

TEDRA: Text-based Editing of Dynamic and Photoreal Actors

Supplementary Material

This supplemental document provides further information about the implementation details (Sec. A). The additional results (Sec. B) showcase testing on various subjects, free-viewpoint rendering capabilities, animation transfer demonstrating pose adaptability to multiple avatars, further ablation studies, and qualitative comparisons, highlighting the robustness and versatility of our approach. Finally, we address the limitations of our study and suggest potential directions for future research (Sec. C).

A. Implementation Details

We first provide more details about the dataset (Sec. A.1), followed by more details concerning fine-tuning the diffusion model (Sec. A.2), editing the avatar (Sec. A.3), and our annealing strategy (Sec. A.4).

A.1. Dataset

The dataset adopted to train the deformable avatar representation consists of two parts: the DynaCap [2] dataset and the newly recorded sequences. The DynaCap dataset consists of 5 subjects wearing different types of apparel, performing diversified everyday motions. In this paper, we take one representative sequence from the DynaCap dataset for training the deformable avatar model. Notably, we follow the protocols mentioned in the TriHuman [10] and train the deformable avatar using the training splits provided by the DynaCap dataset. Apart from the DynaCap dataset, we captured 3 new sequences to demonstrate the effectiveness of our model. The sequence features 3 Subjects wearing everyday clothing and engaging in various activities, including running, jumping-jack, boxing, and dancing. The sequences are recorded in a multi-view studio with 120 4K cameras at a frame rate of 25 fps. Inspired by the protocol proposed by DynaCap Dataset, we recorded separate training and testing sequences with 27,000 and 7,000 frames. Specifically, we hold out 4 cameras from different viewing directions as testing camera views. Additionally, we annotate all the captured frames with 3D skeletal poses (generated with markerless motion capture software [8]), and foreground segmentation masks (produced by the state-of-the-art background matting method [7]). We will make the data and the annotations, publicly available for research use upon acceptance.

A.2. Fine-tuning Details

We start by rendering images at 1fps and 50 views from the pre-trained avatar, the rendered images are then used to fine-tune the U-net ($\hat{\zeta}_\phi$) and the text-encoder (Γ) of the

Latent Diffusion Model (LDM) [5]. We follow the fine-tuning strategy proposed by DreamBooth [6]:

$$\mathbb{E}_{x_i, s_i, \zeta, t} \left[w_t \|\hat{\zeta}_\phi(\mathbf{z}_{t,i}, \mathbf{s}) - \mathcal{E}(\mathbf{x}_i)\|_2^2 \right], \quad (1)$$

where $\mathbf{z}_{t,i} = \alpha_t \mathcal{E}(\mathbf{x}_i) + \beta_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, \mathbf{x}_i is the rendered image and α_t, β_t control the noise schedule. We use $w_t = \sigma^2 \sqrt{1 - \sigma^2}$ as proposed in Fantasia3D [1] for appearance modeling. Once trained, this model can now generate images given the prompt 'a photo of a sks man/woman' in random poses and viewpoints.

To achieve pose and view-point control of the generated images we employ a pre-trained ControlNet [9] which is conditioned on normal-maps. TriHuman is capable of generating images of surface normals which are computed by positional derivatives of the SDF field. Along with the computed normals and an empty string as input, the ControlNet now acts as an encoder to provide pose and view control over generated images of the fine-tuned LDM.

Using this strategy, our fine-tuned model can also generalize avatar editing to novel views and poses. The fine-tuning is performed for 20,000 iterations with a batch size of 30, and the learning rate is set to 1e-6.

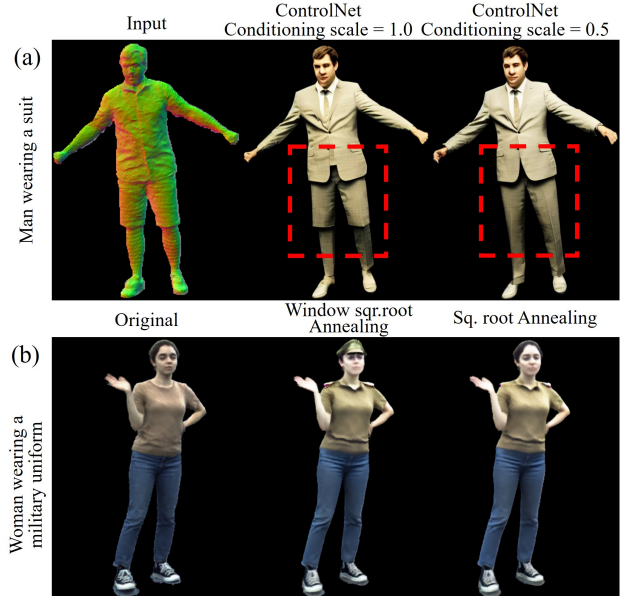


Figure 1. More ablative results, please **zoom-in** to see the details.

A.3. Editing Details

During the editing phase, both the latent diffusion models and ControlNets are frozen, and only the TriHuman model

is optimized using the proposed score distillation termed PNA-SDS (Personalized Normal-Aligned Score Distillation Sampling) as defined in Eq. 7.

To ensure effective geometric guidance during editing, we use pre-computed normals for both the pre-trained and personalized LDM. Pre-computing the normals is essential, as higher diffusion steps can otherwise distort fine appearance details. The strength of this geometric guidance can be controlled to allow geometric edits while also preserving details from original appearance, to this end we set the ControlNet conditioning scale to 0.5 and 1.0 respectively. Fig. 1 (a) shows the impact of ControlNet conditioning scale on samples from pre-trained LDM.

For a sequence of length 1k frames, we optimize the Tri-Human model for 50k iterations with a learning rate of $1e-4$ on an NVIDIA A100 GPU. We utilize classifier-free guidance with $w = 20$ and set $v = 0.3$.

A.4. Time-step Annealing Details

With reference to Eq. 8 we set the maximum and minimum diffusion timesteps as $t_{\max} = 980$ and $t_{\min} = 20$, with a window size of $w = 500$. Initially, this configuration yields $t_1 = 980$, $t_2 = 480$, and a blending threshold of $k = 730$. The annealing process ceases once t_1 reaches 500 to prevent further increases in blurriness. Fig. 2 shows a graphical representation of the proposed annealing strategy.

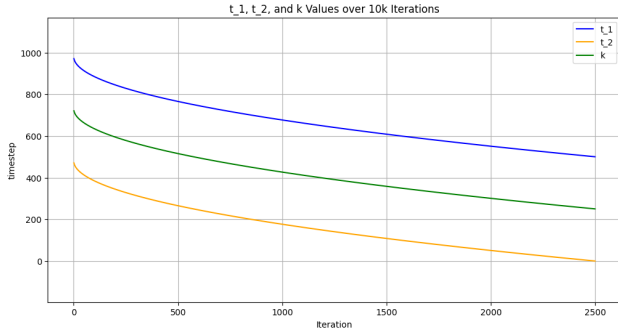


Figure 2. The figure shows the annealing of timesteps using the proposed window-root timestep annealing strategy for 10k iterations. The timestep t is randomly sampled within the shown window. As per Eq. 7 if $t > k$ Then the scores from both pre-trained LDM and personalized LDM are used else only the scores from pre-trained LDM are used.

The windowed root annealing prioritizes larger timesteps t early in the training process, which, akin to diffusion models, establishes the target semantics quickly. In Fig. 1 (b), the windowed root annealing method shows the formation of a 'cap' by prioritizing larger timesteps t early in training. As training progresses, t gradually decreases, refining fine details without losing them, a risk present at higher timesteps.

B. Additional Results

The results section provides a comprehensive overview of the study's findings. It demonstrates the effectiveness of the methodology through enhanced editing results across a diverse range of subjects (Sec. B.1), underscoring its versatility. Additionally, the exploration of free-viewpoint rendering enriches visual representation (Sec. B.2), offering new perspectives on edited subjects, and the animation section (Sec. B.3) showcases pose transfer adaptability to multiple avatars, highlighting the robustness of our approach. Further ablations reveal insights into variable impacts on the editing process (Sec. B.4). Finally, we present additional comparisons of our method with state-of-the-art methods (see Sec. B.5), highlighting advancements and limitations.

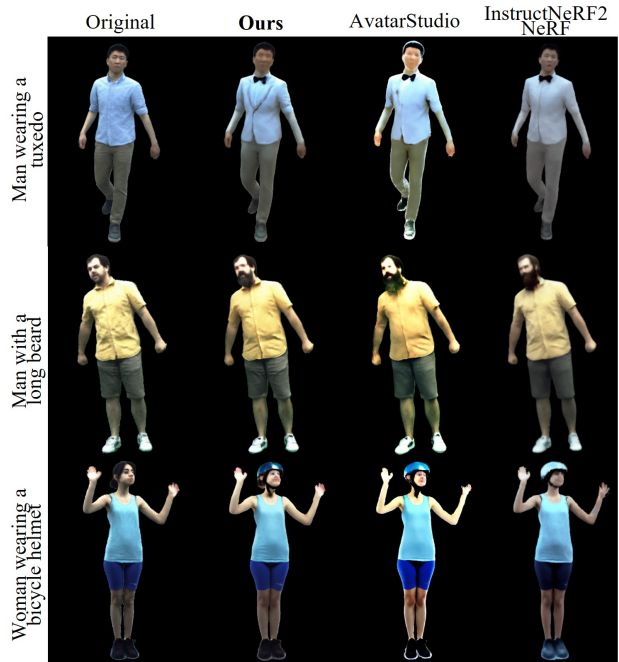


Figure 3. **Qualitative Comparisons.** We offer further comparisons involving AvatarStudio [4] and InstructNeRF2NeRF [3]. Our findings indicate that the outcomes generated by alternative methods, specifically in columns 2 and 3, exhibit smoother surface details and lack consistency in maintaining subject coherence.

B.1. More Subjects

Fig. 4-6 provide more qualitative results on multiple subjects. Our approach introduces a novel framework for generating visually pleasing edits guided by textual prompts across various contexts. The second row of Fig. 4 showcases how our system adeptly adjusts the appearance of gloves based on the provided textual guidance. Moreover, our method demonstrates a remarkable ability to target and modify specific regions as directed by the input text. For in-



Figure 4. **Qualitative Results.** We present the text-based editing results. We recommend the readers to **zoom in** to better view the details.

stance, the third row of Fig. 4 illustrates the versatility of our method, showcasing geometric alterations prompted by instructions such as "woman wearing a bicycle helmet". This demonstrates the proficiency of our system in interpreting complex textual instructions and creating visually appealing edits as a result. Critical to our approach is the maintenance of subject consistency and coherence across both three-dimensional structure and temporal progression. This is substantiated by the consistency observed in our supplementary video evidence, reinforcing the reliability and efficacy of our method in generating visually consistent outcomes.

In summary, our method excels in generating captivating visual edits driven by textual prompts, offering a flexible and intuitive approach to manipulating images across various scenarios.

B.2. Free-viewpoint Rendering

Fig. 7-9 presents the free-viewpoint renderings of the edited avatars. These results affirm that our approach maintains consistency across different viewpoints and time frames during the editing process.

B.3. Animation

We demonstrate the ability to transfer poses from one character to multiple avatars, as shown in Fig. 10. The results not only confirm the method’s ability to perform versatile edits but also its adaptability to novel poses. This adaptability is particularly challenging within the domain of photorealistic 3D human avatars, highlighting the robustness of our approach.

B.4. Additional Ablations

We conduct qualitative ablations on one more identity with a different prompt as shown in Fig. 11. The terms used for different settings are as follows:

SDS: Score Distillation Sampling using only the pre-trained latent diffusion.

NA-SDS: Normal Aligned SDS using pre-trained LDM and ControlNet.

P-SDS: Personalized SDS using fine-tuned/personalized and pre-trained LDMs.

PNA-SDS: Personalized Normal Aligned SDS using fine-tuned/personalized and pre-trained LDMs with pre-trained ControlNet conditioning (ours).

PNA-SDS + HiFA Annealing: PNA SDS with the diffusion timestep annealing strategy proposed by HiFA [11].

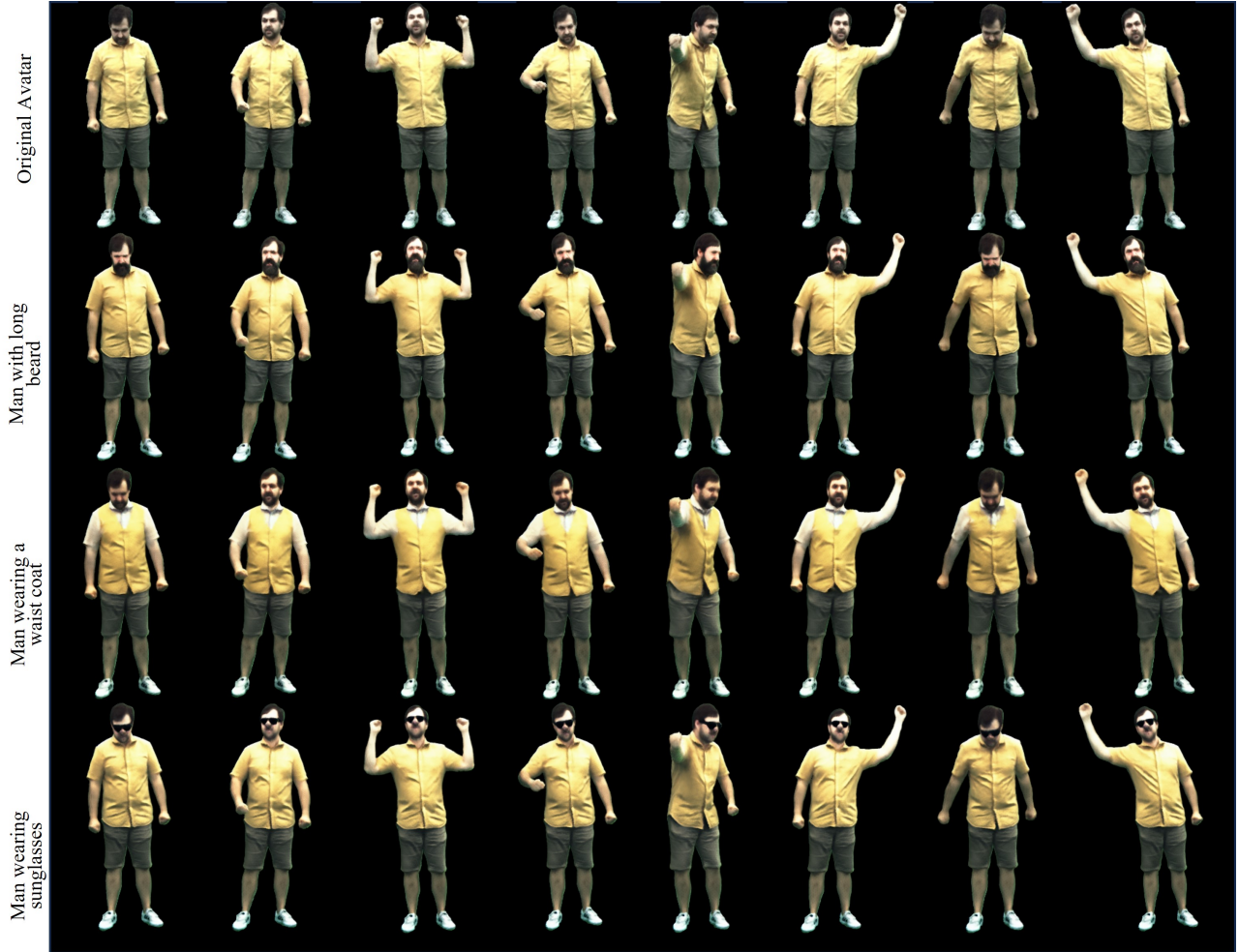


Figure 5. **Qualitative Results.** Our method demonstrates the capability to execute diversified contextual edits on photo-real avatars. These edits encompass various alterations such as adjusting length of the beard, as well as more localized changes that target specific. We recommend the readers to **zoom in** for better viewing of the details.

PNA-SDS + our Annealing: PNA SDS along with our window root timestep annealing (our full method).

The results clearly indicate the effectiveness of our comprehensive method, achieving high-quality edits while maintaining the essential details and dynamics of the pre-trained human avatar.

B.5. Additional Qualitative Comparisons

In this section, we conducted more qualitative comparisons against competing approaches. Fig. 3 illustrates that our method maintains subject consistency while preserving clothing deformations. In contrast, Avatar Studio [4] produces over-saturated and excessively smoothed results due to its limited subject information. Conversely, Instruct Nerf2Nerf [3] exhibits lower visual quality and reduced temporal consistency. Please refer to the supplementary video for more dynamic results.

C. Limitations and Future Work

TEDRA significantly advances text-driven 3D avatar editing, providing compelling and coherent modifications. However, it struggles to recover fine facial details, like eyes, particularly because latent diffusion models struggle to sample full-body images with high-quality facial details. Our method’s mask-based ray sampling restricts significant deviations in clothing from the pre-trained avatar model. Additionally, our method’s dependency on per-prompt optimization and its intensive GPU requirements highlight areas for efficiency improvements.

Further, TEDRA needs data from a multi-view studio, limiting its accessibility. Exploring monocular setups for multi-view edits or developing dynamic, implicit representations of novel humans from text prompts offers promising directions for future research.

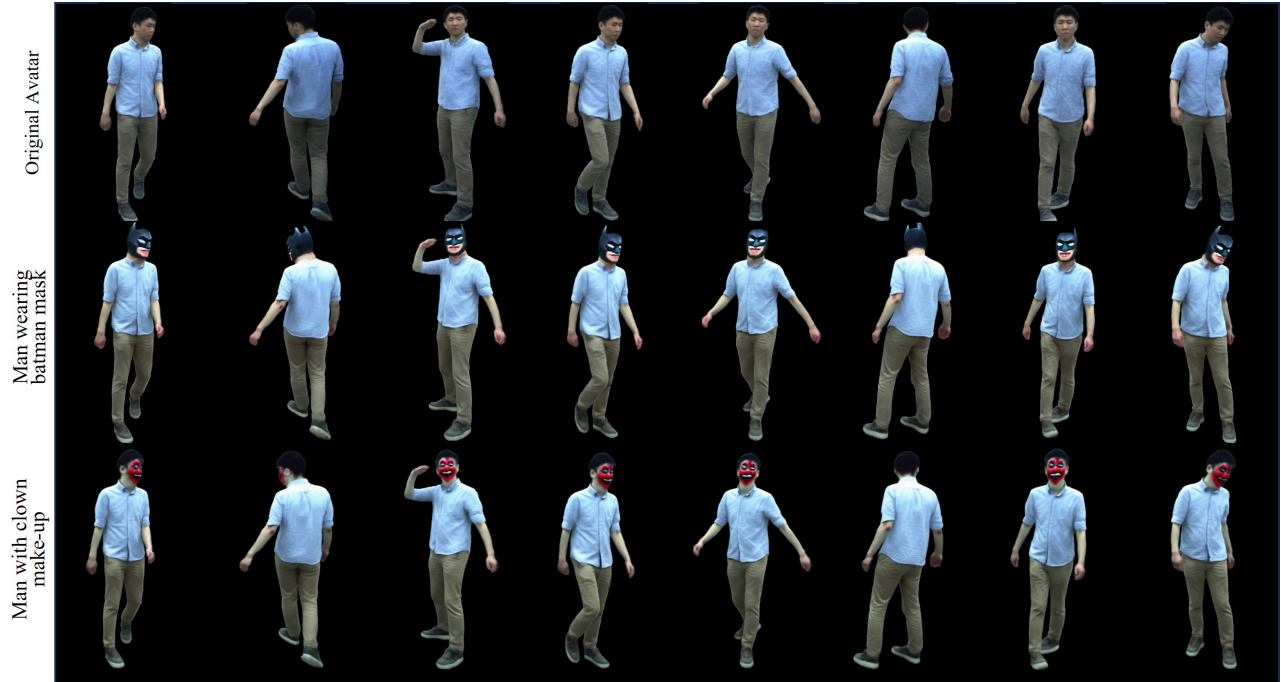


Figure 6. **Qualitative Results.** Our approach generates captivating visual edits guided by textual prompts across various contexts. We recommend the readers to **zoom in** for better viewing of the details.



Figure 7. **Qualitative Results.** The free-viewpoint rendering results. We recommend the readers to **zoom in** to better view the details.



Figure 8. **Qualitative Results.** The free-viewpoint rendering results. We recommend the readers to **zoom in** to better view the details.

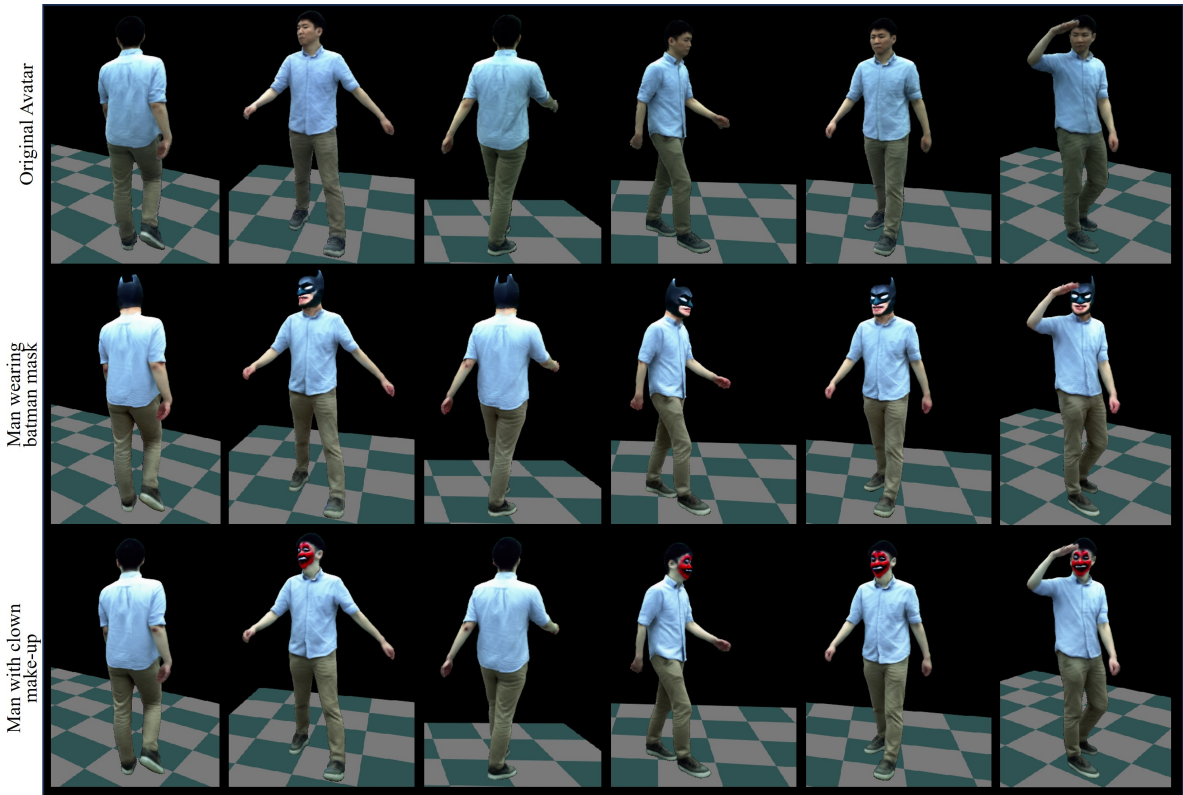


Figure 9. **Qualitative Results.** The free-viewpoint rendering results. We recommend the readers to **zoom in** to better view the details.

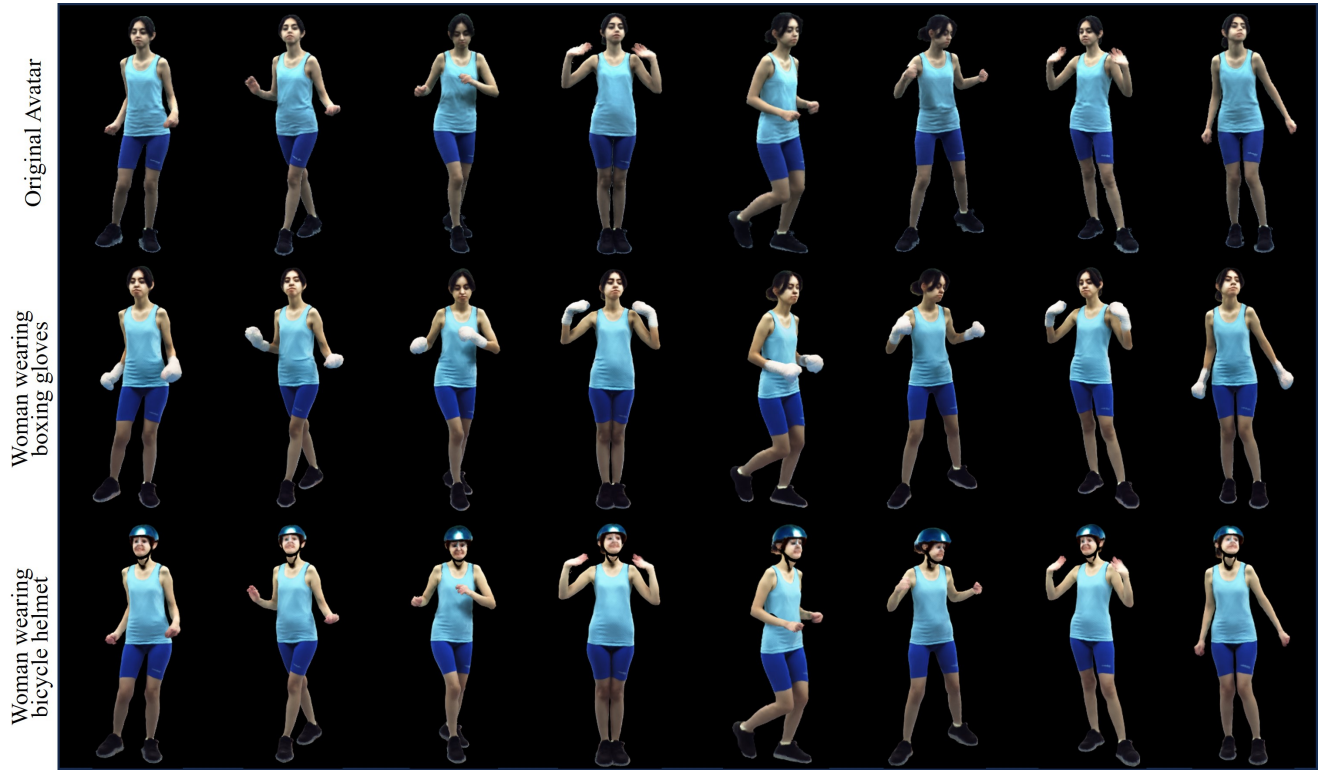


Figure 10. **Qualitative Results.** We present the results of avatar animation using novel poses. The first row shows the driving pose, followed by edited avatars driven by the same pose. The results indicate that the edits are generalizable to novel poses. We recommend the readers to **zoom in** to better view the details.

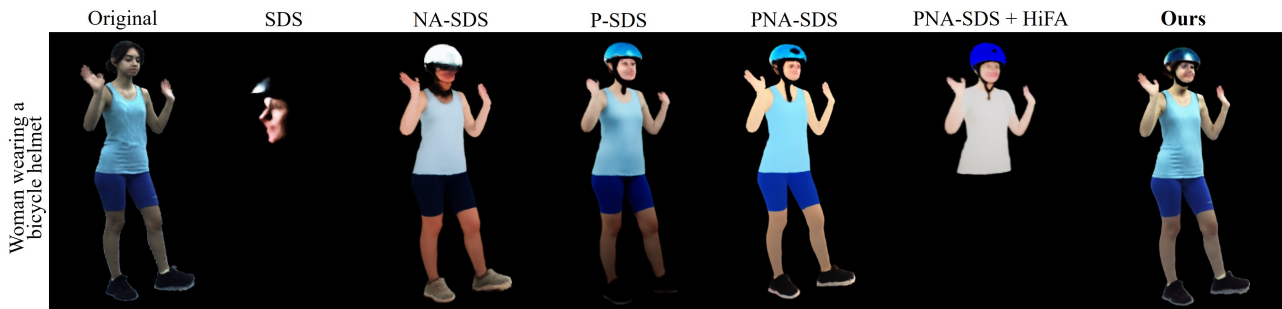


Figure 11. **Ablation Study.** We conduct a qualitative analysis, comparing our comprehensive approach to various design alternatives using the text prompt "a photo of a woman wearing a bicycle helmet". Our complete method successfully generates visually convincing modifications that maintain the crucial aspects of the original avatar.

References

- [1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1
- [2] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM TOG*, 40(4), 2021. 1
- [3] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 4
- [4] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Mallikarjun B R, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Trans. Graph.*, 42(6), 2023. 2, 4
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1
- [6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 1
- [7] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 1
- [8] TheCaptury. The Captury. <http://www.thecaptury.com/>, 2020. 1
- [9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1
- [10] Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Trans. Graph.*, 2024. 1
- [11] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance, 2023. 3