This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ExpPortrait: Expressive Portrait Generation via Personalized Representation

Junyi Wang Yudong Guo* Boyang Guo Shengming Yang Juyong Zhang
University of Science and Technology of China
https://ustc3dv.github.io/ExpPortrait/



Figure 1. ExpPortrait utilizes a personalized head representation for portrait animation, achieving video generation with high consistency and high-fidelity. This stands in contrast to methods like Follow-Your-Emoji [25], which are constrained by low-rank and smooth intermediate representations.

Abstract

While diffusion models have shown great potential in portrait generation, generating expressive, coherent, and controllable cinematic portrait videos remains a significant challenge. Existing intermediate signals for portrait generation, such as 2D landmarks and parametric models, have limited disentanglement capabilities and cannot express personalized details due to their sparse or low-rank representation. Therefore, existing methods based on these models struggle to accurately preserve subject identity and expressions, hindering the generation of highly expressive portrait videos. To overcome these limitations, we propose a high-fidelity personalized head representation that more effectively disentangles expression and identity. This representation captures both static, subject-specific global geometry and dynamic, expression-related details. Furthermore, we introduce an expression transfer module to achieve personalized transfer of head pose and expression details between different identities. We use this sophisticated and highly expressive head model as a conditional signal to train a diffusion transformer (DiT)-based generator to synthesize richly detailed portrait videos. Extensive experiments on self- and cross-reenactment tasks demonstrate that our method outperforms previous models in terms of identity preservation, expression accuracy, and temporal stability, particularly in capturing fine-grained details of complex motion.

*Corresponding author.

1. Introduction

Digital portrait generation is a key technology in computer vision and computer graphics, powering a wide range of

applications from cinematic visual effects to lifelike virtual avatars in the metaverse. However, a major limitation of existing methods is the difficulty in achieving fine-grained, subtle expression control without compromising the consistency of the subject's identity. This paper aims to overcome this challenge by achieving precise control over subtle facial expressions while faithfully preserving identity.

Early controllable synthesis methods employed 2D landmarks [18, 34, 36, 42, 47] or parametric 3D models [9, 22, 23, 28] as intermediate motion signals, but both suffer from severe representational limitations. 2D landmarks-based methods are limited by severe signal sparsity, lacking the geometric details needed to define subject identity or subtle micro-expressions, and exhibit poor stability under significant pose variations. 3D parametric models [22, 28] serve as low-rank, linear approximations of the human face. Their predefined blendshapes are insufficient to handle high-frequency nonlinear dynamics (e.g., wrinkles), leading to severe confusion between identity and expression. This fundamentally hinders the generation of high-fidelity, expressive portraits that preserve identity.

While the rapid development of diffusion models has markedly advanced portrait video generation [14, 33, 49], existing methods still inherit the limitations of these flawed representations. Approaches using explicit 2D/3D parameters [3, 16, 30, 55] as control signals consequently struggle to capture high-frequency details or achieve accurate, identity-adaptive expression transfer. To address this, some studies [4, 45, 46, 53] have attempted to extract implicit motion features from the driving image via motion extractors, aiming to implicitly utilize learned motion information. Yet a major drawback is that the learned features are weakly controllable and insufficiently disentangled, leading to identity leakage and expression drift. Therefore, achieving robust identity preservation without compromising expressive fidelity remains an open and pressing problem.

As shown in Fig. 1, given that existing parametric head models fail to provide a superior intermediate proxy for expressive portrait generation, we aim to improve the information density and controllability of this intermediate signal to better serve this task. Our goal is to construct a personalized head representation that captures unique identity structures and dynamic expression details with high fidelity. We first use SMPL-X [28] mesh parameters as prior information. Based on this, we establish a disentangled relationship between identity and expression by optimizing two complementary offset fields: we learn a static per-vertex offset field for each subject to capture their unique high-frequency identity geometry. Meanwhile, we learn a dynamic per-vertex offset field designed to capture nonlinear skin deformations (e.g., wrinkles) corresponding to specific expressions. This constitutes a highly structured and detailed personalized head representation of identity and expression, significantly enhancing the expressive power of the head model.

However, applying this personalized, highly detailed head model to cross-identity expression transfer still faces compatibility challenges. The difficulty lies in how to adapt one subject's expression offset field to another subject's identity. We further construct an identity-adaptive expression transfer module to address this issue. This module takes the target subject's neutral identity mesh and target expression parameters as input. It employs a lightweight geometry MLP that predicts per-vertex dynamic offsets by conditioning on both the target's neutral geometry and an encoded representation of the driving signals. Finally, it outputs a mesh that presents the corresponding dynamic expression details, thus achieving vivid expression transfer without compromising identity. In summary, our contributions include the following aspects:

- We propose a personalized head representation with high disentanglement between identity and expression, capable of representing high-frequency geometric details and nonlinear facial deformations.
- We design an identity-adaptive expression transfer module to address the inherent incompatibility problem of per-subject learning, enabling the accurate transfer of dynamic facial expression details to the target subject while preserving identity.
- Based on the aforementioned personalized head model and expression transfer module, a diffusion generation model is trained, demonstrating state-of-the-art performance in identity preservation, expression richness, and controllability on both self- and cross-reenactment tasks.

2. Related Work

Face Reenactment. Most face reenactment techniques [8, 12, 27, 48] utilize a Generative Adversarial Network (GAN)-based generator [11], which uses facial cues or attributes, such as facial landmarks, as control signals. These approaches employ warping-based methods that attempt to learn implicit intermediate keypoints, which capture the facial motion between the source and target faces, thereby providing an alternative pathway for motion reenactment and transfer. However, these landmark representations tend to retain the target identity's facial structure, leading to identity drift and unnatural deformations when handling large pose differences or cross-identity reenactment. To improve geometric localization and reduce identity leakage, some studies have utilized 3D morphable parametric face models [1, 22, 56], which can disentangle identity and expression, to preserve identity consistency across different subjects. Methods such as [5, 9, 31] enhance face models with detail displacements or emotion consistency losses to capture finer, detailed expressions. Other methods attempt to encode facial expressions into a latent vector and

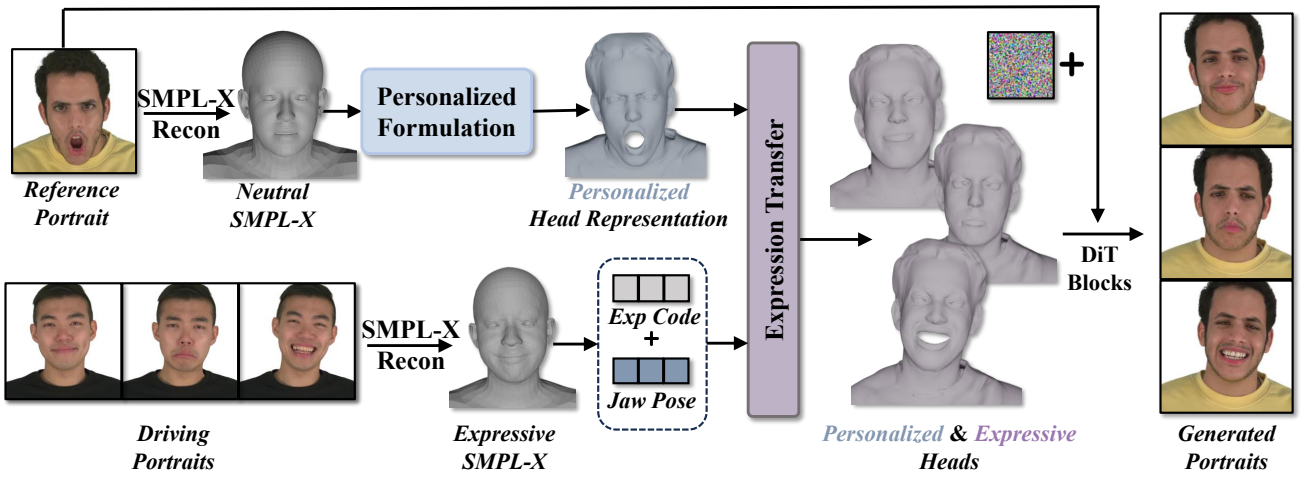


Figure 2. Our framework. To address the limited decoupling capability and insufficient expressiveness of current parametric head representations, we propose a personalized head representation. Starting from the SMPL-X base mesh, we perform joint optimization learning of two complementary static and dynamic offset fields. We then construct an identity-adaptive expression transfer module to achieve cross-identity expression transfer. Using our head representation as a control signal, we guide a diffusion model for highly consistent and expressive portrait video generation.

inject it into the generator network, or utilize Neural Radiance Fields (NeRF) and 3D Gaussians for head reconstruction and animation [7, 13, 17, 26, 43], but the consistency and rendering quality are still not fully satisfactory. The low-dimensional template subspace of the parametric face representations relied upon by these methods restricts their expressive power, hindering the capture of subject-specific anatomical structures or dynamic wrinkles and thus failing to achieve a sufficient balance between consistency and expressiveness.

Diffusion-based Portrait Generation. Diffusion models [2, 14, 15] have demonstrated remarkable capabilities in generating diverse and high-fidelity results, significantly impacting portrait video generation. Within this domain, a primary challenge lies in effectively controlling the subject's motion while preserving their identity. A significant body of work conditions these diffusion models on explicit human representations to guide the generation. These control signals range from 2D sparse landmarks to rendered parameters from 3D parametric models. For instance, [3, 21, 25, 37] utilize landmarks as spatial cues, while [32, 40, 41] employ 3DMM-derived maps to steer a latent diffusion model, prioritizing temporally consistent reenactment. While this explicit paradigm affords strong controllability, it is often constrained by the expressivity ceiling of the chosen representation. Sparse 2D controls can under-determine the full facial geometry and subtle expressions. Similarly, the low-rank subspaces of 3D templates struggle to model highly person-specific shapes or exaggerated expressions, leading to unnatural deformations. A parallel research thrust explores implicit motion control [10, 46, 53]. These methods typically employ a motion encoder to extract driving dynamics directly from video frames, injecting these latent features into the diffusion backbone. While this implicit method can generate highly realistic results, it raises significant concerns regarding incomplete disentanglement. The learned latent motion features may inadvertently capture appearance cues from the driver, leading to identity leakage and compromising the separation of motion from appearance.

3. Methodology

Given a reference portrait image I_R and a driving video sequence $I_D = \{I_{D_i}\}_{i=1}^{F}$, our task is to generate a video $I_G = \{I_{G_i}\}_{i=1}^{F}$ where the subject from I_R reenacts the poses and expressions of the subject in I_D, and F is the number of frames. The result should preserve the reference identity with high fidelity while faithfully and expressively transferring the motion.

To establish a faithful mapping from the driving frames to the reference identity, we first build a personalized head representation that captures the subject's identity and expression space (Sec. 3.1). We then introduce an identity-dependent expression transfer module to robustly transfer poses and expressions across identities (Sec. 3.2). Finally, we fine-tune a pretrained video diffusion model [38], conditioning it on our personalized, detail-rich head representation to synthesize the final high-fidelity video (Sec. 3.3).

The overall pipeline is shown in Fig. 2.

3.1. Personalized Head Representation

Parametric representations, such as SMPL-X and FLAME, provide effective initializations for portrait generation and reenactment. Their blendshape structure promotes identity consistency and expressive control. However, as low-dimensional statistical models, they fail to capture high-frequency, subject-specific details. We therefore exploit SMPL-X's 3D consistency by using its pre-tracked parameters as a coarse base layer, and augmenting them with decoupled, high-fidelity identity and expression details.

3.1.1. Representation Formulation

As shown in Fig. 3, our representation formulation process is as follows: Let $\mathcal{M}$ denote the function that returns mesh vertices given SMPL-X parameters. We first obtain a canonical neutral mesh for a given identity using only its shape parameters $\beta \in \mathbb{R}^{300}$, with the expression coefficients ψ and jaw poses ω set to zero:

$$V = \mathcal{M}(\beta)\big|_{\mathcal{P}} \in \mathbb{R}^{N\times 3}, \tag{1}$$

where $\mathcal{P}$ selects the head-and-shoulders vertex subset with N vertices. Because this mesh is too sparse to capture person-specific geometry and subtle expressions, we upsample it to a dense mesh. Let $\mathcal{B}$ be a fixed barycentric interpolation operator that maps the coarse topology to a high-resolution topology with N_s vertices ($N_s \gg N$):

$$V^s = \mathcal{B}(V) \in \mathbb{R}^{N_s\times 3}. \tag{2}$$

Decoupled Detail Fields. We model high-frequency details with two displacement fields on the dense mesh:

$$\Delta_g^s \in \mathbb{R}^{N_s\times 3}, \qquad \Delta_f^s(i) \in \mathbb{R}^{N_s\times 3}.$$

Here, Δ_g^s is a global per-vertex, identity-dependent offset that captures expression-invariant geometric details, while $\Delta_f^s(i)$ captures expression-dependent facial details for frame i. To make the decoupled detail fields optimizable even from a single reference image, we constrain Δ_g^s to mainly deform the non-facial regions (e.g., hair and clothing regions), and $\Delta_f^s(i)$ to affect only the facial region. In this way, the detailed canonical mesh for frame i is:

$$\widetilde{V}^s(i) = V^s + \Delta_g^s + \Delta_f^s(i). \tag{3}$$

Posing and Projection. Let $\mathcal{T}_i$ denote the SMPL-X Linear Blend Skinning (LBS) operator at frame i, which applies a rigid transformation to each vertex based on its skinning weights. We obtain dense skinning weights for the high-resolution mesh by barycentrically interpolating the base SMPL-X weights. LBS is then applied to the detailed canonical vertices:

$$\widetilde{V}_p^s(i) = \mathcal{T}_i\Big(\widetilde{V}^s(i)\Big) \in \mathbb{R}^{N_s\times 3}. \tag{4}$$

Finally, given camera parameters $\mathbf{c}$, we project the posed vertices to the image plane using the projection operator Π.

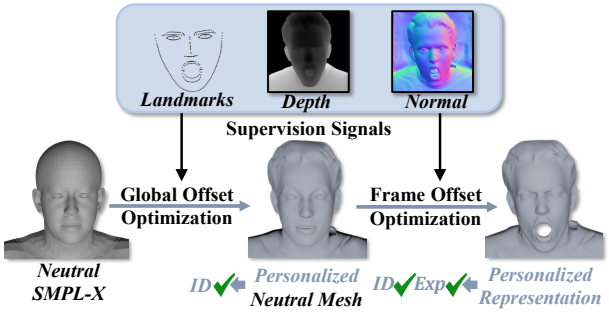


Figure 3. An illustration of our optimization pipeline for transforming a generic SMPL-X mesh into our highly detailed, personalized head representation.

3.1.2. Optimization

To supervise the 3D representation in image space, we render geometric buffers from the posed dense mesh. A differentiable renderer $\mathcal{R}$ [20] takes $\widetilde{V}_p^s(i)$ and camera $\mathbf{c}$ as input, and outputs per-frame renderings:

$$(\hat{N}_i,\ \hat{D}_i) = \mathcal{R}\Big(\widetilde{V}_p^s(i),\ \mathbf{c}\Big),$$

where $\hat{N}_i$ is the rendered normal map and $\hat{D}_i$ is the rendered depth map.

Supervision Signals. We impose sparse landmarks and pixel-wise dense labels as supervision signals. Let $L_{3D}(i)$ be the selected landmarks from the posed mesh, the landmark loss is defined as:

$$\mathcal{L}_{\text{ldmk}} = \big\|\Pi(L_{3D}(i),\ \mathbf{c}) - L_{2D}(i)\big\|_2^2, \tag{5}$$

where $L_{2D}(i)$ is detected with a keypoint detector [18]. For dense supervision, we compare the rendered predictions $(\hat{N}_i, \hat{D}_i)$ against estimated targets (N_i, D_i) from [35]:

$$\mathcal{L}_{\text{normal}} = \|\hat{N}_i - N_i\|_1, \qquad \mathcal{L}_{\text{depth}} = \|\hat{D}_i - D_i\|_1. \tag{6}$$

Geometric Regularization. We apply an ℓ_2 prior on the per-frame expression coefficients:

$$\mathcal{L}_{\text{exp}} = \|\psi_i\|_2^2. \tag{7}$$

To ensure physical plausibility and suppress high-frequency artifacts, we also apply a displacement-magnitude penalty $\mathcal{L}_{\text{dis}}$ and a Laplacian smoothness term $\mathcal{L}_{\text{lap}}$ on both detail fields. The total loss $\mathcal{L}$ is a weighted sum of the above terms.

3.1.3. Disentanglement Principle

A key challenge in our formulation is that the decomposition in Eq. (3) is ill-posed since static identity details could be ambiguously baked into either Δ_g^s or all $\Delta_f^s(i)$ fields. Therefore, we use spatial constraints (facial vs. non-facial) as the initial prior, and on this basis, we introduce an explicit temporal regularizer to force disentanglement and prevent identity information from being leaked into the frame-by-frame dynamics.

Our guiding principle is that for a given video sequence, the static field Δ_g^s should capture the *average* personalized geometry, while the dynamic fields $\Delta_f^s(i)$ should represent the *minimal, zero-mean deviations* from that average. To enforce this, the Δ_g^s is initialized to be non-zero in expected static regions as a prior, and we apply a minimal magnitude penalty to each per-frame offset $\Delta_f^s(i)$. This encourages the optimizer to explain all static and shared geometry using the shared Δ_g^s field, effectively forcing Δ_g^s to represent the true geometric average, while $\Delta_f^s(i)$ models only the expressive dynamics.

3.2. Expression Transfer Module

As shown in Sec. 3.1, personalized head details are learned in an identity-specific manner. Directly applying the expression offsets of one identity to another (even after alignment) leads to identity-expression mismatch. For example, a child should not inherit the deep wrinkle patterns of an elderly subject. To address this, we introduce an identity-adaptive expression transfer module that renders the same driving expression in a way that is anatomically consistent for the target identity, avoiding artifacts and ambiguity.

As shown in Fig. 4, the module has two parts: (i) an expression encoder that encodes driving signals, and (ii) a vertex-wise geometry MLP that predicts per-vertex dynamic offsets conditioned on the encoded expression and the target identity geometry.

Driving Signals Encoder. Let the driving expression coefficients over F frames be $\boldsymbol{\psi} \in \mathbb{R}^{F\times 100}$ and the jaw poses be $\boldsymbol{\omega} \in \mathbb{R}^{F\times 3}$. We feed them into an encoder $\mathcal{E}$ to obtain per-frame conditioning codes:

$$Q \;=\; \mathcal{E}\big(\boldsymbol{\psi},\, \boldsymbol{\omega}\big) \;\in\; \mathbb{R}^{F\times D}, \tag{8}$$

where $Q = \{q_i\}_{i=1}^{F}$ and $q_i \in \mathbb{R}^D$ summarizes the driving motion at frame i.

Details Prediction Network. We use a lightweight MLP to predict expression-dependent per-vertex offsets and add them to the static identity mesh. Consistent with Sec. 3.1, let

$$V_{\text{neutral}} \;=\; V^s + \Delta_g^s \;\in\; \mathbb{R}^{N_s\times 3},$$

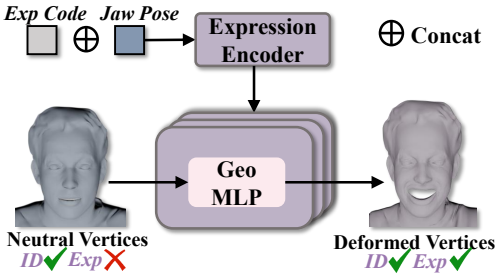


Figure 4. Design of our Expression Transfer Module.

where V^s is the dense canonical mesh and Δ_g^s is the static, identity-dependent offset. The dynamic offsets are predicted by:

$$\Delta_f^s(i) \;=\; \mathcal{G}\big(V_{\text{neutral}},\, q_i\big) \;\in\; \mathbb{R}^{N_s\times 3}. \tag{9}$$

By predicting dynamic offsets conditioned on both the driving code q_i and the target identity's neutral geometry, the module transfers expressions across identities while adapting high-frequency details (e.g., wrinkle placement and intensity) to the target's anatomy, resolving identity-expression incompatibilities and yielding robust expression transfer.

3.3. Video Diffusion Model

To demonstrate the utility of our head representation, we fine-tune a pretrained video generation model [38], implemented as a Diffusion Transformer (DiT) [29] in the latent diffusion framework (LDM) [33].

Control Signals. Given the reference image I_R, we build the personalized neutral head mesh V_{neutral} (Sec. 3.1) and render a reference normal map:

$$N^R \;=\; \mathcal{R}(V_{\text{neutral}},\, \mathbf{c})\,.$$

For the driving sequence $I_D = \{I_{D_i}\}_{i=1}^{F}$, we use the expression transfer module (Sec. 3.2) to obtain $\widetilde{V}^s(i)$ for each frame i, pose it with LBS to $\widetilde{V}_p^s(i)$, and render the corresponding driving normal maps:

$$N_{1:F}^D \;=\; \Big\{\, \mathcal{R}\Big(\widetilde{V}_p^s(i),\, \mathbf{c}\Big) \,\Big\}_{i=1}^{F}.$$

Conditioned Denoising. Following [16, 39, 51], a 3D convolutional pose encoder extracts spatio-temporal features from $N_{1:F}^D$, while a 2D convolutional reference encoder extracts appearance cues from N^R. Let z_0 denote the clean video latent from the Causal VAE [38] and z_t its noised version at timestep t. We patchify the control features from $N_{1:F}^D$ and concatenate them with the patchified tokens of z_t (together with the reference features from N^R) to form the conditioning input c for the DiT backbone.

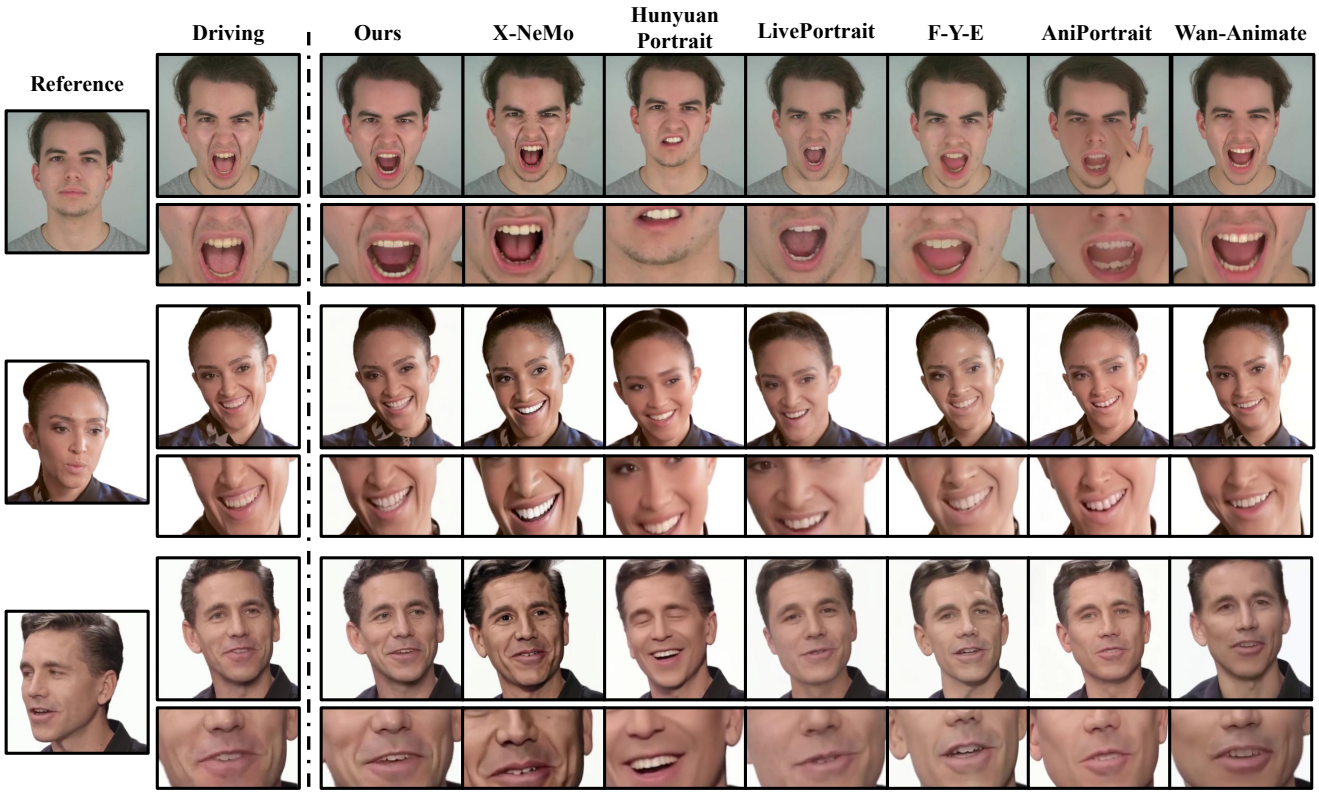


Figure 5. Qualitative results in self-reenactment. Compared to other methods, our method can reveal more details about identity and facial expressions.

We adopt the standard LDM noise-prediction loss. The DiT denoiser ϵ_θ predicts the added noise ϵ:

$$\mathcal{L}_{\text{ldm}} = \mathbb{E}_{z_0,\, \epsilon\sim\mathcal{N}(0,1),\, t}\Big[\, \big\|\epsilon - \epsilon_\theta(z_t,\, t,\, c)\big\|_2^2 \,\Big], \qquad (10)$$

where c encodes the control signals described above. In the cross-driving setting, $N^D_{1:F}$ is rendered from the expression-transferred meshes derived by combining the driving SMPL-X expression parameters with the reference identity's personalized neutral mesh, ensuring identity-consistent yet expressive guidance.

4. Experiments

4.1. Implementation Details

We collect and process 4,000 videos from VFHQ [44], CelebV-HQ [54], and HDTF [52] datasets, totaling approximately 10 hours of footage. All frames are cropped and resized to 512×512. We then apply SMPL-X reconstruction and the joint geometric optimization described in Sec. 3.1 to obtain personalized head models. The expression parameters and personalized meshes extracted from all videos are used to train the identity-adaptive expression transfer module. Subsequently, the full set of personalized, detailed head representations together with the original videos serves as training data for the diffusion model, during which the expression transfer module is frozen. We train for 30 epochs on $4\times$ NVIDIA A800 GPUs with a batch size of 1 per GPU and a learning rate of 10^{-4}.

Given the stringent requirements of film-grade digital humans regarding identity consistency, expressive fidelity, and high visual clarity, we conduct a fair evaluation on two datasets that are not used for training: RAVDESS [24] and NeRSemble [19]. These datasets feature rich facial expressions and high video quality. The RAVDESS test set includes 20 portrait videos, and the NeRSemble test set comprises 80 portrait videos. At inference, we sample frames with a temporal stride to assess generalization to head poses and expressions that differ from the reference image. All metrics are computed at 512×512 resolution.

4.2. Evaluations and Comparisons

To demonstrate the effectiveness of our method, we compare against prior portrait generation methods, including *LivePortrait* [12], an implicit keypoint-based approach; *AniPortrait* [40] and *Follow-Your-Emoji (F-Y-E)* [25], which use FLAME- and keypoint-driven explicit control;

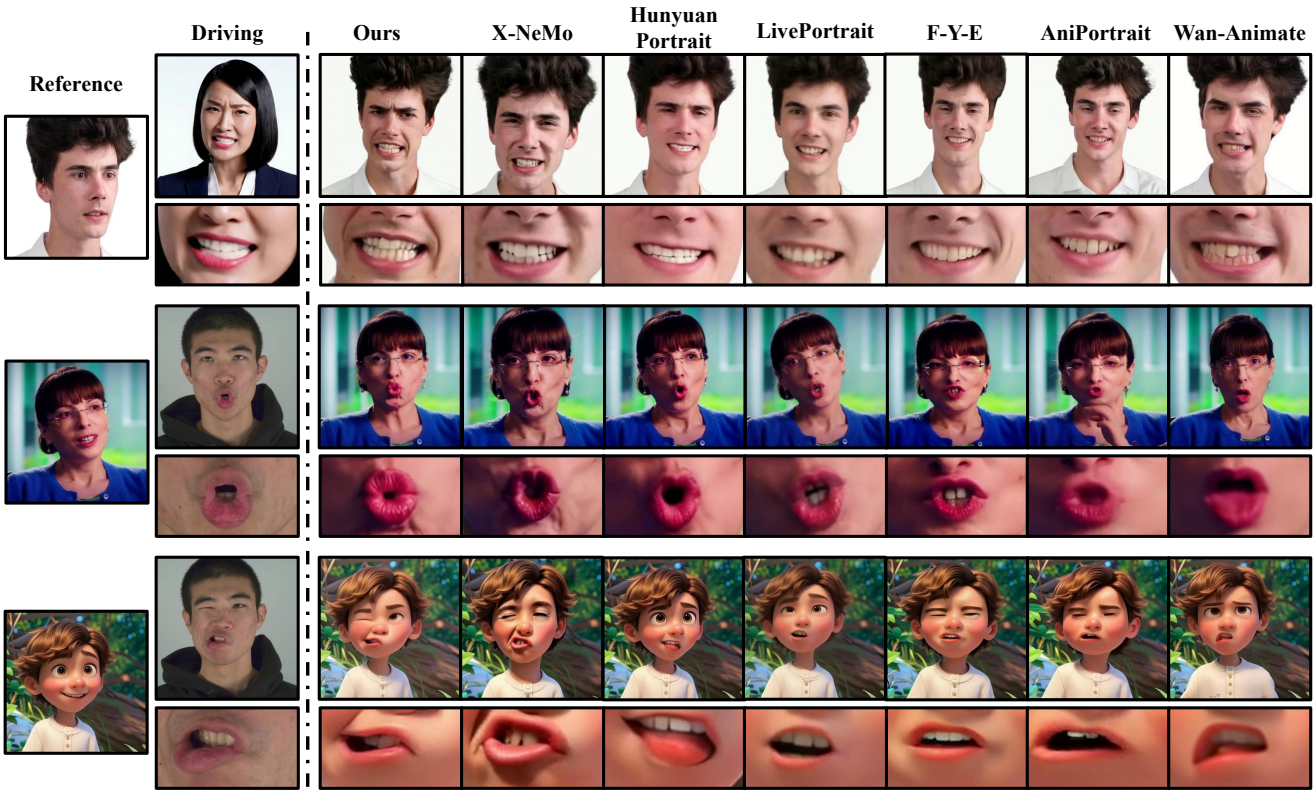


Figure 6. Qualitative results in cross-reenactment. Our method can effectively transfer body posture and facial expressions while ensuring identity consistency.

Table 1. Quantitative comparison. Our method achieves superior numerical results to all the baselines in self- and cross-reenactment tasks, evaluated on an image resolution of 512×512.

Method	Self Reenactment							Cross Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓	AED ↓	APD ↓	CSIM ↑	AED ↓	APD ↓	CSIM ↑
LivePortrait [12]	23.2897	0.82985	0.37339	0.04619	**0.12924**	0.02144	0.82972	0.28641	0.23048	0.72869
AniPortrait [40]	22.2527	0.80838	0.30018	0.04078	0.17581	0.01631	0.76038	0.25256	0.02161	0.62998
Follow-Your-Emoji [25]	25.6872	0.84146	0.23645	0.02905	0.14717	0.01485	0.80284	0.22069	0.02338	0.68373
Hunyuan Portrait [46]	22.9453	0.79333	0.27573	0.03962	0.15593	0.02329	0.81150	0.22753	0.08784	0.67451
X-NeMo [53]	21.5618	0.78051	0.32411	0.04760	0.13671	0.01789	0.83018	**0.17092**	0.02101	0.72164
Wan-Animate [4]	26.5109	0.84360	0.21397	0.02183	0.14415	0.00955	0.82566	0.22156	0.01745	0.72637
Ours	**26.5507**	**0.85908**	**0.18394**	**0.02160**	0.13185	**0.00948**	**0.83506**	0.21092	**0.01277**	**0.72917**

X-NeMo [53], which relies on implicit latent feature control from a driving image; as well as *Wan-Animate* [4] and *Hunyuan Portrait* [46], which utilize explicit pose representations and implicit expression features for control.

Following *Learn2Control* [10], we conduct both qualitative and quantitative evaluations to assess video quality and motion accuracy. For self-reenactment, we report PSNR, SSIM, LPIPS [50], and ℓ_1 distance. Identity preservation is measured by cosine similarity (CSIM) between face recognition embeddings [6]. Motion accuracy is quantified by the Average Expression Distance (AED) and Average Pose Distance (APD). For cross-reenactment, we report AED, APD, and CSIM.

Quantitative Comparison. As demonstrated in Tab. 1, our method achieves state-of-the-art performance, quantitatively outperforming all baselines in both self- and cross-reenactment tasks. In self-reenactment, our approach produces results with significantly higher reconstruction fidelity while more accurately preserving the target's pose and identity. For the more challenging cross-reenactment

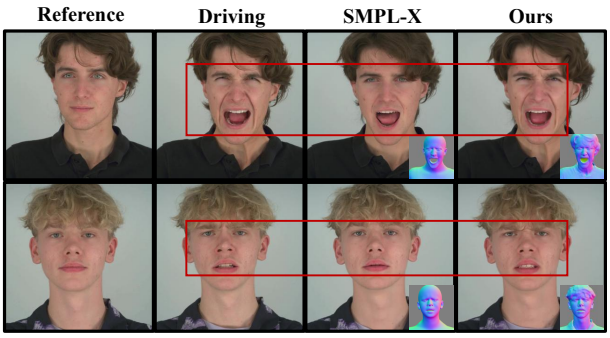


Figure 7. Ablation study between SMPL-X and our head representation.

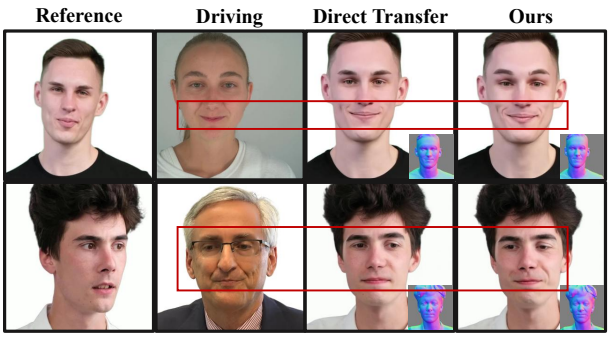


Figure 8. Ablation study between directly performing deformation with offsets and using our expression transfer module.

task, our method demonstrates a superior ability to faithfully transfer motion and expression while robustly preserving the source subject's identity.

Qualitative Analysis. As illustrated in the qualitative comparisons (Figs. 5 and 6), existing methods exhibit distinct limitations that underscore the necessity of our personalized head representation. Explicit-control baselines, such as *AniPortrait* and *F-Y-E*, rely on sparse, low-rank parametric signals. These signals are insufficient for capturing fine-grained details, resulting in animations with limited expressiveness and weaker identity preservation. In contrast, implicit-control approaches, including *LivePortrait*, *Hunyuan Portrait*, and *X-NeMo*, utilize latent motion features to achieve greater expressiveness. However, this paradigm introduces severe trade-offs: the control over head pose is often imprecise, and the motion features frequently entangle expression dynamics with identity attributes. This critical entanglement leads to identity leakage in both self-reenactment (via identity drift) and cross-reenactment (via identity bleed-through from the driver). Notably, while *Wan-Animate* demonstrates that simply scaling up data and model parameters can enhance generalization capabilities, it fails to significantly improve fine-grained controllability or the decoupling of identity and expression. As shown in Fig. 6, *Wan-Animate* still falls short in precise expression control and identity decoupling. Benefiting from our personalized and disentangled 3D head representation, our *ExpPortrait* effectively overcomes these issues. It successfully decouples identity from expression, enabling the simultaneous generation of high-fidelity, high-expressiveness, and highly-controllable portrait animations that faithfully preserve identity while accurately reproducing nuanced expressions.

4.3. Ablation Studies

We assess the efficacy of our personalized head representation via an ablation against a SMPL-X baseline. The baseline uses the standard SMPL-X mesh as the sole 3D head representation, whereas our full model employs the proposed personalized, high-detail head proxy. For a fair comparison, both variants share the same diffusion architecture, training data, and hyperparameters; the only difference is the underlying 3D head representation. As shown in Fig. 7, the SMPL-X baseline struggles to capture fine facial nuances and exhibits limited facial dynamics, often producing rigid or muted expressions. In contrast, our representation delivers markedly better expressiveness and identity fidelity, recovering a wider and more accurate range of personalized identity cues and subtle expression details.

We also conduct an ablation study on our expression transfer module, benchmarking it against a strategy that directly applies expression offsets to perform deformation on the reference subject's head representation. As illustrated in Fig. 8, this direct transfer approach exhibits significant drawbacks: it yields muted facial expressiveness. In contrast, our proposed expression transfer module is capable of generating more vivid and realistic expressions.

5. Conclusion

In this work, we addressed the problem of expressive, controllable, and identity-preserving portrait video generation. We introduced a high-fidelity personalized head representation that disentangles identity-specific static geometry from frame-wise expression details, and an identity-adaptive expression transfer module. Conditioned on this 3D proxy, a DiT–based generator achieved state-of-the-art performance, leading to clear improvements in identity preservation and expression accuracy.

Our method still has limitations. The personalized representation does not explicitly model the inner mouth, hindering the precise generation of the tongue, nor does it capture fine-grained eyeball motion. Nevertheless, we believe this representation provides a promising foundation for natural extensions to full-body animation and other tasks, such as speech-driven animation and broader interactive avatar applications, which we leave for future investigation.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.62272433, No.62402468, No.U25A20390), Anhui Provincial Natural Science Foundation (No.2508085ZD011) and the Fundamental Research Funds for the Central Universities.

References

[1] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 2

[2] Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Human-centric video generation via collaborative multi-modal conditioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2939–2947, 2026. 3

[3] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2410, 2025. 2, 3

[4] 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. 2, 7

[5] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

[6] 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*, pages 4690–4699, 2019. 7

[7] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 3

[8] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 14398–14407, 2021. 2

[9] 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 (ToG)*, 40(4): 1–13, 2021. 2

[10] Xuan Gao, Jingtao Zhou, Dongyu Liu, Yuqi Zhou, and Juyong Zhang. Constructing diffusion avatar with learnable embeddings. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–13, 2025. 3, 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[12] 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. 2, 6, 7

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021. 3

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 2, 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[16] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8153–8163, 2024. 2, 5

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

[18] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 2, 4

[19] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 6

[20] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 4

[21] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[22] 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. 2

[23] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *NeurIPS*, 2023. 2

[24] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 6

[25] 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*, pages 1–12, 2024. 1, 3, 6, 7

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[27] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5

[30] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2

[31] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2

[32] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[34] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, 2013. 2

[35] Fatemeh Saleh, Sadegh Aliakbarian, Charlie Hewitt, Lohit Petikam, Xian Xiao, Antonio Criminisi, Thomas J Cashman, and Tadas Baltrusaitis. David: Data-efficient and accurate vision models from synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5348–5358, 2025. 4

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2377–2386, 2019. 2

[37] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 3

[38] 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. 3, 5

[39] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 2025. 5

[40] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 3, 6, 7

[41] Mengting Wei, Yante Li, Tuomas Varanka, Yan Jiang, and Guoying Zhao. Magicportrait: Temporally consistent face reenactment with 3d geometric guidance. *arXiv preprint arXiv:2504.21497*, 2025. 3

[42] Jiahao Xia, Min Xu, Wenjian Huang, Jianguo Zhang, Haimin Zhang, and Chunxia Xiao. Mitigating knowledge discrepancies among multiple datasets for task-agnostic unified face alignment. In *IJCV*, 2025. 2

[43] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 3

[44] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6

[45] 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*, pages 1–11, 2024. 2

[46] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15909–15919, 2025. 2, 3, 7

[47] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 2

[48] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 2

[49] Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu

Guo, and Jiashi Feng. Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*, 2023. 2

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[51] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. 5

[52] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3661–3670, 2021. 6

[53] 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 *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 7

[54] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 6

[55] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 2

[56] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 2