AVATARBOOST: A UNIVERSAL REAL-TIME VIDEO ENHANCEMENT MODEL FOR 4D AVATARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

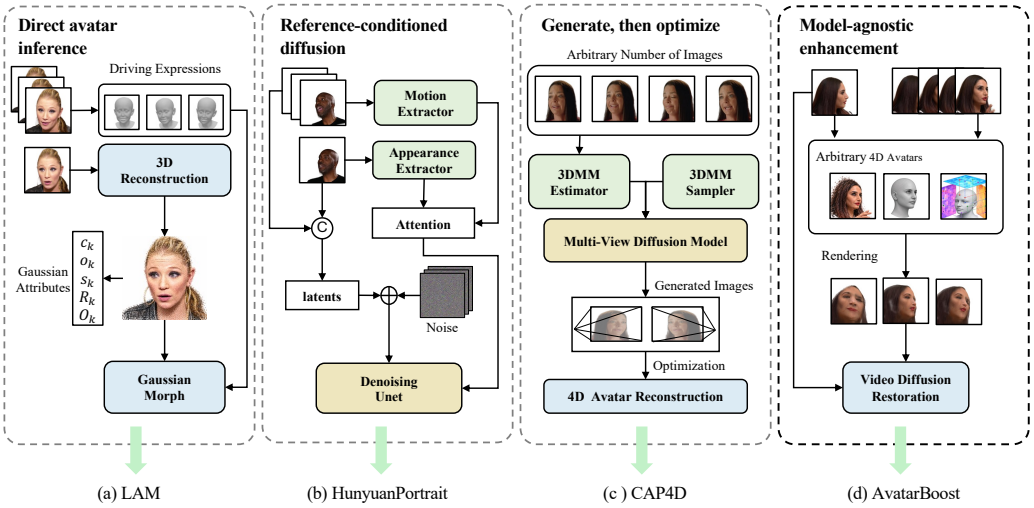
High-quality 4D portrait animation requires photorealistic appearance, consistent novel-view synthesis, and faithful motion control. While explicit 3D representations, such as 3D Gaussians, naturally support efficient rendering and reliable geometric consistency, feed-forward avatar models often produce videos with blurred textures and boundary artifacts. Conversely, video diffusion models can synthesize rich visual details, yet they lack explicit 3D constraints and suffer from costly multi-step sampling. To bridge this gap, we present **AvatarBoost**, a model-agnostic video enhancement framework that boosts the visual quality of existing 4D avatar models. Operating directly on rendered video outputs, AvatarBoost enhances their appearance without altering the underlying avatar representations or animation controls. In this framework, the rendered videos naturally provide viewpoint and motion guidance, which we complement with multiple reference images and their corresponding facial landmark maps to supply rich identity and facial-structure cues. Conditioned on these joint inputs, a video diffusion transformer learns to suppress rendering artifacts and faithfully restore fine details. To ensure practical efficiency, we further distill the enhancement model into a two-step generator using distribution matching distillation (DMD). Extensive experiments on VFHQ and NeRSemble demonstrate that a single AvatarBoost model improves the perceptual quality of both LAM and UIKA without backend-specific fine-tuning, reducing LPIPS by up to 42.1% and FID by up to 64.4% in self-reenactment. AvatarBoost achieves an inference throughput of 20 FPS on a single NVIDIA H800 GPU. An anonymous project page is available at https://anonymousauthor021.github.io/AvatarBoostProject/.

1 INTRODUCTION

Photorealistic 4D avatar animation aims to animate a person's appearance from reference images under specified head poses, facial expressions, and camera viewpoints. Motivated by widespread applications in telepresence, virtual humans, digital content creation, and immersive AR/VR, a convincing avatar must faithfully preserve the subject's identity and fine appearance details across continuous motions and arbitrary viewing angles. However, jointly achieving high visual fidelity, reliable motion control, and low latency remains challenging.

Existing approaches offer different trade-offs among visual quality, multi-view consistency, inference speed, and per-subject fitting, as summarized in Fig. 1. One popular line of work focuses on feed-forward 4D avatar models, which construct animatable 3D representations to enable efficient rendering alongside explicit viewpoint and motion control (Chu & Harada, 2024; He et al., 2025; Wu et al., 2026; Liu et al., 2026; 2025), exemplified by LAM in Fig. 1(a). While their underlying 3D geometry provides a solid basis for multi-view consistency, reconstruction inaccuracies often result in blurred facial textures, missing details, and boundary artifacts. In contrast, direct portrait generators animate appearance directly from motion conditions using GAN-based motion transfer (Wang et al., 2021; Guo et al., 2024) or video diffusion models (Xie et al., 2024; Xu et al., 2025). Among them, video diffusion methods, such as HunyuanPortrait in Fig. 1(b), provide rich appearance priors to generate vivid details; however, they do not explicitly enforce 3D consistency and suffer from heavy multi-step sampling latency. To bridge generative priors with explicit 3D geometry, another line of work adopts a generate-then-optimize paradigm: methods like CAP4D and MVP4D first synthesize multi-view observations using diffusion models and then optimize a subject-specific 4D

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Criterion	LAM	HunyuanPortrait	CAP4D	AvatarBoost
Inference speed				20 FPS
Visual quality				
Multi-view consistency				
Per-subject fitting	—	—		—

Figure 1: **Head-animation pipelines and their trade-offs.** (a) LAM (He et al., 2025) directly reconstructs 3D avatars for fast rendering. (b) HunyuanPortrait (Xu et al., 2025) generates portrait videos directly with diffusion. (c) CAP4D (Taubner et al., 2025b) fits avatars to diffusion-generated multi-view images before rendering. (d) AvatarBoost enhances existing avatar renders with a shared video diffusion model, improving visual quality under geometric guidance without additional avatar fitting. The bottom comparison highlights the trade-offs in quality, multi-view consistency, and efficiency. Rockets/turtles indicate fast/slow computation, faces denote relative strengths/limitations, and dashes indicate no additional fitting.

avatar (Taubner et al., 2025b;a) (Fig. 1(c)). Although the resulting avatars allow fast subsequent rendering, creating each new avatar still demands time-consuming generative sampling and per-subject optimization.

These complementary strengths suggest a natural synergy: using video diffusion as a post-rendering enhancement stage for existing 4D avatar models. Since the rendered video already supplies the correct viewpoint, head pose, and facial expressions, the diffusion model can focus on suppressing rendering artifacts and restoring fine details without learning 3D control from scratch. Operating directly on rendered video outputs provides a model-agnostic interface that avoids costly per-subject avatar fitting while remaining compatible with diverse avatar backends. To ensure practical deployment, the enhancement stage should introduce minimal inference overhead, preserving the efficiency of the underlying rendering pipeline. Our goal is therefore a shared, lightweight video enhancer that unites the explicit control of 4D avatars with the perceptual quality of video diffusion, as illustrated in Fig. 1(d).

We introduce **AvatarBoost**, a general-purpose video enhancement framework for 4D avatarization that combines these complementary strengths through the post-rendering pipeline in Fig. 1(d). By operating directly on rendered RGB sequences, AvatarBoost serves as a model-agnostic post-processor that preserves the upstream 3D representation and motion controls intact, ensuring broad applicability across diverse 4D avatar backends. To guide visual refinement, we adapt a pretrained video diffusion model using the conditioning interface of VACE (Jiang et al., 2025), which integrates the coarse rendered sequence with multiple identity-reference images and their corresponding facial

landmark maps. While the reference images provide complementary appearance and identity cues, the landmark maps explicitly constrain the spatial arrangement of facial features, collectively guiding the diffusion prior to suppress rendering artifacts without compromising motion fidelity. Furthermore, to eliminate the diffusion sampling bottleneck and ensure real-time execution, we distill the conditional model into a two-step generator via Distribution Matching Distillation (DMD) (Yin et al., 2024b;a), while employing Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient adaptation and training.

Our evaluation considers monocular and multi-view data from VFHQ (Xie et al., 2022) and NeRSemble (Kirschstein et al., 2023), examining appearance quality, identity and motion preservation, and the trade-off between restoration quality and computational cost.

Our contributions are summarized as follows:

- We introduce AvatarBoost, a shared video enhancement model designed to support different 4D avatar systems through their rendered outputs, without an additional avatar-fitting stage.
- We develop a conditional video restoration model that combines coarse renders with multiple identity-reference images and their corresponding landmark maps to guide appearance enhancement.
- We develop a few-step distillation scheme for avatar restoration using distribution matching and low-rank adaptation, targeting efficient inference and parameter-efficient training.

2 Related Work

2.1 2D Portrait Animation

Early 2D portrait animation approaches typically animate source images from motion cues using learned keypoints, optical flow warping, or 3DMM priors (Siarohin et al., 2019; Wang et al., 2021; Ren et al., 2021; Doukas et al., 2021), with audio-driven variants extending these representations to talking-head synthesis (Zhang et al., 2023). More recently, diffusion-based frameworks have substantially advanced visual fidelity by leveraging powerful generative video priors (Wei et al., 2024; Xie et al., 2024; Xu et al., 2025). However, because these methods operate primarily in the 2D image space, they do not inherently enforce an explicit 3D geometric structure, making consistent novel-view synthesis difficult under large head rotations. In addition, iterative diffusion sampling incurs substantial inference latency. Unlike these approaches that perform animation entirely in 2D, AvatarBoost retains explicit 3D geometric control by utilizing diffusion strictly as a fast post-rendering refinement module.

2.2 Optimization-based 3D Head Avatars

Optimization-based methods reconstruct subject-specific 3D representations to ensure geometrically consistent novel-view synthesis. Neural-field approaches, such as NerFACE (Gafni et al., 2021), Nerfies (Park et al., 2021), HeadNeRF (Hong et al., 2022), and INSTA (Zielonka et al., 2023), substantially improve appearance modeling beyond conventional face meshes, while PointAvatar (Zheng et al., 2023) employs deformable points to capture detailed head geometry. More recently, frameworks like Gaussian Head Avatar (Xu et al., 2024), GaussianAvatars (Qian et al., 2024), and FlashAvatar (Xiang et al., 2024) leverage 3D Gaussian representations to combine high-fidelity reconstruction with efficient rendering. Despite these advances, constructing each new avatar still demands time-consuming per-subject optimization on captured video sequences. Although generative paradigms such as CAP4D (Taubner et al., 2025b) alleviate capture requirements by synthesizing multi-view observations, they still rely on an expensive subject-specific fitting process. In contrast, AvatarBoost adopts a feed-forward, post-rendering enhancement strategy, using a shared video diffusion model to improve visual quality across diverse avatars without any per-subject optimization.

2.3 Feed-forward and One-shot 3D/4D Head Avatars

Feed-forward methods reconstruct animatable head avatars from one or a few images without test-time optimization. Neural-field approaches, such as GPAvatar (Chu et al., 2024) and Real3DPortrait (Ye et al., 2024), learn generalizable identity and motion priors, while Portrait4D (Deng et al., 2024a) and Portrait4D-v2 (Deng et al., 2024b) leverage synthetic or pseudo multi-view supervision to improve geometric consistency and reenactment quality. To achieve faster rendering, recent works build on 3D Gaussian Splatting (Kerbl et al., 2023): models like GAGAvatar (Chu & Harada, 2024), LAM (He et al., 2025), and UIKA (Wu et al., 2026) enable one-shot head reconstruction and real-time animation via explicit Gaussian representations. Despite their impressive efficiency and explicit control, pure 3D feed-forward models can still produce noticeable artifacts in fine-scale structures (e.g., hair and teeth) or disoccluded regions under large expression deformations and extreme camera viewpoints. AvatarBoost addresses these limitations by introducing a diffusion-based video enhancement stage that enriches visual details while preserving the underlying Gaussian representations and animation controls intact.

2.4 Diffusion-based Portrait Synthesis and Efficient Video Diffusion

Diffusion models enable realistic portrait animation, with frameworks such as AniPortrait (Wei et al., 2024) and EchoMimic (Chen et al., 2024) supporting audio-driven synthesis, and HunyuanPortrait (Xu et al., 2025) transferring facial expressions and head poses from driving video sequences. However, iterative multi-step sampling is computationally expensive, and these 2D approaches do not maintain an explicit animatable 3D representation for consistent novel-view synthesis. To accelerate diffusion inference, recent distillation strategies such as DMD (Yin et al., 2024b) and DMD2 (Yin et al., 2024a) compress sampling trajectories into few-step generators, while Self Forcing (Huang et al., 2025) addresses the training–inference mismatch in autoregressive video diffusion by training on self-generated sequences. Inspired by these advances, we employ distribution matching distillation for fast, two-step enhancement of rendered head videos, combining diffusion-based appearance refinement with the explicit geometric control provided by existing 3D avatars.

3 Method

AvatarBoost improves videos rendered by existing animatable avatar models. The rendered video already follows the desired motion, but it often contains artifacts and lacks fine detail. We therefore treat it as a coarse guide and use a video diffusion model with the conditioning interface of VACE (Jiang et al., 2025) to restore realistic appearance. Reference images of the same person help preserve identity. As illustrated in Fig. 2, AvatarBoost is trained in two stages: supervised enhancement with flow matching, followed by distribution matching distillation. The enhancement stage learns to enhance coarse avatar renderings using identity references and their landmark maps, while distillation reduces inference to two denoising steps.

3.1 Training Data Construction

We build training pairs from single-view talking-head videos. Given a T-frame video $\mathbf{y}^\star$, we reconstruct an avatar from its first frame using the upstream model. We then animate and render the avatar using the pose, expression, and camera parameters estimated for each target frame, obtaining a coarse video $\mathbf{x} = (x_1, \ldots, x_T)$. The pair $(\mathbf{x}, \mathbf{y}^\star)$ provides supervision for video enhancement. We also collect K reference images $\{I_k\}_{k=1}^K$ of the same identity and render their facial landmark maps $\{\ell_k\}_{k=1}^K$. Interleaving the two gives a reference bank $\mathcal{B} = (I_1, \ell_1, \ldots, I_K, \ell_K)$ with $Q = 2K$ entries. Finally, a frozen text encoder maps a text prompt to an embedding h.

3.2 Context-Conditioned Video Diffusion Model

Encoding the references. All visual inputs are encoded by a frozen video VAE with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, which performs $4\times$ temporal and $8\times$ spatial downsampling. Each entry B_q in the reference bank is encoded independently as a single-frame video latent $\mathbf{r}_q = \mathcal{E}(B_q)$, preserving one dedicated latent slot per reference. Here, the RGB references supply identity and detailed appearance

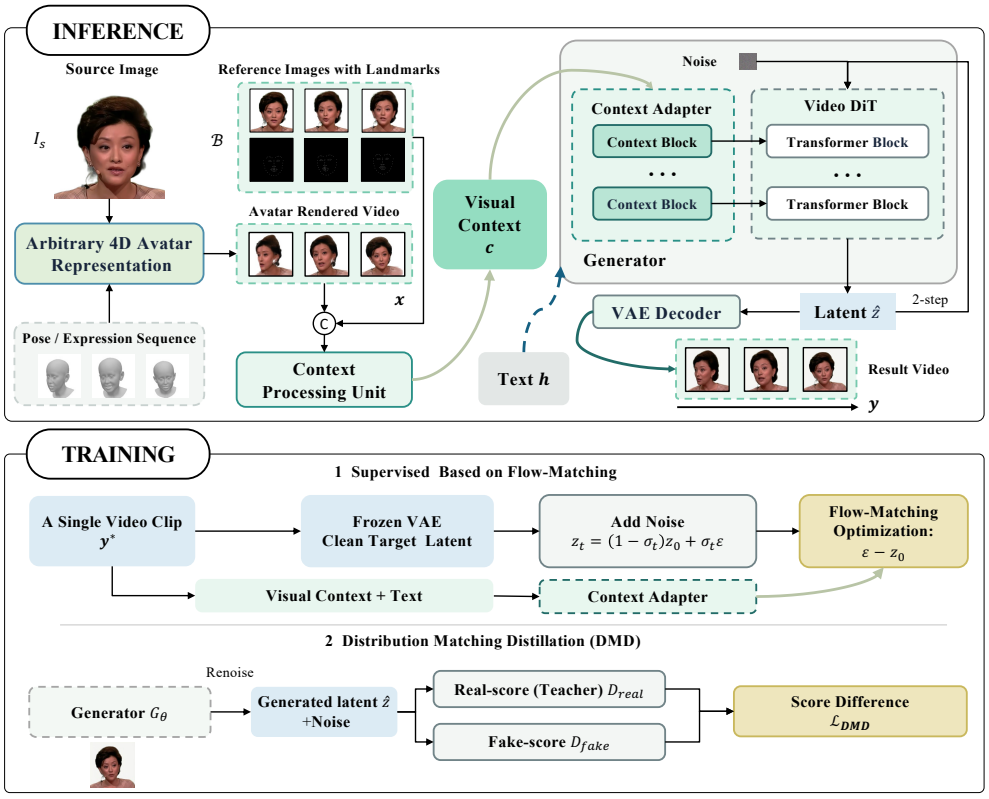


Figure 2: **Overview of AvatarBoost. Inference:** Given pose or expression controls, an existing animatable avatar model renders a coarse video. A context adapter conditions a video diffusion transformer (DiT) on this rendering and on a bank of reference images paired with landmark maps. The distilled DiT then generates an enhanced video in only two denoising steps, followed by VAE decoding. **Training:** We first train the model on supervised enhancement with a flow-matching objective, using paired samples built from single-view video clips. We then apply Distribution Matching Distillation (DMD), which alternates between updating the generator, guided by a frozen real-score teacher and a fake-score model, and updating the fake-score model on generated samples.

information, while the corresponding landmark maps encode the spatial layout of facial features to provide explicit geometric guidance.

Encoding the rendered video. Following VACE (Jiang et al., 2025), we inject the rendered video through a masked-video conditioning interface. An optional spatial-temporal mask $\mathbf{m}$ separates the rendering into an invariant region $(1-\mathbf{m}) \odot \mathbf{x}$ and a region to be refined $\mathbf{m} \odot \mathbf{x}$. The two complementary components are independently encoded by $\mathcal{E}$ and concatenated along the channel dimension together with the downsampled mask, yielding the composite visual feature $\mathbf{u}$. By default, we set $\mathbf{m} = \mathbf{1}$, allowing the diffusion model to refine the entire rendered frame under structural guidance.

Visual context and generator. We zero-pad each reference latent $\mathbf{r}_q$ along the channel dimension to match that of $\mathbf{u}$, yielding a padded reference representation $\mathbf{c}_q^{\text{ref}}$. We then concatenate all references and the rendered feature along the temporal dimension:

$$\mathbf{c} = \left[\mathbf{c}_1^{\text{ref}}; \ldots; \mathbf{c}_Q^{\text{ref}}; \mathbf{u}\right]_{\text{time}}. \tag{1}$$

Let L denote the temporal length of $\mathcal{E}(\mathbf{y}^\star)$ (where $L = n+1$ for $T = 4n+1$ input frames). The resulting visual context $\mathbf{c}$ encompasses $Q+L$ temporal frames: Q static reference frames followed by L frames from the rendered sequence. A context adapter maps $\mathbf{c}$ into residual features, which are directly injected into selected transformer blocks of the video DiT. The text prompt embedding

$\mathbf{h}$ is incorporated through standard cross-attention conditioning. We denote the joint system of the context adapter and the video transformer as the generator G_θ. Accordingly, the noisy input latent is extended to length $Q + L$ to maintain frame-wise temporal alignment with the visual context.

3.3 Stage 1: Supervised enhancement with Flow Matching

In the first stage, we train G_θ to reconstruct the target video conditioned on $(\mathbf{c}, h)$. The clean target latent is $\mathbf{z}_0 = [\mathbf{r}_1; \ldots; \mathbf{r}_Q; \mathcal{E}(\mathbf{y}^\star)]_{\text{time}}$, where the reference latents are prepended to match the layout of $\mathbf{c}$. We sample a diffusion time τ and Gaussian noise $\boldsymbol{\epsilon}$, form the noisy latent $\mathbf{z}_\tau = (1 - \sigma_\tau)\mathbf{z}_0 + \sigma_\tau \boldsymbol{\epsilon}$, where σ_τ is the noise level, and train G_θ to predict the flow velocity $\boldsymbol{\epsilon} - \mathbf{z}_0$:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\Big[w(\tau)\,\big\|G_\theta(\mathbf{z}_\tau, \tau; \mathbf{c}, h) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\big\|_2^2\Big], \tag{2}$$

where $w(\tau)$ is the scheduler's weighting defined in Appendix A. The rendering already provides the motion, so the model mainly has to learn to suppress rendering artifacts and restore appearance details using the references. From a predicted velocity, the clean latent can be recovered as $\widehat{\mathbf{z}} = \mathbf{z}_\tau - \sigma_\tau\, G_\theta(\mathbf{z}_\tau, \tau; \mathbf{c}, h)$. In the rest of this section, we describe all networks by these clean-latent predictions.

3.4 Stage 2: Distribution Matching Distillation

The Stage-1 model needs many denoising steps. To speed up inference, we distill it into a few-step generator with Distribution Matching Distillation (DMD). DMD does not require each generated video to match a specific ground-truth video. Instead, it trains the generator so that the distribution of its outputs matches the distribution modeled by a pretrained diffusion teacher.

Model roles. Three networks are involved. (i) The *generator* G_θ, initialized from Stage 1, is the network being distilled. (ii) The *real-score model* D_{real}, which serves as the teacher, is a frozen, pretrained text-to-video model. It receives only the noisy latent and the text embedding, with no visual context, and represents the distribution of realistic videos. (iii) The *fake-score model* D_{fake} has the same architecture and conditioning as G_θ. It is trained online on generated samples and represents the generator's current output distribution. The fake and teacher models estimate scores with different conditioning, corresponding to $p_{\text{fake},\tau}(\cdot \mid \mathbf{c}, h)$ and $p_{\text{real},\tau}(\cdot \mid h)$, respectively. The difference between the two score models' predictions tells the generator how to change its outputs to look more realistic.

Few-step rollout. Inspired by the on-policy training principle of Self Forcing (Huang et al., 2025), we optimize the generator using samples produced by its own denoising rollout. Starting from Gaussian noise, the generator runs through a fixed sequence of decreasing diffusion timesteps $\tau_1 > \cdots > \tau_N$ with corresponding noise levels σ_{τ_i}. At each step, it predicts a clean latent. Unless it is the last step, it then adds fresh noise to this prediction to reach the next noise level. The final prediction $\widehat{\mathbf{z}}$ is the generated sample.

Generator update. Following DMD (Yin et al., 2024b), the generator is trained to minimize the KL divergence between its output distribution p_{fake} and the real distribution p_{real}. Both distributions are compared after diffusion, at a randomly sampled noise level τ. We add noise to the generated sample $\widehat{\mathbf{z}}$ to obtain $\widetilde{\mathbf{z}}_\tau$ and query both score models. The real-score model uses classifier-free guidance with scale γ, giving $\widehat{\mathbf{z}}_{\text{real}} = \widehat{\mathbf{z}}^+ + \gamma(\widehat{\mathbf{z}}^+ - \widehat{\mathbf{z}}^-)$. Here $\widehat{\mathbf{z}}^+$ and $\widehat{\mathbf{z}}^-$ are its predictions with the prompt embedding h and a negative-prompt embedding. The fake-score model predicts $\widehat{\mathbf{z}}_{\text{fake}} = D_{\text{fake}}(\widetilde{\mathbf{z}}_\tau, \tau; \mathbf{c}, h)$. The score difference is proportional to the difference between the two clean-latent predictions, yielding the following DMD objective and approximate gradient:

$$\begin{aligned} \mathcal{L}_{\text{DMD}} &= \mathbb{E}_\tau\Big[D_{\text{KL}}\big(p_{\text{fake},\tau} \,\|\, p_{\text{real},\tau}\big)\Big], \\ \nabla_\theta \mathcal{L}_{\text{DMD}} &\approx \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\left[w_\tau \left(\widehat{\mathbf{z}}_{\text{fake}} - \widehat{\mathbf{z}}_{\text{real}}\right) \frac{\partial \widehat{\mathbf{z}}}{\partial \theta}\right]. \end{aligned} \tag{3}$$

where the weight $w_\tau = 1/\operatorname{mean}\big|\widehat{\mathbf{z}} - \widehat{\mathbf{z}}_{\text{real}}\big|$ is computed separately for each sample and keeps the update size similar across noise levels (Yin et al., 2024b). A gradient descent step therefore moves $\widehat{\mathbf{z}}$ toward the real-score prediction and away from the fake-score prediction.



Figure 3: **Qualitative comparisons on NeRSemble, VFHQ, and in-the-wild examples.** Insets show target frames for self-reenactment or driving frames for cross-identity reenactment. AvatarBoost reduces rendering artifacts and restores facial and silhouette details.

Fake-score model update. The generator's output distribution changes during training, so D_{fake} is updated to track this evolving distribution. We draw a new generated sample $\bar{\mathbf{z}}$ without gradient tracking. We then train D_{fake} with the same flow-matching objective as Eq. (2), using $\bar{\mathbf{z}}$ as the clean target and computing the loss only over video slots. The real-score model is not used in this update. We alternate between the generator update and the fake-score update throughout distillation.

3.5 Inference

Of the three diffusion networks in DMD, only the distilled generator G_θ is retained at inference. We render the avatar under the desired motion and camera controls. We then build the visual context $\mathbf{c}$ from the rendering and the reference bank (Sec. 3.2) and encode the text prompt into h. Starting from Gaussian noise with $Q + L$ slots, the generator performs the two-step rollout of Sec. 3.4 using base schedule indices 1000 and 500, mapped to model timesteps by the shifted schedule in Appendix A. Finally, we discard the reference slots and decode the remaining latents with the VAE decoder $\mathcal{D}$ to obtain the enhanced video.

4 Experiments

We evaluate AvatarBoost under self- and cross-identity reenactment tasks as a modular enhancement stage for existing state-of-the-art feed-forward Gaussian avatars: LAM and UIKA. A single AvatarBoost checkpoint trained on LAM-generated data is used to enhance both models without further backend-specific fine-tuning.

4.1 Experimental Setup

Data and training. We use VFHQ (Xie et al., 2022) and NeRSemble (Kirschstein et al., 2023) with 1,000 training clip sets (400 and 600, respectively) and 40 test clip sets (20 per dataset). Training and test identities are disjoint. Paired LAM renders and real videos provide supervision. Both stages

use 81-frame clips, with supervised enhancement and distillation trained for 50 and 10 epochs, respectively.

Baselines. We compare LAM (He et al., 2025) and UIKA (Wu et al., 2026) with their enhanced counterparts, alongside Portrait4D-v2 (Deng et al., 2024b) and HunyuanPortrait (Xu et al., 2025). GAGAvatar (Chu & Harada, 2024) is additionally included in qualitative comparisons.

Evaluation. We measure reconstruction fidelity with PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018); identity similarity with CSIM (Deng et al., 2022); expression and head-pose errors with AED and APD (Deng et al., 2019); landmark alignment with AKD (Bulat & Tzimiropoulos, 2017); and distributional quality with FID (Heusel et al., 2017). Cross-identity reenactment has no paired ground truth, so we omit PSNR, SSIM, and LPIPS in this setting. Preprocessing, training settings, and evaluation details are provided in Appendix A.

4.2 Qualitative Comparisons

Figure 3 presents self- and cross-identity reenactment examples from NeRSemble, VFHQ, and additional in-the-wild data. The insets show target frames for self-reenactment and driving frames for cross-identity reenactment. Compared with the original LAM and UIKA outputs, AvatarBoost reduces visible rendering artifacts and improves facial contours and head or shoulder boundaries in the displayed examples. These results illustrate how render-guided restoration can enhance appearance while following the motion supplied by the upstream avatar.

4.3 Quantitative Comparisons

Table 1 shows that AvatarBoost improves both backends. In self-reenactment, LPIPS decreases by about 40% and FID by 42–64%, with both enhanced models reaching FID values near 36.3. Both outperform all baselines on PSNR, LPIPS, AED, APD, and FID, and UIKA + AvatarBoost ranks first on six of eight metrics. In cross-identity reenactment, enhancement improves expression, pose, and landmark accuracy and FID over both backends, although CSIM decreases. Figure 3 provides qualitative examples of source-identity preservation. These results demonstrate cross-backend gains in perceptual quality and motion fidelity while highlighting the remaining challenge of identity preservation.

Table 1: **Self- and cross-identity reenactment.** Higher (↑) or lower (↓) is better. Red/blue: best/second-best within each setting. AKD is in pixels.

(a) Self-reenactment

Method	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD (px)↓	FID↓
HunyuanPortrait	17.823	0.729	0.220	0.868	0.059	0.058	20.600	45.267
Portrait4D-v2	14.784	0.692	0.292	0.889	0.062	0.070	21.144	62.683
LAM	18.088	0.777	0.235	0.735	0.087	0.076	4.280	102.088
UIKA	18.814	0.790	0.174	0.882	0.067	0.062	5.272	62.246
LAM + AvatarBoost	19.310	0.787	0.136	0.820	0.055	0.052	5.177	36.377
UIKA + AvatarBoost	20.961	0.808	0.105	0.885	0.052	0.049	4.541	36.281

(b) Cross-identity reenactment

Method	CSIM↑	AED↓	APD↓	AKD (px)↓	FID↓
HunyuanPortrait	0.580	0.162	0.131	54.393	143.695
Portrait4D-v2	0.751	0.152	0.106	25.798	135.478
LAM	0.560	0.165	0.143	11.867	146.258
UIKA	0.685	0.148	0.131	10.365	141.960
LAM + AvatarBoost	0.474	0.144	0.130	11.155	138.328
UIKA + AvatarBoost	0.607	0.146	0.126	10.354	140.311

4.4 Ablation Study

We evaluate supervised enhancement, reference conditioning, and distillation using LAM on VFHQ under self-reenactment (Fig. 4 and Table 2). We compare the pretrained model, the full 50-step

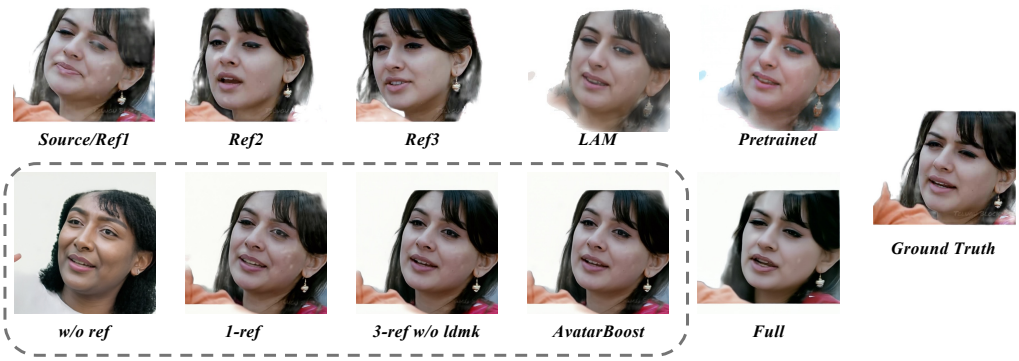


Figure 4: **Qualitative ablation on VFHQ.** The dashed box groups two-step variants. Reference conditioning improves identity preservation; the 50-step Full model better corrects severe rendering errors.

enhancement model, and two-step variants with no references, one reference pair, three RGB references without landmarks, or three reference–landmark pairs. Detailed configurations and reference selection are provided in Appendix A.5.

Table 2: **Ablation of reference conditioning and distillation on VFHQ.** Full uses 50 denoising steps; distilled variants use two. Red/blue: best/second-best across all variants.

Variant	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	FID↓
LAM	0.413	0.461	0.201	0.110	7.292	121.609
Full (50-step)	0.202	0.721	0.074	0.060	6.342	67.616
Distilled variants						
- w/o ref	0.462	0.148	0.148	0.081	7.679	178.274
- 1-ref	0.375	0.549	0.091	0.062	8.299	148.416
- 3-refs w/o ldmk	0.287	0.759	0.079	0.077	7.232	113.774
AvatarBoost (3-ref)	0.286	0.762	0.075	0.070	7.202	104.535

Effectiveness of supervised enhancement. The *Pretrained* model leaves rendering artifacts and introduces color shifts. Supervised enhancement improves artifact removal and motion following. However, the *w/o ref* variant generates a different identity in the displayed example despite following the rendered motion, resulting in substantially lower CSIM. This highlights the importance of identity references.

Effectiveness of reference images. A single reference provides limited appearance information and can propagate its artifacts to the output. Multiple references supply complementary identity cues and improve perceptual quality. Adding landmark maps yields small metric gains with little visible change (Table 2), suggesting that facial-structure cues help the model use reference appearance.

Effectiveness of distillation. The two-step model trades some restoration quality for speed and may retain or amplify upstream errors. In Fig. 4, LAM renders open eyes despite the subject's heavy eye shadow; the distilled enhancer exaggerates this error, while Full corrects it more effectively. Reference images partially mitigate this failure. Nevertheless, all reference-conditioned distilled variants improve LPIPS, CSIM, AED, and APD over LAM, although the single-reference variant performs worse on AKD and FID.

5 Conclusion

We presented AvatarBoost, a model-agnostic video diffusion framework for enhancing renders from existing 4D avatar models. Conditioned on rendered videos, identity references, and their facial landmark maps, it reduces artifacts and restores appearance details without subject-specific fine-tuning or changes to upstream models and animation controls. Distribution matching distillation enables two-step inference. Experiments on LAM and UIKA using VFHQ and NeRSemble demonstrate improved perceptual quality and motion fidelity in self- and cross-identity reenactment.

However, identity preservation remains challenging in cross-identity reenactment, and the two-step generator may amplify severe rendering errors. Future work could use the learned restoration priors to refine upstream avatars and improve robustness.

AI USE STATEMENT

We used generative AI tools for manuscript drafting and translation, literature review, feedback on experimental design, implementation and debugging, result analysis, and visualization. AI-assisted work was reviewed and revised under author supervision. The authors take responsibility for the accuracy, originality, and integrity of the final submission.

ETHICS STATEMENT

AvatarBoost enhances the realism of animated portrait videos, supporting applications such as telepresence and digital content creation. This capability also poses risks of impersonation, non-consensual reenactment, and deceptive media. Generated videos should be clearly identified as synthetic, and the use of a person's likeness should respect their consent and privacy. The framework may inherit biases from pretrained models and training data, and enhancement does not guarantee faithful identity preservation. Use of the datasets and pretrained models should comply with their respective licenses and access conditions.

REPRODUCIBILITY STATEMENT

Section 4 describes the evaluation tasks, datasets, baselines, and metrics. Appendix A documents data preprocessing, reference selection, model initialization, training hyperparameters, LoRA configurations, noise schedules, and loss conventions; Appendix A.5 specifies the ablation variants. These details complement the two-stage training procedure described in the method section. Our anonymous project page provides an additional resource for this work.

REFERENCES

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230, 000 3D facial landmarks). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1021–1030. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.116. URL https://doi.org/10.1109/ICCV.2017.116.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable Gaussian Head Avatar. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6a14c7f9fb3f42645cfa6bd5aa446819-Abstract-Conference.html.

Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. doi: 10.1109/TPAMI.2021.3087709. URL https://doi.org/10.1109/TPAMI.2021.3087709.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 285–295. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPRW.2019.00038. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/AMFG/Deng_Accurate_3D_Face_Reconstruction_With_Weakly-Supervised_Learning_From_Single_Image_CVPRW_2019_paper.html.

Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4D: Learning one-shot 4D head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4D-v2: Pseudo multi-view data creates better 4D head synthesizer. In *European Conference on Computer Vision*, pp. 316–333. Springer, 2024b.

Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. HeadGAN: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14398–14407, 2021.

Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8649–8658, 2021.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. LivePortrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

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 *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–13, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A real-time NeRF-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20374–20384, 2022.

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 *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. URL https://arxiv.org/abs/2503.07598.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. doi: 10.1145/3592455.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813.

Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. AvatarArtist: Open-domain 4D avatarization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10758–10769, 2025.

Hongyu Liu, Xuan Wang, Zijian Wu, Yating Wang, Ziyu Wan, Yue Ma, Runtao Liu, Boyao Zhou, Yujun Shen, and Qifeng Chen. Avatarpointillist: Autoregressive 4d gaussian avatarization. *arXiv preprint arXiv:2604.04787*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

Shenhan Qian. VHAP: Versatile head alignment with adaptive appearance priors, sep 2024.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024.

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. PIRenderer: Controllable portrait image generation via semantic neural rendering. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B. Lindell. MVP4D: Multi-view portrait video diffusion for animatable 4D avatars. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025a.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5318–5330, 2025b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10039–10049, 2021.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. URL https://www.cns.nyu.edu/~lcv/ssim/.

Huawei Wei, Zejun Yang, and Zhisheng Wang. AniPortrait: Audio-driven synthesis of photorealistic portrait animations. *arXiv preprint arXiv:2403.17694*, 2024.

Zijian Wu, Boyao Zhou, Liangxiao Hu, Hongyu Liu, Yuan Sun, Xuan Wang, Xun Cao, Yujun Shen, and Hao Zhu. UIKA: Fast universal head avatar from pose-free images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient Gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1802–1812, 2024.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-Portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian Head Avatar: Ultra high-fidelity head avatar via dynamic Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. HunyuanPortrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15909–15919, 2025.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3D-Portrait: One-shot realistic 3D talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661, 2023.

Yufeng Zheng, Yifan Wang, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. PointAvatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21057–21067, 2023.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4574–4584, 2023.

A Implementation and Reproducibility Details

This appendix provides additional implementation and training details.

A.1 Dataset Preprocessing

We use VFHQ (Xie et al., 2022) and NeRSemble (Kirschstein et al., 2023) for training and testing, comprising in-the-wild single-view portrait videos and synchronized multi-view facial performances, respectively. We use 1,000 training clip sets: 600 from NeRSemble and 400 from VFHQ, together with 40 test clip sets, 20 from each dataset. Training and test identities are disjoint. LAM supplies the coarse rendered videos, and the corresponding real videos provide the enhancement targets. We crop the portrait region, remove the background, and estimate camera and FLAME(Li et al., 2017) parameters with VHAP (Qian, 2024). The specified input resolutions are 512×512 for VFHQ and 480×832 for NeRSemble, in height-by-width order; cropping and resizing preserve aspect ratio. The first frame supplies the reconstruction input, and subsequent frames provide driving motion and, for self-reenactment, the target appearance. For VFHQ (Xie et al., 2022), the three identity-reference images include the source image and two other images from the same video. For NeRSemble (Kirschstein et al., 2023), the three references exclude the source image and are sampled from alternative viewpoints. Each RGB reference is paired with its landmark map, giving six reference slots in the full model.

A.2 Training Details

We implement both training stages in PyTorch and optimize the trainable parameters using AdamW (Loshchilov & Hutter, 2019) with a batch size of one, and each training video clip contains 81 frames. The learning rates for the generator and the fake-score estimator are 2×10^{-6} and 4×10^{-7}, respectively, and the fake-score estimator is updated five times per generator update. Enhancement fine-tuning and distillation run for 50 and 10 epochs, taking approximately 4 and 2 days, respectively, on four NVIDIA RTX 5880 GPUs with 48 GB of memory each. Thus, the full training pipeline does not require data-center-class GPUs.

A.3 LoRA Setting

We use LoRA (Hu et al., 2022) for parameter-efficient adaptation and apply bfloat16 precision and gradient checkpointing to reduce memory usage. Supervised restoration uses rank-32 adapters, whereas distillation uses rank-8 adapters with LoRA alpha set to 8 for both the generator and the fake-score model. The restoration model and the fake-score model are initialized from Wan2.1-VACE-1.3B. The frozen teacher uses Wan2.1-T2V-14B, and the distilled generator incorporates the supervised restoration context adapter (Team Wan et al., 2025).

The Stage-1 LoRA configuration targets the context adapter's attention projections (`q`, `k`, `v`, and `o`) and feed-forward projections (`ffn.0` and `ffn.2`). Its default LoRA alpha equals its rank, giving $\alpha/r = 1$. During distillation, the base weights of both the DiT and the context adapter are frozen; rank-8 LoRA modules are added to matching attention projections (`q`, `k`, `v`, `o`, `k_img`, and `v_img`, where present) in both components. The context adapter is therefore adapted through LoRA rather than fully fine-tuned. The rank-8 configuration uses zero dropout and $\alpha/r = 1$. The distillation configuration uses one clip per process, with an optimizer step after each corresponding loss evaluation and no gradient accumulation. The teacher guidance coefficient is $\gamma = 3$ in the convention of Sec. 3.4; the fake-score model uses no classifier-free guidance. Text prompts are read from the sample metadata, and a fixed negative prompt is used for the teacher.

A.4 Noise Schedule and Loss Conventions

Diffusion time τ and noise level σ_τ are related by $\tau = 1000\sigma_\tau$. For a base schedule value $u \in [0, 1]$, the shifted schedule is $\sigma(u) = su/[1 + (s-1)u]$. With generator shift $s = 5$, base indices 1000 and 500 correspond to $u = 1$ and $u = 1/2$, hence noise levels 1 and $5/6$ and model timesteps 1000 and approximately 833.3. The final clean prediction is decoded without an additional network evaluation at time zero.

Stage 1 samples uniformly from a 1,000-entry training schedule. Let $a_j = \exp[-2(\tau_j/1000 - 1/2)^2]$ and $a_{\min} = \min_j a_j$. Its normalized scheduler weight is

$$w(\tau_j) = \frac{a_j - a_{\min}}{\frac{1}{1000}\sum_{k=1}^{1000}(a_k - a_{\min})}. \tag{4}$$

The Stage-1 error includes reference and video slots. During distillation, both losses exclude reference slots, while all three models receive the entire noisy sequence. The normalizer in Eq. (3) is computed over all slots before masking, and non-finite normalized updates are sanitized in the implementation.

For Gaussian corruption with $\alpha_\tau = 1 - \sigma_\tau$, a clean prediction $\widehat{\mathbf{z}}_D$ corresponds to the score estimate $(\alpha_\tau \widehat{\mathbf{z}}_D - \widetilde{\mathbf{z}}_\tau)/\sigma_\tau^2$. Before additional weighting, the gradient of the noise-perturbed KL objective contains $\alpha_\tau^2/\sigma_\tau^2$ multiplying the clean-prediction difference and the transposed generator Jacobian. Equation (3) instead specifies the normalized surrogate used for optimization. This distinction is necessary because the normalization and reference-slot mask change the update.

A.5 Ablation Configurations

We conduct all ablations on VFHQ under self-reenactment, using LAM renders as input. *Pretrained* applies the pretrained video diffusion model with text prompts and reference images. *Full* uses the supervised enhancement model with 50 denoising steps. All remaining variants use two-step distilled models: *w/o ref* removes reference images and their landmark maps while retaining rendered-video and text conditions; *1-ref* uses the source image and its landmark map; *3-refs w/o ldmk* uses three RGB references without landmark maps; and *AvatarBoost* uses three reference images paired with their landmark maps. Frames temporally adjacent to target frames are excluded from reference selection.

B Multi-view and Temporal Consistency

To show that AvatarBoost is more than a per-frame 2D image enhancer, we apply it to novel-view videos. Specifically, we use LAM He et al. (2025) to reconstruct a 4D avatar from a single-view image and render it from camera viewpoints that differ substantially from the input view. The rendered motion sequences show the expression dynamics, and the corresponding point clouds show the underlying 3D geometry. In this setting, all reference images come from the single-view source video. AvatarBoost must therefore carry texture information from the captured view over to viewpoints it has never observed. We use two subjects that also appear in the earlier visual comparison . As shown in Figure 5, AvatarBoost stays stable under large camera movements.

We further present additional cross-identity reenactment results to examine identity preservation, since AvatarBoost yields lower CSIM scores than the original backends in the quantitative comparison (Table 1). As illustrated in Figure 6, the enhanced results remain visually consistent with the source identity across diverse driving expressions and head poses.

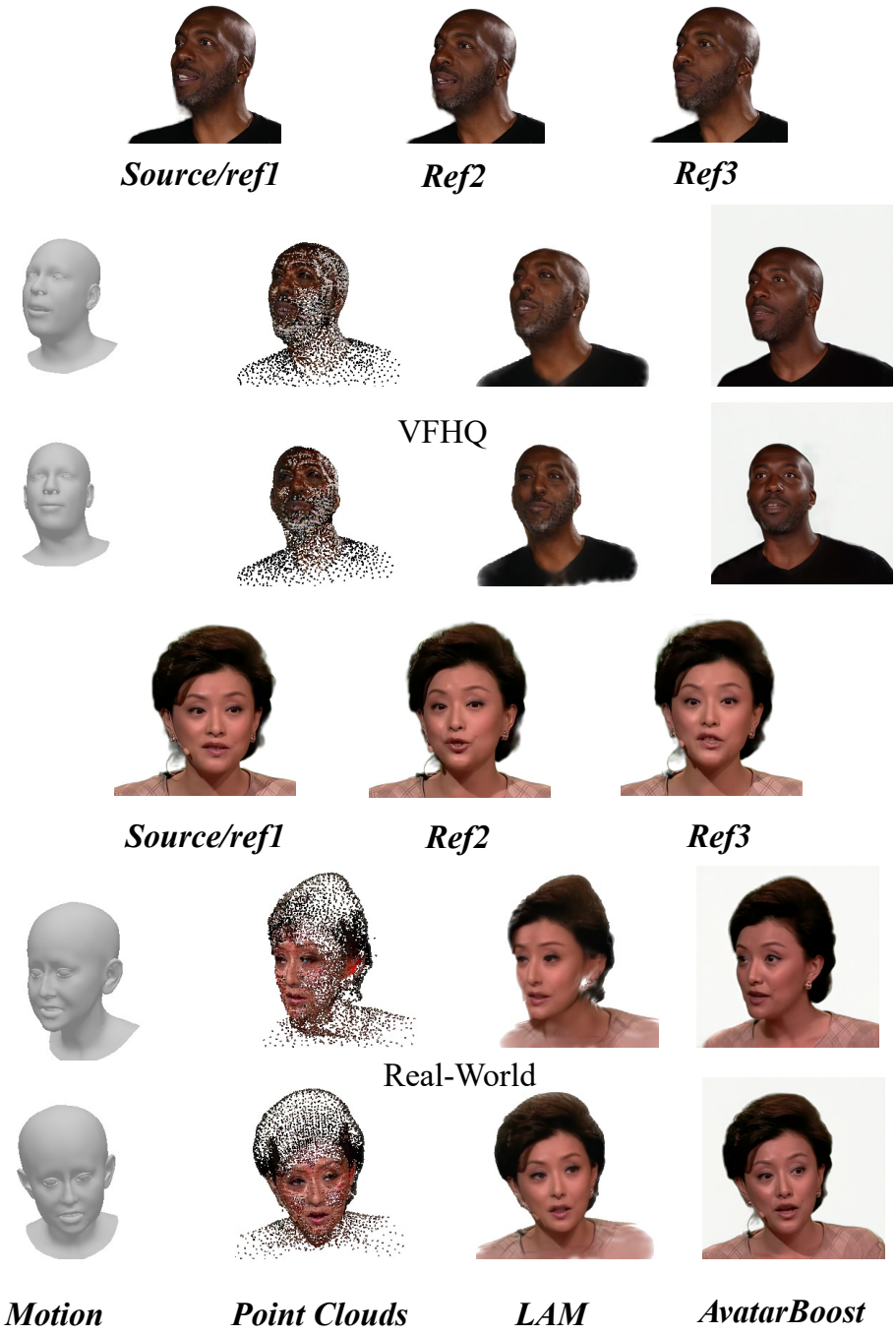


Figure 5: **The enhancement performance of Avatar in novel view synthesis task.**

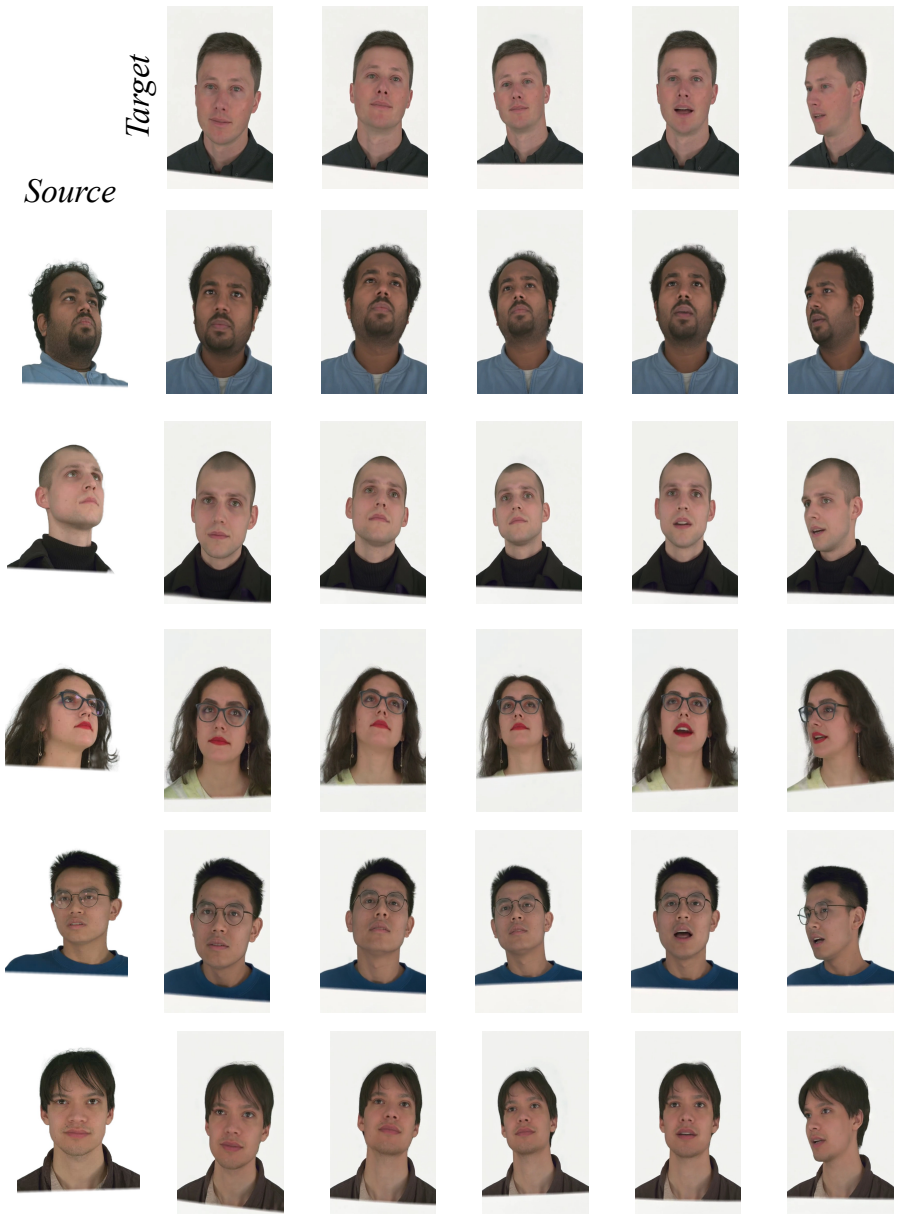


Figure 6: **The enhancement performance of AvatarBoost in NeRSemble dataset.**