UNIPORTRAIT4D: UNIFYING 4D PORTRAIT RESHOOTING AND ANIMATION

Anonymous authors
Paper under double-blind review

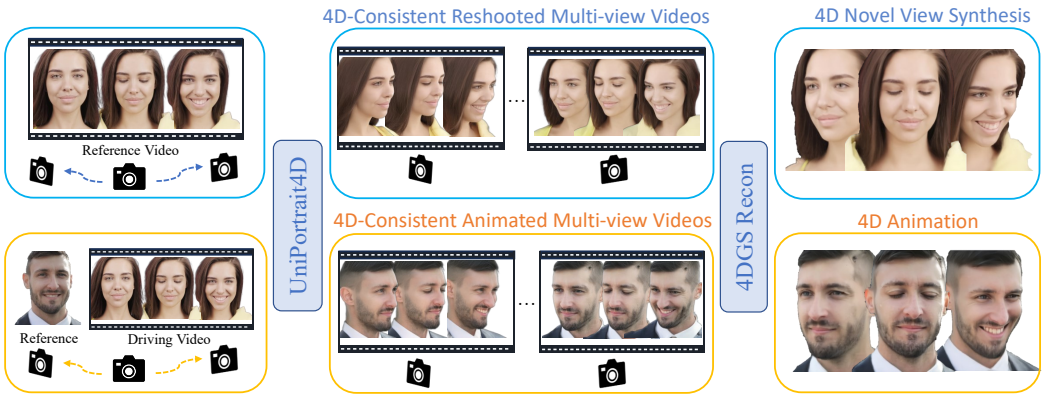

Figure 1: **4D reshooting and animation results by UNIPORTRAIT4D.** Given a monocular portrait video, it synthesizes synchronized, multi-view-consistent videos from novel viewpoints. Given a reference portrait and a driving video, UNIPORTRAIT4D transfers expressions while preserving the reference identity. These generated views can further be reconstructed into a 4D Gaussian representation for real-time free-viewpoint rendering.

ABSTRACT

Portrait animation and reshooting (novel-view synthesis) have been extensively studied for video generation, yet existing approaches typically treat them as separate problems: expression-driven animation is often restricted to a fixed viewpoint, while reshooting relies on limited multi-view training data with insufficient identity diversity. We present UNIPORTRAIT4D, a unified video DiT that supports both synchronized multi-view reshooting from a monocular portrait video and 4D portrait animation. To unify 4D portrait reshooting and animation, we leverage FLAME-based rendering as a unified camera representation for multi-view generation. For reshooting, we additionally incorporate point-cloud renderings to improve multi-view consistency. For animation, we employ a high-capacity latent representation for facial expressions and adversarially train it with gradient reversal to reduce driver-identity predictability while preserving fine-grained facial motion. Benefiting from the 4D consistency of our video DiT, the generated multi-view videos can be reconstructed into a 4DGS representation, enabling real-time free-viewpoint rendering. Extensive experiments demonstrate that our method achieves strong reference-identity preservation in cross-identity animation and high multi-view consistency in reshooting.

1 INTRODUCTION

Creating 4D portrait assets has broad applications in immersive VR, telepresence, digital-human creation, and film production. In particular, generating such assets directly from a single image or monocular video would substantially reduce capture requirements and make personalized 4D

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

content more accessible. A practical 4D portrait should support both controllable facial expressions and free-viewpoint observation. A model that jointly supports expression editing and novel-view synthesis would therefore greatly simplify the creation and manipulation of 4D portraits from casual monocular inputs.

Previous work supports novel-view rendering and animation by optimizing animatable head avatars (Qian et al., 2024; Saito et al., 2024). While effective with dense multi-camera capture, monocular reconstruction is ill-posed because much of the head is unobserved. Diffusion-based approaches reduce this ambiguity by synthesizing additional views and expressions before avatar distillation (Taubner et al., 2025b;a; Xu et al., 2026a). Yet FLAME-driven avatars (Li et al., 2017) cannot faithfully represent out-of-template dynamics such as fine mouth-interior and tongue motion. We instead reconstruct a dynamic 4D Gaussian representation from generated multi-view videos (Wang et al., 2025), preserving such evidence beyond FLAME. This requires reconstruction-grade cross-view and temporal consistency because generation errors become baked into the 4DGS. To meet this requirement, our reshooting branch uses a hybrid condition that combines detailed point renderings with complete FLAME geometry. These complementary geometric cues improve the spatiotemporal consistency of the generated multi-view videos, enabling downstream 4DGS reconstruction.

However, a reshooting-only solution still faces a generalization bottleneck: synchronized multi-view portrait datasets provide dense camera coverage but far fewer identities than monocular video corpora (Kirschstein et al., 2023; Xie et al., 2022). We therefore extend the task domain from reshooting alone to both reshooting and animation, and present UNIPORTRAIT4D, a unified framework for *free-view reshooting* and *free-expression animation*. This extension naturally brings abundant monocular animation videos into training, broadening the renderer's identity and motion coverage and improving robustness on in-the-wild inputs. Conversely, camera-supervised reshooting teaches cross-view correspondences that improve the multi-view consistency of portrait animation.

Animation poses an additional identity-preservation challenge. Expressive implicit motion latents expand the controllable expression space (Zhao et al., 2025; Shi et al., 2026), but high-capacity latents learned by self-reconstruction can encode driver identity together with motion. We retain a high-dimensional expression latent and apply identity-adversarial gradient-reversal training to reduce decodable driver-identity information while preserving detailed motion. This improves reference-identity preservation under cross-identity driving without sacrificing expressive animation.

Our contributions are summarized as follows:

- We introduce, one single video DiT that supports both synchronized reshooting from a monocular portrait video and implicit-motion cross-identity animation.
- We propose a domain-adversarial post-training strategy for the expression encoder that uses reversed identity gradients to selectively suppress identity-related features in the high-dimensional expression latent, reducing identity leakage while preserving expressive facial-motion capacity.
- We propose a hybrid conditioning architecture that combines detailed point renders with completed FLAME geometry for consistent novel-view synthesis, including unseen regions, and downstream 4DGS reconstruction.

2 RELATED WORK

Video reshooting and 4D head generation. Video reshooting changes an input video's camera trajectory while preserving its appearance and dynamics. General-purpose and portrait-specific methods formulate it as video-conditioned generative re-rendering guided by camera trajectories or explicit geometric evidence, including reprojections and point clouds (Bai et al., 2025a; Lyu et al., 2026; Zhang et al., 2025; Yu et al., 2025; Ren et al., 2025; Hu et al., 2025; Cai et al., 2026; Lin et al., 2026). Multi-camera generation further exchanges information across target views to improve synchronization (Bai et al., 2025b). For 4D head generation, portrait-focused methods use synthesized views and expressions to construct animatable 3D or 4D head representations from sparse or monocular inputs (Taubner et al., 2025b;a; Xu et al., 2026a; Deng et al., 2024a;b; Tran et al., 2024b;a; Gao et al., 2025). Our distinction is the capability combination: synchronized video reshooting that preserves the motion of a monocular input sequence and cross-identity animation driven by an implicit motion latent are supported within the same DiT.

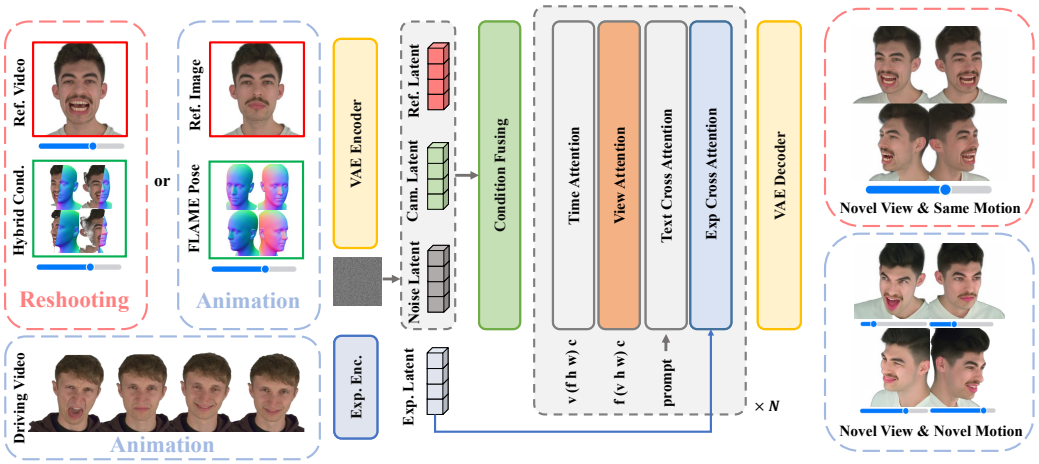

Figure 2: **Overview of UNIPORTRAIT4D.** A shared video DiT performs both portrait reshooting and animation. In reshooting mode, VHAP (Qian, 2024) recovers the FLAME parameters, while Face Anything (Kocasarı et al., 2026) reconstructs a dynamic point cloud. The resulting FLAME normal and point cloud are rendered from the target cameras and injected as spatial conditions. Together with reference tokens, they are concatenated with noise to serve as conditions and processed by the DiT with time attention, view attention and text cross-attention. In animation mode, an identity-adversarial expression encoder extracts a high-capacity expression latent, which is injected through additional cross-attention.

Portrait animation and identity leakage. Portrait animation uses either explicit controls or learned implicit motion features. Landmarks, 3DMM coefficients, and motion fields provide interpretable or spatially aligned guidance (Siarohin et al., 2019; Zhao & Zhang, 2022; Ma et al., 2024; Zhang et al., 2024; Lin et al., 2025), but their expressiveness is bounded by the tracker or parametric basis, making subtle mouth-interior motion and facial wrinkles difficult to capture. Implicit representations learn richer motion directly from images (Xie et al., 2024; Guo et al., 2024; Shi et al., 2026), but their higher capacity also permits driver identity to leak into the motion condition. Existing remedies include cross-identity augmentation, identity-altered driving data, explicit separation, and compact motion bottlenecks (Xie et al., 2024; Yang et al., 2024; Tan et al., 2025; Xu et al., 2026b); these either reduce expression capacity or do not directly enforce identity invariance. We instead employ a high-dimensional expression latent and use an ArcFace identity regressor with gradient reversal (Deng et al., 2019; Ganin et al., 2016) to suppress decodable identity information while preserving expression utility.

3 METHOD

3.1 OVERVIEW

Let V_s denote a source video, $\mathcal{C} = \{C_v\}_{v=1}^{V}$ a set of target cameras, I_r a reference portrait, and V_d a driving video. We define multi-view reshooting and portrait animation as

$$\hat{V}_{\text{re}} = \mathcal{S}(V_s, \mathcal{C}), \qquad \hat{V}_{\text{ani}} = \mathcal{A}(I_r, V_d), \tag{1}$$

The reshooting mapping $\mathcal{S}$ should preserve the appearance and frame-wise performance of V_s while changing only its camera trajectory. The animation mapping $\mathcal{A}$ should preserve the identity and appearance of I_r while transferring the facial expression and motion of V_d.

As illustrated in Figure 2, UNIPORTRAIT4D realizes both mappings with a shared video diffusion transformer G_θ. For reshooting, Face Anything (Kocasarı et al., 2026) reconstructs a dynamic point cloud and VHAP (Qian, 2024) estimates FLAME geometry and pose. Both are rendered into the target views as point renders and global-coordinate normal maps, then processed with the source-video features by time attention, view attention, and text cross-attention. For animation, a

domain-adversarial expression encoder extracts an expressive latent from V_d with reduced driver-identity leakage and injects it through expression cross-attention. The generated reshooting views further support 4DGS reconstruction and real-time rendering.

Section 3.2 details the hybrid geometry conditioning and view attention for consistent reshooting, while Section 3.3 describes the domain-adversarial motion encoder for expressive, identity-preserving animation.

3.2 Hybrid Geometry-Conditioned Video Reshooting

Existing video reshooting methods can generate visually plausible novel-view videos, but outputs from multiple target cameras are often not mutually consistent in 3D. This discrepancy limits their use as synchronized observations for downstream reconstruction. We improve multi-view 3D consistency with two complementary strategies: a hybrid condition concatenates detailed point renders with complete, global-coordinate FLAME normals, and view attention directly exchanges information among the generated cameras using these geometric cues.

Hybrid point-and-normal condition injection. We provide the denoiser with complementary target-view evidence that balances detail and coverage. Given a monocular source video, Face Anything (Kocasarı et al., 2026) estimates a temporally tracked 4D facial point cloud $\mathcal{Q}_t = \{(\mathbf{x}_{t,i}, \mathbf{a}_{t,i})\}_{i=1}^{K}$, where $\mathbf{x}_{t,i}$ is a 3D point and $\mathbf{a}_{t,i}$ stores its source color. Rasterizing $\mathcal{Q}_t$ with target camera C_v produces an RGB point render $P_{t,v}$ and a binary validity mask $M_{t,v}$:

$$(P_{t,v}, M_{t,v}) = \mathcal{R}(\mathcal{Q}_t, C_v). \tag{2}$$

Point renders preserve source detail but cover only visible regions, whereas global-coordinate FLAME normals provide complete geometry and view-consistent soft correspondences. Let E_c denote the video VAE encoder and E_{mask} the mask-to-latent alignment operation. We concatenate the encoded conditions with the noisy video latent:

$$\tilde{z}_{t,v} = \text{Concat}\left[z_{t,v},\, E_c(P_{t,v}),\, E_{\text{mask}}(M_{t,v}),\, E_c(N_{t,v})\right]. \tag{3}$$

The conditions are concatenated along the channel dimension, whereas the source-video references are appended along the token-length dimension to provide appearance and motion context. Details of E_{mask} are provided in Appendix B.

During training, we use double reprojection (Yu et al., 2025) to construct point-cloud renderings aligned with the target-view ground-truth frames; the complete procedure is given in Appendix C.

View attention. For latent tokens with v views, f frames, spatial size $h \times w$, and c channels, full attention has complexity $\mathcal{O}((vfhw)^2)$. We instead alternate time attention on (v, fhw, c) and view attention on (f, vhw, c), reducing the cost to $\mathcal{O}(v(fhw)^2 + f(vhw)^2)$ while propagating information through the full 4D space. The view-attention blocks inherit the self-attention design and pretrained weights of Wan2.2-DiT (Wan et al., 2025). This initialization allows the base model's temporal-consistency prior to serve as a strong starting point for learning cross-view consistency.

Downstream 4DGS reconstruction. To assess downstream reconstruction, we feed 18 synchronized outputs and their target-camera parameters to FreeTimeGS (Wang et al., 2025). The views cover frontal and profile directions together with elevated and depressed cameras. The reconstructed 4DGS supports real-time free-viewpoint rendering.

3.3 Domain-Adversarial Motion Encoding

Why additional identity suppression is needed. Cross-identity portrait animation rarely has paired data in which the same motion is performed by different identities. Most methods therefore adopt self-supervised self-reenactment, using frames of the same identity as both source and driving supervision. Under this objective, the model must implicitly separate identity from expression, and the learned expression latent can retain driving-identity cues that compromise identity consistency during cross-identity reenactment. Prior methods mitigate this leakage either by compressing the expression code through a low-capacity bottleneck or by applying data augmentations that discourage identity encoding. However, a narrow bottleneck also limits the representation of subtle and complex

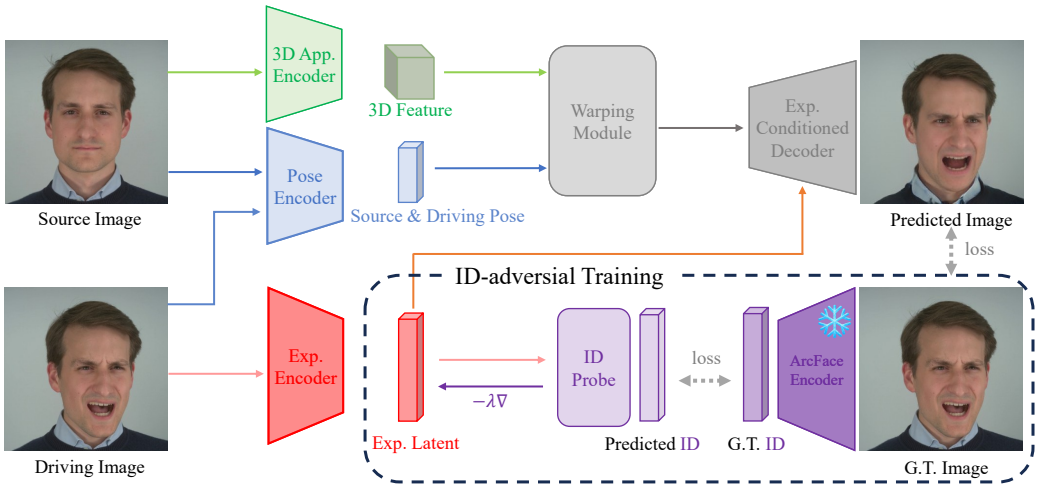

Figure 3: **Identity-adversarial expression learning with gradient reversal.** Pose and expression encoders are trained through self-supervised reconstruction. A gradient-reversal layer (GRL) enables adversarial training between the expression encoder and an identity probe, encouraging the expression latent to suppress identity information while preserving expression details for reconstruction.

expressions, whereas augmentation alone may not sufficiently suppress identity information. We therefore employ an expressive 512-dimensional latent and introduce a domain-adversarial objective that explicitly suppresses its decodable identity information while preserving expression utility.

Self-supervised motion encoder initialization. Following DeX-Portrait (Shi et al., 2026), we initialize pose and expression encoders by reconstructing driving frames from source–driving pairs sampled from the same video. The expression encoder F_ϕ maps an aligned driving face to a 512-dimensional latent; architectural details are provided in Appendix B.

Identity-probe training. We freeze F_ϕ and train a six-layer MLP probe D_ψ to predict the ArcFace identity embedding $\mathbf{f}_{\text{id}}$ using

$$\mathcal{L}_{\text{id}} = 1 - \cos(D_\psi(e), \mathbf{f}_{\text{id}}), \qquad e = F_\phi(I_d). \tag{4}$$

The probe reaches a mean cosine similarity above 0.5, confirming that the initial high-capacity expression latent contains readily decodable driver-identity information.

Identity-adversarial encoding with gradient reversal. Our key extension is a gradient-reversal layer GRL_λ placed between F_ϕ and D_ψ, defined by

$$\text{GRL}_\lambda(e) = e, \qquad \frac{\partial\, \text{GRL}_\lambda}{\partial e} = -\lambda \mathbf{I}. \tag{5}$$

After probe initialization, we jointly optimize D_ψ, F_ϕ, and the animation generator G_ω. During this stage, we evaluate Equation 4 with $e = \text{GRL}_\lambda(F_\phi(I_d))$. The forward pass through the GRL is unchanged, so D_ψ continues to minimize $\mathcal{L}_{\text{id}}$, whereas F_ϕ receives the reversed identity gradient and learns to make driver identity harder to decode. Simultaneously, F_ϕ and G_ω retain the base generation objective $\mathcal{L}_{\text{gen}}$ so that identity suppression does not collapse expression content; its components are detailed in Appendix B. The final joint objective is

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{id}}. \tag{6}$$

During adversarial fine-tuning, the identity probe minimizes $\mathcal{L}_{\text{id}}$, while the expression encoder receives its reversed gradient scaled by λ. In this way, identity-related features embedded in the expression latent are suppressed.

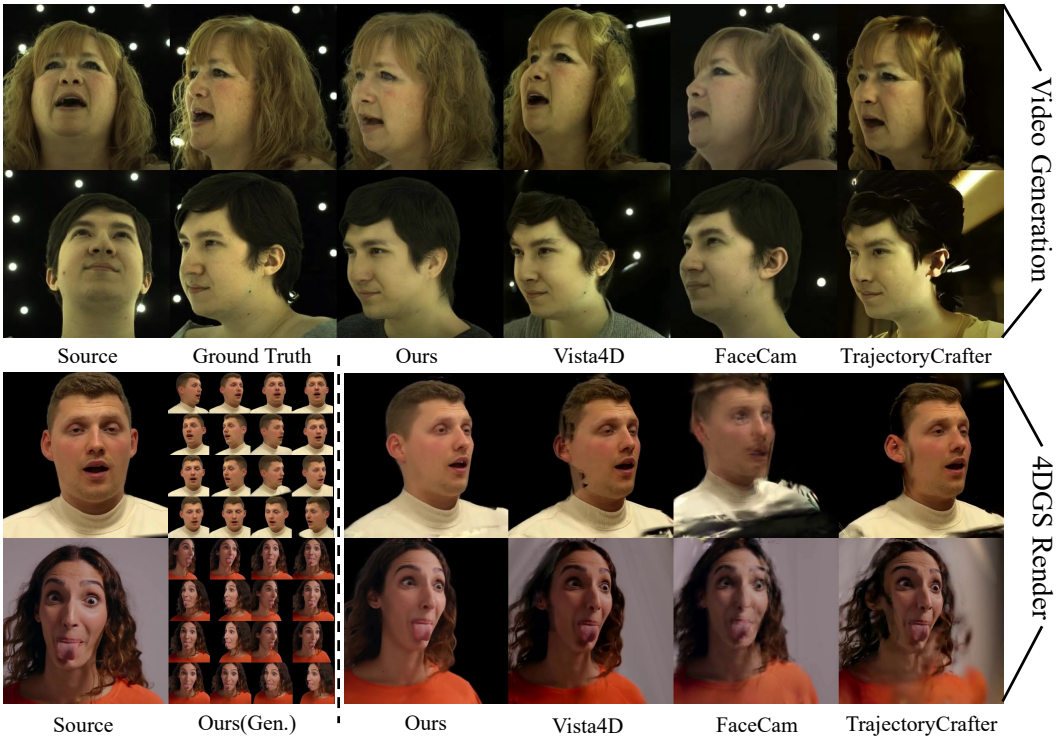

Figure 4: Qualitative comparison of portrait reshooting and downstream 4DGS reconstruction. **Top:** Target-view generation compared with ground truth. **Bottom:** 4DGS reconstructions obtained from the generated multi-view videos.

4 Experiments

4.1 Experimental Setup

Benchmarks. For reshooting, we use the multi-view AVA-256 benchmark (Martinez et al., 2024) with 50 held-out identities and identity-disjoint training, validation, and test partitions. We additionally use manually selected monocular portrait videos from Pexels (Pexels, 2026) to qualitatively evaluate generated novel views under in-the-wild motion, illumination, and background variation. For portrait animation, we evaluate on FEED (Drobyshev et al., 2024), which contains extreme and asymmetric facial actions, and on the same in-the-wild source. We use both *self-reenactment*, which provides paired ground truth, and *cross-identity reenactment*, which evaluates reference-identity preservation and driving-motion transfer. No evaluation identity is used for model training or checkpoint selection.

Baselines. For reshooting, we compare against FaceCam (Lyu et al., 2026), Vista4D (Lin et al., 2026), and TrajectoryCrafter (Yu et al., 2025). FaceCam controls the target camera through facial landmarks, whereas Vista4D and TrajectoryCrafter use point-cloud guidance. For animation, we compare against Wan-Animate (Cheng et al., 2025), Wan-Animate-2 (Wang et al., 2026), and X-NeMo (Zhao et al., 2025). X-NeMo and Wan-Animate encode driving expressions into implicit motion latents, whereas Wan-Animate-2 directly conditions on driving video frames without intermediate motion representations.

Metrics. Following 4DAnyone (Jin et al., 2026), we report three complementary evaluations for reshooting. First, Generated-to-GT fidelity compares generated videos directly with captured ground truth using PSNR, SSIM, and LPIPS. Second, 4DGS-to-GT fidelity compares held-out renderings of a 4DGS reconstructed from the generated views with the captured videos, directly measuring downstream reconstruction accuracy. Third, generated-view consistency (GVC) fits a 4DGS to 16 of



Figure 5: Qualitative comparison of portrait animation on extreme expressions and in-the-wild inputs.

Table 1: Portrait video reshooting on 50 held-out AVA-256 identities. We separately evaluate generated-view fidelity, downstream 4DGS fidelity to captured ground truth, and internal generated-view consistency (GVC).

	Generated-to-GT			4DGS-to-GT			GVC		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
TrajectoryCrafter (Yu et al., 2025)	17.44	0.5462	0.4048	16.87	0.5314	0.4821	19.78	0.6531	0.3395
FaceCam (Lyu et al., 2026)	18.93	0.6255	0.3876	14.41	0.6087	0.4013	15.95	0.6472	0.3402
Vista4D (Lin et al., 2026)	19.49	0.6931	0.2936	18.92	0.6768	0.3079	21.67	**0.7920**	0.2010
UNIPORTRAIT4D	**21.40**	**0.7740**	**0.2664**	**20.76**	**0.7589**	**0.2796**	**24.33**	0.7872	**0.1970**

18 generated views and compares its renderings at the remaining two cameras with the corresponding generated videos, measuring internal cross-view explainability. For animation, we evaluate self-reenactment with the same image metrics and cross-identity reenactment with identity similarity (CSIM), average pose distance (APD), and average expression distance (AED), where APD and AED are computed from MediaPipe Face Mesh landmarks (Kartynnik et al., 2019). Detailed protocols are provided in Appendix A.

4.2 COMPARISON

Table 2: Portrait animation. Pixel-aligned metrics are reported for self-reenactment, while cross-identity pairs are evaluated for reference-identity preservation and driver-motion transfer.

	Self-reenactment			Cross-identity reenactment		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	APD ↓	AED ↓
Wan-Animate (Cheng et al., 2025)	26.94	0.833	0.109	0.618	1.465	0.0256
Wan-Animate-2 (Wang et al., 2026)	22.81	0.751	0.160	0.622	2.751	0.0299
X-NeMo (Zhao et al., 2025)	20.91	0.710	0.192	0.602	1.820	0.0274
UNIPORTRAIT4D	**27.20**	**0.841**	**0.107**	**0.630**	**1.320**	**0.0241**

Reshooting and Multi-view Consistency. As shown in Figure 4, TrajectoryCrafter suffers from weak point-cloud guidance, FaceCam exhibits side-view misalignment due to inaccurate landmark conditioning, and Vista4D struggles to recover facial geometry under large viewpoint changes.

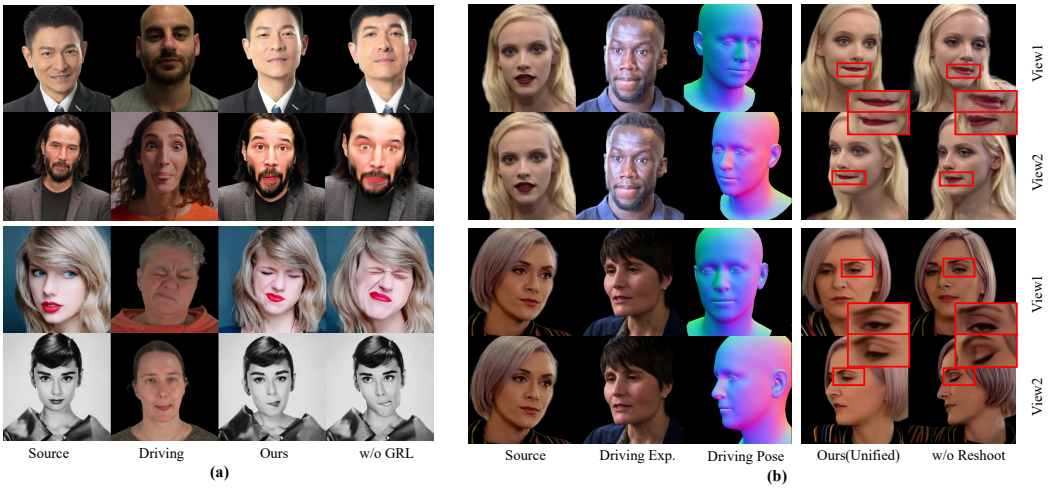

Figure 6: Qualitative ablations. **(a)** GRL improves identity consistency while preserving expression transfer. **(b)** Unified training of reshooting and animation improves the cross-view consistency of multi-view animation, yielding better alignment of the eyes and mouth across viewpoints.

In contrast, our method preserves source details and accurate camera control. While competing methods generate target views independently due to memory constraints, our method jointly generates synchronized views with cross-view attention, improving multi-view coherence and downstream 4DGS reconstruction. Table 1 reports the best Generated-to-GT and 4DGS-to-GT results across PSNR, SSIM, and LPIPS, alongside the best PSNR and LPIPS and competitive SSIM for generated view consistency, demonstrating both target-view fidelity and multi-view consistency.

Cross-identity reenactment. As shown in Fig. 5, X-NeMo transfers expressive facial motion, but its pose control is less accurate. Wan-Animate restricts expression transfer to 20 linear bases, limiting complex motions, while Wan-Animate-2 improves identity consistency through cross-identity training but may suffer from imperfect motion alignment. Our method combines high-dimensional motion latents with adversarial identity suppression, achieving accurate expression transfer and identity preservation, consistent with the quantitative results in Table 2.

4.3 Ablation Studies

Table 3: Core ablations. **(a)** GRL improves identity preservation while retaining motion accuracy. **(b)** Unified training improves animation quality and multi-view consistency. **(c)** Hybrid point-cloud and FLAME conditioning improves reshooting. P.C. denotes point-cloud renderings.

(a) Domain-Adversarial Motion Encoding

Variant	CSIM ↑	AED ↓	APD ↓
w/o GRL	0.555	0.0242	1.789
with GRL	**0.703**	**0.0224**	**1.770**

(b) Unified Training for Multi-view Animation

Variant	CSIM ↑	AED ↓	APD ↓	MEt3R ↓
w/o Reshoot	0.658	0.0236	1.764	0.2896
Unified	**0.664**	**0.0226**	**1.473**	**0.2515**

(c) Reshooting Condition Architecture

Variant	PSNR ↑	SSIM ↑	LPIPS ↓
FLAME only	20.46	0.690	0.260
P.C. only	22.66	0.680	0.180
P.C. + FLAME	**23.38**	**0.770**	**0.170**

Domain-adversarial motion encoding. We first compare the same 512-dimensional motion encoder trained with and without GRL. As shown in Fig. 6(a), GRL improves reference-identity preservation during cross-identity animation. Table 3(a) further shows comparable APD and AED. These results support reduced identity leakage without an evident loss of pose or expression accuracy.

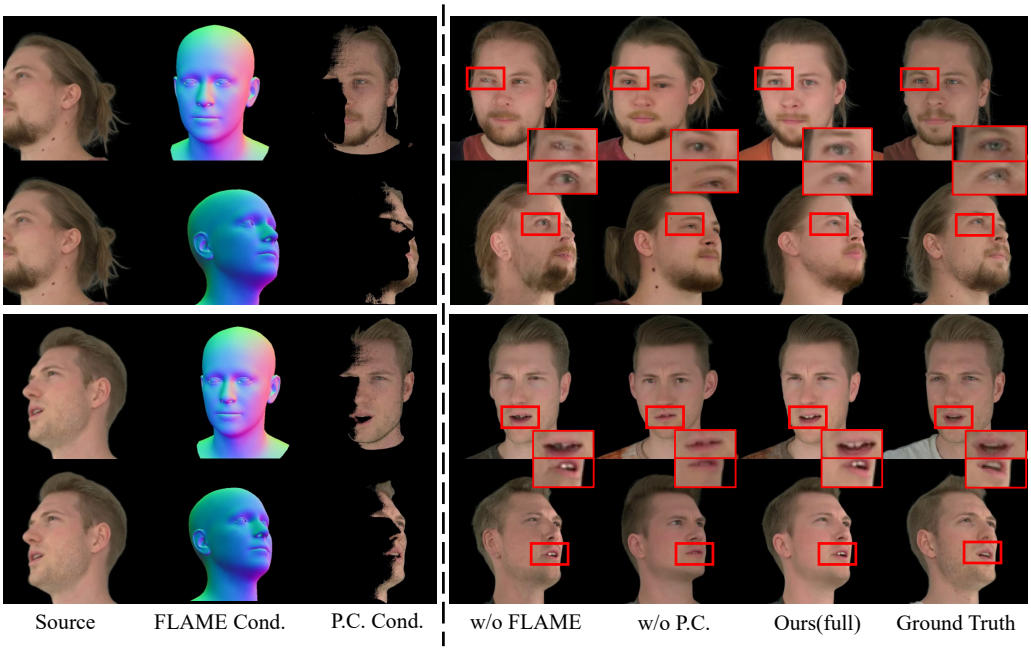

Figure 7: Condition-architecture ablation. Detailed point-render evidence and complete FLAME geometry provide complementary constraints for target-view fidelity and cross-view consistency.

Unified training. We compare the unified model with the model without reshooting training. Table 3(b) reports identity consistency and motion transfer accuracy together with MEt3R (Asim et al., 2025), which measures cross-view consistency between synchronized views and evaluates camera-supervised reshooting improves the cross-view consistency of multi-view animation.

Condition architecture. We define regions observed in the monocular input video as visible regions and all remaining areas as unseen regions. We find that point-cloud conditioning alone produces inconsistent predictions in unseen regions, whereas FLAME-only conditioning causes a severe loss of fine-grained detail in visible regions. As shown in Fig. 7, our hybrid point-cloud and FLAME conditioning injects complementary 3D-consistency cues into both visible and unseen regions, alleviating both failure modes. Table 3(c) reports the corresponding quantitative comparison.

5 Limitations

For reshooting, errors in the point cloud and FLAME estimation can propagate into generated views, especially under heavy occlusion, loose hair, or viewpoints with little source visibility. For animation, the lack of cartoon training data limits generalization to stylized portraits. Video generation also requires multiple denoising steps, so the complete generation-and-reconstruction pipeline is slow; distillation or a more efficient backbone could reduce this cost.

6 Conclusion

We introduced UniPortrait4D, a single video DiT for synchronized multi-view portrait-video reshooting and implicit-motion cross-identity animation. For reshooting, hybrid point-cloud and FLAME conditioning improves novel-view fidelity and supports downstream 4DGS reconstruction. For animation, domain-adversarial post-training makes the high-capacity expression latent less identity-predictive while retaining detailed facial motion. Experiments demonstrate that UniPortrait4D achieves strong cross-view consistency in reshooting and accurate, identity-preserving expression transfer in animation.

AI ASSISTANCE.

We used large language models to assist with language editing and manuscript organization, including improvements to grammar, clarity, and presentation. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript, technical claims, and references.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1021–1030. IEEE, 2017.

Hongrui Cai, Junjie Luo, Zhihong Fu, Shengnan Zhu, Jiawei Wen, Wanquan Feng, Songtao Zhao, and Qian He. Scaling4d: Pushing the frontier of video novel view synthesis through large-scale monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11184, 2026.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

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, 2019.

Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7119–7130. IEEE, 2024a.

Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pp. 316–333. Springer, 2024b.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8498–8507. IEEE, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Xuan Gao, Jingtao Zhou, Dongyu Liu, Yuqi Zhou, and Juyong Zhang. Controlling avatar diffusion with learnable gaussian embedding. *arXiv preprint arXiv:2503.15809*, 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.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Yudong Jin, Tao Xie, Qihang Zhang, Zehong Shen, Zhen Xu, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Yinghao Xu. 4danyone: Create anyone in 4d from a casual monocular video. *arXiv preprint arXiv:2608.20335*, 2026.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116. IEEE, 2020.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

Umut Kocasarı, Simon Giebenhain, Richard Shaw, and Matthias Nießner. Face anything: 4d face reconstruction from any image sequence. In *European Conference on Computer Vision*, pp. 269–287. Springer, 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 Trans. Graph.*, 36(6):194–1, 2017.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. Vista4d: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32671–32682, 2026.

Yukang Lin, Hokit Fung, Jianjin Xu, Zeping Ren, Adela SM Lau, Guosheng Yin, and Xiu Li. Mvportrait: Text-guided motion and emotion control for multi-view vivid portrait animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26242–26252. IEEE, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Weijie Lyu, Ming-Hsuan Yang, and Zhixin Shu. Facecam: Portrait video camera control via scale-aware conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30607–30617, 2026.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024.

Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024.

Pexels. Pexels. `https://www.pexels.com/`, 2026. Accessed: 2026-07-08.

Shenhan Qian. VHAP: Versatile head alignment with adaptive appearance priors, sep 2024. URL `https://github.com/ShenhanQian/VHAP`.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20299–20309. IEEE, 2024.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 130–141. IEEE, 2024.

Yuxiang Shi, Zhe Li, Yanwen Wang, Hao Zhu, Xun Cao, and Ligang Liu. Dex-portrait: Disentangled and expressive portrait animation via explicit and latent motion representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40009–40019, 2026.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Shuai Tan, Bill Gong, Bin Ji, and Ye Pan. Fixtalk: Taming identity leakage for high-quality talking head generation in extreme cases. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24–36. IEEE, 2025.

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, 2025a.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5318–5330. IEEE Computer Society, 2025b.

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. *arXiv preprint arXiv:2405.16204*, 2024a.

Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10336–10348. IEEE, 2024b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, et al. Wan-animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10034–10044. IEEE, 2021.

Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21750–21760. IEEE, 2025.

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.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665. IEEE, 2022.

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

Chao Xu, Xiaochen Zhao, Xiang Deng, Jingxiang Sun, Donglin Di, Zhuo Su, and Yebin Liu. Geodiff4d: Geometry-aware diffusion for 4d head avatar reconstruction. *arXiv preprint arXiv:2602.24161*, 2026a.

Zhufeng Xu, Xuan Gao, Feng-Lin Liu, Haoxian Zhang, Zhixue Fang, Yu-Kun Lai, Xiaoqiang Liu, Pengfei Wan, and Lin Gao. Im-animation: an implicit motion representation for identity-decoupled character animation. *arXiv preprint arXiv:2602.07498*, 2026b.

Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2062. IEEE, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shengkai Zhang, Nianhong Jiao, Tian Li, Chaojie Yang, Chenhui Xue, Boya Niu, and Jun Gao. Hellomeme: Integrating spatial knitting attentions to embed high-level and fidelity-rich conditions in diffusion models. *arXiv preprint arXiv:2410.22901*, 2024.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3647–3656. IEEE, 2022.

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. In *International Conference on Learning Representations*, volume 2025, pp. 77070–77088, 2025.

A Evaluation Metrics and Protocols

Generated-to-ground-truth fidelity. Generated-to-GT directly compares each generated target-view video with the captured video at the same camera using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). PSNR and SSIM measure pixel-level and structural agreement, respectively, while LPIPS measures perceptual distance. Higher PSNR and SSIM and lower LPIPS indicate better fidelity. We use the same paired image metrics for self-reenactment animation.

4DGS-to-ground-truth fidelity. We fit FreeTimeGS (Wang et al., 2025) to 16 of the 18 generated view videos, render the remaining two held-out cameras, and compare these renderings directly with the captured videos using PSNR, SSIM, and LPIPS. This protocol measures downstream reconstruction accuracy and penalizes generated views that are mutually consistent but geometrically inconsistent with the captured scene. As an oracle reference, we reconstruct FreeTimeGS directly from captured multi-view videos using the same cameras, view split, optimization budget, and evaluation code.

Generated-view consistency. GVC uses the same 16-view 4DGS reconstruction and two held-out cameras as 4DGS-to-GT, but compares the held-out 4DGS renderings with the corresponding generated videos rather than the captured videos. Outputs that cannot be represented coherently by a shared 4D representation score poorly under this protocol. GVC therefore measures internal cross-view consistency among generated views rather than fidelity to the captured scene. We use fixed view splits and identical reconstruction budgets for all methods.

Cross-identity animation metrics. We report ArcFace cosine identity similarity (CSIM) between generated frames and the reference identity. To evaluate motion transfer, we use MediaPipe Face Mesh (Kartynnik et al., 2019) to extract facial landmarks from both the generated and driving frames and derive their head poses. Average pose distance (APD) measures the distance between the estimated head poses, while average expression distance (AED) measures the distance between the normalized facial landmarks. Thus, higher CSIM indicates stronger reference-identity preservation, whereas lower APD and AED indicate more faithful transfer of the driving pose and expression.

B Implementation Details

Condition encoding. We use the video VAE encoder as E_c and apply it separately to the point-render and FLAME-normal sequences. Because the binary validity mask should preserve hard visibility labels, we align it without VAE encoding. Specifically, E_{mask} resizes each mask to the latent spatial resolution and, for every latent time step, folds the four masks corresponding to the VAE's $4\times$ temporal downsampling into the channel dimension. This produces a four-channel mask aligned with the point and normal latents.

Inherited self-supervised motion trainer. Following DeX-Portrait (Shi et al., 2026), the base trainer reconstructs a driving frame from a source–driving pair sampled from the same video, without pose or expression annotations. Following prior volumetric portrait animators (Wang et al., 2021; Guo et al., 2024; Drobyshev et al., 2024), the appearance encoder combines 2D and 3D CNNs to encode the source portrait into a volumetric feature of spatial size $16 \times 64 \times 64$. A ConvNeXt (Liu et al., 2022) pose encoder predicts the global similarity transform $P = [sR \mid \mathbf{t}]$ and warps this feature by $P_d P_s^{-1}$. In parallel, a FAN (Bulat & Tzimiropoulos, 2017) expression encoder extracts a 512-dimensional latent from a MediaPipe-aligned 224×224 face crop. Random rotation and cross-view sampling reduce pose leakage into this latent, while masking the eye and mouth regions in the pose input using MediaPipe landmarks (Kartynnik et al., 2019) reduces expression leakage into the pose estimate. The original and warped 3D features are concatenated, processed by a 3D residual block, and flattened into 2D appearance features. A StyleGAN2-like generator (Karras et al., 2020) then decodes these features, with the expression latent injected through AdaIN (Huang & Belongie, 2017).

Motion-trainer losses. The base generation objective combines pixel reconstruction, perceptual, component-aware perceptual, and image-adversarial losses:

$$\mathcal{L}_{\text{gen}} = 10\mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + 100\mathcal{L}_{\text{cLPIPS}} + \mathcal{L}_{\text{GAN}}. \tag{7}$$

Motion encoder training with GRL. The facial encoder receives 224×224 crops and produces 512-dimensional expression features. Training-time crop augmentation uses scale in $[0.76, 1.4]$, rotation in $[-45°, 45°]$, and temporally consistent color jitter. The identity probe is a six-layer, 512-wide MLP with leaky-ReLU activations and dropout. We first freeze the expression encoder and optimize only this probe to validation convergence. We then jointly optimize the probe, expression encoder, reconstruction pathway, and image discriminator; the probe minimizes identity-regression loss while the GRL sends its reversed gradient to the expression encoder. We set $\lambda = 5$. Finally, we freeze the resulting encoder, reinitialize the probe, and train it again to convergence for the reported leakage diagnostic. Training uses batch size 8, BF16 precision, expression-encoder learning rate 10^{-5}, probe learning rate 10^{-4}, and gradient-norm clipping at 1.

Unified generator. The base network is a 30-block, 5B-parameter video diffusion transformer with 3D patch size $(1, 2, 2)$ and a 3D causal VAE. We jointly train on monocular VFHQ (Xie et al., 2022) and synchronized multi-view NeRSemble (Kirschstein et al., 2023) for approximately 40,000 optimization steps. Training uses 512×512 resolution, BF16 precision, learning rate 10^{-5}, and gradient checkpointing. View attention is inserted after self-attention in every block. Expression cross-attention is inserted every two blocks. Text cross-attention and feed-forward layers inherited from the base model are frozen.

Condition dropout and guidance. Expression features are dropped with probability 0.1. During inference, we apply classifier-free guidance to the expression condition with a guidance scale of 4.5.

Inference and 4DGS reconstruction. We generate 18 synchronized views spanning frontal and bilateral profile directions together with elevated and depressed viewpoints. FreeTimeGS reconstruction from these views takes approximately 20 minutes per sequence. The resulting representation supports interactive real-time rendering.

C Double-Reprojection Details

We adapt the double-reprojection construction of TrajectoryCrafter (Yu et al., 2025) to create point-render conditions that are strictly aligned with their supervision frames. Let I_t be a target frame, $\mathcal{Q}_t$ its estimated colored point cloud in the target-camera coordinate system, and K_t its intrinsic matrix. We sample a relative camera transformation ΔT and first render the point cloud from this perturbed view:

$$(I'_t, D'_t, M'_t) = \Phi(\Delta T\, \mathcal{Q}_t, K_t), \tag{8}$$

where I'_t, D'_t, and M'_t are the rendered color, depth, and validity mask, respectively, and Φ denotes z-buffered point rasterization.

We then lift the valid pixels of the intermediate rendering back into a point cloud and return it to the original camera with the inverse transformation:

$$\mathcal{Q}'_t = \Phi^{-1}(I'_t, D'_t, M'_t, K_t), \qquad (P_t^{\text{dr}}, M_t^{\text{dr}}) = \Phi(\Delta T^{-1}\mathcal{Q}'_t, K_t). \tag{9}$$

Although the round trip restores the original camera, visibility changes in the perturbed view leave holes and boundary artifacts in P_t^{dr}. It therefore resembles an imperfect target-view point render while remaining pixel-aligned with I_t, allowing us to use $(P_t^{\text{dr}}, M_t^{\text{dr}})$ as the spatial condition and I_t as supervision. Finally, we apply the same crop transform A_t to both branches and update the intrinsics as $K'_t = A_t K_t$, ensuring that the condition and target frame share the same pixel grid.

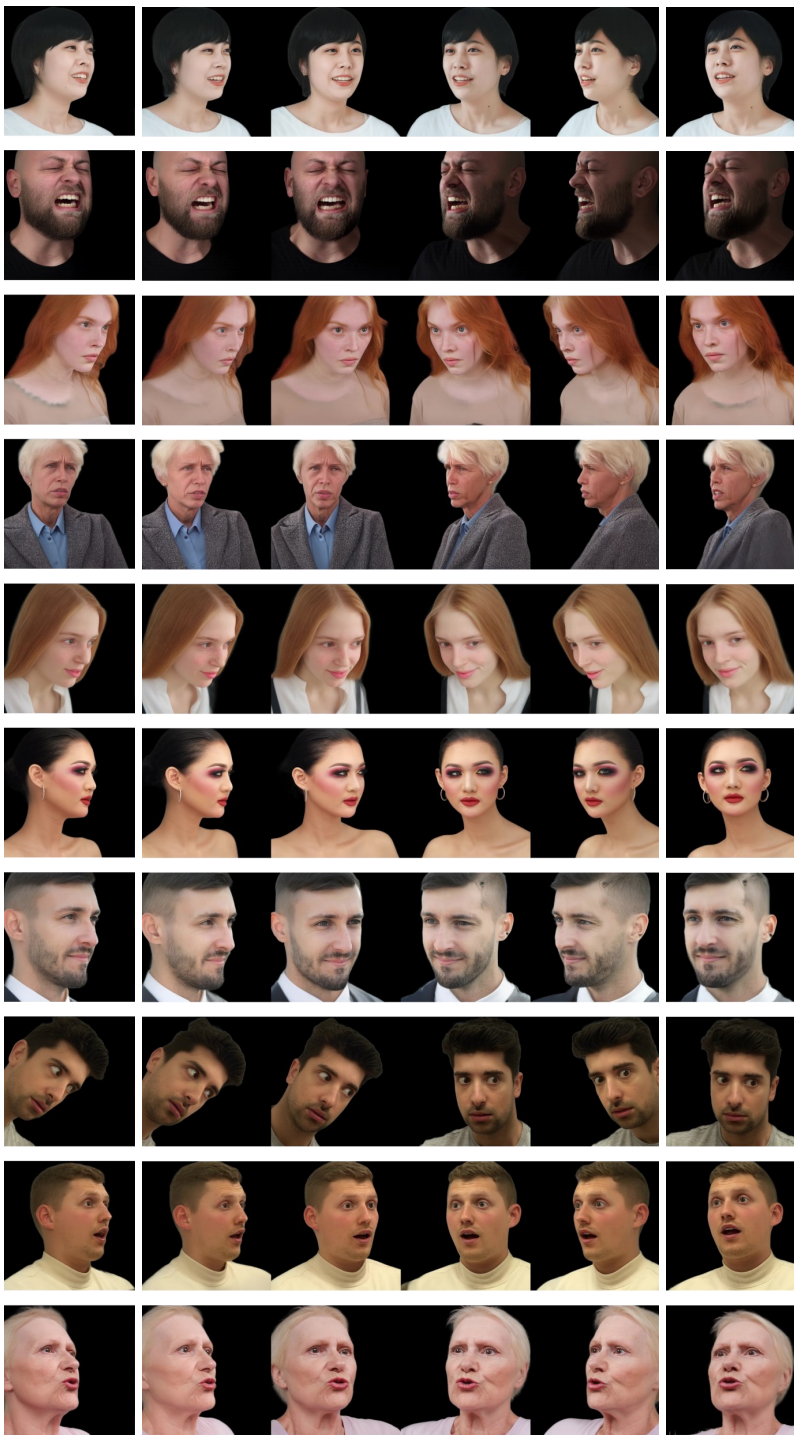

Figure 8: More results for video reshooting and 4DGS rendering. Left: source video; middle four columns: generated novel-view videos; right: corresponding 4DGS renderings.