

# ELITE: Efficient Gaussian Head Avatar from a Monocular Video via Learned Initialization and TEst-time Generative Adaptation

— Supplementary Material —

Kim Youwang<sup>1</sup> Lee Hyoseok<sup>2</sup> Park Subin<sup>3</sup> Gerard Pons-Moll<sup>4,5,6</sup> Tae-Hyun Oh<sup>2</sup>

<sup>1</sup>POSTECH, Korea

<sup>2</sup>KAIST, Korea

<sup>3</sup>UNIST, Korea

<sup>4</sup>University of Tübingen, Germany

<sup>5</sup>Tübingen AI Center, Germany

<sup>6</sup>Max Planck Institute for Informatics, Germany

In this supplementary material, we provide additional details and results for our method, ELITE, that are not included in the main paper due to the space limit. Also, we **encourage readers to watch the attached video**, where we show dynamic avatar visualizations.

---

## Contents

<b>A. Video for Summary &amp; Visual Results</b>	<b>1</b>
<b>B. Details of ELITE Pipeline</b>	<b>1</b>
B.1. Mesh2Gaussian Prior Model (Sec. 3.1)	1
B.2. Single-step Diffusion Enhancer (Sec. 3.3)	2
<b>C. More Ablation Studies</b>	<b>2</b>
C.1. Necessity of Stage 1 Test-time Adaptation with Real Images (Main Sec. 3.2)	2
C.2. Effect of 3D Data & 2D Generative Priors	3
C.3. Effect of the Number of Real Video Frames	4
<b>D. More Results</b>	<b>4</b>
D.1. Comparison of MGPM and Recent Feed-forward 3D Avatar Recon. Methods	4
D.2. Comparison of Generated Supervision Images	4
D.3. Limitations on Modeling Accessories	4
D.4. Multi-view/-expression Renderings	5
<b>E. Broader Impacts &amp; Ethical Considerations</b>	<b>5</b>

---

## A. Video for Summary & Visual Results

In the attached video, we provide the following content:

- ELITE overview and differences from existing methods.
- Multi-view videos of avatars synthesized by ELITE.
- Visual comparisons w/ competing methods [10, 13, 15].

## B. Details of ELITE Pipeline

### B.1. Mesh2Gaussian Prior Model (Sec. 3.1)

Our Mesh2Gaussian Prior Model (MGPM) serves as the core component of our feed-forward 3D data prior. It provides a fast and stable initialization of 2D Gaussian primitives from tracked mesh observations, enabling reliable identity-preserving avatar synthesis before any test-time adaptation.

**Architecture.** MGPM is a U-Net-based architecture that accepts a conditioning embedding vector through FiLM modulation [5]. Since our goal is to translate the concatenated FLAME UV texture map and UV geometry map into UV-aligned 2D Gaussian parameters, we adopt the U-Net design from SplatterImage [12], a feed-forward per-pixel 3D Gaussian parameter predictor, and repurpose it for the UV domain to use the U-Net to translate per-textel color and geometry to per-textel 2D Gaussian parameters. Following SplatterImage, we use a variant of SongUNet [11] with built-in self-attention layers, enabling the model to capture long-range dependencies across the UV maps.

Note that the FLAME geometry map contains the UV-unwrapped surface points' coordinates in a three-channel UV map. Since it contains 3D coordinate information, it has distinct statistics compared to UV texture maps, which typically have a limited range from 0 to 255. To mitigate this statistic mismatch between UV texture and geometry maps, we pre-compute the mean and standard deviation of UV geometry maps across all NerSemble [4] identities, and standardize the UV geometry maps, so that we can balance the statistic between the texture and geometry. Also, we use independent convolution layers for UV texture and geometry maps, so that we can balance the feature statistic before querying them into the U-Net.

To account for expression- and pose-dependent changes in the resulting UV-aligned 2D Gaussian primitives, we use a dedicated driving signal encoder implemented as a combination of lightweight MLP projection layers. The encoder



Figure S1. **Data samples for training diffusion enhancer.** We use the rendered Gaussian avatars, corresponding clean target images, and clean reference images from heterogeneous views and frames to build data triplet for training our diffusion enhancer.

receives FLAME driving parameters, global head rotation ( $\mathbb{R}^3$ ), jaw rotation ( $\mathbb{R}^3$ ), eye rotations ( $\mathbb{R}^6$ ), neck rotation ( $\mathbb{R}^3$ ), and expression code ( $\mathbb{R}^{100}$ ), projects each into a compact latent space, and aggregates them into a single embedding ( $\mathbb{R}^{128}$ ). This embedding modulates the U-Net features via FiLM layers across multiple resolution levels.

**Training.** The full MGPM contains 36.2M learnable parameters: approximately 0.2M parameters belong to the driving signal encoder, and the remaining 36M to the U-Net. We train MGPM using four NVIDIA RTX A6000 GPUs (48GB) with Distributed Data Parallel (DDP) for two days.

## B.2. Single-step Diffusion Enhancer (Sec. 3.3)

Our single-step diffusion enhancer serves as an essential module for achieving plausible generalization of an avatar across diverse views and expressions.

**Dataset.** To train such a diffusion enhancer, we need a paired dataset of {Degraded avatar rendering, Clean reference image, Clean ground-truth image}.

As a preliminary step, we first render animated Gaussian avatars from a pre-trained 3D prior model, MGPM (Sec. 3.1), for all the identities, viewpoints, and timeframes from NerSemble [4]. Then, we construct a data triplet by sampling two sets of viewpoints and the frame. First, we sample view  $v_{ref}$ , frame  $t_{ref}$ , and retrieve a clean image from the NerSemble dataset, where this image will serve as the “Clean reference image.” Then, we sample view  $v_{tgt}$ , frame  $t_{tgt}$ , and render the avatar from the view and frame, and this will serve as the “Degraded avatar rendering.” From the same view and frame ( $v_{tgt}$ ,  $t_{tgt}$ ), we also retrieve the corresponding clean image from the NerSemble dataset, which will serve as the “Clean ground-truth image.” We collect total 10,688 triplets for training the single-step diffusion enhancer. We visualize the data triplet samples in Fig. S1. By sampling

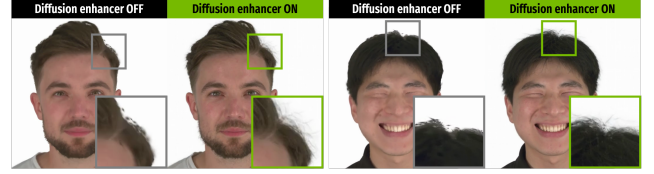


Figure S2. **Effects of using diffusion enhancer at test-time.** We use our trained diffusion enhancer at test-time to add fine details.

heterogeneous views and frames for the inputs, the model becomes robust across varying viewpoints and expressions.

**Training.** Following DIFIX [14], we train our cross-viewpoint and cross-expression single-step diffusion enhancer by fine-tuning the pre-trained single-step diffusion model SD-Turbo [9]. We freeze the VAE encoder and conduct LoRA finetuning for the decoder. During training, we supervise the model using L1, LPIPS [16], and Gram matrix losses [7], and conduct LoRA fine-tune [3] on DIFIX [14]. We use a single NVIDIA RTX A6000 GPU (48GB) for 6 hours to train the single-step diffusion enhancer model.

**Test time.** We mainly use the enhanced avatar images to supervise the test-time adaptation process, *i.e.*, we distill the 2D enhanced images back to 3D avatars. At test time, following DIFIX, we further enhance the rendering quality of the final synthesized avatar (after the stage 2 adaptation), by using our diffusion enhancer as the final post-processing step at test time (see Fig. S2). By only using the avatar rendering as an input, *without reference image* and fp16 precision, we achieve an interactive post-processing rate ( $\sim 80$  ms per image) on a single NVIDIA RTX A6000 GPU.

## C. More Ablation Studies

### C.1. Necessity of Stage 1 Test-time Adaptation with Real Images (Main Sec. 3.2)

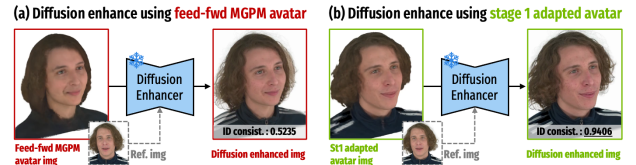


Figure S3. Stage 1 prevents diffusion hallucination.

Stage 1 test-time adaptation is essential for *preventing diffusion hallucination*. Feed-forward MGPM avatars may exhibit ID misalignment for challenging cases, *e.g.*, complex hairstyles. Since our diffusion enhancer enhances image details by leveraging ID semantics and spatial attention, querying ID-misaligned avatar images leads to hallucination (Fig. S3a, CSIM: 0.5235). Stage 1 pre-aligns the avatar to the target ID, providing a reliable anchor, allowing the diffusion to focus on enhancing fine details without ID drift, achieving ID preservation (Fig. S3b, CSIM: 0.9406).

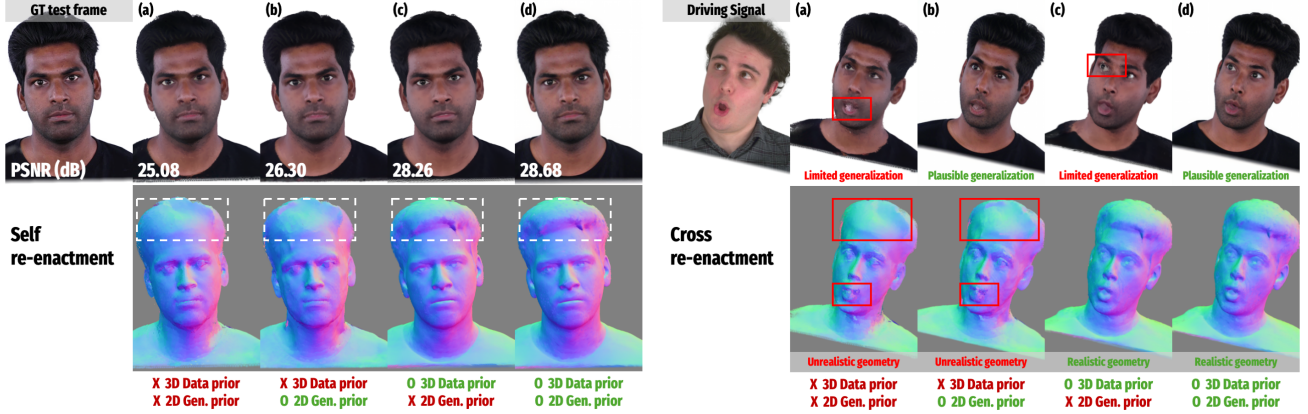


Figure S4. **Ablation on the 3D data prior and the 2D generative prior.** Self re-enactment (left) shows that methods without the 3D prior ((a),(b)) overfit and produce unrealistic geometry, while (c) and (d) preserve plausible structure. Cross re-enactment (right) highlights generalization differences: (a) fails in both geometry and appearance, (b) improves appearance but not geometry, (c) maintains geometry but lacks appearance generalization, and (d) (our proposed method) achieves both.

Table S1. **Ablation on the 3D data prior and the 2D generative prior (Self Re-enactment).** Our hybrid 3D data & 2D generative prior approach achieves the highest reconstruction performance on self re-enactment task, and achieves the most plausible appearance and geometry results on cross re-enactment (Fig. S4-right).

Variants	3D Data Prior	2D Gen. Prior	PSNR ( $\uparrow$ )	SSIM ( $\uparrow$ )	LPIPS ( $\downarrow$ )	CSIM ( $\uparrow$ )
(a)	X	X	22.23	0.8334	0.0910	0.6827
(b)	X	✓	23.19	0.8392	0.0891	0.7128
(c)	✓	X	24.63	0.8521	0.0816	0.7113
(d) Ours	✓	✓	25.220	0.8771	0.0732	0.7396

## C.2. Effect of 3D Data & 2D Generative Priors

We analyze the contribution of each prior by evaluating four system variants: (a) an *overfitting baseline* without the 3D data prior or the 2D generative prior, (b) a *2D generative prior* variant without the 3D data prior, (c) a *3D data prior* variant without the 2D generative prior, and (d) our *hybrid* model combining both priors. Variant (b) discards a 3D data prior (MGPM) and initializes using a template mesh; it adapts avatar to real images (Stage 1 *w/o* 3D prior) and uses generated images for Stage 2 adaptation. Variant (c) is identical to ELITE *w/o* Stage 2; it initializes via MGPM and adapts using only real images.

Fig. S4-left shows self re-enactment results, evaluated on held-out frames for which full metrics can be computed. For all the variants, we use three input frames for supervising the test-time adaptation. Quantitative comparisons for the self re-enactment PSNR, SSIM, LPIPS, and CSIM are provided in Table S1. Since the held-out frames are visually similar to the training data (speech-driven frames with limited pose variation), all the methods achieve comparable PSNR values. However, geometry quality differs significantly: methods (a) and (b), which lack a 3D data prior and optimize directly

from a template mesh, overfit to RGB observations and converge to flattened, unrealistic facial geometry. In contrast, methods (c) and (d) benefit from the 3D prior and faithfully preserve plausible facial structure. Because the held-out frames are close to the training distribution, the influence of the 2D generative prior is less noticeable in this setting.

Fig. S4-right further evaluates cross re-enactment, where each avatar is driven by novel and challenging poses and expressions. This setting exposes clear differences in generalization performance. Variant (a) shows limited generalization in appearance due to the absence of any prior and producing noticeable geometric collapses. Variant (b) leverages the 2D generative prior and therefore plausibly generalizes to unseen poses and expressions, yet still suffers from unrealistic geometry because it lacks the 3D prior. Variant (c) produces realistic geometry thanks to the learned 3D prior, but its RGB appearance does not generalize well to out-of-distribution poses when trained solely on real monocular data. Finally, our hybrid approach (d), using both priors, achieves faithful geometry and appears to have strong view/expression generalization simultaneously, producing the most plausible re-enactment results.

Overall, this ablation confirms three key observations: (1) without a 3D data prior, monocular reconstruction easily overfits and produces inaccurate geometry even when the rendered appearance seems plausible; (2) without a 2D generative prior, appearance-space generalization to unseen poses and expressions remains limited; and (3) combining both priors yields a complementary effect, enabling ELITE to achieve realistic geometry and plausible re-enactment quality across both seen and unseen driving signals.



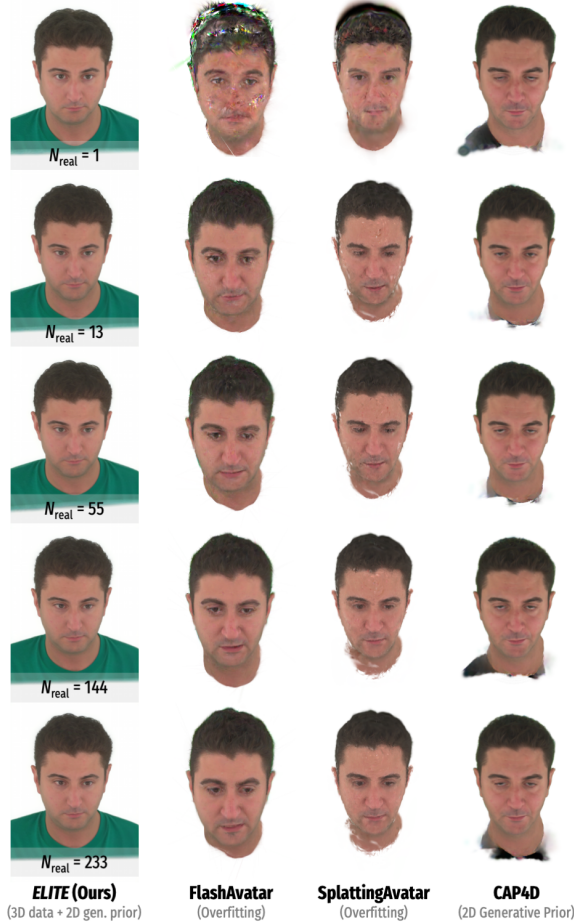


Figure S5. Effect of the number of real frames.

### C.3. Effect of the Number of Real Video Frames

Figure S5 compares the cross re-enactment quality as we vary the number of real supervision frames  $N_{\text{real}}$ . Although self re-enactment metrics (e.g., PSNR) improve with more real frames (Sec. 4.3 & Fig. 10 in the main paper), we observe that ELITE already produces stable and high-quality cross re-enactment results even with a single supervision frame. We attribute this robustness to our 3D data prior, which provides strong initialization, and to our generative adaptation stage, which supplies synthetic multi-view supervision regardless of  $N_{\text{real}}$ . In contrast, overfitting-based methods, FlashAvatar [15] and SplattingAvatar [10], show limited generalization to unseen expressions when  $N_{\text{real}}$  is small, as they rely solely on limited observations. CAP4D [13] benefits from synthetic views but still suffers from identity drift and limited expression fidelity. Overall, ELITE maintains strong cross-view and cross-expression generalization even under extremely sparse supervision.

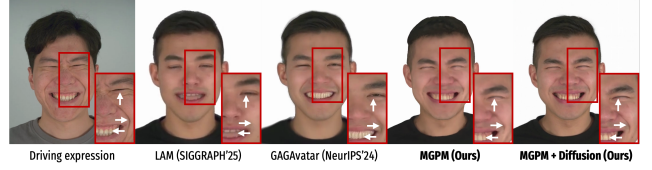


Figure S6. Comparison with recent feed-forward methods.

## D. More Results

### D.1. Comparison of MGPM and Recent Feed-forward 3D Avatar Recon. Methods

We compare feed-forward avatars of our MGPM and recent feed-forward 3D avatar reconstruction models [1, 2] in Fig. S6. LAM [2] animates a canonical avatar via explicit LBS, limited in modeling expression-aware effects, e.g., wrinkles. GAGAvatar [1] shows grid artifacts that may be attributed to its 3D point-based MLP, limited in expressing details. Our MGPM is a convolutional U-Net with expression-aware feature modulation, which models complex expression-aware effects with superior details.

### D.2. Comparison of Generated Supervision Images

In Fig. S7, we qualitatively compare the uncured sets of supervision images produced by CAP4D and our method.

Since CAP4D synthesizes each image by performing full diffusion denoising from pure noise, its outputs frequently exhibit severe artifacts (e.g., distorted facial regions, inconsistent geometry, or implausible textures) and suffer from noticeable identity drift. In contrast, our single-step diffusion enhancer is grounded on the rendered Gaussian avatar, providing strong geometric and appearance cues that guide the single-step generation process. As a result, our generated images preserve identity much more faithfully and contain significantly fewer visual artifacts. Moreover, by avoiding multi-step diffusion sampling, our method achieves **60× faster** generation while delivering cleaner and more reliable supervision for test-time adaptation.

### D.3. Limitations on Modeling Accessories

Our method has room for improvement in modeling accessories such as eyeglasses. Because the underlying 3D data prior model, MGPM, is trained on NerSemble [4], and we filtered out few identities with accessories to focus on pure head geometry and appearance, ELITE did not have a chance to learn explicit geometry priors for glasses. As a result, while the RGB appearance partially follows the glasses in input frames, the rendered normal maps reveal that no corresponding 3D structure is reconstructed (see Fig. S8), meaning the glasses are effectively baked into the texture space rather than modeled as geometry. Extending the prior to jointly learn facial and accessory geometry remains an important direction for future work.





(a) Uncurated Set of Generated Images from CAP4D



(b) Uncurated Set of Generated Images from **ELITE** (Ours)

Figure S7. **Uncurated comparison of generated supervision images.** (a) CAP4D produces images via full denoising from pure noise, leading to severe artifacts and identity drift, whereas (b) our rendering-grounded single-step enhancer generates identity-preserving, artifact-free images with significantly higher consistency, with  $60\times$  faster generation speed.

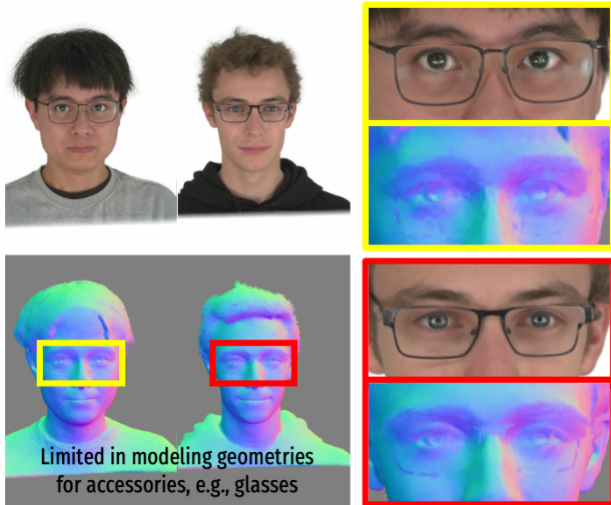


Figure S8. **Limitation in modeling accessories.** Although the RGB appearance from ELITE follows the eyeglasses in the input, the normal maps show no corresponding geometry, indicating that the glasses are baked into the texture.

#### D.4. Multi-view/-expression Renderings

In Figs. S9 & S10, we show multi-view rendered images and normal renderings of the Gaussian avatars synthesized from our method. We use our held-out test identities from the NerSemble-V2 dataset and test identities from the INSTA dataset. For all the identities, we use three images from the videos as test-time supervision for avatar adaptation. Overall, our method synthesizes high-fidelity, authentic Gaussian avatars with faithful appearances and geometries

that generalize across diverse expressions and viewpoints.

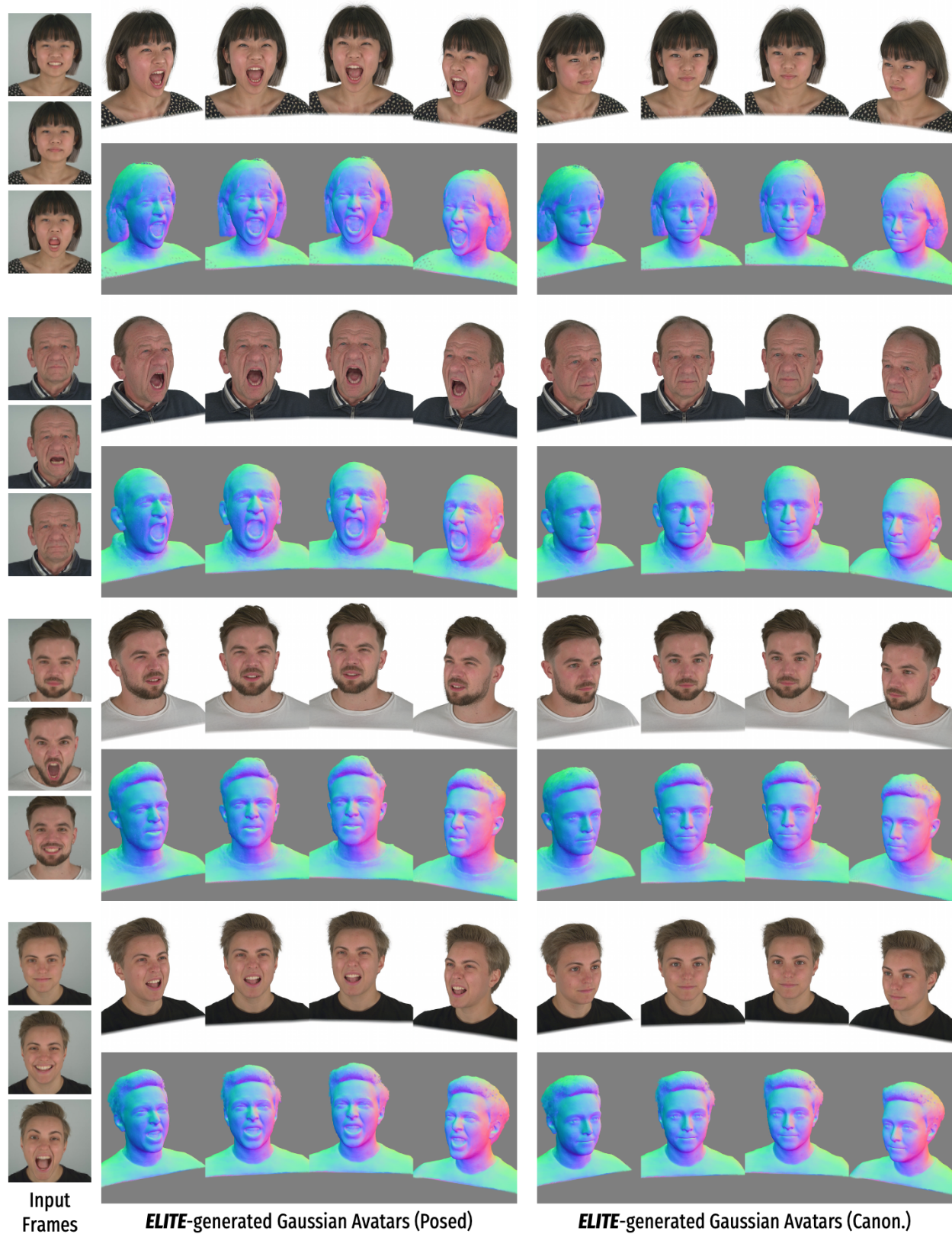
#### E. Broader Impacts & Ethical Considerations

**Societal Impact.** The primary goal of ELITE is to enabling accessible high-fidelity avatar synthesis for applications in telepresence, mixed reality, and we recognize the potential risks associated with misuse. To mitigate these risks, we advocate for the community’s ongoing efforts in avatar fingerprinting [6] and digital media forensics [8] to support the detection of synthetic media. To promote transparency and reproducibility, we plan to release our code and models strictly for research purposes.

**Data Considerations.** ELITE utilizes open-sourced academic datasets (NerSemble-V2, INSTA) to learn geometric and appearance priors. While ELITE demonstrates plausible generalization across various identities, we are aware of the importance of continued improvements in dataset diversity.

#### References

- [1] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 4
- [2] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *ACM Transactions on Graphics (SIGGRAPH)*, 2025. 4
- [3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Internat-*



**ELITE-generated Gaussian Avatars (Posed)** **ELITE-generated Gaussian Avatars (Canon.)**

Figure S9. Multi-view, Multi-expression Renderings of ELITE-generated Gaussian Avatars.





Figure S10. Multi-view, Multi-expression Renderings of ELITE-generated Gaussian Avatars.



*tional Conference on Learning Representations (ICLR)*, 2022. 2

- [4] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 1, 2, 4
- [5] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 1
- [6] Ekta Prashnani, Koki Nagano, Shalini De Mello, David Luebke, and Orazio Gallo. Avatar fingerprinting for authorized use of synthetic talking-head videos. In *European Conference on Computer Vision (ECCV)*, 2024. 5
- [7] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022. 2
- [8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint, 1803.09179*, 2018. 5
- [9] Axel Sauer, Dominik Lorenz, A. Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024. 2
- [10] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 4
- [11] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 1
- [12] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1
- [13] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 4
- [14] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difx3d+: Improving 3d reconstructions with single-step diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2
- [15] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 4
- [16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2