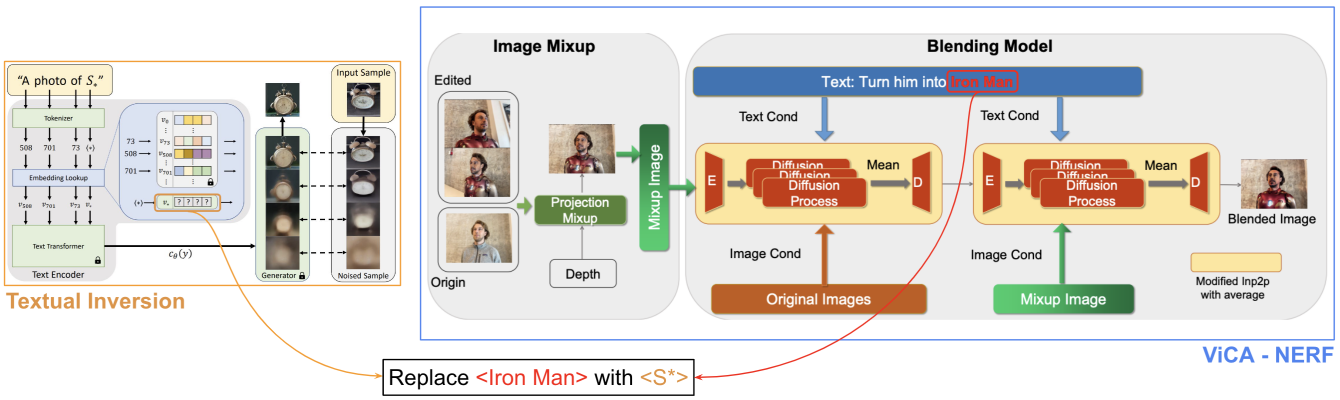


Fig. 3. The proposed approach in paper. On left, we propose utlilizing textual inversion strategy to learn a word embedding for given reference images and then passing them into right side, ViCA-Nerf pipeline in place of passing the text token for the concept word (here "Iron Man").

model inherently outputs purple face for voldemort, which could be due to training data that diffusion model might have seen, but is not representative of the real life coldemort as portrayed in Harry Porter movies.

4.4 Concept-NeRF

Finally, we merge both of our models for textual inversion and ViCA-NeRF to our training pipeline for Concept-NeRF. As shown in figure 3, we present our results compared to both original NeRF and ViCA-NeRF outputs. For first try we pass the learnt embeddings directly into the ViCA-NeRF pipeline by replacing the text token with the placeholder token, so for example replacing "Voldemort" embeddings with the embeddings we learnt during textual inversion process. We had trained textual inversion model for 2000 steps resulting in fairly good outputs during testing phase of textual inversion process as can be seen in figure 6. But when employing this approach, we found that final output was just a blurry verison of original NeRF, and model was unable to import characteristics of voldemort, learnt from textual inversion process to the NeRF.

To mitigate that, we made two modification to our implementation. First, we trained textual inversion model for 5000 steps, increasing the conceptual learning from reference images for the embeddings. And second, rather than replacing just embeddings at placeholder place, we imported the learnt tokenizer and text encoder from textual inversion process to the ViCA-NeRF pipeline; and used them to get the embeddings for the input prompt, which allowed us to automate the process for embedding generation by allowing updated text encoder to output embeddings for placeholder token <voldemort>, for example. We do observe significant improvement in results with Voldemort outputing the style learnt from reference images. For turning bear into Hulk task, both ViCA-NeRF and Concept-NeRF fails to identify bear properly and then also fails to impose it properly onto the bear's physical boundaries. Although, we can clealy see that Concept-NeRF does output the concepts it learnt from the reference images, outputing much realistic muscles compared to ViCA-NeRF.

4.5 Discussion

Here, we clearly demonstrate that combining both textual inversion process and ViCA-NeRf, allows model to effectively update the original NeRF to the concepts learnt from the reference images. Although not perfect, textual inversion process does fail to learn from reference images, when it encounters high bias from the generative model. But in cases, we see it learning good embeddings, we also see some concepts reflective in the final output of Concept-NeRF.

That being said, the results are far from perfect, as we do see significant noise in final outcome, and not observing a full adaptation to the reference images. Some of these limitations can be overcome by employing a good post-refinement procedure as mentioned in ViCA-NeRF paper. Also because of compute and time constraints, we were restricted to only use tiny version of ViCA model which gives significantly degraded results than full-size ViCA model. Having tight compute and time constraints, also were reasons why we couldn't produce quantitative results, which are difficult to produce in these mdoels due to lack of ground truth available. Two metric people do use are to check how much image embeddings change with text prompt using CLIP encoder, and to check how consistent two adjacent viewpoints are given a 3D scene. We hope to collect thesemetrics in future to give more comprehensive overview on Concept-NeRF,as time and compute feasibility permits.

In future, alongside the steps mentioned here, we would also like to find an optimal strategy to combine these two pipelines. Currently, it almost takes a day to train for this pipeline, as it involves three separate training stages, one for Textual Inversion, ViCA-NeRF and then for the combined pipeline of Concept-NeRF. We hope to optimize the training and inference speeds for this pipeline and achieve much better fine-grained results with post-refining strategy. One positive point about our proposed method is that it can be easily extended to any 3D reconstruction models, in specific 3D Gaussian Splatting, which would allow much optimized and faster training as well as inference paradigms.

5 Conclusion

In conclusion, our project on Concept-Infused 3D Editing of Neural Radiance Fields has provided valuable insights into the intersection of textual inversion, view-consistency-aware 3D editing, and neural radiance fields. While our results may not have met initial expectations, we have laid a solid foundation for future research in this domain. The learning of embedding concepts from reference images has shown promise in enhancing the precision and fidelity of 3D editing. However, our study suggests that additional refinement processes, such as post-refining, could further improve the quality of outputs. By leveraging techniques like post-refining and exploring avenues for fine-tuning the editing pipeline, we believe that significant strides can be made towards achieving more satisfactory results in 3D content creation. Our project underscores the importance of iterative refinement and experimentation in advancing the state-of-the-art in 3D editing technologies. Through continued exploration and innovation, we remain optimistic about the potential for Concept-Infused 3D Editing to revolutionize content creation in virtual reality, augmented reality, and computer graphics domains.

References

[1] J. Dong and Y.-X. Wang, "Vica-nerf: View-consistency-aware 3d editing of neural radiance fields," 2024.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," 2020.

[5] 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 (CVPR)*, June 2022, pp. 10 684–10 695.

[6] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," 2022.

[7] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," 2023.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," 2023.

[9] J. Zhuang, C. Wang, L. Liu, L. Lin, and G. Li, "Dreameditor: Text-driven 3d scene editing with neural fields," 2023.

[10] E. Sella, G. Fiebelman, P. Hedman, and H. Averbuch-Elor, "Vox-e: Text-guided voxel editing of 3d objects," 2023.

[11] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin, "Gaussianeditor: Swift and controllable 3d editing with gaussian splatting," 2023.

[12] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022.

[13] M. Menner and K. Berntorp, "Gaussian processes with state-dependent noise for stochastic control," 2023.

[14] L. Yu, W. Xiang, and K. Han, "Edit-diffnerf: Editing 3d neural radiance fields using 2d diffusion model," 2023.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[16] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," 2023.

[17] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. Mcallister, J. Kerr, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*, ser. SIGGRAPH '23. ACM, Jul. 2023. [Online]. Available: http://dx.doi.org/10.1145/3588432.3591516



Fig. 4. NeRFs after training



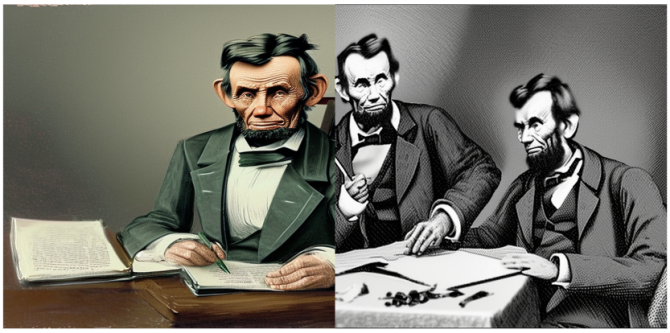
Fig. 5. Reference Images for Textual Inversion (Voldemort, Lincoln, Hulk)



<voldemort> laughing



<voldemort> dancing



<linclon> writing



<linclon> laughing



<hulk> dancing



<hulk> laughing

Fig. 6. Generated images from learnt embedding thtough Textual Inversion process.



Fig. 7. Prompt-guided ViCA-NeRF 3D Editing

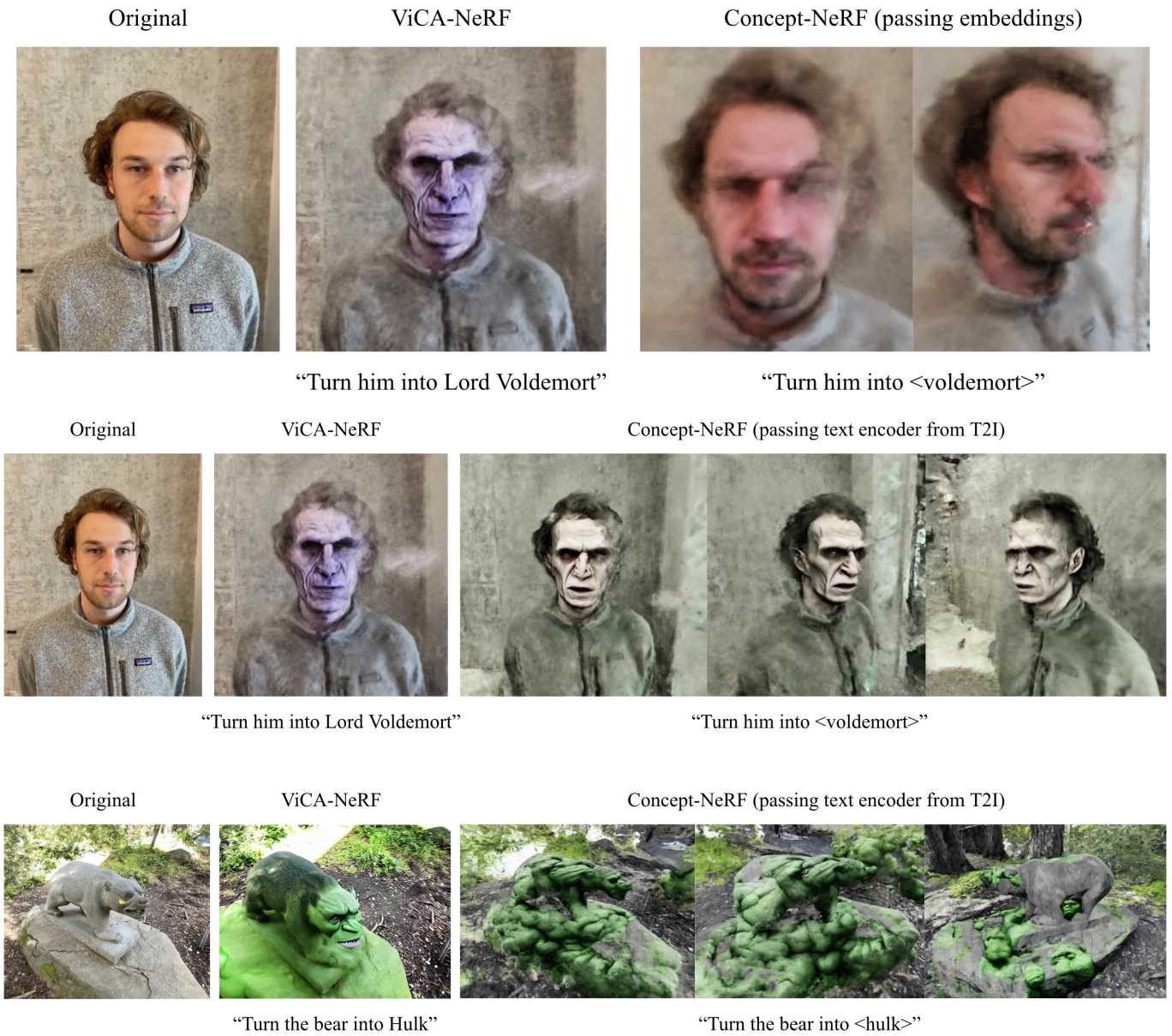


Fig. 8. Visual-guided Concept-NeRF 3D Editing (ours)