VIDEO DIFFUSION AVATARS: TAMING VIDEO DIFFUSION MODELS INTO CONTROLLABLE 3D AVATARS

Anonymous authors
Paper under double-blind review



Figure 1: **Video Diffusion Avatars (VDAs)**: given a single image, VDAs produce photo-realistic renderings with independent control over camera, upper-body pose, and facial expression, inheriting hair dynamics, shadows, reflections and refractions from its VDM prior.

ABSTRACT

Video diffusion models (VDMs) trained on internet-scale data encode a rich prior on appearance, motion, and physics that is difficult to replicate with explicit 3D rendering and costly 3D data capture. We introduce Video Diffusion Avatars (VDAs), a method that creates photo-realistic, multi-view consistent avatars from a single image by fine-tuning a VDM for precise control over camera viewpoint, facial expression, and upper-body pose. Departing from 3D morphable models (3DMMs), VDAs use a learned expression space that lifts the expressiveness ceiling of 3DMMs, *e.g.* for the tongue and mouth interior. However, learned expression codes exhibit heavy viewpoint bias, as our DISTRACT benchmark reveals. We show that this bias can be removed through disentanglement, enabling their use as driving signals for multi-view consistent avatars. Camera and upper-body pose are controlled jointly via proxy-mesh renderings, giving independent control over expression, pose, and viewpoint. A camera-triplet training strategy makes the approach efficient: our models train for 34 hours on a single A100 using only public multi-view data. VDAs outperform state-of-the-art single-image 3D avatars on novel-view reenactment, reducing face LPIPS by 29% and improving perceptual quality (JOD) by 66% on BecomingLit. A static 3DGS reconstruction probe further confirms that our generations are the most multi-view consistent. Built on a video prior, VDAs produce effects beyond the reach of explicit renderers, including physically plausible hair dynamics, shadows, reflections, and refractions.

1 INTRODUCTION

Virtual avatars promise to reshape communication, gaming, and visual effects. Many applications demand exact control over facial expression, upper-body pose, and camera viewpoint, and the smallest artifacts read as uncanny. In-the-wild use raises the bar further: accessories, complex lighting, and hair dynamics must be handled convincingly from a single reference image.

The dominant paradigm builds explicit volumetric avatars, typically 3DGS (Kerbl et al., 2023) or voxels (Lombardi et al., 2021; Cao et al., 2022), with expressions driven by 3DMMs (Blanz & Vetter,

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.

1999; Li et al., 2017). This approach faces two fundamental limits. First, no large-scale 3D training data exists, and, additionally, effects such as shadows, reflections, refractions, and physically plausible hair dynamics are hard to render explicitly. Second, 3DMM-driven expression control suffers fidelity loss from 2D-to-3D registration. Alternatively, learned expression spaces (Drobyshev et al., 2022; Cheng et al., 2025; Zhao et al., 2025) avoid this ceiling but, as our proposed DISTRACT benchmark reveals, carry a heavy viewpoint bias that must be addressed before they can drive multi-view consistent avatars.

Video diffusion models (VDMs), trained on internet-scale data, offer a rich prior on 3D geometry, photo-realistic appearance, motion, and physics. However, existing VDM-based avatar methods either build 2D avatars without camera control (Tian et al., 2024; Xie et al., 2024; Guo et al., 2024; Zhao et al., 2025), or use the VDM only to synthesize multi-view images that are then distilled into a 3DGS avatar (Taubner et al., 2025b;a), trading off rendering quality against speed.

We introduce Video Diffusion Avatars (VDAs), and depart from the two most common building blocks for virtual avatars: 3DMM-control and 3DGS rendering. Following DH-STAR (Zielonka et al., 2026), we define an avatar as a model that, conditioned on a single identity image, renders that person under any requested facial expression, upper-body pose, and camera viewpoint. In our approach the VDM *is* the avatar: no 3D asset is ever built, and the prior is fully preserved. VDAs rests on three cornerstones. *Expression control:* we remove the viewpoint bias from learned expression encoders with an explicit disentangler, obtaining higher expression fidelity than FLAME-driven control. *Camera and pose control:* a proxy mesh combining a FLAME head with a SAM-3D-Body (Yang et al., 2026b) torso jointly encodes camera viewpoint and upper-body pose, while exposing separate control for both. *Training:* camera triplets from curated multi-view data (Kirschstein et al., 2023; Martinez et al., 2024), with reference, driver and target from different cameras, teach multi-view consistency in 34 hours on a single A100.

VDAs outperform all state-of-the-art single-image 3D avatars on BecomingLit (Schmidt et al., 2025), Multiface (Wuu et al., 2022), and VFHQ-test (Xie et al., 2022), improving LPIPS by 29% and JOD by 66% over the strongest baseline. A 3DGS optimization against our generations outperforms prior diffusion-based avatars (Taubner et al., 2025b;a), camera-controlled VDMs (Lyu et al., 2026), and 3DGS-based avatars (Kirschstein et al., 2026; Wu et al., 2026). Our contributions are:

(a) VDAs replace the explicit 3D asset with a camera- and expression-controlled VDM, inheriting shadows, reflections, refractions, and hair dynamics from its prior.
(b) We expose viewpoint bias in learned expression encoders with the DISTRACT benchmark, debias it, and obtain higher expression fidelity than FLAME-driven control.
(c) Camera and pose conditioning via a Plücker-colored proxy mesh, that separates control over camera viewpoint from neck pose.
(d) Affordable training strategy on public multi-view data using camera triplets in 34 hours on a single A100, which preserves performance on challenging in-the-wild inputs.

2 Related Work

Facial expression spaces and 3D morphable models. 3DMMs (Li et al., 2017; Blanz & Vetter, 1999) are the dominant expression parameterization, but frequently miss microexpressions and are lacking mouth interior, limiting the fidelity of any system built on them. However, feed-forward 3DMM regressors such as DECA (Feng et al., 2021), EMOCA (Danecek et al., 2022), and robust tracking (Giebenhain et al., 2025) are widely available and make 3DMM-based control practical for in-the-wild applications. Analogous body models include SMPL (Loper et al., 2015) and SAM-3D-Body (Yang et al., 2026b). Implicit neural head representations (Giebenhain et al., 2023; 2024a; Zheng et al., 2022) go beyond the linear 3DMM expression space. Alternatively, several methods learn an expression latent directly from 2D data: MegaPortraits (Drobyshev et al., 2022) relies on cycle-consistency, while WAN-Animate (Cheng et al., 2025) and X-NeMo (Zhao et al., 2025) learn expression representations end-to-end for video-driven animation. In Section 3.1 we compare the qualities of such learned expression spaces and reveal a heavy viewpoint bias that must be addressed for multi-view consistent avatars.

Avatars. Most recent avatar systems use 3DGS (Kerbl et al., 2023), either for generalizable single-image creation (Kirschstein et al., 2026; He et al., 2025; Wu et al., 2026) or for per-identity optimization from multi-view captures (Xu et al., 2024; Qian et al., 2024; Giebenhain et al., 2024b). A separate line uses video diffusion to generate multi-view images that are then distilled into a 3DGS avatar: CAP4D (Taubner et al., 2025b) conditions an image-diffusion model (MMDM) on FLAME

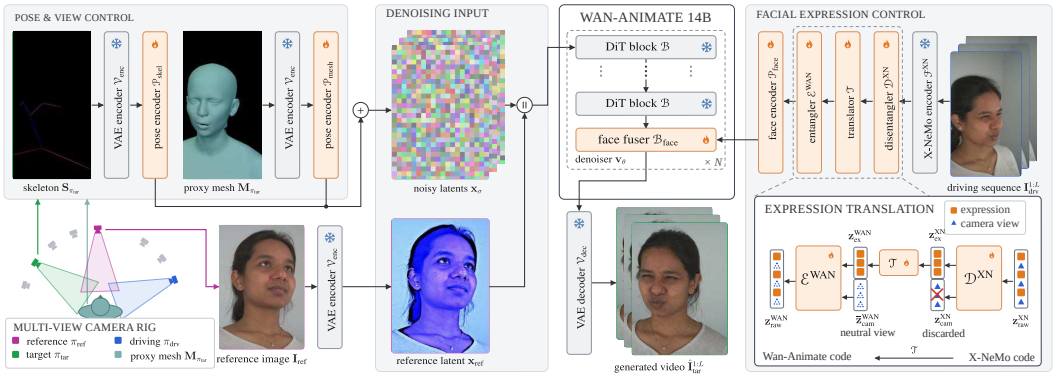

Figure 2: **Video Diffusion Avatar architecture**: Camera, upper-body pose, and facial expression are injected as independent control signals into WAN-Animate-14B. Disentangled X-NeMo expression codes are translated into WAN-Animate's native conditioning format with a viewpoint-neutral camera code, while a proxy-mesh rendering encodes camera and pose. Camera triplets from multi-view data teach multi-view consistency.

renderings, and MVP4D (Taubner et al., 2025a) uses a video-diffusion model instead. Both trade the VDM's photorealistic prior for the rendering speed of 3DGS. We refer to DH-STAR (Zielonka et al., 2026) for a broader survey.

Video diffusion models for avatars. VDMs such as WAN-Animate-14B (Wan Team et al., 2025; Cheng et al., 2025) synthesize temporally coherent, photorealistic human video at unprecedented quality. Diffusion-based portrait-animation methods (Guo et al., 2024; Tian et al., 2024; Xie et al., 2024; Zhao et al., 2025; Li et al., 2025) condition these models on signals ranging from sparse 2D keypoints to dense 3DMM-derived renderings, but typically produce 2D avatars without explicit camera control. FaceCam (Lyu et al., 2026) adds camera control to a VDM, but lacks expression control. DeX-Portrait (Wu et al., 2025) controls neck pose and expression, but not the camera. FactorPortrait (Tang et al., 2025) is concurrent work that controls expression, head pose, and viewpoint in a VDM. However, it trains on proprietary data using limited-quality 3DGS renderings as pseudo ground truth, and no code or model is available.

3 Video Diffusion Avatars (VDAs)

Following DH-STAR (Zielonka et al., 2026), an avatar is a function $\mathcal{A}$ that renders a person, given an identity specification, under control signals for expression, body pose, and camera viewpoint. Our avatar takes a single reference image $\mathbf{I}_{\text{ref}}$ as identity input and renders a video $\hat{\mathbf{I}}_{\text{tar}}^{1:L}$ conditioned on per-frame disentangled expression codes $\mathbf{z}_{\text{ex}}^{1:L}$ (Section 3.1), and a proxy mesh $\mathbf{M}_{\pi_{\text{tar}}}$ and skeleton $\mathbf{S}_{\pi_{\text{tar}}}$ rendered from a target camera π_{tar}, describing camera viewpoint and upper-body pose (Section 3.3):

$$\hat{\mathbf{I}}_{\text{tar}}^{1:L} = \mathcal{A}\big(\mathbf{z}_{\text{ex}}^{1:L},\, \mathbf{M}_{\pi_{\text{tar}}}^{1:L},\, \mathbf{S}_{\pi_{\text{tar}}}^{1:L} \mid \mathbf{I}_{\text{ref}}\big). \tag{1}$$

Figure 2 illustrates the architecture and how each signal enters our fine-tuned WAN-Animate-14B backbone: expression through dedicated face-fuser blocks (Section 3.2), camera and pose through the rendered proxy mesh and skeleton (Section 3.3). Unlike a 3D avatar, which builds an explicit 3D representation and renders it, $\mathcal{A}$ synthesizes the video directly, so multi-view consistency has to be learned rather than built in. Section 3.4 shows how camera triplets from public multi-view data enable extremely efficient training (34 hours on a single GPU), while preserving the video prior.

3.1 Expression Control and the DISTRACT Benchmark

3DMMs such as FLAME limit the detail and range of representable expressions, motivating the use of learned expression encoders. A good encoder must preserve expression while discarding identity and camera viewpoint. We introduce the **DISTRACT** benchmark, which directly measures how strongly an encoder's code is *distracted* by these confounders.

DISTRACT evaluates two invariance axes on NeRSemble (Kirschstein et al., 2023): *viewpoint invariance* (same identity and timestep, different camera) and *identity invariance* (same expression across different identities, verified by manual inspection). Full details are in Appendix B.

Most encoders are heavily viewpoint-biased (Table 1 and Figure 3): X-NeMo reaches only 18%



Figure 3: Nearest-neighbour retrieval in X-NeMo's expression space: viewpoint dominates, yet expression information is retained.

Table 1: DISTRACT results (%) on raw codes and after a 2-layer MLP probe. Err-X: how often distractor type X outranks the true positive.

	Encoder	Latents R-1↑	Latents Err A/B/C/D↓	MLP Probe R-1↑	MLP Probe Err A/B/C/D↓
View Inv.	DECA (Feng et al., 2021)	61	33/18/18/11	85	12/06/01/00
	PD-FGC (Wang et al., 2023)	08	90/42/79/43	90	08/03/00/01
	WAN-Animate (Cheng et al., 2025)	21	72/34/59/33	94	05/01/01/00
	X-NeMo (Zhao et al., 2025)	18	78/35/70/34	98	02/00/00/00
Identity Inv.	DECA (Feng et al., 2021)	34	51/37/25/24	32	58/51/24/16
	PD-FGC (Wang et al., 2023)	07	91/48/81/43	46	44/40/10/18
	WAN-Animate (Cheng et al., 2025)	14	74/32/62/26	66	22/23/10/06
	X-NeMo (Zhao et al., 2025)	14	78/37/68/37	76	18/17/04/05

Rank-1 on viewpoint invariance. Yet a small MLP probe on the *same frozen* latents raises this to 98%, confirming that expression information survives but is entangled with viewpoint, thus motivating an explicit disentanglement module.

Given an image $\mathbf{I}$, a learned expression encoder produces an entangled code $\mathbf{z}_{\text{raw}} \in \mathbb{R}^d$. Our *disentangler* $\mathcal{D}$ splits it into an expression latent $\mathbf{z}_{\text{ex}}$ and a camera latent $\mathbf{z}_{\text{cam}}$, and an *entangler* $\mathcal{E}$ reconstructs the original:

$$\mathcal{D}(\mathbf{z}_{\text{raw}}) = (\mathbf{z}_{\text{ex}}, \mathbf{z}_{\text{cam}}), \qquad \mathcal{E}(\mathbf{z}_{\text{ex}}, \mathbf{z}_{\text{cam}}) = \hat{\mathbf{z}}_{\text{raw}} \approx \mathbf{z}_{\text{raw}}, \tag{2}$$

both implemented as dedicated MLP heads. The architecture is encoder-agnostic; we train separate instances for X-NeMo and WAN-Animate codes. Training uses a reconstruction loss, a supervised contrastive loss (Khosla et al., 2020) on $\mathbf{z}_{\text{ex}}$, and a cycle-consistency loss; details are in Appendix G.

3.2 Latent Expression Injection

We drive VDAs using X-NeMo expression codes, which outperformed the WAN-Animate codes on DISTRACT after the MLP probe. Concretely, we replace the view-biased codes $\mathbf{z}_{\text{raw}}^{\text{WAN}}$ with our disentangled counterpart $\mathbf{z}_{\text{ex}}^{\text{XN}}$. Since the two latent spaces are unrelated, and we aim to minimally disturb the pretrained weight space, we bridge this gap using latent translation, depicted in Figure 2. An MLP *translator* $\mathcal{T}$ maps the disentangled X-NeMo expression latent into WAN-Animate's disentangled expression space, and the *entangler* $\mathcal{E}^{\text{WAN}}$ combines it with a viewpoint-neutral camera code $\bar{\mathbf{z}}_{\text{cam}}^{\text{WAN}}$, the mean of the camera latents of $\mathcal{D}^{\text{WAN}}$ over the training set:

$$\mathbf{z}_{\text{ex}}^{\text{WAN}} = \mathcal{T}(\mathbf{z}_{\text{ex}}^{\text{XN}}), \qquad \mathbf{z}_{\text{raw}}^{\text{WAN}} = \mathcal{E}^{\text{WAN}}(\mathbf{z}_{\text{ex}}^{\text{WAN}}, \bar{\mathbf{z}}_{\text{cam}}^{\text{WAN}}). \tag{3}$$

The result is a code in WAN-Animate's native format, now stripped of viewpoint bias. WAN-Animate's face encoder $\mathcal{P}_{\text{face}}$ maps it to face tokens, which enter the denoiser $\mathbf{v}_\theta$ through face-fuser blocks $\mathcal{B}_{\text{face}}$ interleaved with its DiT blocks $\mathcal{B}$. Since the DiT has not seen view-neutralized codes during pretraining, and our disentanglement is not trained with the DiT in-the-loop, we fine-tune $\mathcal{T}$, $\mathcal{E}^{\text{WAN}}$, $\mathcal{D}^{\text{XN}}$, $\mathcal{P}_{\text{face}}$, $\mathcal{B}_{\text{face}}$, and the pose encoders (Section 3.3), while we keep $\mathcal{B}$ frozen to preserve as much prior as possible.

3.3 Pose Conditioning

WAN-Animate relies on 2D skeleton prediction from ViTPose (Xu et al., 2022) for pose control, which cannot be transferred to novel viewpoints for novel-view synthesis. For 3D-consistent skeleton control, we replace the 2D detections with skeletons derived from SAM-3D-Body (Yang et al., 2026b), a single-image 3D body estimator. To reduce the noise of SAM-3D-Body's per-frame pose estimates, we introduce *pelvis-anchored kinematic fitting* (Appendix F.1), which exploits the seated capture setup and assumes a constant pelvis position to remove jitter and drift. Importantly, we retain WAN-Animate's skeleton conditioning format for better prior preservation. We note that at inference, the seated assumption can be relaxed to plain temporal smoothing.

However, we find that skeleton conditioning alone is too ambiguous for precise camera control and that VDAs benefits from a more precise and correlated signal. We therefore render a proxy mesh $\mathbf{M}$ combining a FLAME head with a SAM-3D-Body torso, that jointly encodes camera viewpoint and body/neck pose. We color each pixel by the direction of its camera ray, *i.e.* the directional part of its Plücker coordinates, blended with Lambertian shading. Crucially, head pixels express this direction in the head's local frame and torso pixels in a pelvis-aligned frame (Appendix F.2). The

Table 2: Quantitative results (BecomingLit/Multiface per cell). MMDM is CAP4D's diffusion component. Cross-reenactment on BecomingLit only; colours rank within each dataset.

	Self-reenactment ($n_{\text{BecomingLit}}$=25 / $n_{\text{Multiface}}$=22)						Cross-reenact. (n=13)		
Method	PSNR ↑	LPIPS ↓	JOD ↑	ID ↑	Expr ↑	AKD ↓	ID ↑	Expr ↑	AKD ↓
LAM	18.95/19.83	0.408/0.415	2.00/1.92	0.295/0.334	0.670/0.528	0.103/0.114	0.314	0.623	0.091
FlexAvatar	20.55/19.45	0.327/0.364	2.28/2.33	0.624/0.645	0.890/0.850	0.086/0.092	0.570	0.700	0.130
UIKA	20.37/20.76	0.327/0.332	2.34/2.41	0.532/0.571	0.865/0.805	0.076/0.072	0.493	0.775	0.084
MMDM	20.85/21.02	0.283/0.270	1.56/1.35	0.555/0.608	0.853/0.840	0.058/0.063	0.522	0.780	0.069
MVP4D	20.16/21.64	0.282/0.273	1.85/1.64	0.490/0.531	0.831/0.828	0.063/0.060	0.480	0.755	0.074
CAP4D	20.73/21.40	0.322/0.303	2.40/2.49	0.562/0.608	0.852/0.826	0.059/0.059	0.529	0.770	0.070
Ours	22.43/22.56	0.202/0.201	3.99/3.49	0.699/0.719	0.920/0.862	0.055/0.057	0.623	0.808	0.066

color contrast between head and torso thus directly encodes neck pose, keeping camera viewpoint and neck rotation separable. Rendering from the target camera π_{tar} yields $\mathbf{S}_{\pi_{\text{tar}}}$ and $\mathbf{M}_{\pi_{\text{tar}}}$, which are encoded by $\mathcal{P}_{\text{skel}}$ and $\mathcal{P}_{\text{mesh}}$ (the latter zero-initialized) and added to the noisy video latents.

3.4 Training with Camera Triplets

We train VDAs on public multi-view video datasets of human faces. Each training sample is a *camera triplet* $(\pi_{\text{ref}}, \pi_{\text{drv}}, \pi_{\text{tar}})$ of three *different* cameras of the same multi-view video clip. The reference image $\mathbf{I}_{\text{ref}}$ is seen from π_{ref}, the driving clip $\mathbf{I}_{\text{drv}}^{1:L}$ from π_{drv}, and the target clip $\mathbf{I}_{\text{tar}}^{1:L}$ from π_{tar}. The expression codes are extracted from the driving video with the frozen X-NeMo image encoder $\mathcal{F}^{\text{XN}}$ and our disentangler:

$$\mathbf{z}_{\text{ex}}^{1:L} = \mathcal{D}^{\text{XN}}\big(\mathcal{F}^{\text{XN}}(\mathbf{I}_{\text{drv}}^{1:L})\big). \tag{4}$$

Similarly, the proxy mesh is estimated from $\mathbf{I}_{\text{drv}}^{1:L}$, with the exception of FLAME's identity parameters which come from π_{ref}, and rendered from π_{tar} to obtain $\mathbf{M}_{\pi_{\text{tar}}}^{1:L}$ and $\mathbf{S}_{\pi_{\text{tar}}}^{1:L}$. Because the π_{drv} and π_{tar} differ, viewpoint information inside $\mathbf{z}_{\text{ex}}$ is actively misleading and training suppresses it (Section 4.2).

We fine-tune with the rectified-flow matching loss. Given the target latent $\mathbf{x}_0 = \mathcal{V}_{\text{enc}}(\mathbf{I}_{\text{tar}}^{1:L})$ and Gaussian noise $\boldsymbol{\epsilon}$, the noised latent is $\mathbf{x}_\sigma = (1-\sigma)\,\mathbf{x}_0 + \sigma\,\boldsymbol{\epsilon}$, and the denoiser $\mathbf{v}_\theta$ predicts the velocity $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{x}_0$:

$$\mathcal{L} = \mathbb{E}\big[w(\sigma)\,\big\|\mathbf{v}_\theta\big(\mathbf{x}_\sigma,\,\sigma;\,\mathbf{z}_{\text{ex}}^{1:L},\,\mathbf{M}_{\pi_{\text{tar}}}^{1:L},\,\mathbf{S}_{\pi_{\text{tar}}}^{1:L},\,\mathbf{x}_{\text{ref}}\big) - \mathbf{v}\big\|^2\big], \tag{5}$$

where $\mathbf{x}_{\text{ref}} = \mathcal{V}_{\text{enc}}(\mathbf{I}_{\text{ref}})$ is concatenated to the noisy latents and excluded from the loss, and $w(\sigma)$ is a bell-shaped weighting peaking at $\sigma = 0.5$. At inference, $\mathcal{A}$ integrates $\mathbf{v}_\theta$ from $\sigma = 1$ to 0 and decodes the result with the VAE decoder $\mathcal{V}_{\text{dec}}$. No auxiliary losses are used. For more training details we refer to Appendix E.

4 Experiments

Implementation details. We fine-tune WAN-Animate-14B on curated multi-view data from NeRSemble (Kirschstein et al., 2023) and AVA-256 (Martinez et al., 2024). Training takes **34 hours on a single A100 GPU**. The proxy-mesh head comes from monocular pixel3DMM (Giebenhain et al., 2025) FLAME tracking; the body and neck from SAM-3D (Yang et al., 2026b). At inference we perform 20 denoising steps. Full hyperparameters are in Appendix E.

4.1 Evaluation Setup

Baselines. We compare against the single-image 3DGS avatars LAM (He et al., 2025), FlexAvatar (Kirschstein et al., 2026), and UIKA (Wu et al., 2026), and against CAP4D (Taubner et al., 2025b) and MVP4D (Taubner et al., 2025a), which distill FLAME-conditioned image- and video-diffusion outputs, respectively, into a 3DGS avatar. For CAP4D, we report both the full pipeline (**CAP4D**) and its diffusion model (**MMDM**). For computational reasons, we evaluate only MVP4D's video-diffusion model (MMVDM, labeled **MVP4D** in tables and figures). The camera-controlled VDM FaceCam (Lyu et al., 2026) lacks expression control, so we include it only in the multi-view consistency probe. Details in Appendix C.



Figure 4: Self-reenactment on BecomingLit and Multiface.

Table 3: Quantitative results on the VFHQ-test (Xie et al., 2022) dataset.

Method	PSNR ↑	LPIPS ↓	ID ↑	Expr ↑	JOD ↑	AKD ↓
LAM	20.4	0.050	0.605	0.791	3.0	0.037
FlexAvatar	22.2	0.043	0.804	0.835	3.3	0.038
UIKA	19.8	0.053	0.736	0.842	3.0	0.043
MMDM	22.2	0.040	0.793	0.834	3.3	0.034
MVP4D	22.2	0.040	0.749	0.824	3.4	0.031
Ours	24.9	0.024	0.852	0.931	5.0	0.028

Table 4: 3DGS distillation ($n = 13$): Evaluated against 16 G.T. cameras.

Method	PSNR ↑	LPIPS ↓	ID ↑	Expr ↑
UIKA	21.96	0.244	0.697	0.868
FlexAvatar	23.28	0.203	0.753	0.941
MMDM	23.28	0.242	0.722	0.871
MVP4D	21.99	0.275	0.687	0.889
FaceCam	18.86	0.292	0.731	0.922
Ours	23.69	0.192	0.869	0.968

Novel-View Reenactment Task. Our evaluation must simultaneously measure how well an avatar preserves identity and expression *and* how multi-view consistent it is. We therefore make the protocol as viewpoint-challenging as possible: each evaluation triplet consists of a single reference image $\mathbf{I}_{\text{ref}}$, a driving video, and a target video from *different* camera viewpoints, similar to the camera-triplet setup of training (Section 3.4). We evaluate on BecomingLit (Schmidt et al., 2025) (25 self-reenactment and 13 cross-reenactment clips) and Multiface (Wuu et al., 2022) (22 self-reenactments), since no unified cross-method split exists for NeRSemble or AVA-256. Those held-out datasets eliminate training-data leakage for all baselines. Details are in Appendix B.

3DGS Distillation. To explicitly measure multi-view consistency, we additionally fit a static 3DGS scene from each diffusion-based method's generations along a closed ellipse camera trajectory with a frozen expression, and evaluate the reconstruction. We choose the static setting because dynamic 3DGS optimization is far less standardized and can itself introduce fitting flaws that would confound the comparison. We render the fit at the 16 real rig cameras, none of which is part of the generated trajectory, and compare against the captured images. Details are in Appendix B. Metric definitions are given in Appendix A.

4.2 Results

Human perception of faces is highly sensitive to temporal artifacts such as flicker, identity drift, and unnatural motion, which single frames cannot convey. We therefore strongly encourage viewing the supplementary video alongside the figures below.

Self-reenactment. VDAs outperform every baseline on every metric of both datasets (Table 2), as well as on VFHQ-test (Xie et al., 2022) (Table 3, Figure 17 in the appendix), an in-the-wild single-view benchmark testing generalization beyond the multi-view training distribution. Qualitatively (Figure 4), our results retain sharper identity detail and transfer subtle expressions and micro-expressions, as well as rare expressions outside the FLAME space, which limits the baselines.

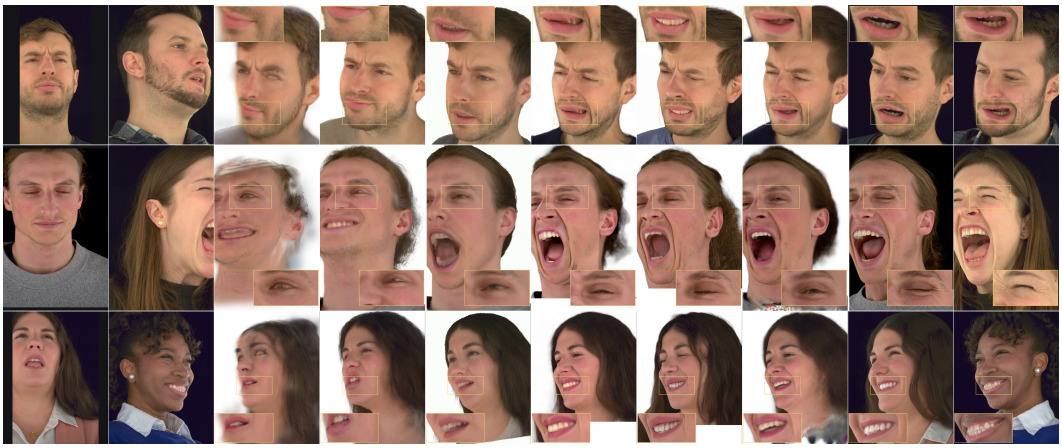

Figure 5: Cross-reenactment on BecomingLit. The reference identity is driven by a different person's expression and rendered to the target camera; the last column shows the driver from that camera.

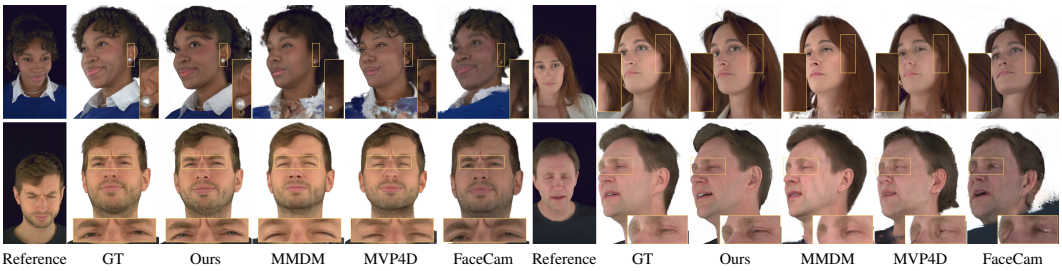

Figure 6: Multi-view consistency probe: each method's generation along a novel-view ellipse is fit into a static 3DGS and re-rendered at a rig camera (vs. GT). Quantitative results in Table 4.



Figure 7: In-the-wild novel-view reenactment: the camera sweeps a wide arc, isolating viewpoint consistency from expression transfer.

Cross-reenactment. Results appear in Figure 5 and the rightmost block of Table 2. Without a ground-truth target, pixel metrics are not applicable. Thus we measure identity similarity to the reference and expression and keypoint distance to the driver as seen from the target camera.

Multi-view consistency probe. The 3DGS probe (Figure 6, Table 4) measures multi-view consistency quantitatively: since the rig cameras lie off the generated trajectory, a fit only scores well if the generations agree across views and stay faithful to the subject. Our fits outperform every baseline on every metric and are significantly sharper than all baseline fits, whose view-inconsistent generations blur the reconstruction (Figure 6), confirming the multi-view consistency of VDAs.

Generalization. Figure 7 combines cross-reenactment with a wide camera sweep on challenging in-the-wild references: cast shadows, reflections and refractions in glasses, and human-like characters far outside every method's training data. Figure 8 orbits the camera to the back of the head, beyond the viewpoints of our training captures. Together, these results show that VDAs generalize to rendering effects, subjects, and camera poses absent from the training data. This capacity stems from the internet-scale video prior, which our conditioning is designed to preserve. The supplementary video further shows clothing dynamics, dangling earrings, and view-dependent lighting that follow the head, effects that explicit 3D renderers struggle to reproduce.

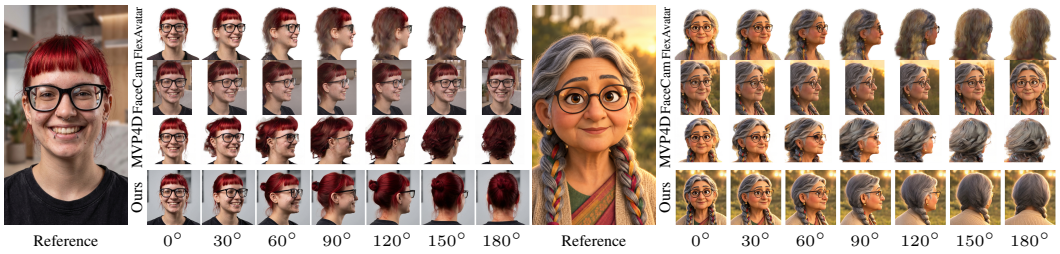

Figure 8: 360° novel-view synthesis from a single reference image: the camera orbits from the front (0°) to the back of the head (180°). Rows: FlexAvatar, FaceCam, MVP4D, and ours.

Table 5: **Self-reenactment ablations** on BecomingLit v3 (n=25), the only ablation dataset.

	Ablated knob				Self-reenactment, BecomingLit v3 (n=25)						
Method	Fine-tune	Expr. code	Pose cond.	Disent.	PSNR ↑	LPIPS ↓	JOD ↑	ID ↑	Expr. ↑	AKD ↓	APD ↓
WAN-A.	none	WAN native	2D skel.	–	16.87	0.2582	3.897	0.5825	0.8279	0.0813	6.06
WAN-A. disent.	none	WAN native	2D skel.	✓	17.03	0.2584	3.934	0.6044	0.8386	0.0791	5.72
Ours, WAN-A. expr.	all but enc.	WAN native	skel. + mesh	–	21.52	0.2098	4.438	0.7012	0.9115	0.0532	4.06
+ frozen fusers	all but enc., fusers	WAN native	skel. + mesh	–	19.67	0.2806	4.052	0.6534	0.8175	0.0808	6.03
Ours, no disent. †	full	X-NeMo	skel. + mesh	×	22.06	0.2048	4.608	0.6927	0.9099	0.0550	3.94
self-driven †	full	X-NeMo	skel. + mesh	×	20.69	0.2207	4.613	0.6866	0.8744	0.0608	4.51
Ours, no expr. codes	full	zeroed	skel. + mesh	–	21.61	0.2052	4.609	0.6896	0.8796	0.0561	3.96
Ours, no proxy mesh	full	X-NeMo	2D skel.	✓	18.34	0.2503	4.196	0.6840	0.9124	0.0722	4.93
Ours, rand. corpus	full	X-NeMo	skel. + mesh	✓	21.71	0.2487	4.462	0.6881	0.8954	0.0661	4.75
Ours (full)	full	X-NeMo	skel. + mesh	✓	22.43	0.2021	4.638	0.6989	0.9201	0.0549	3.97

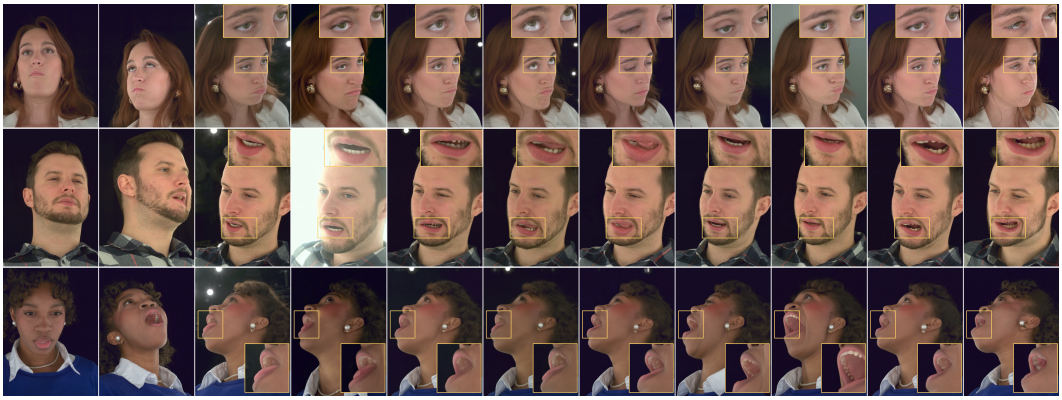

Figure 9: Qualitative ablation on BecomingLit v3; columns match Table 5.

Ablations. Table 5 and Figure 9 analyze our design choices on BecomingLit self-reenactment. For more details on the exact settings, we refer to Appendix C. Notably, the frozen WAN-Animate base model performs poorly at novel-view self-reenactment, and replacing its expression code $\mathbf{z}_{\text{raw}}^{\text{WAN}}$ with the viewpoint-neutral code improves the metrics only slightly.

Expression codes. Our X-NeMo codes outperform WAN-Animate's native expression code in image quality and expression fidelity, validating our choice. Nevertheless, the native-code variant remains competitive, showing that our recipe works with different latent expression spaces and can adopt stronger encoders in the future. Without any expression code, the model misses subtle expressions (Figure 9), which are not captured by FLAME. *Data and pose.* Training on a random instead of our curated corpus degrades every metric, and conditioning only on a ViT-Pose skeleton, WAN-Animate's native pose signal, without the proxy mesh costs 4.1 dB PSNR and raises AKD from 0.055 to 0.072, while expression fidelity stays high (0.912 vs. 0.920). *Disentanglement.* Without disentangler and camera triplets (self-driven: $\pi_{\text{drv}} = \pi_{\text{tar}}$), performance drops markedly. Without the disentangler but with camera triplets, the model is almost on par with our full approach. Remarkably, triplet supervision alone enforces view-invariant expression codes, raising DISTRACT rank-1 accuracy from 0.244 to 0.792 during training, whereas self-driven training stalls at 0.486 (Figure 14, Appendix J).

Table 6: FactorPortrait comparison (face-region metrics, Appendix A).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	JOD ↑	AKD ↓	APD ↓	ID ↑	Expr ↑
FactorPortrait (Tang et al., 2025)	24.71	0.712	0.0208	7.02	0.0329	2.37	0.747	0.946
Ours	24.75	0.717	0.0190	7.14	0.0322	2.94	0.803	0.969



Figure 10: Comparison to FactorPortrait on its released examples.

FactorPortrait comparison. FactorPortrait (Tang et al., 2025) is concurrent work that also controls expression, pose, and viewpoint in a VDM. No code or model is available and the authors confirmed that none can be shared, so we run our method on the self-reenactment examples released with their paper. These examples follow FactorPortrait's rather specific data distribution, which favors FactorPortrait. Nevertheless, VDAs slightly improve image quality and clearly improve identity preservation and expression fidelity (Table 6). Figure 10 further shows that FactorPortrait degrades under strong viewpoint changes.

5 Limitations and Future Work

Unlike a 3DGS avatar, which renders in real time once built, VDAs pay the cost of VDM inference for every clip (about 12 minutes for 77 frames at 576×1024 on one A100), and long sequences are prone to temporal drift. Autoregressive video and world models (Liu et al., 2026; Samuel et al., 2026; Zhu et al., 2026; Zhao et al., 2026), including a real-time variant of WAN-Animate-2 (Wang et al., 2026), have shown that long, consistent generation can be made fast. Motivated by this, we distill our model with Self Forcing (Huang et al., 2025) into a four-step auto-regressive VDM for temporal consistency as a prototype (Appendix K). This unoptimized implementation generates 4.8 frames per second at 288×512 on one A100 and stays temporally stable over more than a thousand frames, at the cost of a small loss in quality. Furthermore, lighting is baked into the reference image, and we rely on a fixed expression space rather than training it end-to-end. Our recipe is anchored to WAN-Animate-14B, although we expect it to transfer to other backbones, as well.

6 Conclusion

We showed that an avatar does not have to be a 3D asset: by making a video diffusion model controllable in expression, pose, and camera viewpoint, VDAs turn its internet-scale prior into a single-image avatar that is multi-view consistent and more faithful than both 3DGS-based and diffusion-based avatars. Central to this is expression control through learned latent spaces: our recipe works with different expression spaces, and those of image encoders capture expressions with higher fidelity than 3DMMs, but must first be disentangled from camera viewpoint for the best results. Training is affordable and accessible, taking 34 hours on a single GPU with public multi-view data only. Because the prior is preserved rather than distilled into 3D, VDAs render what explicit renderers struggle with (shadows, reflections, refractions, hair dynamics, and clothing that deforms with motion) and generalize to subjects, effects, and viewpoints absent from the training data. We see this as evidence that controllable video models are a practical foundation for photorealistic avatars.

Ethics Statement

VDAs turn a single photograph into a photo-realistic, controllable video of the depicted person. The same capability enables impersonation and non-consensual synthetic media. We train and evaluate only on multi-view datasets whose participants gave informed consent and which their authors released for research use (NeRSemble (Kirschstein et al., 2023), AVA-256 (Martinez et al., 2024), BecomingLit (Schmidt et al., 2025), Multiface (Wuu et al., 2022)) and on the public VFHQ benchmark (Xie et al., 2022).

Reproducibility Statement

We will release our full code, covering data preprocessing, training, inference, and evaluation, together with the trained model checkpoints. All training data (NeRSemble (Kirschstein et al., 2023), AVA-256 (Martinez et al., 2024)) and evaluation data (BecomingLit (Schmidt et al., 2025), Multiface (Wuu et al., 2022), VFHQ (Xie et al., 2022)) are public research datasets. Section 3 describes the architecture and training objective, Appendix E lists the training and inference hyperparameters, Appendix F details the proxy-mesh construction, and Appendices G and H.1 the disentangler. Appendices A and B define the evaluation protocol, including the DISTRACT benchmark, and every metric; Appendix C describes how each baseline and ablation was run.

AI Use Statement

In this work, we used generative AI tools to implement methods, as coding agents that followed detailed, author-written specifications, and to generate synthetic data, namely a small set of AI-generated portrait images that serve as reference inputs for qualitative examples; these images are used neither for training nor for any quantitative evaluation, but merely to demonstrate interesting and challenging inputs. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to assist with translation, or to clean and reformat datasets, and the rest of the required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, including the scripts that build the figures and tables (under careful supervision), to draft and edit parts of the paper, and to search the literature. For the writing, the authors wrote detailed bullet-point outlines, used generative AI to turn them into prose, and then revised the text manually over several iterations. We have reviewed all AI-assisted work: the authors reviewed and tested the generated code, our own evaluation pipeline produces every number in the paper and the authors checked each number against its source files, and the authors read and revised all AI-assisted text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 187–194, 1999.

Chen Cao, Tomas Ren, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Tomas Simon, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. In *ACM Trans. Graph. (SIGGRAPH Asia)*, volume 41, 2022.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Radek Danecek, Michael J. Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *CVPR*, pp. 8498–8507, 2024.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 40(4), 2021.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *CVPR*, 2023.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *CVPR*, 2024a.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia Conference Papers*, 2024b.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025.

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 *ACM SIGGRAPH Conference Papers*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, pp. 167283–167308, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 42(4), 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Flexavatar: Learning complete 3d head avatars with partial supervision. In *CVPR*, 2026.

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 (Proc. SIGGRAPH Asia)*, 36(6), 2017.

Zhiyuan Li, Chi-Man Pun, Chen Fang, Jue Wang, and Xiaodong Cun. Personalive! expressive portrait image animation for live streaming. *arXiv preprint arXiv:2512.11253*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. In *ACM Trans. Graph. (SIGGRAPH)*, volume 40, 2021.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6), 2015.

Weijie Lyu, Ming-Hsuan Yang, and Zhixin Shu. Facecam: Portrait video camera control via scale-aware conditioning, 2026.

Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 40(4), 2021.

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 (NeurIPS), Datasets and Benchmarks Track*, 2024. introduces the Ava-256 dataset.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR*, pp. 20299–20309, 2024.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. FASTAR: Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. In *ICML*, 2026.

Jonathan Schmidt, Simon Giebenhain, and Matthias Nießner. Becominglit: Relightable gaussian avatars with hybrid neural shading. In *NeurIPS*, 2025.

Liam Schoneveld, Zhe Chen, Davide Davoli, Jiapeng Tang, Saimon Terazawa, Ko Nishino, and Matthias Nießner. Sheap: Self-supervised head geometry predictor learned via 2d gaussians. In *ICCV*, 2025.

Jiapeng Tang, Kai Li, Chengxiang Yin, Liuhao Ge, Fei Jiang, Jiu Xu, Matthias Nießner, Christian Häne, Timur Bagautdinov, Egor Zakharov, and Peihong Guo. FactorPortrait: Controllable portrait animation via disentangled expression, pose, and viewpoint, 2025.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B. Lindell. MVP4D: Multi-view portrait video diffusion for animatable 4D avatars. In *SIGGRAPH Asia 2025 Conference Papers*, 2025a. doi: 10.1145/3757377.3763889.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *CVPR*, 2025b.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions. In *ECCV*, 2024.

Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 43(6), 2024.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, pp. 17979–17989, 2023.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Xindi Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, Gang Cheng, Hai Xu, and Bang Zhang. Wan-Animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026.

Yilin Wu et al. Dex-portrait: Disentangled and expressive portrait animation via explicit and latent motion representations. *arXiv preprint arXiv:2512.15524*, 2025.

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 *CVPR*, 2026.

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, pp. 21875–21895, 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 *CVPR 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 Conference Papers*, 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 *CVPR*, 2024.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *ICLR*, 2026a.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, pp. 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, pp. 22963–22974, 2025.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal Forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *CVPR*, 2022.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal Forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *ICML*, 2026.

Wojciech Zielonka, Tobias Kirschstein, Timo Bolkart, Simon Giebenhain, Vanessa Sklyarova, Xiang Deng, Donglai Xiang, Shunsuke Saito, Yebin Liu, Matthias Nießner, and Justus Thies. How to build digital humans? from priors to photorealistic avatars. *Computer Graphics Forum*, 45(2), 2026. Eurographics State-of-the-Art Report (STAR).

A Evaluation Metrics

We report PSNR and LPIPS over a face-only crop. We also report JOD, a perceptual video-quality metric computed via pyfvvdp (Mantiuk et al., 2021) that captures temporal, video-level quality beyond per-frame image metrics. We further report face-specific semantic metrics: identity preservation as ArcFace (Deng et al., 2019) cosine similarity to the ground-truth target, and expression similarity as the cosine similarity between the DECA (Feng et al., 2021) FLAME expression codes of the generated and ground-truth frames, averaged over the clip. We deliberately score expression in a *third-party* space rather than in X-NeMo's (Zhao et al., 2025) own motion embedding: our method is driven by X-NeMo codes, so a cosine similarity in that same space would be a metric we have a stake in. For pose accuracy we report keypoint distance normalized by inter-ocular distance (AKD/IOD), and in the ablations additionally APD, the average rotational distance (in degrees) between the head poses estimated with the FAN3D landmark detector on the generated and ground-truth frames.

To make these and other pixel-aligned metrics meaningful, we align the 3DMM present in every compared model to the ground-truth 3DMM before computing them, removing an otherwise-unresolved ambiguity. On the multi-view benchmarks this alignment is possible because a ground-truth target view exists. On monocular VFHQ-test it does not, so there we follow the standard protocol for that dataset and align on detected facial keypoints instead, landmark-cropping every method's output to a common head scale and centre before scoring. That alignment is what makes the pixel metrics comparable across methods whose outputs live in different frames, including methods that render into the reference frame rather than reprojecting to a target camera. Multi-view consistency is measured separately with the 3DGS probe (Section 4.2). Cross-reenactment has no ground-truth target image, so the pixel metrics do not apply there. Identity is instead measured against the reference, and expression and pose against the driver's own target-view video — the driver has the correct expression and pose for that camera, only the wrong identity, which is exactly what those two metrics need.

B Evaluation Protocol Details

DISTRACT. DISTRACT is built on the FLAME-tracked sequences of NeRSemble (Kirschstein et al., 2023) (all 16 cameras, sentence-reading sequences excluded). Each query has one positive and four distractors. In the *viewpoint-invariance* task (500 queries), query and positive are two cameras of the same identity, sequence and timestep. In the *identity-invariance* task, the positive is the nearest frame of another identity in FLAME expression and jaw parameters; of 456 mined sets, we kept the 126 whose positive we manually verified to show the query's expression. Distractors are random frames that share with the query its identity and camera (A), only its identity (B), only its camera (C), or neither (D). Codes are compared by cosine distance; Rank-1 is the fraction of queries for which no distractor is closer than the positive, and Err-X the fraction for which distractor X is. The MLP probe (two layers, supervised contrastive loss (Khosla et al., 2020), all views of one frame form a class) is fitted on each encoder's NeRSemble embeddings, which include the benchmark frames; it is thus transductive and measures how much view-invariant expression information the frozen latents contain, not generalization to unseen subjects.

Reenactment evaluation. BecomingLit (Schmidt et al., 2025) and Multiface (Wuu et al., 2022) are part of neither our training data nor that of any compared method; we avoid NeRSemble and AVA-256 because train/test splits differ between methods. Each case is a camera triplet, and all methods receive the same reference image, driving video and target calibration; expression is estimated from the driving video only, and multi-view tracking only places the head under π_{tar}. BecomingLit has 25 self-reenactment cases from 13 identities, with a near-frontal reference camera and driving and target cameras 35–101° apart in azimuth (median 77°). Its 13 cross-reenactment pairs form one cycle over the identities; ID is measured against the reference identity, and Expr and AKD against the driving identity's own video from π_{tar}. Multiface has 22 cases from 11 identities, with driving and target cameras 40.5–90.1° apart. VFHQ-test (Xie et al., 2022) adds 44 in-the-wild clips.

3DGS consistency probe. The probe has one case per BecomingLit identity (13), frozen at one frame, with the frontal photo as the only reference and the expression held fixed. The generative methods follow 81 poses on a closed ellipse around the head (azimuth $\pm 60°$, elevation $+5°$ to $-30°$); ours and MMDM generate the first 77. For each, we fit a static 3DGS scene (Kerbl et al., 2023) (20k iterations, identical settings) to the generated frames with the exact trajectory cameras, render it at the 16 real rig cameras, which lie 0.2–11.6° (median 2.3°) from the nearest trajectory pose, and compare with the captured photos; PSNR and LPIPS are face-region metrics. MVP4D runs MMVDM from a static camera with the trajectory folded into the head pose, in order to produce smooth renderings. Its multi-view variant reaches 23.40 dB PSNR, but flickers significantly, and sits still below ours on all metrics. UIKA and FlexAvatar are rendered directly; FlexAvatar uses its feed-forward mode, since its per-subject fitting would have access to the evaluated facial expression.

C Baseline Details

Baselines. All methods receive the same reference image, driving video and target calibration; expression is estimated monocularly from the driving video, and our multi-view FLAME tracking only places the head under π_{tar}. LAM (He et al., 2025) and FlexAvatar (Kirschstein et al., 2026) are driven by pixel3DMM (Giebenhain et al., 2025) FLAME2020 fits; FlexAvatar uses its per-subject inversion (200 steps). UIKA (Wu et al., 2026) is driven by a monocular VHAP fit. CAP4D (Taubner et al., 2025b) and MVP4D (Taubner et al., 2025a) use pixel3DMM fits converted to CAP4D's FLAME basis: **CAP4D** is the full pipeline, **MMDM** samples the target camera directly, and **MVP4D** is its video-diffusion model MMVDM at the target camera, since fitting MVP4D's 3DGS avatar per clip is computationally prohibitive. FaceCam (Lyu et al., 2026) re-renders a given video under a new camera and takes neither a reference identity nor an expression signal, so it appears only in the 3DGS probe.

Ablation variants. *WAN-A.* runs the frozen WAN-Animate-14B with its own motion encoder and a ViT-Pose (Xu et al., 2022) skeleton detected on the SAM-3D-Body mesh rendered into π_{tar}. *WAN-A. disent.* passes the same motion codes through $\mathcal{D}^{\text{WAN}}$ and $\mathcal{E}^{\text{WAN}}$, re-entangled with the mean camera latent of π_{tar} over the other identities; it uses no X-NeMo code and no fine-tuning.

All fine-tuned rows are trained for $5{,}000 + 250$ steps. *WAN-A. expr.* feeds WAN-Animate's native code to a frozen $\mathcal{P}_{\text{face}}$; $+$ *frozen fusers* also freezes $\mathcal{B}_{\text{face}}$. *No disent.* † replaces $\mathcal{D}^{\text{XN}}$ by a translator from raw X-NeMo to raw WAN-Animate codes, and *self-driven* † additionally reads the expression code from the target camera. *No expr. codes* zeroes the expression code, so expression enters only through the FLAME head of the proxy mesh. *No proxy mesh* removes the proxy mesh and, at test time, uses the skeleton of the training-free rows. *Rand. corpus* draws clips and windows at random from the same sources.

D FactorPortrait Comparison Details

FactorPortrait (Tang et al., 2025) has released neither code nor a model, so we compare on the self-reenactment videos from its project page, which contain the reference still, FactorPortrait's output, and the ground truth, but neither camera parameters nor a separate driving video. Both methods are scored against the same ground-truth frames with the face-region metrics of Appendix A, without 3DMM alignment. The setup favors FactorPortrait: its authors chose the clips, which come from its own capture system, whereas we have to estimate camera, head, and body by monocular tracking, and our reference image is extracted from a compressed web video.

E Implementation Details

We report the configuration of the model evaluated as "Ours" in the BecomingLit self-reenactment experiments and the ablations (Tables 2 and 5).

Data and preprocessing. The corpus contains 4,696 clips of 81 frames from 640 identities: 1,846 clips from AVA-256 (Martinez et al., 2024) and 2,850 from NeRSemble (Kirschstein et al., 2023), mostly expression and emotion sequences. Each clip fixes a camera triplet (Section 3.4); reference and driving cameras are predominantly near-frontal, whereas target cameras cover a wide range of viewpoints. The FLAME (Li et al., 2017) head of the proxy mesh comes from monocular pixel3DMM (Giebenhain et al., 2025) tracking of the driving video and is placed per frame by a similarity transform onto the dataset's multi-view head tracking; torso and neck come from SAM-3D-Body (Yang et al., 2026b), stabilized by pelvis-anchored kinematic fitting (Appendix F.1). The skeleton projects the keypoints of the same fitted body into π_{tar} and is drawn with WAN-Animate's own renderer.

Training and inference. We fine-tune 861.5 M parameters of WAN-Animate-14B (Cheng et al., 2025) without LoRA (Table 7); the DiT blocks $\mathcal{B}$, the VAE, and the image and text encoders stay frozen. In our implementation, $\mathcal{T}$ and $\mathcal{E}^{\text{WAN}}$ with the fixed camera code $\bar{\mathbf{z}}_{\text{cam}}^{\text{WAN}}$ form a single MLP, and $\mathcal{P}_{\text{mesh}}$ is initialized to zero, so the proxy mesh has no effect at initialization. At inference, we integrate $\mathbf{v}_\theta$ with 20 Euler steps at 576×1024.

Table 7: Fine-tuning configuration. Wall-clock times include periodic validation.

Trainable module	Parameters	Learning rate
Face fusers $\mathcal{B}_{\text{face}}$	839.0 M	5×10^{-6}
Face encoder $\mathcal{P}_{\text{face}}$	17.8 M	10^{-5}
Translator ($\mathcal{T}$, $\mathcal{E}^{\text{WAN}}$ merged)	3.1 M	10^{-5}
Expression head of $\mathcal{D}^{\text{XN}}$	0.8 M	10^{-5}
$\mathcal{P}_{\text{skel}}$, $\mathcal{P}_{\text{mesh}}$	0.3 M each	5×10^{-5}
	Stage 1	Stage 2
Resolution	288×512	576×1024
Steps (wall-clock)	5,000 (25.8 h)	250 (8.1 h)
AdamW, weight decay 10^{-3}, one 81-frame clip per step (77 supervised), gradient clipping 1.0, one A100 (80 GB).		

F Proxy-Mesh Conditioning Details

F.1 Pelvis-Anchored Kinematic Fitting

SAM-3D-Body (Yang et al., 2026b) predicts each frame independently. Aligning these per-frame meshes to the tracked FLAME head with a per-frame similarity transform fits the face well, but any error in that fit propagates down the kinematic chain: shoulders and hips swing although the subject sits still, most visibly from oblique cameras. Temporal smoothing suppresses the high-frequency jitter but not the low-frequency drift.

Subjects in multi-view face captures are seated, so their pelvis is stationary. We exploit this by re-expressing the sequence with a *single* similarity transform $\mathcal{W}$ and per-frame MHR joint angles $\boldsymbol{\phi}_t$ (6D rotation parameterization), solved jointly over all T frames:

$$\begin{aligned}\min_{\mathcal{W},\{\boldsymbol{\phi}_t\}} \sum_{t=1}^{T} &\big\|\mathcal{W}\big(\mathbf{k}(\boldsymbol{\phi}_t)\big) - \mathbf{l}_t\big\|^2 + \lambda_{\text{pel}}\big\|\mathcal{W}\big(\mathbf{j}_{\text{pel}}(\boldsymbol{\phi}_t)\big) - \tilde{\mathbf{j}}_{\text{pel}}\big\|^2 \\ &+ \lambda_{\text{sh}}\big\|\mathcal{W}\big(\mathbf{j}_{\text{sh}}(\boldsymbol{\phi}_t)\big) - \tilde{\mathbf{j}}_{\text{sh}}\big\|^2 + \lambda_{\text{prior}}\|\boldsymbol{\phi}_t - \hat{\boldsymbol{\phi}}_t\|^2 + \lambda_{\text{temp}}\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\|^2,\end{aligned} \tag{6}$$

where $\mathbf{k}$ are the MHR face keypoints corresponding to the FLAME landmarks $\mathbf{l}_t$, $\mathbf{j}_{\text{pel}}$ and $\mathbf{j}_{\text{sh}}$ the pelvis and shoulder joints with temporal medians $\tilde{\mathbf{j}}$, and $\hat{\boldsymbol{\phi}}_t$ the SAM-3D per-frame estimate. The pelvis term encodes the stationary-pelvis assumption; since the pelvis lies outside the conditioning crop, a light shoulder pin additionally holds the visible torso in place. Because the MHR forward kinematics is differentiable, we optimize with Adam for 400 iterations, using $\lambda_{\text{pel}}{=}1$, $\lambda_{\text{sh}}{=}0.3$, $\lambda_{\text{prior}}{=}10^{-3}$, $\lambda_{\text{temp}}{=}0.03$.

We validate the fitting on NeRSemble EXP-1 sequences, whose large head rotations (37–56°) would expose an over-constrained fit (Figure 11). The pelvis term alone reduces the median pelvis drift of per-frame alignment from 52–79 mm to 6–12 mm; the shoulder pin then reduces the remaining shoulder wander from 11.2 to 0.07 mm and frame-to-frame jitter by 14×, at an unchanged face residual (3.04 vs. 3.06 mm). Relative to per-frame alignment, which is optimal for the face by construction, the face residual rises by at most 0.8 mm: genuine head motion is still tracked, and the constraint only removes motion the seated subject never performed.

F.2 Pose-Aware Proxy-Mesh Coloring

For a pixel $\mathbf{u}$ with unit world-space ray direction $\mathbf{d}(\mathbf{u})$, we color the proxy mesh as

$$\mathbf{c}(\mathbf{u}) = \tfrac{1}{2}\, g(\mathbf{u}) + \tfrac{1}{2} \cdot \tfrac{1}{2}\big(\mathbf{R}^\top \mathbf{d}(\mathbf{u}) + \mathbf{1}\big), \tag{7}$$

where g is Lambertian shading (key light 25° above the camera) and $\mathbf{R}$ is the local frame of the body part covering $\mathbf{u}$. For the head, $\mathbf{R}_{\text{h}}$ is the rotation that places the FLAME head in the scene. For the torso, $\mathbf{R}_{\text{b}}$ has its up axis from the hip midpoint to the neck and its lateral axis between the hips, orthogonalized. Head and torso are rendered separately and the head is composited over the torso, so the two regions never interpenetrate.

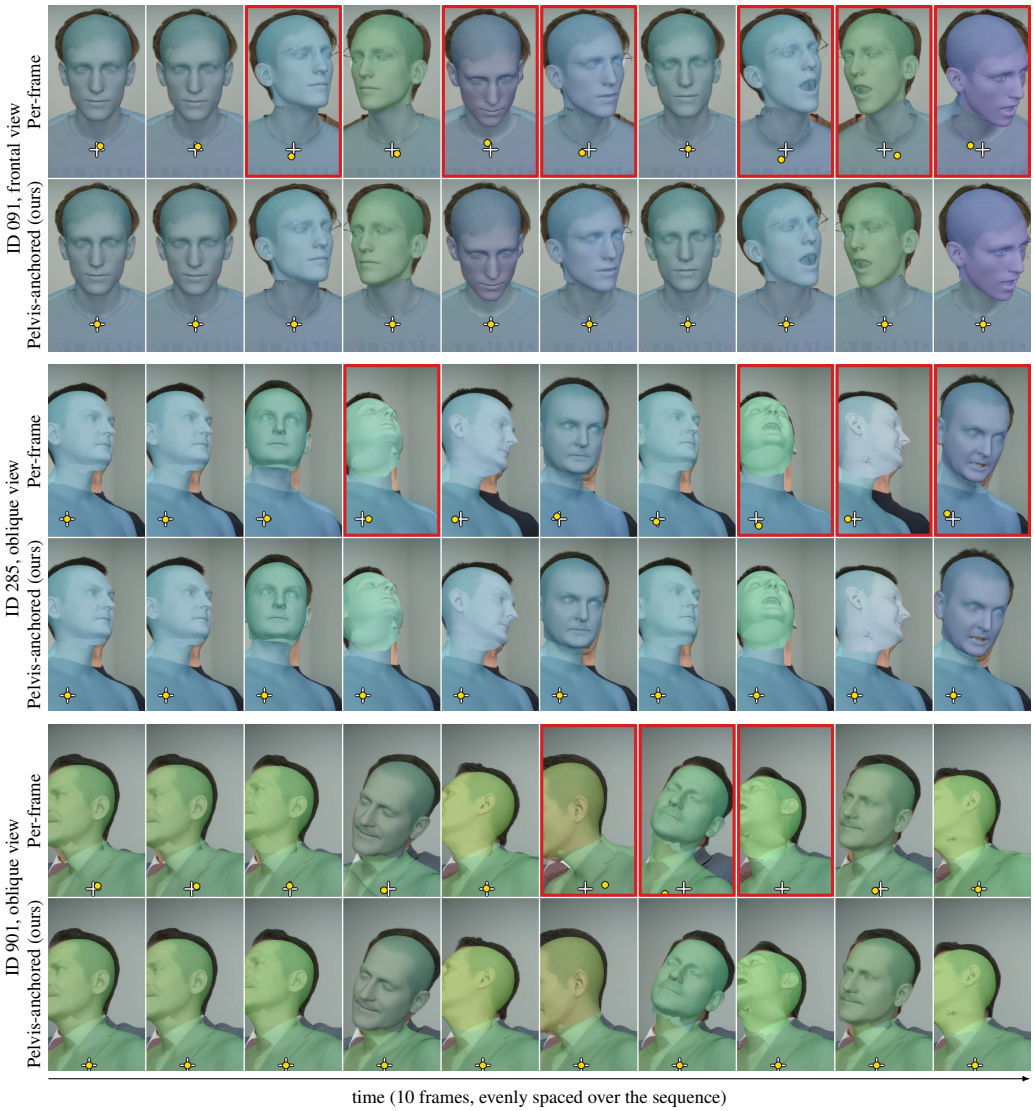

Figure 11: **Pelvis-anchored kinematic fitting stabilizes the proxy-mesh body.** Proxy-mesh conditioning renders (Appendix F.2) overlaid on the ground-truth frames of three NeRSemble EXP-1 sequences with large head rotations. For each subject, the top row aligns every per-frame SAM-3D estimate independently to the tracked FLAME head; the bottom row uses pelvis-anchored kinematic fitting. Both rows share the same head. The crosshair marks the shoulder midpoint's median position over the sequence, the yellow dot its position in the shown frame; red boxes mark frames whose shoulder midpoint lies more than 45 px (at 720×1280) from that median. Under per-frame alignment the torso drifts, although the subject is seated; pelvis-anchored fitting holds it in place while the head, and with it the head's color, follows the performance. The three camera views show how the ray-direction colors depend on the viewpoint.

The choice of frame determines what the color can encode. Camera-space directions depend only on pixel position and carry no pose information; world-space directions encode the camera but not the body; and a single shared frame for both parts leaves a neck rotation visible only through the silhouette. With per-part local frames, a camera move shifts both regions' colors together, while a neck rotation changes the head's color relative to the torso. This ties the neck-control signal directly to the RGB content of the conditioning. Figure 11 illustrates this: with the torso held in place by pelvis-anchored fitting, its color stays constant over the sequence, while the head's color changes with every head rotation.

G Disentanglement Module Details

The cycle-consistency loss encourages a clean and stable factorization. Given two entangled codes $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we first disentangle them as

$$\mathcal{D}^{\text{XN}}(\mathbf{a}) = (\mathbf{a}_{\text{ex}}, \mathbf{a}_{\text{cam}}), \quad \mathcal{D}^{\text{XN}}(\mathbf{b}) = (\mathbf{b}_{\text{ex}}, \mathbf{b}_{\text{cam}}). \tag{8}$$

We then swap the camera latents and pass the resulting codes through the entangler and disentangler:

$$(\hat{\mathbf{a}}_{\text{ex}}, \hat{\mathbf{b}}_{\text{cam}}) = \mathcal{D}^{\text{XN}}\big(\mathcal{E}^{\text{XN}}(\mathbf{a}_{\text{ex}}, \mathbf{b}_{\text{cam}})\big), \qquad (\hat{\mathbf{b}}_{\text{ex}}, \hat{\mathbf{a}}_{\text{cam}}) = \mathcal{D}^{\text{XN}}\big(\mathcal{E}^{\text{XN}}(\mathbf{b}_{\text{ex}}, \mathbf{a}_{\text{cam}})\big). \tag{9}$$

The cycle-consistency loss enforces recovery of the original factors after this swap-and-reconstruct operation:

$$\mathcal{L}_{\text{cycle}} = \sum_{k \in \{\text{ex,cam}\}} \Big(\|\hat{\mathbf{a}}_k - \mathbf{a}_k\| + \|\hat{\mathbf{b}}_k - \mathbf{b}_k\| \Big). \tag{10}$$

We additionally apply a supervised contrastive loss to the expression latent $\mathbf{z}_{\text{ex}}^{\text{XN}}$. Samples sharing the same expression label are treated as positive pairs, while samples with different expression labels are treated as negatives. The objective encourages expression latents from the same class to be close in the embedding space and those from different classes to be far apart:

$$\mathcal{L}_{\text{con}} = -\sum_{p \in \mathcal{P}} \log \frac{\exp\big(\text{sim}(\mathbf{z}_{\text{ex}}^i, \mathbf{z}_{\text{ex}}^p)/\tau\big)}{\sum_{a \in \{\mathcal{P} \cup \mathcal{N}\}} \exp\big(\text{sim}(\mathbf{z}_{\text{ex}}^i, \mathbf{z}_{\text{ex}}^a)/\tau\big)}, \tag{11}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the positive and negative candidate samples for anchor i, respectively.

H Disentanglement Training Details

H.1 Disentanglement Training

Model. Let $\mathbf{z} \in \mathbb{R}^d$ be the entangled code of a frozen, pretrained expression encoder (X-NeMo, $d = 512$). The disentangler D consists of independent MLP heads that read the same code in parallel,

$$D(\mathbf{z}) = \big(f_{\text{id}}(\mathbf{z}),\ f_{\text{cam}}(\mathbf{z}),\ f_{\text{ex}}(\mathbf{z})\big) = (\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{cam}}, \mathbf{z}_{\text{ex}}), \tag{12}$$

with $\mathbf{z}_{\text{ex}} \in \mathbb{R}^{d_{\text{ex}}}$, $\mathbf{z}_{\text{cam}} \in \mathbb{R}^{d_{\text{cam}}}$ and $\mathbf{z}_{\text{id}} \in \mathbb{R}^{d_{\text{id}}}$. An optional fourth head f_{res} yields a small residual $\mathbf{z}_{\text{res}} \in \mathbb{R}^{d_{\text{res}}}$ that absorbs what none of the three supervised factors explains (lighting, crop, background). The entangler E concatenates the latents and maps them back into the encoder's space,

$$E(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{cam}}, \mathbf{z}_{\text{ex}}) = g\big([\mathbf{z}_{\text{id}};\ \mathbf{z}_{\text{cam}};\ \mathbf{z}_{\text{ex}};\ \mathbf{z}_{\text{res}}]\big) = \hat{\mathbf{z}}. \tag{13}$$

All heads $f_\bullet$ and g are three-layer ReLU MLPs with a linear output layer; no latent is normalized, since the norm of $\mathbf{z}$ carries expression intensity and must survive the round trip.

Supervision from multi-view video. We train on NeRSemble (Kirschstein et al., 2023), where every image is indexed by identity, sequence, frame, and camera, (p, s, t, c). Codes $\mathbf{z}$ are pre-computed once for all images of the training identities, so training operates on cached codes only. Three label functions provide the supervision: (i) the *expression* label $y^{\text{ex}} = (p, s, t)$ – all 16 cameras of a frame show the same expression. The human-verified cross-identity pairs of the DISTRACT identity-invariance task (Section 3.1) are merged into the same class (transitive closure), so an expression class may span several identities; (ii) the *viewpoint* label $y^{\text{cam}} = c$, the physical camera (16 classes); (iii) the *identity* label $y^{\text{id}} = p$.

Supervised contrastive losses. Given a batch $\mathcal{B} = \{\mathbf{z}_i\}_{i=1}^N$ with labels y_i and a head f, we apply the supervised contrastive loss (Khosla et al., 2020) to the ℓ_2-normalized head outputs $\mathbf{u}_i = f(\mathbf{z}_i)/\|f(\mathbf{z}_i)\|_2$, with temperature-scaled similarities $s_{ij} = \mathbf{u}_i^\top \mathbf{u}_j/\tau$:

$$\mathcal{L}_{\text{con}}(f, y; \mathcal{B}) = -\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s_{ip})}{\sum_{a \in N(i)} \exp(s_{ia})}, \tag{14}$$

where $N(i) = \{a \neq i\} \setminus \mathcal{I}(i)$ is the set of samples anchor i is contrasted against, $P(i) = \{p \in N(i) : y_p = y_i\}$ its positives, $\mathcal{A} = \{i : P(i) \neq \emptyset\}$ the anchors with at least one positive, and $\mathcal{I}(i)$ an optional per-anchor ignore set (empty unless stated otherwise). Batches are class-balanced: we draw n classes and k samples per class. Each head gets its own loss and its own batch,

$$\mathcal{L}^{\star}_{\text{con}} = \mathcal{L}_{\text{con}}(f_{\star}, y^{\star}; \mathcal{B}_{\star}), \qquad \star \in \{\text{ex}, \text{cam}, \text{id}\}. \tag{15}$$

Positives of f_{ex} are the same expression seen from *different* cameras (and identities), whereas positives of f_{cam} are the same camera across *different* frames and identities; the three losses thus pull each factor into its own head and push the other factors out of it.

Reconstruction and residual penalty. The entangler must reproduce the original code,

$$\mathcal{L}_{\text{rec}} = \big\| E(D(\mathbf{z})) - \mathbf{z} \big\|_1, \tag{16}$$

so the split loses no information; here and below $\|\cdot\|_1$ denotes the mean absolute error, *i.e.* the ℓ_1 norm divided by the dimension. When a residual head is used, an ℓ_1 penalty $\mathcal{L}_{\text{res}} = \|\mathbf{z}_{\text{res}}\|_1$ keeps it from becoming a shortcut for the reconstruction; its weight $\lambda_{\text{res}}(e) = \lambda_{\text{res}} \min(1, e/50)$ is ramped in linearly over the first 50 epochs e.

Cycle consistency. Reconstruction and contrastive terms alone do not force the entangler to actually *use* each latent for its designated factor. We therefore swap latents between two random codes $\mathbf{z}^a$ and $\mathbf{z}^b$ of the batch, re-entangle, and require the re-disentangled latents to match:

$$\hat{\mathbf{z}}^{a \leftarrow b} = E\big(\mathbf{z}^a_{\text{id}}, \mathbf{z}^b_{\text{cam}}, \mathbf{z}^a_{\text{ex}}\big), \quad \hat{\mathbf{z}}^{b \leftarrow a} = E\big(\mathbf{z}^b_{\text{id}}, \mathbf{z}^a_{\text{cam}}, \mathbf{z}^b_{\text{ex}}\big), \tag{17}$$

$$\begin{aligned} \mathcal{L}_{\text{cyc}} = \tfrac{1}{4}\Big(& \|f_{\text{ex}}(\hat{\mathbf{z}}^{a \leftarrow b}) - \mathbf{z}^a_{\text{ex}}\|_1 + \|f_{\text{cam}}(\hat{\mathbf{z}}^{a \leftarrow b}) - \mathbf{z}^b_{\text{cam}}\|_1 \\ & + \|f_{\text{ex}}(\hat{\mathbf{z}}^{b \leftarrow a}) - \mathbf{z}^b_{\text{ex}}\|_1 + \|f_{\text{cam}}(\hat{\mathbf{z}}^{b \leftarrow a}) - \mathbf{z}^a_{\text{cam}}\|_1 \Big) \\ + \tfrac{1}{2}\Big(& \|f_{\text{id}}(\hat{\mathbf{z}}^{a \leftarrow b}) - \mathbf{z}^a_{\text{id}}\|_1 + \|f_{\text{id}}(\hat{\mathbf{z}}^{b \leftarrow a}) - \mathbf{z}^b_{\text{id}}\|_1 \Big). \end{aligned} \tag{18}$$

Gradients flow through both the swapped latents and the re-encoding. The residual fed to E in Eq. (17) is detached and, per batch, either kept (each code retains its own $\mathbf{z}_{\text{res}}$), exchanged between a and b, or zeroed, with equal probability, so the cycle cannot be closed through the residual. All cycle terms are switched on at epoch $e_{\text{cyc}} = 35$, once the heads have stabilized.

expression nor leak into $\mathbf{z}_{\text{ex}}$, and vice versa.

Total objective. With e the current epoch, $\gamma_e = \mathbf{1}[e \geq e_{\text{cyc}}]$ gating the cycle terms, and $\star$ running over the three heads, the training loss is

$$\begin{aligned} \mathcal{L} = \lambda_{\text{rec}}\, \mathcal{L}_{\text{rec}} + \sum_{\star} \lambda_{\star}\, \mathcal{L}^{\star}_{\text{con}} + \lambda_{\text{res}}(e)\, \mathcal{L}_{\text{res}} \\ + \gamma_e \Big(\lambda_{\text{cyc}}\, \mathcal{L}_{\text{cyc}} \Big). \end{aligned} \tag{19}$$

Implementation details. For the X-NeMo disentangler we use $d = d_{\text{ex}} = 512$, $d_{\text{cam}} = 32$, no identity or residual head, and MLP width 512. Loss weights are $\lambda_{\text{rec}} = \lambda_{\text{ex}} = \lambda_{\text{cam}} = 1$, $\lambda_{\text{cyc}} = \lambda^{\text{ex}}_{\text{cyc}} = \lambda^{\text{cam}}_{\text{cyc}} = 0.1$ and $\tau = 0.07$ throughout; contrastive batches contain 32×4 (expression) and 8×16 (camera) codes. We optimize with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, halved every 200 epochs) for 1000 epochs with 64 reconstruction codes per step, holding out 10% of the identities for validation, and select the checkpoint with the lowest validation reconstruction error. Table 8 lists the configurations of all disentanglers.

Table 8: Disentangler configurations. *X-NeMo*: the checkpoint that produces $\mathbf{z}_{\text{ex}}$ for all VDA training and inference. *WAN-Animate*: the disentangler of WAN-Animate's native expression code used for training-free viewpoint control.

	X-NeMo (deployed)	WAN-Animate (`epoch_0850`)
d / d_{ex} / d_{cam}	512 / 512 / 32	512 / 512 / 51
d_{id} / d_{res}	– / –	128 / 32
MLP width (3 layers)	512	512
λ_{rec}	1	100
λ_{ex} / λ_{cam} / λ_{id}	1 / 1 / –	1 / 1 / 1
λ_{res} (ramp 50 ep.)	–	10
λ_{cyc} ($e_{\text{cyc}} = 35$)	0.1	10
$\lambda_{\text{cyc}}^{\text{ex}}$ / $\lambda_{\text{cyc}}^{\text{cam}}$	0.1 / 0.1	0.1 / 0.1
τ (all heads)	0.07	0.07
$n \times k$ expression	32×4	32×4
$n \times k$ camera	8×16	8×16
$n \times k$ identity	–	16×8
viewpoint label y^{cam}	camera	camera
batch (recon. codes)	64	64
lr / weight decay	10^{-3} / 10^{-4}	10^{-3} / 10^{-4}
lr schedule	step $\times 0.5$ / 200	step $\times 0.5$ / 200
epochs (selected)	1000 (700)	1000 (850)

I Training-Free Use of the Disentangled Codes

The disentangler of Section 3.1 splits an expression encoder's raw code into an expression latent $\mathbf{z}_{\text{ex}}$ and a camera latent $\mathbf{z}_{\text{cam}}$, and the entangler maps any pair back into the encoder's native code space. Both generators below stay frozen: we only change the code they receive.

X-NeMo. Keeping $\mathbf{z}_{\text{ex}}$ fixed and sweeping $\mathbf{z}_{\text{cam}}$ through the codes of other camera views rotates the generated head while the expression is preserved, both for a constant and for an extreme driving expression (Figure 12). X-NeMo cannot distinguish a camera change from a neck rotation, so the camera code moves the head rather than the viewpoint. This is why our avatar model takes camera and body pose from the proxy mesh instead (Section 3.3).

WAN-Animate. WAN-Animate's native expression codes carry the viewpoint of the frame they are extracted from. When the target view differs from the expression source, rendering with the raw codes produces subtle but visible expression artifacts (parted lips, a shifted mouth shape) that vary along the camera path. Re-entangling our disentangled expression latent with the target view's camera code removes them (Figure 13). The remaining artifacts of this training-free variant, and the coarse camera control it offers, motivate the camera-triplet fine-tuning of Section 3.

J Emergent View Invariance

Camera-triplet supervision alone makes the expression code view-invariant. We compare the two runs marked † in Table 5, which differ only in whether the driving expression code is read from a different camera than the reconstructed frame (cross-view) or from the same camera (self-driven), and track DISTRACT viewpoint-invariance rank-1 accuracy of the expression code at the face-encoder output during training (Figure 14). Both start at 0.244; cross-view training climbs to 0.792, whereas self-driven training stalls at 0.486. Since the proxy mesh already fixes the target camera, viewpoint information in the expression code is never needed; under cross-view supervision it is actively misleading, and training suppresses it. The gain carries over to generation quality (Table 5).

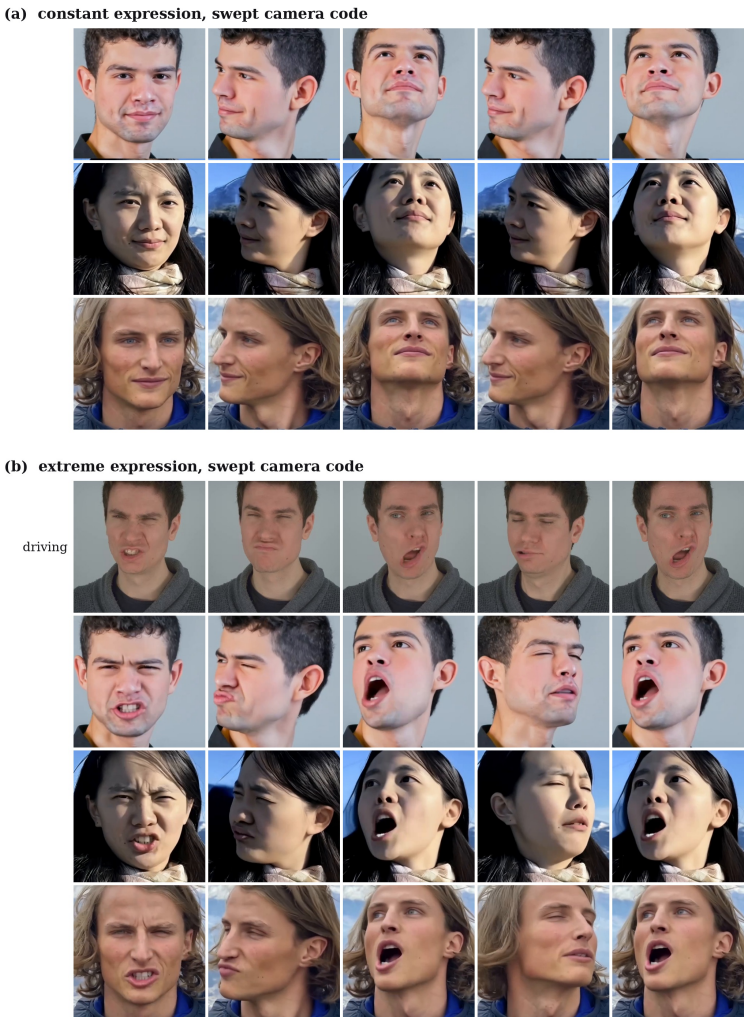

Figure 12: Training-free head-pose control in X-NeMo. The disentangled expression code is kept fixed while the camera code is swept through other views, (a) with a constant expression and (b) with an extreme driving expression (top row: driving frames). The head rotates and the expression is preserved; X-NeMo does not separate camera from neck motion, so the camera code moves the head rather than the viewpoint.

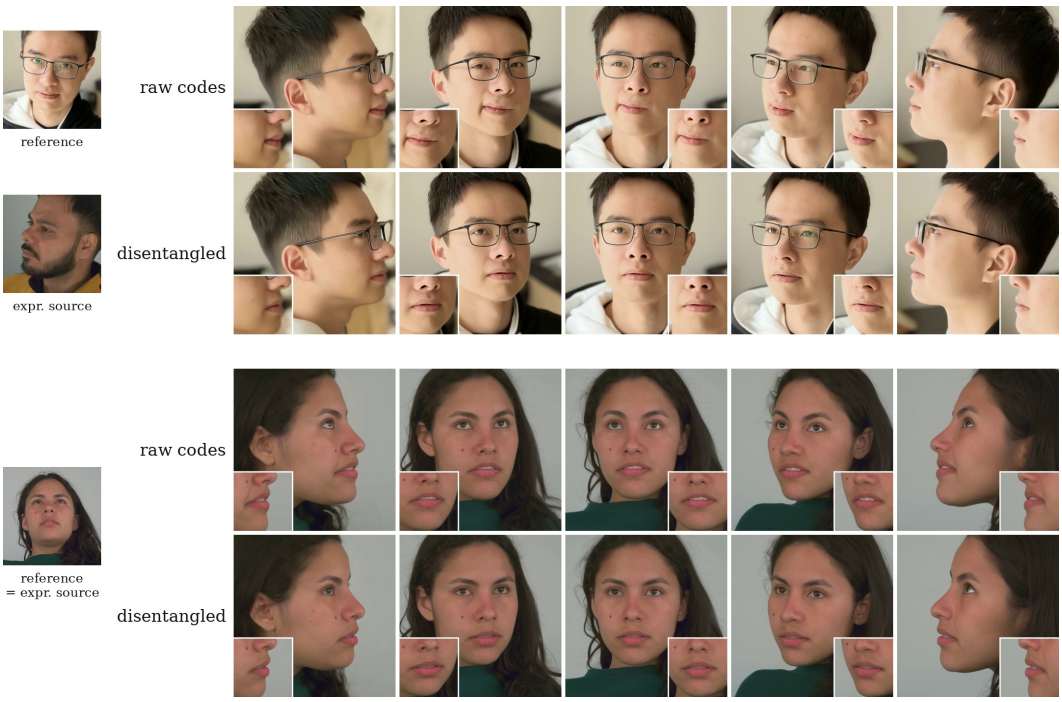

Figure 13: Frozen WAN-Animate rendering five views of a camera orbit, driven by its raw expression codes or by our disentangled codes (expression latent re-entangled with the target view's camera code). Insets magnify the mouth. Top: the expression source is a profile view of another person. Bottom: the expression is taken from the reference image itself. Raw codes carry the source viewpoint, which shows up as expression artifacts once the target view differs; the disentangled codes preserve the source expression.

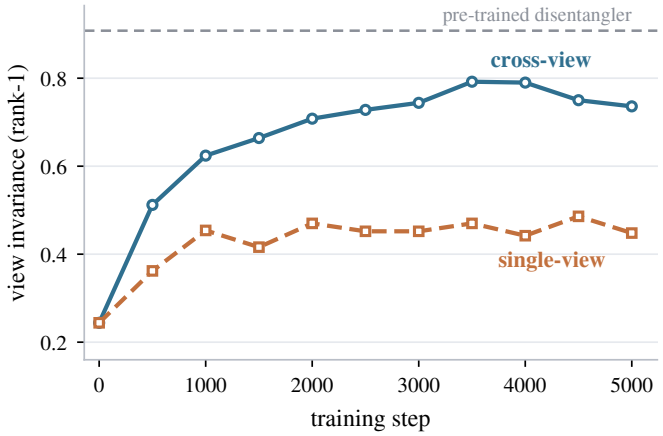

Figure 14: **Cross-view supervision creates view invariance; single-view supervision does not.** DISTRACT viewpoint-invariance rank-1 of the expression code at the face-encoder output, over 500 triplets, for two runs that differ only in the camera the driving expression code is read from. Both begin at 0.244, WAN-Animate's pre-trained face encoder on the shared translator initialisation. Cross-view training reaches 0.79, while single-view training stalls below 0.49. The final gap is 0.288.

K Streaming Distillation Prototype

As a proof of concept for fast and long generation, we distill our model with Self Forcing (Huang et al., 2025) into a causal, four-step student. The student shares our architecture, weights, and conditioning; we only train rank-128 LoRA adapters on the transformer and the face-fuser blocks, and make attention block-causal over blocks of four latent frames (16 video frames). Training follows the two stages of Self Forcing, an ODE initialization (Yin et al., 2025) on 1,500 teacher trajectories followed by distribution matching distillation (Yin et al., 2024b;a) on self-generated videos, at 288×512 and takes about 60 A100 GPU hours. At inference, the student streams the video block by block with a rolling key–value cache and two frame sinks (Xiao et al., 2024; Yang et al., 2026a), the reference latent and the first generated frames, so that memory is constant in the video length.

On one A100, the student generates 4.8 frames per second at 288×512: a 41-second video (1029 frames) takes about four minutes, whereas our full model needs about 12 minutes for a single 77-frame window at 576×1024. On 41-second videos of 11 held-out subjects, identity and appearance remain stable to the last frame, and expression fidelity stays constant while chunked generation in overlapping windows degrades over time. The student loses some identity and expression fidelity relative to our full model and runs below real time; training at higher resolution and system-level optimizations are natural next steps.

L Full Quantitative Results

Table 9 reports all evaluated encoders on DISTRACT without the MLP probe.

Table 9: All 8 evaluated encoders on the DISTRACT benchmark (n=500 for Task 1, n=126 for Task 2; entries in %), without the MLP probe. Columns as in Table 1. Classical 3DMM regressors (top) are consistently less viewpoint-biased than the learned reenactment/portrait-animation encoders below the rule.

	View Inv.		Identity Inv.	
Encoder	R-1↑	Err A/B/C/D↓	R-1↑	Err A/B/C/D↓
DECA (Feng et al., 2021)	61	33/18/18/11	34	51/37/25/24
EMOCA (Danecek et al., 2022)	42	47/22/30/16	32	56/34/43/21
SHeaP (Schoneveld et al., 2025)	54	41/17/07/03	17	69/52/31/21
EMOPortraits (Drobyshev et al., 2024)	23	70/37/51/28	14	76/42/58/36
PD-FGC (Wang et al., 2023)	08	90/42/79/43	07	91/48/81/43
VOODOO-XP (Tran et al., 2024)	15	84/36/55/28	06	94/56/71/39
WAN-Animate (Cheng et al., 2025)	21	72/34/59/33	14	74/32/62/26
X-NeMo (Zhao et al., 2025)	18	78/35/70/34	14	78/37/68/37

M Additional Qualitative Results

We encourage the reader to watch our supplementary video. Full 8-row self-reenactment comparisons on BecomingLit (Figure 15) and Multiface (Figure 16); Figure 4 in the main paper shows a subset of these rows. Figure 17 shows qualitative comparisons on VFHQ-test.

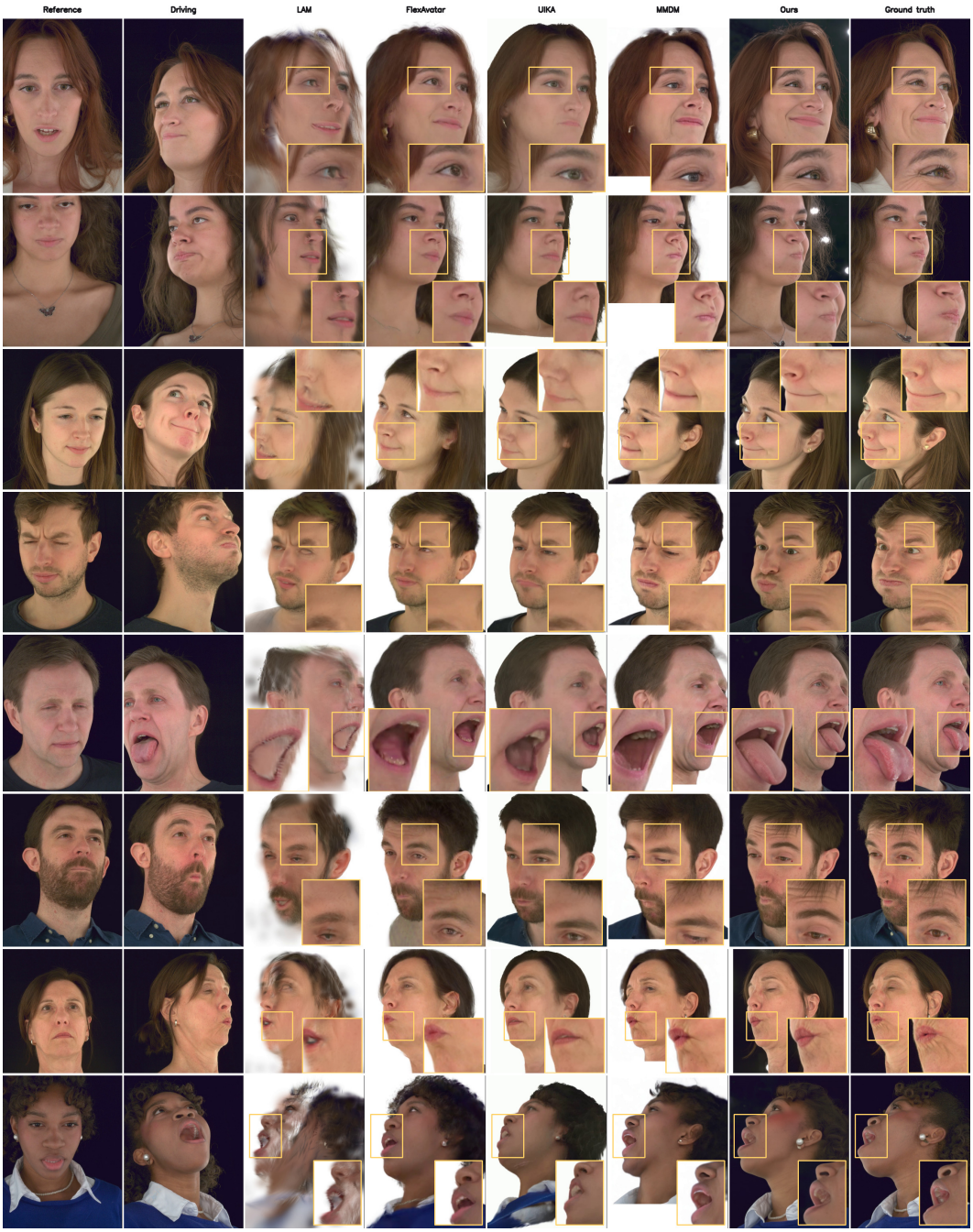

Figure 15: BecomingLit self-reenactment, qualitative comparison of our method against LAM (He et al., 2025), FlexAvatar (Kirschstein et al., 2026), UIKA (Wu et al., 2026), and MMDM (the morphable multi-view diffusion component of CAP4D (Taubner et al., 2025b), rendered directly at the target camera – not the full CAP4D avatar pipeline). One row per identity (8 of the benchmark's 13), each frame chosen to maximize expression diversity within that identity's own driving clips (farthest-point sampling over X-NeMo expression embeddings (Zhao et al., 2025)), then narrowed from an initial per-identity diverse set down to these 8 through several rounds of manual review.



Figure 16: Multiface self-reenactment, qualitative comparison of our method against LAM (He et al., 2025), FlexAvatar (Kirschstein et al., 2026), UIKA (Wu et al., 2026), and MMDM (the morphable multi-view diffusion component of CAP4D (Taubner et al., 2025b), rendered directly at the target camera – not the full CAP4D avatar pipeline). One row per identity (8 of the benchmark's 11), each frame chosen to maximize expression diversity within that identity's own driving clips (farthest-point sampling over X-NeMo expression embeddings (Zhao et al., 2025)), then narrowed from an initial per-identity diverse set down to these 8 through two rounds of manual review.

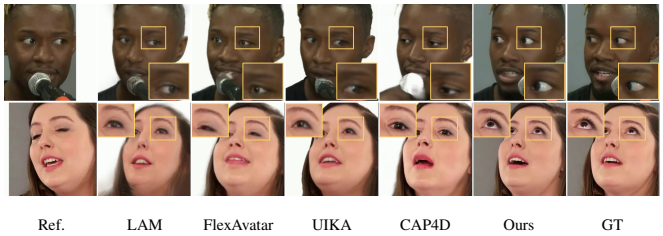

Figure 17: Qualitative comparisons on the in-the-wild VFHQ test set (Xie et al., 2022), against SotA 3D avatars.