TAOHEADVERSE: LEARNING EXPRESSIVE FULL-HEAD AVATARS BEYOND REPRESENTATION CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized full-head avatars for mobile XR and telepresence require faithful animation and efficient rendering. Creating such avatars from one or a few portraits remains challenging because it requires completing unseen regions, preserving source identity under cross-identity driving, and modeling person-specific expression responses. We present TaoHeadVerse, a feed-forward framework that jointly estimates motion and predicts an explicit mesh-bound Gaussian avatar. To address representation confounding in joint avatar-motion learning and mixed-source training, Neutral Identity Anchoring limits identity leakage into expression components, while Adversarial Domain Alignment discourages canonical features from encoding dataset-level acquisition attributes. The predicted asset stores identity-adapted mesh blendshapes and Gaussian appearance bases, enabling animation through basis mixing without per-frame neural avatar decoding. Experiments on VFHQ, AVA256, and RenderMe-360 show improved reconstruction, identity preservation, and full-head completion with competitive expression fidelity. Given precomputed driving controls, four concurrent avatars achieve animation and rendering throughput above 60 FPS on a shared 1024×1024 canvas on both mobile devices.

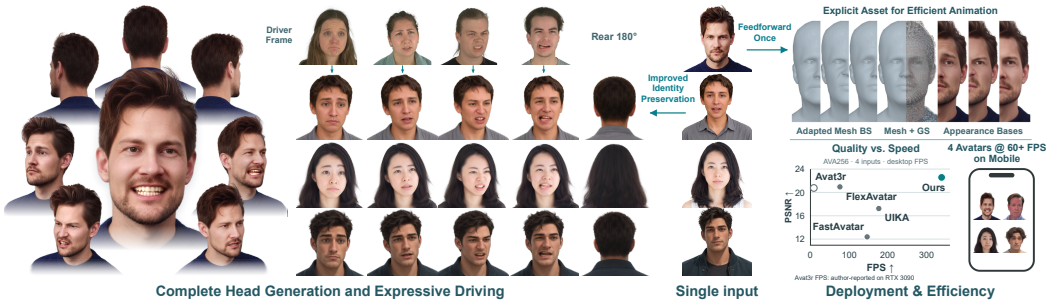


Figure 1: **TaoHeadVerse reconstructs explicit full-head avatars from one or a few portraits for expressive and identity-preserving animation.** Four concurrent avatars run above 60 FPS on mobile hardware; the inset shows the AVA256 quality–speed trade-off.

1 INTRODUCTION

Applications such as mobile XR and telepresence require personalized full-head avatars that combine faithful appearance with real-time animation on resource-constrained devices. Creating and animating such avatars from one or a few casual portraits is challenging because it requires completing unseen head regions, accurately estimating driving expressions, and efficiently representing person-specific expression responses.

Recent Gaussian avatar methods (He et al., 2025; Kirschstein et al., 2025; Ji et al., 2026; Wu et al., 2026b) often rely on externally estimated motion, whose errors can propagate into reconstruction and animation. Joint estimation can better align motion with appearance, as explored by Personalized Face Modeling (Chaudhuri et al., 2020) and FFAvatar (Nguyen et al., 2026). Yet joint reconstruction alone does not prevent a form of *representation confounding*. Self-reenactment supervision cannot determine whether static identity cues reside in the neutral asset or in expression-dependent components. An avatar may thus perform accurate self-reenactment yet exhibit source-identity drift under cross-identity driving. We address this failure by explicitly anchoring identity in the neutral asset.

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.

A second form of representation confounding arises when training across heterogeneous data sources. Monocular collections provide identity and expression diversity, whereas multiview datasets such as AVA256 (Martinez et al., 2024) provide broader coverage of side- and rear-head regions. Correlations between acquisition conditions and full-head supervision may bias canonical features toward source-specific cues, weakening their source invariance. Such shortcuts may limit how effectively multiview supervision benefits avatars reconstructed from frontal portraits. FlexAvatar (Kirschstein et al., 2026) addresses mixed-source bias using source-specific tokens that absorb acquisition-specific information. In contrast, our approach discourages canonical features from encoding acquisition-related cues.

We introduce **TaoHeadVerse**, a feed-forward framework that jointly estimates motion and generates an explicit, mesh-bound Gaussian full-head asset (Li et al., 2017; Kerbl et al., 2023; Qian et al., 2024). Neutral Identity Anchoring (NIA) anchors static identity in the neutral asset through zero-expression supervision, while Adversarial Domain Alignment (ADA) makes canonical features less predictive of acquisition conditions (Tzeng et al., 2015). The resulting asset stores identity-adapted mesh blendshapes and person-specific Gaussian appearance bases, enabling animation through basis mixing, skinning, mesh-bound Gaussian transformation, and rasterization without per-frame neural avatar decoding.

Experiments on VFHQ, AVA256, and RenderMe-360 demonstrate improved reconstruction, identity preservation, and full-head completion. On VFHQ, TaoHeadVerse improves self-reenactment PSNR/SSIM over FlexAvatar by up to 2.50 dB/0.036 and better preserves source identity under cross-reenactment while maintaining competitive expression fidelity. Matched ablations show higher cross-identity similarity with NIA and lower rear-view reconstruction error with ADA, while the explicit assets enable real-time mobile animation and rendering.

Our contributions are: (1) **Neutral Identity Anchoring (NIA)**, a zero-expression identity constraint that mitigates identity leakage in joint avatar–motion learning; (2) **Adversarial Domain Alignment (ADA)**, which reduces acquisition-attribute predictability in canonical features for mixed-source full-head learning; and (3) **feed-forward explicit dynamic assets** with identity-adapted mesh blendshapes and Gaussian appearance bases, enabling real-time mobile animation without per-frame neural decoding.

2 Related Work

Head avatar reconstruction spans per-subject optimization of neural fields and Gaussian representations (Gafni et al., 2021; Zheng et al., 2022; Zielonka et al., 2023; Qian et al., 2024; Xiang et al., 2024) and feed-forward prediction from sparse portraits (Chu & Harada, 2024; Kirschstein et al., 2025; Nguyen et al., 2026; Kirschstein et al., 2026). TaoHeadVerse follows the feed-forward setting. Methods such as LAM (He et al., 2025), FastAvatar (Wu et al., 2026a), and UIKA (Wu et al., 2026b) use FLAME (Li et al., 2017)-based controls typically estimated separately from avatar reconstruction, allowing fitting errors, template mismatch, and neutral-point offsets to propagate into the avatar. Personalized Face Modeling (Chaudhuri et al., 2020) and FFAvatar (Nguyen et al., 2026) estimate personalized appearance together with per-frame motion but, unlike TaoHeadVerse with NIA, do not explicitly address identity allocation between neutral and expression-dependent components.

Sparse-input full-head reconstruction requires learned priors or complementary supervision. PanoHead (An et al., 2023) and SphereHead (Li et al., 2024) learn generative 360-degree priors, while PanoLAM (Li et al., 2025b) transfers such priors through synthetic multiview training. FaceLift (Lyu et al., 2025) and other methods synthesize or complete missing non-frontal observations (Taubner et al., 2025; Xin et al., 2025; Zhang et al., 2025), making reconstruction dependent on the learned prior or pseudo-observation quality. FlexAvatar (Kirschstein et al., 2026) instead learns from mixed monocular and multiview supervision using source-specific tokens to absorb capture bias, whereas TaoHeadVerse adversarially reduces acquisition-attribute predictability in the shared canonical representation.

Efficient animation requires balancing person-specific dynamics against per-frame computation. LAM (He et al., 2025) and UIKA (Wu et al., 2026b) use parametric rigs without neural decoding, whereas expression-conditioned networks, including FastGHA, predict per-Gaussian deformations from externally estimated expression controls at a recurring inference cost (Ji et al., 2026; Kirschstein et al., 2025; 2026; Xu et al., 2024). Explicit dynamic bases retain efficient linear mixing: Personalized

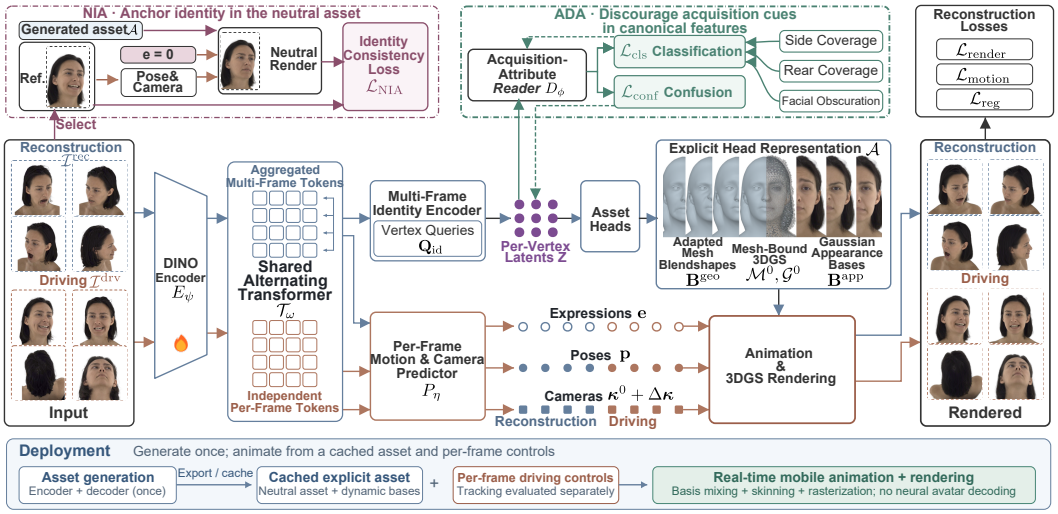


Figure 2: **Overview of TaoHeadVerse.** Reconstruction views jointly define the identity asset, while driving frames independently predict motion. NIA anchors identity in the neutral asset, and ADA suppresses acquisition cues in canonical features.

Face Modeling (Chaudhuri et al., 2020) predicts mesh and albedo corrections, while Gaussian Blendshapes (Ma et al., 2024) and RGBAvatar (Li et al., 2025a) optimize person-specific Gaussian bases from video. Pruned local blendshapes (Zhan et al., 2026) extend this design to full-body mobile animation. TaoHeadVerse instead predicts identity-adapted mesh blendshapes and Gaussian appearance bases from sparse portraits for feed-forward full-head reconstruction and animation without per-frame neural decoding.

3 METHOD

Figure 2 outlines TaoHeadVerse. We present joint avatar–motion estimation and explicit asset construction, then NIA for identity anchoring and ADA for mitigating acquisition-related confounding.

3.1 END-TO-END AVATAR AND MOTION LEARNING

During training, sampled frames from the same identity are split into reconstruction images $\mathcal{I}^{\text{rec}} = \{I_s^{\text{rec}}\}_{s=1}^{S}$ for avatar construction and driving frames $\mathcal{I}^{\text{drv}} = \{I_t^{\text{drv}}\}_{t=1}^{T}$. A DINOv2-initialized encoder (Oquab et al., 2024) E_ψ is fine-tuned end-to-end to map each image to patch tokens. A shared Transformer $\mathcal{T}_\omega$ processes them under two attention masks:

$$\mathbf{T}^{\text{rec}} = \mathcal{T}_\omega(E_\psi(\mathcal{I}^{\text{rec}}); \mathbf{M}_{\text{rec}}), \qquad \mathbf{T}_t^{\text{drv}} = \mathcal{T}_\omega\big(E_\psi(I_t^{\text{drv}}); \mathbf{M}_{\text{drv}}\big). \tag{1}$$

The reconstruction mask $\mathbf{M}_{\text{rec}}$ alternates within-view and global cross-view attention, allowing information exchange among reconstruction images. The driving mask $\mathbf{M}_{\text{drv}}$ instead isolates each driving frame.

Canonical vertex queries $\mathbf{Q}_{\text{id}}$ cross-attend only to $\mathbf{T}^{\text{rec}}$, aggregating the reconstruction views into canonical per-vertex features $\mathbf{Z}$. From $\mathbf{Z}$, asset heads predict the neutral mesh, canonical mesh-bound Gaussians, and appearance bases, while the geometric bases are adapted to the predicted neutral mesh. Together, they constitute the explicit avatar $\mathcal{A}$ detailed in Section 3.2.

Within the shared motion predictor P_η, learned motion queries $\mathbf{Q}_{\text{mot}}$ cross-attend to a contextualized reconstruction representation $\mathbf{T}_s^{\text{rec}}$ or an isolated driving representation $\mathbf{T}_t^{\text{drv}}$ to estimate per-frame controls $\mathbf{u} = (\mathbf{e}, \mathbf{p}, \Delta\boldsymbol{\kappa})$: expression, pose, and a residual to the tracked camera intrinsics $\boldsymbol{\kappa}^0$:

$$\mathbf{u}_s^{\text{rec}} = P_\eta(\mathbf{Q}_{\text{mot}}, \mathbf{T}_s^{\text{rec}}), \qquad \mathbf{u}_t^{\text{drv}} = P_\eta(\mathbf{Q}_{\text{mot}}, \mathbf{T}_t^{\text{drv}}). \tag{2}$$

P_η denotes the motion-token extractor together with the expression, pose, and camera heads. We use $E = 51$ ARKit blendshapes (excluding `tongueOut`) on our FLAME-based head mesh. The expression vector $\mathbf{e}$ controls facial, jaw, and eye motion, while $\mathbf{p}$ controls head pose. Both reconstruction

and driving frames are reconstructed during training, but only $\mathbf{T}^{\text{rec}}$ contributes to $\mathcal{A}$; at inference, driving frames may depict a different identity. Appendix A.3 details the predictor architecture.

3.2 Explicit Full-Head Avatar Representation

Given the canonical per-vertex features $\mathbf{Z}$, we construct an explicit avatar comprising a neutral mesh $\mathcal{M}^0$, canonical mesh-bound Gaussians $\mathcal{G}^0$, and geometric and appearance bases:

$$\mathcal{A} = \left(\mathcal{M}^0, \mathcal{G}^0, \mathbf{B}^{\text{geo}}, \mathbf{B}^{\text{app}}\right).$$

A vertex head predicts identity-specific offsets from the neutral FLAME template (Li et al., 2017) to obtain the neutral vertex coordinates $\mathbf{V}^0$ of $\mathcal{M}^0$. For each face f, its three vertex features are concatenated into $\widetilde{\mathbf{z}}_f^{\text{G}}$, from which a Gaussian head predicts the position, scale, rotation, opacity, and SH coefficients of the bound Gaussians. Their positions and covariances follow mesh deformation through mesh binding (Qian et al., 2024).

To capture person-specific expression geometry, Mesh Blendshape Adaptation (MBA) rescales each ARKit blendshape along three coordinate axes using identity-specific factors. The scale factors are estimated by a closed-form facial-graph fit weighted by the blendshape-induced edge displacements. This yields $\mathbf{B}^{\text{geo}} = \{\mathbf{B}_j^{\text{geo}}\}_{j=1}^{E}$, where each $\mathbf{B}_j^{\text{geo}} \in \mathbb{R}^{|\mathbf{V}^0|\times 3}$ is the identity-adapted per-vertex displacement field for expression channel j.

In parallel, a shared appearance head predicts Gaussian Appearance Bases (GAB) $\mathbf{B}^{\text{app}} = \{\mathbf{B}_j^{\text{app}}\}_{j=1}^{E}$ from the same face-level features, where $\mathbf{B}_j^{\text{app}} \in \mathbb{R}^{|\mathcal{G}^0|\times 3}$ represents per-Gaussian degree-zero SH corrections for expression channel j. GAB captures expression-dependent appearance while keeping higher-order SH coefficients fixed. Implementation details of MBA and GAB are provided in Appendices A.2 and A.1.

Let $\mathbf{c}^{0,\text{DC}} \in \mathbb{R}^{|\mathcal{G}^0|\times 3}$ denote the neutral degree-zero SH coefficients predicted by the Gaussian head. At frame t, the same expression $\mathbf{e}_t$ jointly drive geometric deformation and appearance variation:

$$\mathbf{V}_t = \text{LBS}\Big(\mathbf{V}^0 + \textstyle\sum_{j=1}^{E} e_{t,j}\mathbf{B}_j^{\text{geo}},\, \mathbf{p}_t\Big), \qquad \mathbf{c}_t^{\text{DC}} = \mathbf{c}^{0,\text{DC}} + \textstyle\sum_{j=1}^{E} e_{t,j}\mathbf{B}_j^{\text{app}}. \tag{3}$$

3.3 Mitigating Representation Confounding

Reconstruction alone does not determine how identity is allocated between the neutral asset and expression-dependent bases, while heterogeneous training data can introduce acquisition-specific cues into canonical features.

Neutral Identity Anchoring. NIA encourages identity cues to reside in the neutral asset by matching an observed frame to its zero-expression render in ArcFace feature space. For a selected near-frontal frame $I_i \in \mathcal{I}^{\text{rec}} \cup \mathcal{I}^{\text{drv}}$, let $(\mathbf{p}_i, \Delta\boldsymbol{\kappa}_i)$ denote its predicted pose and intrinsics residual. We form the zero-expression render $I_i^{\mathbf{0}} = \mathcal{R}(\mathcal{A}; \mathbf{0}, \mathbf{p}_i, \boldsymbol{\kappa}^0 + \Delta\boldsymbol{\kappa}_i)$, where $\mathcal{R}$ denotes the Gaussian renderer. We compute a shared alignment transform $\mathcal{C}$ from the projected facial landmarks and apply it to both images. Using a frozen, normalized ArcFace encoder (Deng et al., 2019a) φ, we define

$$\mathcal{L}_{\text{NIA}} = 1 - \left\langle \varphi\big(\mathcal{C}(I_i^{\mathbf{0}})\big)\,, \varphi(\mathcal{C}(I_i))\right\rangle. \tag{4}$$

ArcFace remains frozen, while gradients propagate through its input to the neutral rendering path. Since ArcFace is optimized for identity discrimination, its embeddings prioritize identity-related information over expression-dependent variation. We therefore use ArcFace similarity to supervise the neutral asset using sampled frames with varying expressions. NIA is restricted to near-frontal frames to reduce pose and alignment errors.

Adversarial Domain Alignment. Training datasets differ in acquisition coverage and facial visibility, which can encourage shortcut learning from dataset-specific supervision patterns. We therefore introduce Adversarial Domain Alignment (ADA) to suppress these acquisition cues in the canonical features and improve complete head reconstruction. A dataset-level lookup table assigns each sample b binary attributes $\mathbf{a}_b$ and an availability mask $\mathbf{m}_b$. Index $k \in \{1, 2, 3\}$ denotes rear-head coverage, side-head coverage, or facial obscuration, with $m_{b,k} = 1$ marking an available label. These attributes characterize the supervision provided by each acquisition setup and remain fixed throughout training.

A global reader D_ϕ predicts attribute logits from the canonical vertex features $\mathbf{Z}$. Let $z_{b,k} = D_{\phi,k}(\mathbf{Z}_b)$, $z^{\mathrm{sg}}_{b,k} = D_{\phi,k}(\mathrm{sg}[\mathbf{Z}_b])$ (sg denotes stop-gradient), and $\mathcal{V} = \{(b,k) \mid m_{b,k} = 1\}$. We use $\langle\cdot\rangle_{\mathcal{V}}$ to denote the mean over valid sample–attribute pairs and BCE_w for binary cross-entropy with positive-class weight w:

$$\mathcal{L}_{\mathrm{cls}} = \left\langle \mathrm{BCE}_{w_k}(z^{\mathrm{sg}}_{b,k}, a_{b,k}) \right\rangle_{\mathcal{V}}, \quad \mathcal{L}_{\mathrm{conf}} = \left\langle \mathrm{BCE}_1(z_{b,k}, \tfrac{1}{2}) \right\rangle_{\mathcal{V}}. \tag{5}$$

Here, w_k is the positive-class weight for attribute k. The stop-gradient restricts $\mathcal{L}_{\mathrm{cls}}$ to updating D_ϕ, whereas D_ϕ is held fixed when $\mathcal{L}_{\mathrm{conf}}$ updates the upstream feature network to make acquisition attributes less predictable.

Appendix A.4 details the reader architecture, acquisition attributes, and class-weight construction, and provides a qualitative PCA visualization in Figure 6. Optimization schedules are provided in Appendix A.5. The reader is discarded after training and adds no animation-time cost.

3.4 Optimization and Real-Time Animation

While the reader D_ϕ is trained by a separate objective $\mathcal{L}_{\mathrm{cls}}$, the generator and motion heads are optimized with

$$\mathcal{L} = \mathcal{L}_{\mathrm{render}} + \mathcal{L}_{\mathrm{motion}} + \mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{n}}\mathcal{L}_{\mathrm{NIA}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}. \tag{6}$$

The rendering term combines RGB, alpha, and SSIM (Wang et al., 2004) losses with frozen DINOv2 (Oquab et al., 2024) and SAM 2.1 (Ravi et al., 2024) feature losses following PercHead (Oroz et al., 2026); the motion term uses landmark-reprojection, expression-sparsity, and expression-bound losses, while the regularization term constrains geometry and Gaussian parameters. We set $\lambda_{\mathrm{n}} = 0.1$. Appendix A.5 provides additional loss details, individual weights, and activation schedules, including the schedule for λ_{conf}.

At deployment, the exported asset is animated using per-frame expression $\mathbf{e}$, pose $\mathbf{p}$, and camera parameters. These controls can be extracted from a video clip by our motion predictor or supplied by the user. Each frame requires only basis mixing, LBS, mesh-bound Gaussian transformation, and rasterization; graph fitting and neural asset decoding are performed only during asset generation.

4 Experiments

We evaluate TaoHeadVerse on portrait animation, multiview reenactment, single-image full-head reconstruction, component ablations, and desktop and mobile efficiency. Since static figures cannot fully convey temporal animation quality, we strongly encourage readers to watch the supplementary video with sound enabled.

4.1 Experimental Setup

Protocols and baselines. We evaluate portrait animation on VFHQ-Test (Xie et al., 2022), sparse-view reenactment on AVA256 (Martinez et al., 2024), and single-image full-head reconstruction on RenderMe-360 (Pan et al., 2023). For VFHQ and AVA256, we follow the FlexAvatar (Kirschstein et al., 2026) protocols, including test splits, inputs, source–driver pairings, preprocessing, and evaluation; all test identities are held out from training. Reproduced baselines use the same inputs and targets, while author-reported results, protocol differences, and optional fitting are marked explicitly. Section 4.4 defines the RenderMe-360 protocol.

Training. TaoHeadVerse-Public uses only the public datasets listed in Table 6 of Appendix A.4, whereas TaoHeadVerse-Full additionally uses synthetic, rendered MetaHuman, and studio-captured multiview data. Unless otherwise specified, "Ours" in figures and tables denotes TaoHeadVerse-Full. Both variants share the same architecture, optimizer, and training schedule; ADA class weights are computed separately from each variant's sampling distribution. All ablations use the Full setting. We train for 700,000 steps with AdamW (Loshchilov & Hutter, 2019) and a global batch size of 16; Appendix A.5 provides complete training details.

Table 1: VFHQ-Test comparison. Self-reenactment evaluates held-out appearance, identity and motion; cross-reenactment evaluates source identity and driving-motion agreement. Bold/underline denote the best/second-best results under our protocol; author-reported or fitted results are excluded from ranking.

	Self-reenactment								Cross-reenactment				
Method	PSNR↑	SSIM↑	LPIPS↓	CSIM_{Arc} ↑	CSIM_{Ada} ↑	AED↓	APD↓	AKD↓	CSIM_{Arc} ↑	CSIM_{Ada} ↑	AED↓	Ref-AED↓	APD↓
Reported in prior work													
GPAvatar* (Chu et al., 2024; Chu & Harada, 2024)	21.04	0.807	0.150	0.772	–	0.132	0.189	4.226	0.564	–	0.255	–	0.328
Portrait4D-v2* (Deng et al., 2024; Chu & Harada, 2024)	21.34	0.791	0.144	0.803	–	0.117	0.187	3.749	0.656	–	0.268	–	0.273
GAGAvatar* (Chu & Harada, 2024)	21.83	0.818	0.122	0.816	–	0.111	0.135	3.349	0.633	–	0.253	–	0.247
LAM* (He et al., 2025)	22.65	0.829	0.109	0.822	–	0.102	0.134	2.059	0.651	–	0.250	–	0.356
FFAvatar-4D* (Yao et al., 2026)	21.82	0.843	0.108	0.817	–	0.109	0.149	–	–	–	–	–	–
Our evaluation without fitting													
UIKA (Wu et al., 2026b)	20.11	0.792	0.127	0.751	0.753	0.116	0.022	8.656	0.655	0.660	0.244	0.130	0.040
FastAvatar (Wu et al., 2026a)	16.70	0.782	0.169	0.652	0.659	0.117	0.021	6.056	0.507	0.512	0.240	0.132	0.050
FlexAvatar (Kirschstein et al., 2026)	22.48	0.818	0.122	0.762	0.759	0.090	0.011	4.449	0.596	0.597	**0.221**	0.117	**0.025**
TaoHeadVerse-Public	<u>24.78</u>	<u>0.851</u>	<u>0.097</u>	**0.842**	**0.847**	<u>0.081</u>	**0.010**	<u>3.812</u>	**0.774**	**0.782**	0.243	<u>0.113</u>	0.031
TaoHeadVerse-Full	**24.98**	**0.854**	**0.093**	<u>0.838</u>	<u>0.843</u>	**0.080**	**0.010**	**3.723**	<u>0.766</u>	<u>0.774</u>	<u>0.233</u>	**0.111**	<u>0.029</u>
Additional baseline with fitting													
FlexAvatar (w/ fitting) (Kirschstein et al., 2026)	23.32	0.836	0.099	0.830	0.844	0.077	0.010	3.646	0.663	0.672	0.225	0.119	0.025

* Author-reported values copied from prior comparison tables may follow different evaluation protocols. In particular, their APD values may use different pose estimators, parameterizations, normalization, or aggregation; we therefore do not compare APD across the reported and reproduced blocks.

Metrics. Self-reenactment uses PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) for image fidelity; AKD, AED, and APD for motion fidelity; and CSIM for identity preservation. CSIM_{Arc} and CSIM_{Ada} use fixed evaluation networks independent of the NIA identity encoder; AVA256 reports CSIM_{Arc} only. In cross-reenactment, CSIM is computed against the source, whereas AED and APD are computed against the driver. On VFHQ, we also report reference-relative AED (Ref-AED), which compares generated and driving expression changes after subtracting each trajectory's estimate at the first valid frame. This reduces appearance-dependent offsets in the pretrained 3DMM evaluator. We report it alongside raw AED. Appendix A.6 provides the formal definition and evaluator details. Full-head reconstruction additionally uses mask IoU and a two-pixel-tolerance boundary F-score.

4.2 Portrait Animation

Table 1 shows that, over FlexAvatar, TaoHeadVerse-Public/Full improve self-reenactment PSNR by 2.30/2.50 dB and SSIM by 0.033/0.036, while Public raises cross-reenactment CSIM_{Arc}/CSIM_{Ada} from 0.596/0.597 to 0.774/0.782. TaoHeadVerse-Full has slightly higher raw AED than FlexAvatar (0.233 vs. 0.221), but lower Ref-AED (0.111 vs. 0.117), suggesting closer agreement in reference-relative expression changes, while the higher raw AED may reflect absolute expression mismatch and/or evaluator-dependent offsets. Figure 3(a) qualitatively illustrates improved source-identity preservation while maintaining expressive motion transfer.

4.3 Multi-View Reenactment

Following FlexAvatar (Kirschstein et al., 2026), we evaluate the same eight held-out AVA256 identities from one or four source images. Self-reenactment uses held-out expressions and calibrated target cameras, while cross-reenactment uses fixed circular identity pairings over 631 driving frames. Because FlexAvatar supports multiple inputs only through fitting, we compare one-input methods without fitting and report four-input feed-forward and fitted results separately. Similar to FlexAvatar, our optional fitting variant optimizes the predicted per-vertex latents for 200 steps while keeping all network weights fixed.

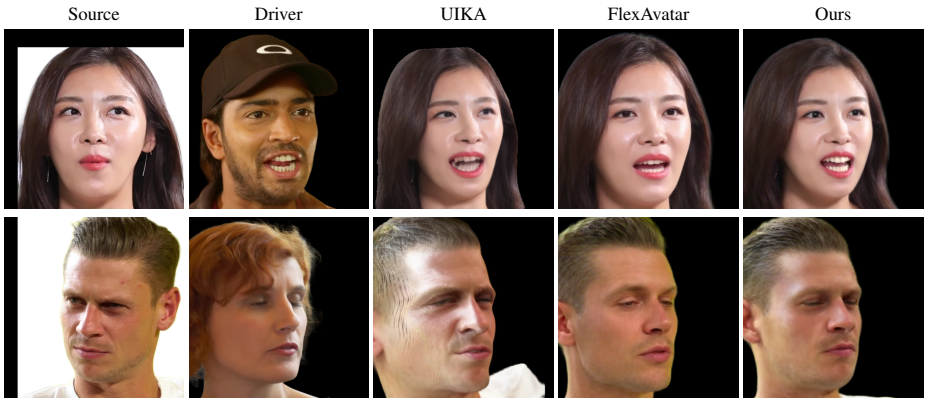
Table 2 shows consistent gains over the feed-forward baselines. With one input, TaoHeadVerse-Public improves PSNR from 19.3 to 21.7 and cross-reenactment CSIM_{Arc} from 0.513 to 0.540 over FlexAvatar. With four inputs, TaoHeadVerse-Full reaches 22.6 PSNR and 0.207 LPIPS, while TaoHeadVerse-Public obtains the highest feed-forward cross-reenactment CSIM of 0.562. Figure 3(b) further illustrates improved source-identity preservation without fitting.

4.4 Full-Head Reconstruction

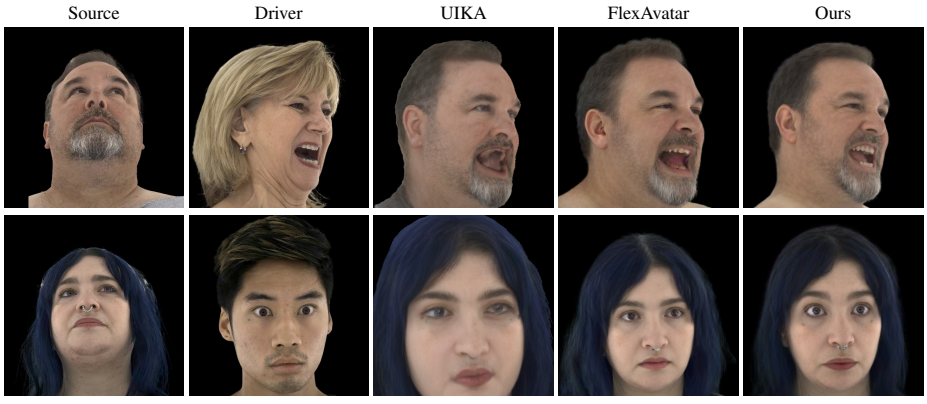
We evaluate 20 RenderMe-360 subjects (Pan et al., 2023) using one frontal RGB input per identity. We compare with SphereHead+PTI (Li et al., 2024; Roich et al., 2022), FaceLift (Lyu et al., 2025),

Table 2: AVA256 comparison with one or four inputs. Reproduced results use the shared 631-frame protocol; four-input Avat3r values follow the FlexAvatar evaluation. Bold marks the best feed-forward result per input count; fitted results are listed separately.

		Self-reenactment							Cross-reenactment		
Method	Inputs	PSNR↑	SSIM↑	LPIPS↓	CSIM_{Arc} ↑	AED↓	APD↓	AKD↓	CSIM_{Arc} ↑	AED↓	APD↓
One input: feed-forward without fitting											
UIKA (Wu et al., 2026b)	1	13.7	0.571	0.458	0.458	0.161	0.050	28.49	0.413	0.229	0.063
FastAvatar (Wu et al., 2026a)	1	12.4	0.576	0.492	0.170	0.224	0.127	27.87	0.147	0.252	0.127
FlexAvatar (w/o fitting) (Kirschstein et al., 2026)	1	19.3	0.693	0.276	0.580	0.125	0.041	11.33	0.513	0.215	0.060
TaoHeadVerse-Public (ours, w/o fitting)	1	**21.7**	**0.735**	0.239	**0.683**	**0.103**	**0.030**	5.97	**0.540**	0.212	**0.044**
TaoHeadVerse-Full (ours, w/o fitting)	1	21.6	0.734	**0.237**	0.669	**0.103**	0.031	**5.81**	0.525	**0.203**	0.045
Four inputs: feed-forward without fitting											
Avat3r (Kirschstein et al., 2025)	4	20.8	0.715	0.310	0.616	–	–	5.66	–	–	–
UIKA (Wu et al., 2026b)	4	17.3	0.671	0.301	0.625	0.134	0.049	16.51	0.497	0.224	0.070
FastAvatar (Wu et al., 2026a)	4	12.4	0.580	0.495	0.162	0.231	0.123	28.05	0.111	0.246	0.122
TaoHeadVerse-Public (ours, w/o fitting)	4	**22.6**	0.746	0.212	**0.721**	0.101	**0.030**	5.67	**0.562**	0.211	**0.044**
TaoHeadVerse-Full (ours, w/o fitting)	4	**22.6**	**0.749**	**0.207**	0.712	**0.099**	**0.030**	**5.50**	0.551	**0.204**	0.045
Four inputs: with fitting											
FlexAvatar (w/ fitting) (Kirschstein et al., 2026)	4	21.0	0.732	0.219	0.769	0.097	0.029	6.06	0.550	0.216	0.062
TaoHeadVerse-Full (w/ fitting)	4	23.1	0.757	0.205	0.783	0.093	0.027	5.43	0.606	0.214	0.051



(a) VFHQ



(b) AVA256

Figure 3: Cross-identity reenactment on (a) VFHQ-Test and (b) AVA256. The expanded VFHQ comparison including GAGAvatar and additional AVA256 examples appear in Appendix A.7.1.

and FlexAvatar (Kirschstein et al., 2026). To account for monocular scale ambiguity and align reconstructed assets with the calibrated camera coordinate system, all methods undergo identical four-view head-mask registration. Evaluation covers 59 non-input cameras per identity (1,180 views, including the registration views). We report identity-averaged LPIPS, mask IoU, and boundary F-score at 512×512. Appendix A.6.2 details the protocol.

TaoHeadVerse-Full achieves the best results across all view bins and metrics, while TaoHeadVerse-Public outperforms the other baselines on all three aggregate metrics (Table 3). Figure 4 shows strict-rear examples; additional results appear in Figure 9 of Appendix A.7.1.

Table 3: RenderMe-360 reconstruction with shared four-view head-mask registration. † additionally uses per-identity optimization. Bold/underline denote best/second-best results; Pub./Full are TaoHeadVerse variants.

	LPIPS↓					Mask IoU↑					Boundary F↑				
View	Sphere†	FaceLift	Flex	Ours-Pub.	Ours-Full	Sphere†	FaceLift	Flex	Ours-Pub.	Ours-Full	Sphere†	FaceLift	Flex	Ours-Pub.	Ours-Full
All	.2833	.2671	.2542	.1977	**.1746**	.8392	.7737	.8668	.8712	**.9015**	.2180	.2572	.2715	.3986	**.4210**
Front	.2474	.2526	.2140	.1721	**.1483**	.8792	.7611	.8922	.8821	**.9239**	.2950	.3111	.3273	.4721	**.5078**
Side	.2877	.2576	.2464	.1961	**.1748**	.8318	.7822	.8505	.8680	**.8881**	.1733	.2169	.2233	.3393	**.3510**
Rear-side	.3052	.2722	.2688	.2099	**.1909**	.7838	.7744	.8343	.8520	**.8701**	.1584	.2015	.2218	.3317	**.3374**
Strict rear	.2971	.2865	.2908	.2150	**.1870**	.8553	.7784	.8853	.8804	**.9195**	.2334	.2893	.3031	.4381	**.4718**

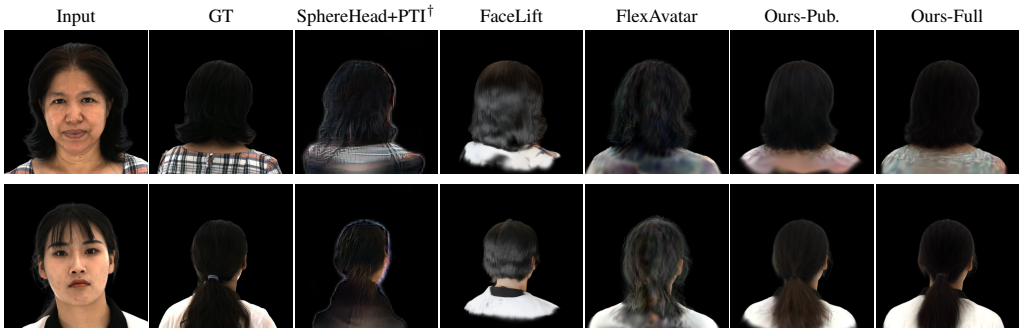


Figure 4: Strict-rear RenderMe-360 reconstruction with shared registration. † additionally uses per-identity optimization.

4.5 Ablation Studies

Table 4 reports six matched comparisons among seven configurations: $A0/A3$, $A1/A3$, $A1^{\mathrm{ID}}/A3$, $A2/A3$, $A3/A4$, and $A4/A5$. Each pair differs in one component, while the initialization, training data, optimization protocol, architecture of the unaffected components, and evaluation targets are held fixed. Identity improvements are interpreted jointly with AED, Ref-AED, and APD to check that they do not arise from suppressing the driving motion.

Joint estimation ($A0 \rightarrow A3$) produces the largest appearance gain, increasing PSNR from 20.54 to 24.72 and reducing LPIPS from 0.134 to 0.097. Compared with the Direct ID control, NIA improves both cross-identity CSIM metrics with comparable expression and pose errors. GAB provides smaller appearance and AED gains.

MBA reduces overall face/mouth AKD by 2.06%/5.90%, with a 1.38% eye increase (Table 8 of Appendix A.7.2), and adds no per-frame cost. Finally, ADA reduces RenderMe-360 rear-side/strict-rear LPIPS by 10.3%/11.3% and improves both silhouette metrics, with a slight cross-identity CSIM trade-off on VFHQ. Appendix A.7.2 provides qualitative comparisons for NIA, ADA, GAB, and MBA in Figures 10, 11, 12, and 13, respectively.

4.6 Efficiency and Mobile Deployment

Because TaoHeadVerse-Public and TaoHeadVerse-Full use the same runtime asset representation and animation pipeline, we profile only the Full variant. Desktop profiling measures animation and rendering time using each baseline's official Python implementation and our Python implementation on an RTX 4090 at 512×512. Mobile WebGPU profiling on iPhone 17 and OnePlus 12 uses three warmed-up 1,000-frame runs with up to eight frames in flight. Reported mobile times are derived from throughput measurements of animation and rendering only.

Table 5 shows that TaoHeadVerse achieves 2.95 ms total desktop latency and 339.10 FPS, outperforming the locally measured baselines. Four concurrent avatars rendered on a single 1024×1024 canvas achieve 84.19 FPS on the iPhone 17 and 68.38 FPS on the OnePlus 12, exceeding 60 FPS on both devices.

Table 4: Matched ablations. (a) VFHQ-Test component comparisons; Direct ID denotes identity supervision applied to predicted-expression renderings, in contrast to the zero-expression renderings used by NIA. (b) ADA-confusion comparison on 20 RenderMe-360 subjects and 1,180 views under the shared four-view head-mask registration protocol.

(a) VFHQ-Test component ablations.

	Components						VFHQ Self					VFHQ Cross-ID				
Variant	Joint	NIA	Direct ID	GAB	MBA	ADA Conf.	PSNR↑	LPIPS↓	CSIM_{Arc} ↑	CSIM_{Ada} ↑	AED↓	CSIM_{Arc} ↑	CSIM_{Ada} ↑	AED↓	Ref-AED↓	APD↓
Fixed tracking ($A0$)		**Y**		**Y**			20.54	0.134	0.765	0.759	0.107	0.755	0.764	0.253	0.123	0.031
w/o NIA ($A1$)	**Y**			**Y**			24.79	0.095	0.821	0.819	0.079	0.680	0.681	0.230	0.114	0.031
Direct ID loss ($A1^{\text{ID}}$)	**Y**		**Y**	**Y**			24.76	0.097	0.835	0.835	0.082	0.715	0.719	0.242	0.112	0.030
w/o GAB ($A2$)	**Y**	**Y**					24.29	0.098	0.829	0.833	0.086	0.772	0.780	0.246	0.118	0.031
w/o MBA ($A3$)	**Y**	**Y**		**Y**			24.72	0.097	0.833	0.836	0.083	0.777	0.775	0.240	0.113	0.031
Base ($A4$, w/o ADA confusion)	**Y**	**Y**		**Y**	**Y**		24.92	0.096	0.839	0.843	0.080	0.775	0.784	0.235	0.112	0.032
Full ($A5$, w/ ADA confusion)	**Y**	**Y**		**Y**	**Y**	**Y**	24.98	0.093	0.838	0.843	0.080	0.766	0.774	0.233	0.111	0.029

(b) Adversarial-confusion ablation on RenderMe-360.

	Front	Rear-side			Strict rear		
Variant	LPIPS↓	LPIPS↓	IoU↑	Bnd. F↑	LPIPS↓	IoU↑	Bnd. F↑
Base ($A4$, w/o ADA confusion)	0.171	0.213	0.849	0.309	0.211	0.897	0.428
Full ($A5$, w/ ADA confusion)	**0.148**	**0.191**	**0.870**	**0.337**	**0.187**	**0.920**	**0.472**

Table 5: Efficiency benchmarks. (a) FP32 desktop runtime at 512×512 and batch size one on an RTX 4090; latency is mean (p95) GPU time and FPS is synchronized throughput. † denotes method-native inputs and ‡ author-reported FPS. (b) Amortized mobile WebGPU runtime; component and total times are measured separately.

(a) Desktop runtime.

Method	Anim. GFLOPs	Animation (ms)	Rendering (ms)	Total (ms)	FPS
Avat3r (Kirschstein et al., 2025)‡	–	–	–	–	∼8
FastGHA (4 inputs) (Ji et al., 2026)‡	–	–	–	–	62
FlexAvatar (Peng et al., 2026)‡	–	–	–	–	∼45
FlexAvatar (Kirschstein et al., 2026)†	182.51	11.41 (12.49)	1.77 (2.07)	13.18 (14.34)	75.86
FastAvatar†	0.04	2.83 (3.45)	3.55 (4.55)	6.38 (7.53)	145.58
UIKA†	0.12	1.48 (1.86)	4.20 (4.97)	5.68 (6.53)	176.10
TaoHeadVerse (Ours)†	0.04	**1.46 (2.86)**	**1.49 (1.98)**	**2.95 (4.67)**	**339.10**

(b) Mobile WebGPU runtime.

Device	SoC	Resolution	Avatars	Animation (ms/frame)	Rendering (ms/frame)	Total (ms/frame)	FPS
iPhone 17	A19	512×512	1	0.91	2.26	3.02	330.91
iPhone 17	A19	1024×1024	4	4.31	11.61	11.88	84.19
OnePlus 12	Snapdragon 8 Gen 3	512×512	1	1.61	2.92	4.14	241.71
OnePlus 12	Snapdragon 8 Gen 3	1024×1024	4	5.73	9.35	14.62	68.38

4.7 Limitations

Expression extraction is not yet real time on mobile devices, although the asset can use platform-native or audio-based controls. MBA uses only three axis-wise scales per expression channel and therefore cannot represent arbitrary vertex corrections. Finally, single-image full-head completion is inherently ambiguous: the model predicts plausible unseen hair and rear-head appearance rather than recovering unobserved ground truth.

5 Conclusion

We presented TaoHeadVerse, a feed-forward framework that jointly estimates motion and reconstructs an explicit full-head avatar from one or a few portraits. Experiments show that Neutral Identity Anchoring improves source-identity preservation under cross-identity driving, while Adversarial Domain Alignment improves full-head reconstruction, particularly in unseen side and rear regions. The public-data variant also outperforms the reproduced feed-forward baselines on key reconstruction and identity metrics, showing that strong overall performance does not require proprietary training data. By storing identity-adapted mesh blendshapes and Gaussian appearance bases in the exported asset, TaoHeadVerse avoids per-frame neural avatar decoding. Given precomputed driving controls, four concurrent avatars achieve animation and rendering throughput above 60 FPS on a shared 1024×1024 canvas on both mobile devices.

AI USE STATEMENT

We used generative AI tools to assist with programming and implementing the proposed method, searching for and summarizing related work, and editing the manuscript for clarity and readability. Generative AI was also used to produce a portion of the training data for TaoHeadVerse-Full, as described in Appendix A.5 and Table 6. These AI-generated training datasets were not used for TaoHeadVerse-Public. The authors reviewed the AI-assisted work and take full responsibility for the implementation, experimental results, citations, and final content of the paper.

ETHICS STATEMENT

This work uses portrait data containing identifiable individuals and may be misused for unauthorized impersonation. Proprietary capture data were collected with informed consent for research use, and public datasets are used in accordance with their respective terms. We recommend that avatar creation and deployment require authorization from the depicted individuals and clear disclosure that the resulting content is synthetic.

Performance may vary across demographic groups and across hairstyles, occlusions, and capture conditions. These potential disparities and robustness limitations warrant evaluation before deployment. Identifiable proprietary capture data will not be released.

REPRODUCIBILITY STATEMENT

We strive to provide comprehensive details to support independent reimplementation of our approach. The main paper presents the overall architecture and avatar parameterization (Sections 3.1–3.2), training constraints and objectives (Sections 3.3–3.4), experimental setup (Section 4.1), the full-head reconstruction protocol (Section 4.4), and ablations of key design choices (Section 4.5). The appendix further provides detailed network architectures, dataset composition and input sampling, loss weights, optimization schedules, and baseline and evaluation configurations. We also report TaoHeadVerse-Public, a variant trained exclusively on the cited public datasets, to help assess the contribution of proprietary training data.

REFERENCES

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20950–20959, June 2023.

Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024. doi: 10.1145/3680528.3687580.

Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2007.06759.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems*, 2024.

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. URL https://openreview.net/forum?id=hgehGq2bDv.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019a. doi: 10.1109/CVPR.2019.00482.

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 Computer Vision and Pattern Recognition Workshops*, 2019b. URL https://arxiv.org/abs/1903.08527.

Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4D-v2: Pseudo multi-view data creates better 4D head synthesizer. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.13570.

Guy Gafni, Justus Thies, Michael Zollhöfer, 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.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3DMM: Versatile screen-space priors for single-image 3D face reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UmOdd5KQ8K.

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*. ACM, 2025. doi: 10.1145/3721238.3730706.

Xinya Ji, Sebastian Weiss, Manuel Kansy, Jacek Naruniec, Xun Cao, Barbara Solenthaler, and Derek Bradley. FastGHA: Generalized few-shot 3D gaussian head avatars with real-time animation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=E7VL9Zl1Nc.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18750–18759, 2022.

Taewoo Kim, Chaeyeon Chung, Sunghyun Park, Gyojung Gu, Keonmin Nam, Wonzo Choe, Jaesung Lee, and Jaegul Choo. K-Hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In *2021 IEEE International Conference on Image Processing*, pp. 1299–1303, 2021.

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

Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3D head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12089–12100, 2025.

Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. FlexAvatar: Learning complete 3D head avatars with partial supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18193–18203, 2026.

Orest Kupyn, Eugene Khvedchenia, and Christian Rupprecht. VGGHeads: 3D multi head alignment with a large-scale synthetic dataset. *arXiv preprint arXiv:2407.18245*, 2024. URL https://arxiv.org/abs/2407.18245.

Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. SphereHead: Stable 3D full-head synthesis with spherical tri-plane representation. In *Computer Vision – ECCV 2024*, pp. 324–341. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73226-3_19. URL https://arxiv.org/abs/2404.05680.

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. RGBAvatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10747–10757, 2025a.

Peng Li, Yisheng He, Yingdong Hu, Yuan Dong, Weihao Yuan, Yuan Liu, Siyu Zhu, Gang Cheng, Zilong Dong, and Yike Guo. PanoLAM: Large avatar model for gaussian full-head synthesis from one-shot unposed image, 2025b. URL https://arxiv.org/abs/2509.07552.

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*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813.

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

Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12691–12701, 2025.

Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3D gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657462. URL https://arxiv.org/abs/2404.19398.

Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83008–83023. Curran Associates, Inc., 2024. doi: 10.52202/079017-2640.

Thuan Hoang Nguyen, Jiahao Luo, Yinyu Nie, Hao Li, Gordon Guocheng Qian, and Jian Wang. FFAvatar: Few-shot, feed-forward, and generalizable avatar reconstruction. *arXiv preprint arXiv:2605.15320*, 2026. URL https://arxiv.org/abs/2605.15320.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Antonio Oroz, Matthias Nießner, and Tobias Kirschstein. PercHead: Perceptual head model for single-image 3D head reconstruction & editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4097–4108, 2026.

Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. RenderMe-360: A large digital asset library and benchmarks towards high-fidelity head avatars. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Cheng Peng, Zhuo Su, Liao Wang, Chen Guo, Zhaohu Li, Chengjiang Long, Zheng Lv, Jingxiang Sun, Chenyangguang Zhang, and Yebin Liu. FlexAvatar: Flexible large reconstruction model for

animatable gaussian head avatars with detailed deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.17717.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2022. doi: 10.1145/3544777.

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*, pp. 1–11, December 2025.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076, 2015.

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. doi: 10.1109/TIP.2003.819861.

Yue Wu, Xuanhong Chen, Yufan Wu, Wen Li, Yuxi Lu, and Kairui Feng. FastAvatar: Towards unified and fast 3D avatar reconstruction with large gaussian reconstruction transformers. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=P7zBSCs4Xt.

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*, 2026b. URL https://arxiv.org/abs/2601.07603.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 656–665, 2022. doi: 10.1109/CVPRW56347.2022.00081.

Shiqi Xin, Xiaolin Zhang, Yanbin Liu, Peng Zhang, and Caifeng Shan. AvatarBack: Back-head generation for complete 3D avatars from front-view images, 2025. URL https://arxiv.org/abs/2508.20623.

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*, pp. 1931–1941, 2024.

Jianjiang Yao, Ke Xian, Renxiang Dai, and Robert Caiming Qiu. FFAvatar: Feed-forward 4D head avatar reconstruction from sparse portrait images. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.30347.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14805–14814, 2023.

Youyi Zhan, He Wang, Tianjia Shao, and Kun Zhou. High-fidelity mobile avatars with pruned local blendshapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32345–32356, 2026.

Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. FATE: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5535–5545, 2025.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D gaussian splatting. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2404.19702.

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://richzhang.github.io/PerceptualSimilarity/.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13545–13555, 2022.

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 Appendix

A.1 Identity Aggregation and Asset Heads

Identity aggregation maps the contextualized reconstruction tokens $\mathbf{T}^{\mathrm{rec}}$ to canonical per-vertex features $\mathbf{Z}$. Asset heads then map these features to the neutral geometry, Gaussian attributes, and appearance bases stored in the exported avatar. For each canonical mesh vertex v at position $\mathbf{x}_v$, a point embedding concatenates $\mathbf{x}_v$ with a 48-dimensional sinusoidal encoding and projects the result to a 1024-dimensional vertex query. The conditioning sequence consists of the reconstruction tokens $\mathbf{T}^{\mathrm{rec}}$ produced by the shared Transformer. Four conditional Transformer blocks, each with eight attention heads, alternate query-to-image cross-attention, self-attention among the mesh queries, and a feed-forward layer. Four learned register tokens and one global token participate in this decoding but are removed before the per-vertex outputs are passed to the asset heads. Thus, identity aggregation returns one feature $\mathbf{z}_v \in \mathbb{R}^{1024}$ for every canonical vertex; driving-frame tokens do not enter this representation path.

Neutral geometry. The vertex head applies a two-hidden-layer MLP of width 512 with SiLU activations to each $\mathbf{z}_v$. Its three outputs are scaled by 0.01 and interpreted as a metric-space vertex offset. The final linear layer is initialized to zero, so the predicted mesh starts from the neutral FLAME template $\bar{\mathbf{V}}$ and learns the identity-specific displacement field

$$\mathbf{V}_v^0 = \bar{\mathbf{V}}_v + H_{\mathrm{V}}(\mathbf{z}_v), \tag{7}$$

where H_{V} includes the fixed output scale of 0.01.

Mesh-bound Gaussian attributes. The head mesh contains 5,768 vertices and 11,392 triangles and is derived from FLAME 2020 with programmatically constructed teeth and mouth-cavity geometry. For a mesh face $f = (a, b, c)$, we concatenate its three decoded vertex features, $\widetilde{\mathbf{z}}_f^{\mathrm{G}} = [\mathbf{z}_a; \mathbf{z}_b; \mathbf{z}_c] \in \mathbb{R}^{3072}$. The Gaussian head uses separate three-hidden-layer MLPs of width 2048 with SiLU activations to jointly predict local anchor offsets, anisotropic scale, 6D rotation, opacity, and SH coefficients up to degree three for all N_f Gaussian anchors attached to face f. Gaussian anchors are placed at triangle centroids across multiple levels of recursive midpoint subdivision. Level 0 consists of the original face; each subsequent level splits every triangle into four using its edge midpoints. We retain the centroids from all levels $i = 0, \ldots, L_f - 1$, yielding $N_f = \sum_{i=0}^{L_f-1} 4^i = (4^{L_f} - 1)/3$ anchors per face. Regular faces use $L_f = 2$ ($N_f = 5$), whereas

168 teeth faces and 976 scalp faces use $L_f = 3$ ($N_f = 21$). The resulting explicit asset contains $(11{,}392 - 168 - 976) \times 5 + (168 + 976) \times 21 = 75{,}264$ active Gaussians, all of which are animated and rendered at each frame. Position is expressed in the local face frame, scale is exponentiated and capped at five, opacity uses a sigmoid, and the 6D rotation is converted to a rotation matrix. During animation, mesh binding maps these local quantities to world space using the deformed face frame and rescales the covariance according to face area.

Gaussian Appearance Bases. The GAB prediction head consumes the same face-level feature $\widetilde{\mathbf{z}}_f^{\mathrm{G}}$. A two-hidden-layer MLP of width 256 with SiLU activations jointly predicts the appearance bases of all N_f anchors on face f. Its output can be arranged as

$$\mathbf{B}_f^{\mathrm{app}} \in \mathbb{R}^{N_f \times E \times 3}, \tag{8}$$

where $\mathbf{B}_{f,n,j}^{\mathrm{app}} \in \mathbb{R}^3$ is the degree-zero SH correction for anchor n on face f and expression channel j. Stacking these corrections over all face–anchor pairs (f, n) gives $\mathbf{B}_j^{\mathrm{app}} \in \mathbb{R}^{|\mathcal{G}^0| \times 3}$ used in Equation 3. Its final layer is zero-initialized. GAB appearance corrections are enabled from training step 300,000. During training, we independently drop the entire appearance correction for each sample–frame pair with probability 0.1, without rescaling retained corrections. A dropped frame uses canonical Gaussian appearance while retaining its expression-driven geometry. This regularization encourages canonical appearance to remain informative without relying on expression-dependent corrections. At inference, the bases are generated once and serialized with the avatar.

A.2 Detailed Formulation of Mesh Blendshape Adaptation

We model geometric dynamics through the underlying mesh, whose deformation is propagated to the positions and covariances of the bound Gaussians. We therefore use MBA to adapt the blendshapes controlling geometric motion, without introducing separate per-Gaussian position, rotation, and scale correction bases. Our template expression basis consists of 51 artist-authored blendshapes corresponding to ARKit controls,[1] excluding `tongueOut`. These blendshapes are defined on the head mesh, including its teeth and oral geometry. MBA personalizes their responses to the predicted neutral geometry without learning additional per-vertex displacement fields.

The adaptation uses three axis-wise scale factors for each blendshape. Although these factors are shared across all vertices of a blendshape, their effect follows its spatial support: zero template displacements remain zero. Moreover, the graph-fitting weights depend on the edge displacements induced by that blendshape, emphasizing the facial structures involved in its response. Thus, each expression channel is adapted independently. The scale factors are estimated in closed form without additional blendshape supervision.

We omit the identity index and use bars to mark template quantities; the predicted identity-specific neutral geometry is denoted by $\mathbf{V}^0$. We compare a common facial graph on the template mesh $\bar{\mathbf{V}}$ and the predicted neutral mesh $\mathbf{V}^0$, including its identity-specific vertex offsets. The mesh includes teeth and oral geometry. The graph retains local brow, eye, and lip structure, refines the inner-lip contour and lip-thickness connections, and adds central lip-to-teeth, nose-wing, and chin-to-jaw edges. We exclude long-range lip-to-nose/chin and brow-to-eyelid anchors, as well as upper-to-lower inner-lip gap edges, from scale estimation. The resulting graph contains L selected edges, shared across identities and blendshapes; all edge indices and sums below refer to this fixed set. Figure 5 visualizes the selected facial-graph edges used for blendshape adaptation.

Let $\mathbf{P}(\mathbf{V})$ extract graph nodes using fixed barycentric landmarks and selected mesh vertices, and let $\mathbf{E}(\mathbf{V})_\ell = \mathbf{P}(\mathbf{V})_{a_\ell} - \mathbf{P}(\mathbf{V})_{b_\ell}$ for edge $\ell = (a_\ell, b_\ell)$. We denote the template and personalized edge vectors by $\mathbf{X} = \mathbf{E}(\bar{\mathbf{V}}) \in \mathbb{R}^{L \times 3}$ and $\mathbf{Y} = \mathbf{E}(\mathbf{V}^0) \in \mathbb{R}^{L \times 3}$. The same linear operator applied to template blendshape $\bar{\mathbf{B}}_j^{\mathrm{geo}}$ gives its edge displacements $\boldsymbol{\Delta}_j = \mathbf{E}(\bar{\mathbf{B}}_j^{\mathrm{geo}})$. For each selected edge, we define the blendshape-specific weights by

$$W_{j,\ell k}^2 = \Delta_{j,\ell k}^2 \mathbf{1}[X_{\ell k} Y_{\ell k} \geq 0]. \tag{9}$$

These nonnegative weights, defined through their squared entries, make each blendshape emphasize edge components that it deforms strongly. The component-wise sign mask excludes oppositely oriented contributions.

[1] `https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation/`

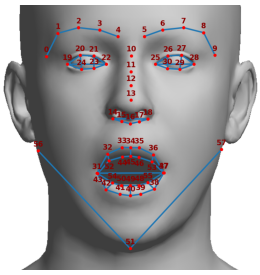


Figure 5: Selected facial-graph edges used by MBA.

We fit three axis-wise scales with a unit-scale prior:

$$\mathbf{s}_j^\star = \arg\min_{\mathbf{s}\in\mathbb{R}^3} \|\mathbf{W}_j \odot [\mathbf{Y} - \mathbf{X}\,\mathrm{Diag}(\mathbf{s})]\|_F^2 + \sum_{k=1}^{3} \lambda_{j,k}(s_k - 1)^2. \tag{10}$$

This objective formalizes the facial-graph fit described in Sec. 3.2. The regularization strength is scaled by identity-independent template energy:

$$A^0_{j,k} = \sum_{\ell=1}^{L} \Delta^2_{j,\ell k} X^2_{\ell k}, \qquad \lambda_{j,k} = \rho A^0_{j,k} + \lambda_{\min}, \tag{11}$$

where $\rho = 0.05$ and $\lambda_{\min} = 10^{-12}$. This energy is computed before applying the identity-dependent sign mask. The solution is separable across axes:

$$s^\star_{j,k} = \frac{\sum_{\ell=1}^{L} W^2_{j,\ell k} X_{\ell k} Y_{\ell k} + \lambda_{j,k}}{\sum_{\ell=1}^{L} W^2_{j,\ell k} X^2_{\ell k} + \lambda_{j,k}}, \qquad k \in \{1,2,3\}. \tag{12}$$

We use unit scale when $A^0_{j,k} < 10^{-12}$. If the sign mask removes all evidence for an axis, the regularizer also returns unit scale. The resulting scales are applied to all vertex displacements of the blendshape:

$$\mathbf{B}^{\mathrm{geo}}_j = \bar{\mathbf{B}}^{\mathrm{geo}}_j \,\mathrm{Diag}\big(\mathrm{sg}[\mathbf{s}^\star_j]\big)\,, \tag{13}$$

where sg denotes stop-gradient. The fit adds no learnable parameters and is evaluated during asset generation. Stopping gradients prevents animation losses from updating the neutral geometry through the scale-fitting path; the neutral geometry remains trainable through the other reconstruction paths.

A.3 Motion and Camera Predictor Architecture

Motion prediction operates independently on the patch tokens of each frame. For the motion outputs, a shared multi-query extractor uses 32 learned expression queries and four learned rigid-pose queries in a two-layer, 16-head Transformer decoder with a feature dimension of 2048. Cross-frame attention is disabled. The resulting query features condition a two-layer, 16-head AdaLN Transformer that performs four iterations of residual parameter refinement. The expression head predicts the 51 ARKit controls used by the avatar, while the pose head produces the configured FLAME head and neck pose parameters.

Camera intrinsics are refined separately from the encoder's frame-local camera token. A four-layer, 16-head AdaLN Transformer performs four refinement iterations over three parameters: normalized log focal length and the normalized horizontal and vertical principal points. These parameters update the tracked camera intrinsics used for rendering. The motion and camera heads share weights across reconstruction and driving frames.

A.4 ADA Reader, Acquisition Attributes, and Class Weighting

Datasets and domain attributes. Table 6 lists the training datasets with their nominal sampling shares and the dataset-level acquisition attributes used to train the domain reader. The attributes describe the supervision available from each acquisition setup rather than the views sampled for

Table 6: Training datasets, nominal sampling shares for each model variant, and dataset-level acquisition attributes used by the domain reader. Sampling shares are obtained by normalizing the effective sampling weights within each variant and do not represent raw dataset sizes. Dashes in the Public and Full columns denote excluded datasets. In the attribute columns, a ✓denotes a positive attribute, a blank denotes a negative attribute, and – denotes unavailable labels excluded from the domain loss. Percentages may not sum to exactly 100 due to rounding.

Dataset	Public (%)	Full (%)	has_back	has_side	facial_obscuration
VFHQ (Xie et al., 2022)	12.50	7.55			
CelebV-Text (Yu et al., 2023)	25.00	15.09			
NeRSemble (Kirschstein et al., 2023)	12.50	7.55		✓	
AVA256 (Martinez et al., 2024)	12.50	7.55	✓	✓	
Cafca-neutral (Buehler et al., 2024)	12.50	7.55	✓	✓	
K-Hairstyle (Kim et al., 2021)	25.00	15.09	✓	✓	✓
Synthetic-Monocular	–	7.55			
Synthetic-Turnaround	–	7.55	–	–	–
MHSD	–	1.89	✓	✓	
StudioCuration	–	22.64	✓	✓	

an individual training example. The Public variant uses the cited public datasets, whereas the Full variant uses all datasets listed in the table.

For each model variant, let p_d denote the nominal sampling share of dataset d under the current variant. The positive-class weight is computed from the unrounded sampling shares of datasets carrying valid attribute labels:

$$w_k = \frac{\sum_{d \in \mathcal{D}_k^-} p_d}{\sum_{d \in \mathcal{D}_k^+} p_d}, \tag{14}$$

where $\mathcal{D}_k^-$ and $\mathcal{D}_k^+$ are the datasets included in the current variant with valid negative and positive labels, respectively. Datasets without attribute labels are excluded from both totals and from the domain-classification loss.

Acquisition-attribute reader. The reader operates on the canonical vertex features $\mathbf{Z} \in \mathbb{R}^{|\mathbf{V}^0| \times 1024}$. After adding a fixed vertex-index encoding, 64 learned 1024-dimensional summary tokens aggregate the vertex features through two interleaved cross-attention and self-attention blocks, each using four attention heads and a dropout rate of 0.1. The resulting summary tokens are flattened and passed to a two-layer classifier: a linear projection from 64×1024 dimensions to 256 dimensions, followed by LayerNorm and SiLU, and a final linear projection to the three attribute logits. Each active iteration performs two updates: first, $\mathcal{L}_{\text{cls}}$ updates only D_ϕ using detached canonical features; second, with D_ϕ fixed, $\mathcal{L}_{\text{conf}}$ updates the upstream feature network. Figure 6 provides a qualitative view of this effect. Without ADA, canonical features are visibly separated by acquisition regime, whereas ADA produces greater overlap across regimes and datasets. Because PCA is fitted independently for the two panels, this comparison concerns cluster separability rather than absolute coordinates.

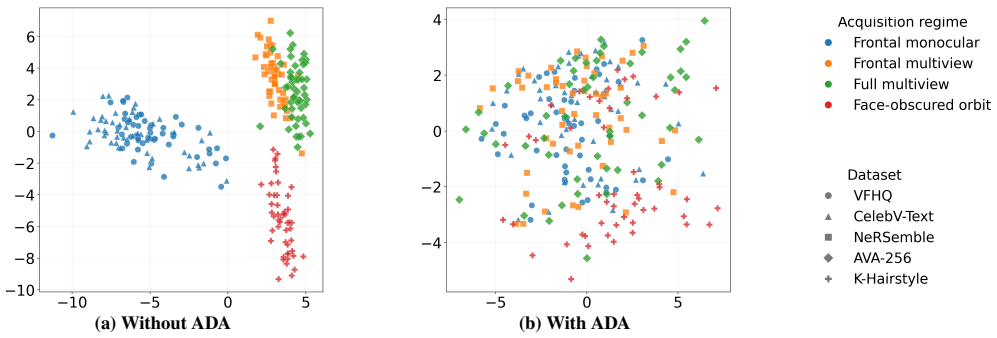


Figure 6: PCA projections of canonical vertex features without and with ADA. Colors indicate acquisition regimes and markers indicate datasets. Because PCA is fitted independently for each panel, axes and coordinates are not directly comparable; the plots provide a qualitative view of acquisition-regime separability.

Table 7: Configured training-loss weights. DINOv2, SAM, and domain confusion reach the listed weights according to the schedules below.

Loss	Weight	Loss	Weight
RGB ℓ_1	1.0	Gaussian scaling	0.1
Alpha ℓ_1	0.1	Gaussian anisotropy (ratio > 10)	0.1
$1 - \text{SSIM}$	0.2	Opacity floor	10^{-3}
DINOv2 feature distance	1.0	Vertex-offset Laplacian	2.0
SAM feature distance	1.0	Landmark SmoothL1	1.0
Neutral ArcFace identity	0.1	Domain classification	1.0
Expression ℓ_1	0.002	Domain confusion	0.002
Expression bound penalty	100.0		

A.5 Implementation, Data, Losses, and Training Procedure

Model scale. The full model contains approximately 878 million trainable parameters. The largest components are the DINOv2-L/14 image backbone (304M parameters) within the complete image encoder (355M), the pose-token extractor (134M), the camera and pose heads (120M and 109M, respectively), and the shared Transformer trunk (76M). The remaining parameters belong to the asset and auxiliary prediction heads.

Optimization and input sampling. We train for 700,000 steps on sixteen 96-GB H20 GPUs for approximately two weeks. We use AdamW (Loshchilov & Hutter, 2019) with a weight decay of 0.01 and default momentum coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$. All parameter groups use a learning rate of 10^{-5}. It is linearly warmed from 10^{-6} during the first 5,000 steps and remains at 10^{-5} thereafter. The per-rank batch size is one, giving a global batch size of 16 with one rank per GPU. Each sample contains eight frames: one to four are selected for avatar reconstruction, and the remainder serve as driving targets. Input images and supervised renderings use a resolution of 518×518, corresponding to a 37×37 patch grid for the DINOv2-L/14 encoder.

K-Hairstyle preprocessing and supervision. K-Hairstyle images contain blurred facial regions. We use VGGHeads (Kupyn et al., 2024) to obtain a coarse initial alignment. During training, the alignment is progressively refined through image reconstruction losses evaluated outside the blurred regions. Blurred facial regions are masked out and do not contribute to these losses.

Non-public datasets. Our synthetic monocular dataset contains 5,336 clips. The synthetic turnaround dataset contains 4,496 model-generated monocular 360-degree turnaround clips. In each clip, the subject completes a full rotation; the viewpoints are observed sequentially within a monocular video rather than captured simultaneously by a synchronized multi-camera system. MHSD comprises rendered MetaHuman[2] data with 142 identities and 32 camera viewpoints. StudioCuration comprises studio-captured real-person data with 1,021 identities and 45 camera viewpoints. These raw dataset sizes are distinct from the nominal sampling shares in Table 6, which are derived from effective sampling weights and do not represent numbers of clips, identities, or frames.

Loss weights. Table 7 lists the individual configured weights within the grouped objective in Equation 6; these entries are not additional multipliers on top of duplicate group weights. The expression sparsity term is

$$\mathcal{L}_{\text{expr}} = \text{mean}_{b,t,j} \, |e_{b,t,j}|, \tag{15}$$

where b, t, and j index samples, frames, and expression channels. The coefficients are the model's refined expression predictions, rather than external tracking targets. The bound penalty applies $\max(0, -e)^2 + \max(0, e - 1)^2$ to each coefficient. Landmark reprojection uses SmoothL1 with transition parameter 0.02.

Activation schedules. The pose head initially refines the tracked head pose. From step 100,000 to step 200,000, the contribution of the tracked pose is linearly reduced to zero, transitioning the model from tracking-based refinement to direct head-pose estimation. GAB corrections are enabled at step

[2] https://dev.epicgames.com/documentation/metahuman

300,000 with the sample–frame dropout described above. ADA is disabled before step 350,000. From that step, the reader uses classification weight 1.0, and the confusion weight increases linearly over 10,000 steps to 0.002:

$$\lambda_{\text{conf}}(n) = 0.002 \, \min\left(1, \max\left(0, \frac{n - 350{,}000}{10{,}000}\right)\right), \tag{16}$$

where n is the global training step. Both updates are performed at every active iteration. The reader uses positive-class weights computed from the sampling shares in Table 6. These weights apply only to the classification branch; the confusion objective uses unit-weight binary cross-entropy. DINOv2 and SAM losses begin at step 400,000 and reach their target weights after a 30,000-step warmup.

NIA frame selection and alignment. Near-frontal selection uses a yaw threshold of 45° and requires a front-view cosine greater than 0.5 (between the head-forward direction and the head-to-camera vector). Projected facial landmarks are converted to five alignment points to construct the 112×112 crops used by the frozen ArcFace encoder for the reference and zero-expression rendering.

A.6 Evaluation and Baseline Protocols

A.6.1 Evaluation Metrics and Evaluator Details

Identity evaluation. CSIM_{Arc} uses an InsightFace[3] `w600k_r50` checkpoint, which is distinct from the MS1MV3 ArcFace R50 encoder used by the NIA training loss. CSIM_{Ada} uses a WebFace12M-pretrained AdaFace (Kim et al., 2022) IR-101 evaluator that is not used during training. AVA256 reports CSIM_{Arc} only.

Reference-relative expression error. The pretrained 3DMM expression estimator (Deng et al., 2019b) can introduce an appearance-dependent offset in cross-identity AED. We therefore additionally report

$$\text{Ref-AED}_t = \frac{1}{d} \left\| [e(Y_t) - e(Y_r)] - [e(B_t) - e(B_r)] \right\|_1, \tag{17}$$

where Y and B are the generated and driving images, $e(\cdot) \in \mathbb{R}^d$ is the frozen expression estimator, and r is the first valid frame of each sequence–camera pair. This reference subtraction mitigates appearance-related offsets.

A.6.2 Full-Head Baseline Implementation Details

Shared four-view head-mask registration. After avatar construction, every method uses head-region masks from calibrated cameras 24, 26, 55, and 28 to fit one shared world-space similarity transform per identity (scale, rotation, and translation). VGGHeads (Kupyn et al., 2024) detects a head box in each registration image; its center is retained while its width and height are expanded by $1.35\times$. The fitting signal is the binary portrait mask restricted to this region. Registration runs for 200 steps at 256×256, with intrinsics scaled to the fitting raster. The same transform is used for registration and final rendering, and the shared evaluation crops remain unchanged. The current 1,180-view aggregate includes the 80 registration views. Asset construction uses one frontal RGB image, followed by registration using masks from four additional calibrated views.

SphereHead+PTI. We use the released 25k SphereHead (Li et al., 2024) checkpoint and optimize each camera-25 portrait for 500 W-inversion steps followed by 350 generator-tuning steps (Roich et al., 2022). These asset-construction stages use no target-view RGB or masks. The subsequent shared registration optimizes the inverse similarity transform of the calibrated camera trajectory through the generator's differentiable opacity output, while keeping the fitted generator fixed.

FaceLift. We use the released FaceLift (Lyu et al., 2025) checkpoints for its multiview diffusion model and GS-LRM-based Gaussian reconstructor (Zhang et al., 2024), with the official background removal and face normalization. We fix the random seed to 4, the classifier-free guidance scale to 3, and use 50 diffusion steps. The six generated canonical views are passed to GS-LRM once to

[3] `https://github.com/deepinsight/insightface`

reconstruct a Gaussian head. The shared registration transforms Gaussian centers and covariances; final rendering applies the equivalent inverse transform to the calibrated camera trajectory while keeping the reconstructed asset fixed. Scale is initialized from the canonical focal gauge; scale, rotation, and translation are all optimized during registration.

FlexAvatar and TaoHeadVerse. FlexAvatar retains its native Pixel3DMM (Giebenhain et al., 2026) input preprocessing and frozen FLEX-2 asset. TaoHeadVerse constructs its asset from the aligned frontal input. Both use the shared registration views, head-mask regions, optimization budget, evaluation crops, and metric implementation described above.

A.7 Additional Experiments and Visualizations

A.7.1 Additional Qualitative Comparisons

Figures 7 and 8 extend the main-paper animation comparisons with the expanded VFHQ qualitative comparison and additional AVA256 examples, respectively. Across these examples, our method better preserves source-specific appearance while following the driving expression, including challenging mouth and eye motion. Figure 9 provides additional strict-rear full-head reconstruction examples, where our predictions retain more coherent rear-head appearance and structure.



Figure 7: Expanded VFHQ cross-identity reenactment comparison including GAGAvatar.

A.7.2 Additional Ablation Results

Figure 10 shows that $A3$ preserves source identity more faithfully than both $A1$ and the direct identity-supervision variant $A1^{\mathrm{ID}}$ under cross-identity driving. Figure 11 shows that $A5$ produces more complete and coherent back-of-head and strict-rear reconstructions than $A4$. Figure 12 shows that $A3$ captures richer expression-dependent appearance dynamics than $A2$, such as the forehead wrinkles produced during eyebrow raising.

Table 8 further breaks down the aggregate MBA results reported in the main paper by facial region and expression intensity. MBA improves face and mouth AKD in every intensity bin, with the largest and most consistent gain in the mouth region. The overall mouth AKD decreases by 5.90%, whereas eye AKD increases slightly by 1.38%, mainly in the mid- and low-intensity bins. Figure 13 complements this analysis qualitatively, showing reduced eyelid distortion for closed-eye driving and fewer lip–teeth artifacts for open-mouth driving.

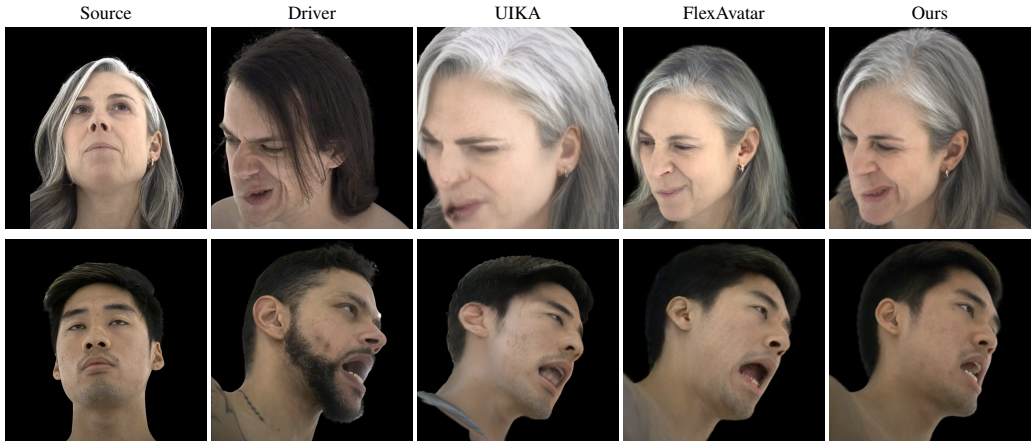


Figure 8: Additional single-source-image cross-identity reenactment results on AVA256. Each row uses the same source image, driving frame, and calibrated target camera; all shown methods run without fitting.

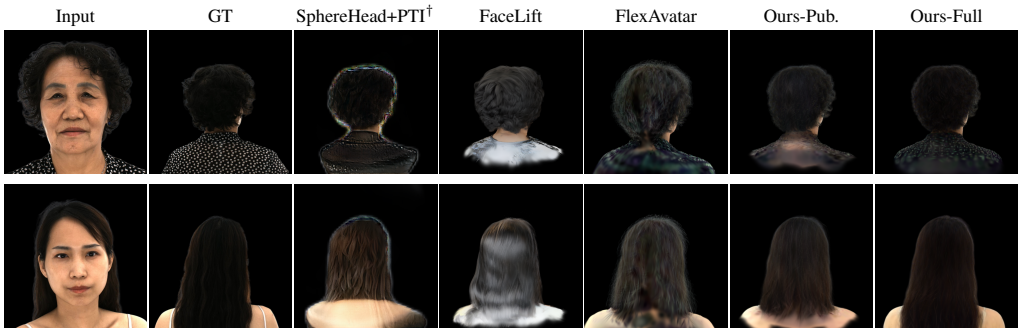


Figure 9: Additional strict-rear RenderMe-360 reconstructions with shared four-view head-mask registration. † additionally uses per-identity optimization.

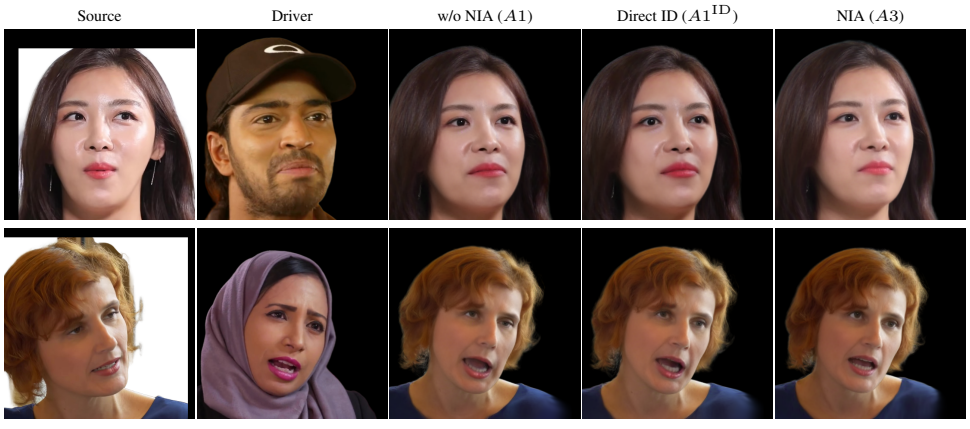


Figure 10: Identity-supervision ablation on VFHQ cross-identity reenactment. All variants use the same source and driving frames.

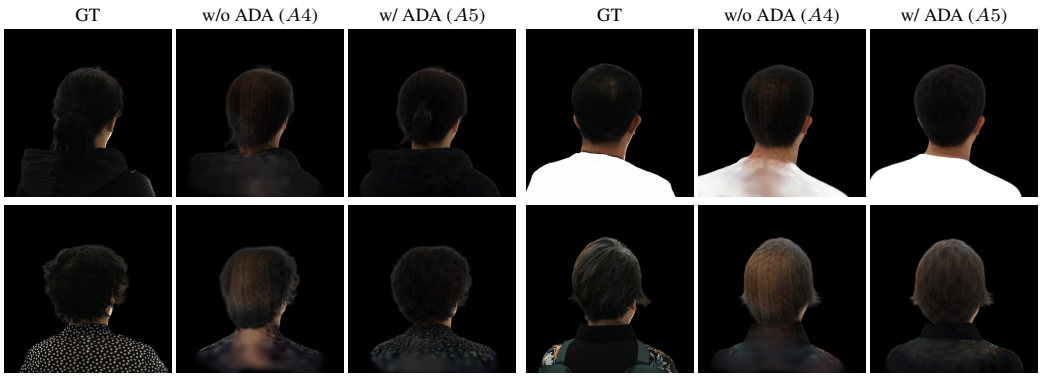


Figure 11: ADA ablation on strict-rear reconstruction. Each triplet shows GT, $A4$/Base without ADA, and $A5$/Full with ADA at the same target view.

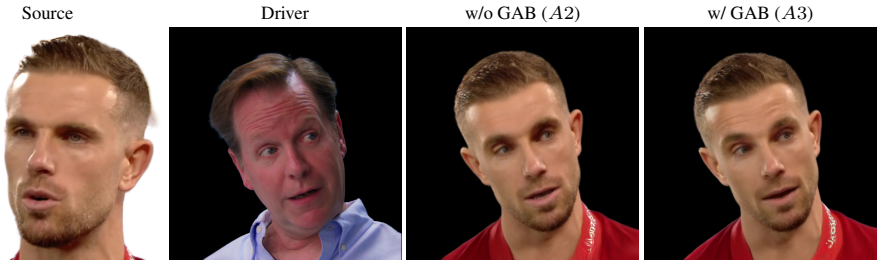
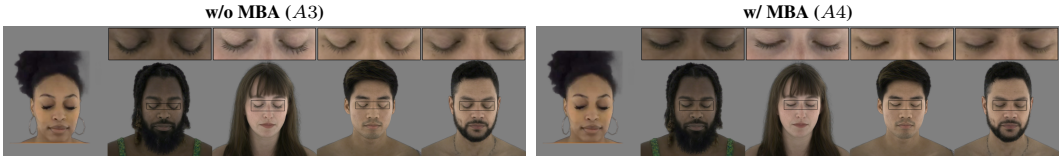


Figure 12: Qualitative ablation of Gaussian Appearance Bases (GAB) on VFHQ cross-identity reenactment. Both variants use the same source and driver.

Table 8: Region-wise AKD for the MBA ablation on VFHQ self-reenactment, stratified by expression intensity. $\Delta = A4 - A3$; negative values indicate improvement. Overall results pool all frames.

	Face AKD				Mouth AKD				Eye AKD			
Expression intensity	A3	A4	Δ	Rel.	A3	A4	Δ	Rel.	A3	A4	Δ	Rel.
High (75–100%)	3.5102	3.3952	−0.115	−3.28%	3.8095	3.5281	−0.281	−7.39%	3.2123	3.1455	−0.067	−2.08%
Mid (25–75%)	3.2040	3.1361	−0.068	−2.12%	3.4212	3.2151	−0.206	−6.02%	2.9535	3.0080	+0.054	+1.84%
Low (0–25%)	3.5588	3.5318	−0.027	−0.76%	3.7545	3.5980	−0.157	−4.17%	3.4092	3.5400	+0.131	+3.84%
Overall	**3.3692**	**3.2998**	**−0.069**	**−2.06%**	**3.6016**	**3.3891**	**−0.213**	**−5.90%**	**3.1321**	**3.1754**	**+0.043**	**+1.38%**



(a) Closed-eye driving frame.



(b) Open-mouth driving frame.

Figure 13: Qualitative comparison of MBA under AVA256 cross-reenactment. Each row uses the same driving portrait; $A3$ disables MBA and $A4$ enables it. Each variant uses its own jointly trained motion estimator, so this comparison evaluates the complete system rather than geometry adaptation alone. $A4$ alleviates eyelid distortion in (a) and lip–teeth artifacts in (b).