# Instant Expressive Gaussian Head Avatar via 3D-Aware Expression Distillation

Kaiwen Jiang[1]† Xueting Li[2] Seonwook Park[2] Ravi Ramamoorthi[1] Shalini De Mello[2] Koki Nagano[2]
[1]University of California, San Diego [2]NVIDIA

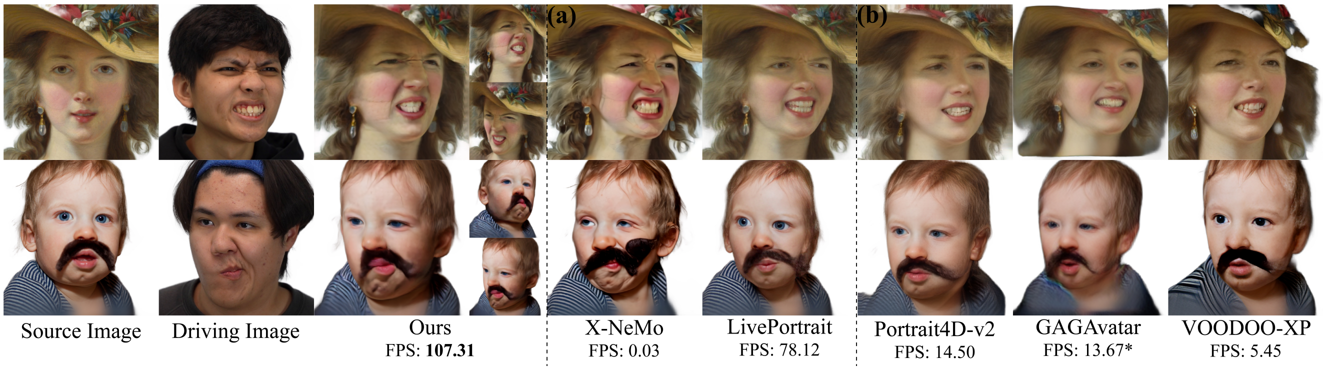

Figure 1. We present an instant feedforward encoder that transforms an in-the-wild source image into an animatable 3D avatar by distilling knowledge from a pre-trained 2D diffusion model. Our method introduces a fast, consistent yet expressive 3D animation representation. Given a driving image, we evaluate both the expression transfer quality and the animation speed (measured as "FPS" on an NVIDIA 6000 Ada GPU) against **(a)** 2D diffusion- or GAN-based methods and **(b)** 3D-aware methods. In the first row, Portrait4D-v2 [16], GAGAvatar [11] and VOODOO-XP [78] fail to faithfully transfer expressions, particularly around the nasal wrinkles. LivePortrait [26] is inaccurate at eyes. In the second row, the baby wears a fake mustache as a decoration. X-NeMo distorts identity and adds a hallucinated mustache. Other methods cannot deal with the asymmetric expression in the driving image well. The FPS marked with $*$ reports inference time excluding time-consuming morphable model fitting optimization required for the method. In contrast, ours not only accurately transfers expressions but also achieves high animation speed and consistent pose control. Insets show our rendered results under different poses.

## Abstract

*Portrait animation has witnessed tremendous quality improvements thanks to recent advances in video diffusion models. However, these 2D methods often compromise 3D consistency and speed, limiting their applicability in real-world scenarios, such as digital twins or telepresence. In contrast, 3D-aware facial animation feedforward methods – built upon explicit 3D representations, such as neural radiance fields or Gaussian splatting – ensure 3D consistency and achieve faster inference speed, but come with inferior expression details. In this paper, we aim to combine their strengths by distilling knowledge from a 2D diffusion-based method into a feed-forward encoder, which instantly converts an in-the-wild single image into a 3D-consistent, fast yet expressive animatable representation. Our animation representation is decoupled from the face's 3D representation and learns motion implicitly from data, eliminating the dependency on pre-defined parametric models that often constrain animation capabilities. Unlike previous computationally intensive global fusion mechanisms (e.g., multiple attention layers) for fusing 3D structural and animation information, our design employs an efficient lightweight local fusion strategy to achieve high animation expressivity. As a result, our method runs at 107.31 FPS for animation and pose control while achieving comparable animation quality to the state-of-the-art, surpassing alternative designs that trade speed for quality or vice versa.*

## 1. Introduction

Creating a digital twin from a single facial image that supports both 3D viewpoint control and animation (4D) is a long-standing goal in computer vision and graphics. Interactive synthesis and animation control of photorealistic digital humans is essential for developing AR/VR, video conferencing, and agentic AI applications.

Achieving such comprehensive 4D control has been historically challenging. With the advent of radiance fields including neural radiance fields (NeRFs) [51] and 3D Gaus-

†This project was initiated and substantially carried out during an internship at NVIDIA.

sians [32], previous 3D-aware face animation work [11, 12, 43, 71, 112] has achieved interactive and consistent animation with photo-realistic view synthesis by using parametric models [4, 20, 40]. However, parametric models inherently limit the animation capability. Follow-up methods [15, 16, 36, 78, 79] therefore resort to learning the animation purely from data. Nevertheless, their representations entangle 3D structure and animation (e.g., global residual triplanes [8] or feature maps), requiring computationally expensive attention mechanisms to repeatedly fuse 3D structure and motion at every animation step. Meanwhile, the introduction of 2D diffusion models [7, 10, 13, 26, 38, 49, 59, 60, 63, 92, 98, 113] into portrait animation has brought the expressivity of achieved facial animation to a whole new level [98, 113]. However, these methods often suffer from 3D inconsistency and remain slow due to the expensive denoising process, preventing them from being used in a real-time system. Fig. 2 shows a quantitative comparison where existing methods fail to excel at *all three criteria*: speed, 3D consistency and expression transfer accuracy.

Different from other slower optimization-based methods [1, 24, 73–75], we seek to design a 3D-aware animation framework that instantly encodes a facial image into an animatable 4D avatar, which supports fast, consistent and detailed animation. Our key insights to solving this problem are twofold. First, we argue that high-quality facial animation has already been effectively learned by 2D diffusion models, and this knowledge can be *distilled* into 3D-aware methods rather than learning from scratch. Our second insight is that achieving efficiency and 3D consistency without sacrificing expressiveness requires a novel animation representation.

Specifically, we build upon 3D Gaussians [32] that makes dynamic deformation more efficient than NeRFs' volumetric fields to propose an expressive yet efficient animation representation. We first encode an input image into triplanes [80] as an intermediate representation, from which we sample feature vectors to decode the Gaussian attributes. For each Gaussian, we encode its motion information in an auxiliary vector which is analogous to learned personalized "PCA bases" in traditional blendshapes animation. This auxiliary vector is then combined with the driving signal to deform each Gaussian on top of the existing 3D structure, updating the 3D representation while keeping the animation representation both efficient and decoupled from the underlying 3D representation. While previous works [45–47, 90] deform the 3D Gaussians in spatial space to model motion, we find it unable to capture expressive facial details. Instead, we propose to deform the Gaussians individually in the high-dimensional feature vector space, rather than in 3D spatial space, which offers better expressivity and is capable of capturing asymmetric expressions, details such as shadow changes and wrinkles (Fig. 1).

Typically, portrait animation methods, including ours, are trained with a self-reenactment objective using datasets that contain multiple expression-varying images per identity. Instead of training on real datasets, we construct a synthetic facial expression dataset by a state-of-the-art facial animation diffusion model [113] for distilling its expressive motion priors in a 3D consistent manner. To mitigate the potential 3D inconsistency issue, we synthesize facial expressions by the diffusion model on real face portraits frontalized by a pre-trained 3D lifting encoder [80] and again use the same 3D lifting encoder to estimate its multi-view synthetic images on the fly during training. In summary, our main contributions include:

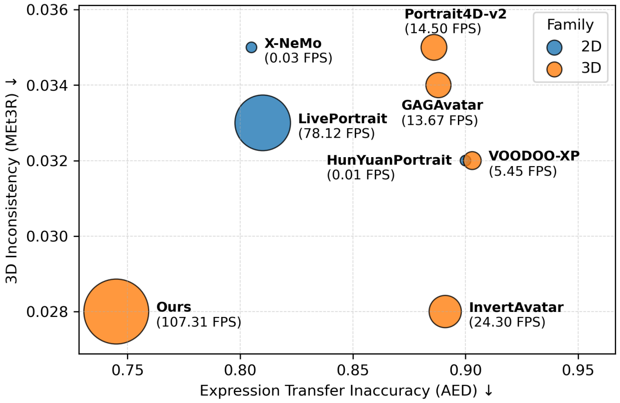

Figure 2. We provide a visualization of the quantitative comparison in terms of 3D inconsistency (measured by MEt3R ↓), expression transfer inaccuracy (measured by AED ↓) and animation speed (measured by FPS ↑, visualized as the size of the circle) with other 2D- or 3D-based baselines, including [11, 16, 26, 78, 98, 112, 113], using the task of cross-reenactment. 2D methods tend to appear on the upper left (better expression transfer accuracy; worse 3D consistency) while 3D methods tend to appear on the lower right (worse expression transfer accuracy; better 3D consistency). Our method is **3-4** orders of magnitude faster than diffusion based models [98, 113] while simultaneously achieving better 3D consistency and expression transfer accuracy.

- We design an expressive yet computationally-efficient animation representation for 3D Gaussians that achieves detailed animation for human faces (Sec. 4.1).
- We propose practical strategies to train such an animation representation by distilling the knowledge of existing diffusion methods into it (Sec. 4.2).
- Our method is the first to simultaneously achieve best 3D consistency, fast inference speed and detailed expressions such as wrinkles, and run orders of magnitude faster than 2D diffusion models during inference (Fig. 1, 2).

## 2. Related Work

**2D and 3D facial portrait animation.** 2D facial portrait animation methods often feature a generative backbone (e.g., GAN [25] or diffusion [28, 69]), which synthesizes driven faces given the control signals. GAN-based methods [5, 17–19, 22, 26, 44, 50, 53, 61, 66, 67, 83–86, 97, 104, 111, 114] feature fast inference using either explicit or implicit expression representations, but are limited by the capability of GAN models. A diffusion backbone–often pre-trained on large-scale internet data–provides much stronger synthesis capability, and has been employed for facial animation [9, 13, 48, 49, 56, 59, 76, 81, 82, 89, 92, 93, 96, 98, 102, 113], showing excellent expression transfer quality. However, the repeated denoising steps trade speed for quality and are thus prohibitive for real-time applications. They are also not 3D consistent. Transforming these diffusion models into single-step or few-step models is a promising direction but still an open problem [30, 64, 65, 105].

Another line of methods builds explicit 3D representations for 3D talking heads, which improve 3D consistency. They typically rely on 3D morphable models (3DMM) [4, 20, 23, 27, 41] or facial motion representations (e.g., rasterized coordinates [115] or facial keypoints) as priors [11, 24, 33, 42, 43, 68, 71, 72, 91, 94, 103, 112] for animating the face. However, morphable models inherently limit facial expressiveness, as their strong statistical priors, which are derived from a finite set of face scans and linear basis representations restrict motion to a narrow, predefined space. Therefore, another group of methods [15, 16, 78, 79] implicitly learns the motion as residual features to the triplanes [8] through data. Notably, Portrait4D [15] also distills from synthetic data. However, in their animation representation, the global residual features coupled with dense attention mechanisms are computationally expensive to infer. We instead propose a local fusion mechanism that individually deforms each 3D Gaussian through a lightweight MLP based on a learned auxiliary vector that encodes all the motion information. We compare to two representative animation representations [11, 16] by training them on our synthetic dataset and demonstrate the superiority of our animation representation design.

**4D representation and rendering.** Aiming at a native 4D representation, many works [21, 54, 55, 58] build upon NeRFs to construct a spatial deformation field, which, however, is computationally expensive. With the advent of Gaussian splatting [32], another family of methods identify the potential of using the spatial deformation of 3D Gaussians to represent the motion. One group of methods [24, 46, 47, 90] learn an implicit representation, such as HexPlanes [6] or coordinate-based MLP, to deform the Gaussians. Another group of methods [24, 27, 73, 109, 116] utilizes an explicit mesh to drive the 3D Gaussians. However, our analysis shows that directly deforming the 3D Gaussians in 3D space provides limited expressiveness for capturing fine-grained facial motions (Sec. 5.2).

Notably, Avat3r [36] and ScaffoldAvatar [1] also use 3D Gaussians to build animatable avatars. However, they are not a feed-forward method from a single image and requires either 3D GAN inversion or mesh tracking. Prior work [74, 75] uses diffusion-generated multi-view images for animated 3D head synthesis, but relies on hours-long optimization to fit a single avatar. We aim to develop a generalizable 3D-consistent framework that requires no tracking, supports instant encoding, and enables real-time expressive animation.

## 3. Preliminaries

**3D avatar encoder and volume rendering.** We choose the state-of-the-art facial 2D-to-3D lifting encoder [80] as our architectural backbone for lifting a single-view image into a 3D avatar. Given a single image $I$, the encoder encodes it into triplanes $\{T_{xy}, T_{yz}, T_{zx}\}$ [8], each of which is $\in \mathbb{R}^{256\times256\times32}$. These planes can then be used for rendering arbitrary viewpoints using volumetric rendering [51]. Specifically, for each queried 3D position $\mathbf{x} = (x, y, z)$, its corresponding feature vector $f(\mathbf{x}) \in \mathbb{R}^{32}$ is retrieved by projecting $\mathbf{x}$ onto each of the three planes via bilinear interpolation and further aggregation by summation. A light-weight non-linear multi-layer perceptron (MLP) decoder then decodes the aggregated features into colors and densities for volume rendering. In this work, we extend this approach by adapting this NeRF-based encoder for encoding a single-view image into a set of **3D Gaussians** as explained later.

**3D Gaussian splatting.** Kerbl et al. [32] provides a differentiable and efficient solution to rendering a set of anisotropic 3D Gaussians into images. Specifically, each 3D Gaussian is parametrized by its position vector $\mu \in \mathbb{R}^3$, scaling vector $\mathbf{s} \in \mathbb{R}^3$, quaternion vector $\mathbf{q} \in \mathbb{R}^4$, opacity $o \in \mathbb{R}$ and color $\mathbf{c} \in \mathbb{R}^3$. The final rendering color of a pixel is calculated by alpha-blending all 3D Gaussians overlapping the pixel. In the following discussion, we use the sub-script $i$ to denote that these quantities belong to the $i^{\text{th}}$ 3D Gaussian.

## 4. Method

**Overview.** Our overall training pipeline is shown in Fig. 3, which consists of (a) reconstruction modules (left) and (b) animation modules (right). Given a source image $I_s$, we train an instant encoder $E$, adapted from [80], to encode $I_s$ into a set of 3D Gaussians [32] for free viewpoint rendering (Sec. 4.1). For animation, we update the set of Gaussians conditioned on an expression from a driving image $I_d$, while preserving the appearance in $I_s$ (Sec. 4.1). We denote the encoding, animation and rendering procedure as $E(I_s, I_d, p)$ to synthesize a 2D image with identity from $I_s$ and expression from $I_d$ and at viewpoint $p$.

For training, we distill a pre-trained diffusion-based portrait animation model [113] by rendering a synthetic dataset from it (Sec. 4.2), in which each identity is represented by multiple images with different expressions. We then perform self-reenactment—using $I_d$ (the driving image) to drive $I_s$ (the source image) of the same identity—and optimize our model by minimizing the reconstruction loss between the reconstructed result and $I_s$, and the driven result and $I_d$ (Sec. 4.2).

During *inference*, we directly input $I_s$ into our encoder $E$ once to reconstruct its 3D representation, and then animate the resulting set of Gaussians according to any given

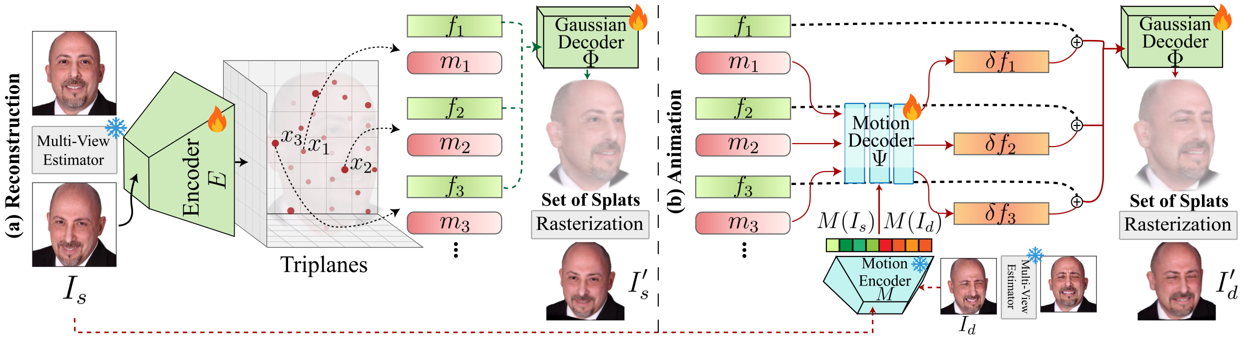

Figure 3. Overview of our **training** pipeline with the two-part self-reenactment task. **(a) Reconstruction**: Given a frontalized source frame with an expression synthesized by a pre-trained diffusion model [113], we first use a multi-view estimator [80] to generate its another viewpoint $I_s$. The encoder $E$ converts $I_s$ into triplanes, from which we sample feature vectors $f_1, f_2, \ldots$ and paired motion basis vectors $m_1, m_2, \ldots$. A Gaussian decoder $\Phi$ maps these features into a set of 3D Gaussians, forming a lifted 3D avatar for $I_s$, which we render at the viewpoint of $I_s$ as $I'_s$. **(b) Animation:** For the synthesized driving frame of the same identity but with a different expression, we similarly obtain its another viewpoint image $I_d$. Both $I_s$ and $I_d$ are input into the motion encoder $M$ to produce motion coefficients $M(I_s)$ and $M(I_d)$. They are concatenated to condition a motion decoder $\Psi$ to predict residual features $\delta f_1, \delta f_2, \ldots$ from paired motion basis vectors. Adding these residuals to the original features and decoding them with $\Phi$ yields an animated set of Gaussians, which we render at the viewpoint of $I_d$ as $I'_d$. The loss is computed between $(I_s, I'_s)$ and $(I_d, I'_d)$. Fire icons denote trainable modules; snow icons denote frozen pre-trained modules.

driving image $I_d$, whose identity may differ from that of $I_s$. Note that our animation pipeline–consisting of compact MLPs–does not require re-encoding the expensive 3D representation from $I_s$ and $I_d$ as in [16, 36, 78], leading to faster inference. The rendering of the Gaussians from arbitrary viewpoints is realized through rasterization [32].

### 4.1. Encoder Design and Animation Representation

**3D Gaussians decoder.** Even though the 3D lifting encoder in [80] is capable of turning a single-view image into a native 3D avatar, its implicit radiance field representation is less ideal for representing dynamics as opposed to 3D Gaussians. 3D Gaussians offer more flexible control over each primitive and fast rendering speed while maintaining 3D consistency [29].

Therefore, we make the minimal change to adapt the architecture of [80] into using 3D Guassians while preserving its strong capacity in faithfully lifting 2D images into 3D. We propose to sample 3D Gaussians from the encoded triplanes, as explored in [3]. We will first explain how we decode 3D Guassians from sampling locations and then clarify how we decide the sampling locations.

We first use 96 channels for the encoded triplanes. Given a sampled location $\mathbf{x}_i$, we project it onto each plane and aggregate along the channel dimension to obtain the feature vector $f_i$ using the first 48 channels and retrieve a vector $\mathbf{m}_i$ using the remaining 48 channels (see Fig. 3). We call $\mathbf{m}_i$ a motion basis vector and explain its details later.

We replace the original NeRF-based decoder in [80] with an MLP $\Phi$ as shown in Fig. 3 with a single hidden layer of 96 units and softplus activation functions to decode the feature vector into a set of attributes for a 3D Gaussian:

$$\Phi(f_i) = \{\mu_i, \mathbf{s}_i, \mathbf{q}_i, o_i, \mathbf{c}_i\}. \quad (1)$$

Therefore, we associate one 3D Gaussian with each sampled location and the aforementioned motion basis vector. The final image is synthesized using the differentiable renderer [32] from the set of 3D Gaussians and the specified camera parameters as shown in Fig. 3. Notably, unlike previous works [11, 16, 78, 112], we do not use the 2D convolution refinement module to improve the rendered image's quality. We denote the rendered result of this encoded set of Gaussians at the viewpoint of $I_s$ as $I'_s$.

To decide where to sample 3D Gaussians from the triplanes, different from [3], we do not have a paired pretrained radiance field model to propose the sampling locations. Instead, we find that simply adapting the original ray shooting and two-pass importance sampling strategy in [8, 51] already gives reasonable performance.

Specifically, given a camera viewpoint, we shoot one ray for one pixel. We uniformly sample locations on each ray to decode 3D Gaussians, and then perform an additional importance sampling based on the opacity of previous decoded Gaussians to decode another set of Gaussians. Notice that the shooting resolution does not need to coincide with the rendering resolution. In practice, we use a sampling resolution of 64 × 64, but a resolution of 512 × 512 for rendering, which greatly improves efficiency. With 48 sampled Gaussians on each ray, this configuration yields about $200K$ Guassians, which is sufficient to render at 512 × 512 resolution [35, 108].

Although this sampling of Gaussians is inherently viewpoint-dependent, we find in practice that during inference, sampling from a fixed frontal viewpoint already produces a sufficiently dense and representative set of Guassians. The same set can then be effectively reused for rendering from novel viewpoints, accelerating inference. During training, the Gaussians are instantiated from the final rendering viewpoint to ensure view consistency.

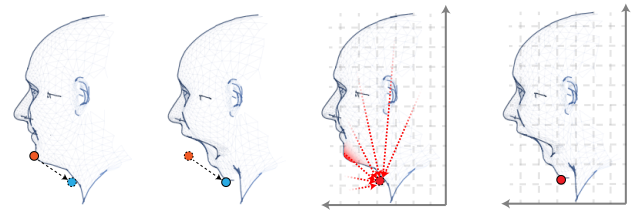
(a) (b)

Figure 4. Conceptual comparison between predicting residual features per Gaussian versus per grid point on the triplanes [16, 78] in the case of realizing the expression of opening the mouth. (a) In our framework, the 3D Gaussian can be transformed independently from the red point to the blue point because its motion basis vector encodes all necessary motion information. (b) In contrast, existing triplanes-based works require aggregating dense global context, to update the features on each grid point. For example, it needs to fuse the shape information from the global context through the attention mechanism to decide whether the mouth will reach the red point and therefore update its geometry or not.

**Feature-space deformation for animation.** Typically, dynamics with Guassians are modeled by deforming the Gaussian attributes directly, i.e., updating their position, scaling and rotation vectors and optionally their color [24, 46, 47, 90]. However, we find that such a design has limited capacity to model expressions and is hard to learn the animation details in the training dataset. We hypothesize that it is because the learning of motion in the low-dimensional 3D space is more difficult compared to learning on a potentially smoother manifold in the high-dimensional feature vector space. We thus propose to deform the feature vector $f_i$ sampled from the triplanes, which encodes information for *all* Gaussian properties and offers a richer deformation space, based on motion signals.

Specifically, we first use the pre-trained motion encoder $M$ in [113] to encode the motion signals in the source image $I_s$ and the driving image $I_d$ into 1D motion coefficients $M(I_s), M(I_d) \in \mathbb{R}^{512}$, respectively (see Fig. 3). Recall that, in Fig. 3, each feature vector $f_i$ is associated with a motion basis vector $\mathbf{m}_i$, which encodes the spatially-varying deformation of the decoded 3D Gaussian and $\mathbf{m}_i$ does not change while animating. The concept is similar to facial muscles or personalized "PCA bases" used in traditional blendshapes animation [39], but it is uniquely adapted for each individual and each facial part to capture personalized differences such as wrinkles. We investigate its semantic meaning in Fig. 5. Altogether, a residual feature $\delta f(I_s \rightarrow I_d) \in \mathbb{R}^{48}$ *due to the motion* from $I_s$ to $I_d$ for each sampled Gaussian is *individually* predicted by a motion decoder $\Psi$ from both the motion basis vector $\mathbf{m}$ and motion coefficients $M(I_s)$ and $M(I_d)$. Formally, the $i^{\text{th}}$ 3D Gaussian is updated as:

$$\Phi(f_i + \delta f_i(I_s \rightarrow I_d)) = \{\mu_i, \mathbf{s}_i, \mathbf{q}_i, o_i, \mathbf{c}_i\}, \qquad (2)$$

**Details of motion decoder $\Psi$.** We adopt a single-layer AdaLN [95] to modulate the motion basis vector $\mathbf{m}$ conditioned on the concatenated motion coefficients $M(I_s)$ and $M(I_d)$ and then pass the modulated vector to another MLP with a single hidden layer of 96 units and softplus activation functions to predict the delta feature vector.

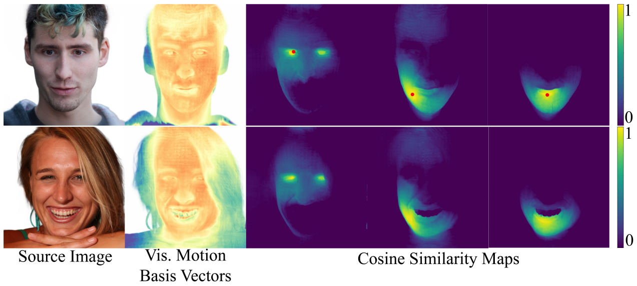

Figure 5. Demonstration of the similarity among motion basis vectors within and across subjects. Given the source images, we render the motion basis vectors of their Gaussian kernels via splatting. For the first-row subject, we select three specific points (red points) and compute the cosine similarity between their motion basis vectors and those of all other locations. We then compute the cosine similarity across subjects between these same points of the first subject and all motion basis vectors of the second subject in the second row. The resulting similarity maps show that our model learns coherent, semantically-meaningful and localized motion basis vectors.

**Discussion.** We highlight the difference between our design of predicting residual features and the design of existing works (i.e., [16, 79]) in Fig. 4 with visual explanations. Our design locally deforms each 3D Gaussian individually without aggregating dense global context, being much more computationally efficient. We provide detailed quantitative comparisons with relevant baselines in terms of speed and animation capabilities in Sec. 5.

## 4.2. Distillation from a Diffusion Model for Training

We adopt the state-of-the-art 2D facial animation diffusion model X-NeMo [113] as our data synthesizer. Specifically, we construct a synthetic dataset for self-reenactment-based training, containing over 60000 real identities from the FFHQ dataset [31], each of which has 8 synthesized expressions with the driving expressions sampled from both the FFHQ dataset and the FEED dataset [19]. These two datasets altogether represent diverse identities and expressions in the real world. To minimize inconsistency and hallucination issues in the diffusion model, we use pre-trained LP3D [70, 80] to frontalize the identity and driving images before synthesizing expressions via X-Nemo. During the training, we apply the same LP3D on these frontalized synthetic portraits on the fly to randomize input and output viewpoints for augmentation and novel view supervision.

**Training procedure.** We use self-reenactment as the training objective. With the synthetic dataset described above, we first sample $I_s$ and $I_d$ from the same identity whose expressions could be same or different, and then randomly sample camera parameters as in [80] and estimate their another viewpoint images using frozen pre-trained LP3D [70, 80] for multi-view supervision, as shown in Fig. 3. When $I_d$ has the same expression with $I_s$, we in-

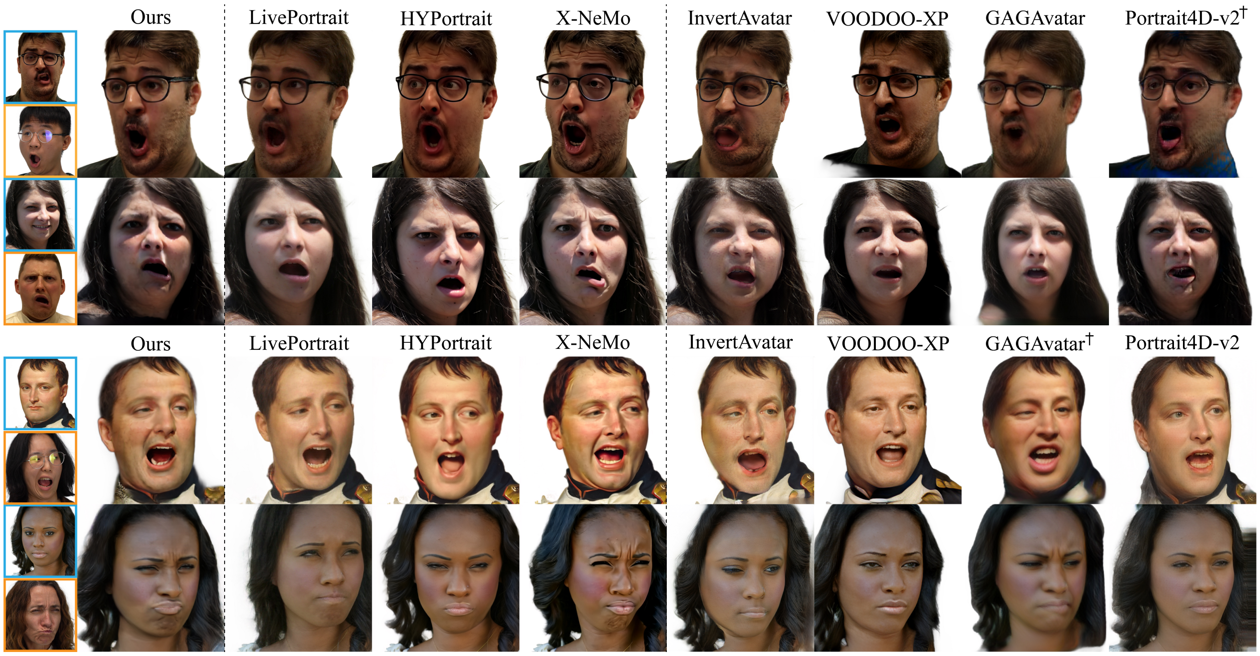

Figure 6. Qualitative comparison between our method and other 2D methods and 3D-aware methods in terms of expression and pose transfer. We denote [98] as "HYPortrait". **Source** images are marked with blue borders at the leftmost column, while the **driving** images are marked with orange borders throughout the paper. † indicates that the methods are trained with our synthetic dataset for distillation from scratch for fair comparisons.

tend to learn zeros residual features. Given the synthesized result $I'_d = E(I_s, I_d, p(I_d))$ and $I_d$, where $p(I_d)$ denotes the viewpoint of $I_d$, we design the loss by comparing them along with an adversarial objective to enhance the image quality as:

$$\begin{aligned}\mathcal{L} =&\lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(I'_d, I_d) + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(I'_d, I_d) + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}(I'_d, I_d)+\\ &\lambda_{\text{Detail}}\mathcal{L}_{\text{Detail}}(I'_d, I_d) + \lambda_{\text{Norm}}\mathcal{L}_{\text{norm}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}},\end{aligned} \tag{3}$$

where $\lambda_{\text{L1}} = 1$; $\mathcal{L}_{\text{LPIPS}}$ denotes the perceptual loss [110] with $\lambda_{\text{LPIPS}} = 1$; $\mathcal{L}_{\text{ID}}$ denotes the identity loss [14] with $\lambda_{\text{ID}} = 0.1$; $\lambda_{\text{Detail}}$ computes an additional L1 loss comparison over the eyes and mouth regions, which are separated out by [106] with $\lambda_{\text{Detail}} = 0.1$; $\lambda_{\text{Norm}}$ regularizes the averaged L2 norm for each predicted residual feature to enforce sparsity as in [87, 91] with $\lambda_{\text{Norm}} = 0.001$; $\mathcal{L}_{\text{adv}}$ is the adversarial loss as in [80], which is further conditioned on the motion coefficients extracted by the motion encoder $M$, with $\lambda_{\text{adv}} = 0.025$. Similarly, we also compute the loss for $I'_s$ and $I_s$.

**Implementation details.** We initialize all our networks including the adversarial discriminator used in $\mathcal{L}_{\text{adv}}$ from random weights, except the motion encoder $M$ from [113] that is pre-trained and frozen, and optimize with the Adam optimizer [34] and batch size 32 and learning rate 0.0001. We gradually increase the rendering resolution from 64 to 512 and introduce the adversarial loss in the middle of training. More details are in the supplementary.

# 5. Results

We provide quantitative and qualitative comparisons and an ablation study here. For more results, please refer to the supplementary and the accompanying video including a real-time demo.

**Metrics.** We conduct experiments for facial animation using the VOODOO-XP test set as in [78] which contains 102 video sequences. It features extreme expressions and wide viewing angles. We extract one out of five consecutive frames for testing to eliminate unnecessary duplication, which in total results in over $20K$ images. We evaluate results on common head regions without background across all methods. We conduct the following experiments: (1) self-reenactment. For each video sequence, we use the first frame as the source frame, and all other frames to drive it. Each method is tasked with reproducing the viewpoint and the expression of the driving frame. (2) cross-reenactment. For each video sequence, we randomly sample another video sequence. We use the first frame of the first video sequence as the source frame, and all frames in the second video sequence to drive it.

Besides the speed measured in FPS on an NVIDIA 6000 Ada GPU, we evaluate performance using the following four aspects: (a) MEt3R [2] for dense 3D inconsistency. Specifically, we only use the driving sequences, which only change the head pose while keeping the same neutral expression throughout the video, (b) face ID consistency [14], (c) SSIM [88] and LPIPS [110] to evaluate the quality of image reconstruction, and (d) the accuracy of expression and pose transfer. For this, we use SMIRK [62] to extract

| | Method | Self-Reenactment | | | | Cross-Reenactment | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MEt3R↓ | SSIM ↑ | LPIPS↓ | AED↓ | MEt3R↓ | ID↑ | EMO↑ | AED↓ | APD↓ | FPS↑ |
| 2D | LivePortrait [26] | 0.032 | 0.8206 | 0.1739 | **0.400** | 0.033 | 0.74 | 0.716 | 0.810 | 0.026 | 78.12 |
| | HYPortrait[98] | 0.028 | 0.8343 | 0.1685 | 0.483 | 0.032 | 0.74 | 0.752 | 0.900 | 0.077 | 0.01 |
| | X-NeMo[113] | 0.032 | 0.8227 | 0.1740 | 0.416 | 0.035 | 0.72 | 0.760 | 0.805 | 0.032 | 0.03 |
| 3D | GAGAvatar[11] | 0.032 | 0.8205 | 0.1914 | 0.501 | 0.034 | 0.77 | 0.654 | 0.888 | **0.025** | 13.67* |
| | InvertAvatar[112] | 0.026 | **0.8456** | **0.1592** | 0.564 | **0.028** | **0.80** | 0.565 | 0.891 | 0.049 | 24.30* |
| | VOODOO-XP[78] | 0.030 | 0.8086 | 0.1966 | 0.560 | 0.032 | 0.77 | 0.699 | 0.903 | 0.028 | 5.45 |
| | Portrait4D-v2[16] | 0.030 | 0.8327 | 0.1709 | 0.545 | 0.035 | 0.79 | 0.589 | 0.886 | 0.029 | 14.50 |
| | GAGAvatar† [11] | 0.027 | 0.8258 | 0.1991 | 0.636 | 0.030 | 0.67 | 0.578 | 0.944 | 0.038 | 13.67* |
| | Portrait4D-v2† [16] | 0.030 | 0.8217 | 0.1952 | 0.682 | 0.034 | 0.76 | 0.448 | 0.972 | 0.026 | 14.50 |
| | Ours | **0.025** | 0.8294 | 0.1864 | 0.496 | **0.028** | 0.75 | **0.771** | **0.745** | 0.028 | **107.31** |

Table 1. Quantitative comparison with other 2D or 3D-aware facial animation baselines on the task of self-reenactment and cross-reenactment. Throughout the paper, we bold the best metric, and underline the second best metric. † indicates that the methods are trained with our synthetic dataset for fair comparisons. The FPS marked with $*$ reports inference time excluding time-consuming morphable model fitting optimization. Full inference time including the optimization for GAGAVatar is 0.41 FPS and InvertAvatar is 0.07 FPS.

the FLAME [40] coefficients for the driving and resulting frames, and measure the averaged distance of the expression coefficients as "AED", and averaged distance of the pose parameters as "APD". Please note that "APD" reveals certain *coarse* 3D shape quality while MEt3R focuses more on dense photometric 3D consistency across views. We also use EmoNet [77] as in [113] to measure the emotion similarity between the driving and resulting frames as "EMO" that is more sensitive to extreme motions.

## 5.1. Comparisons

**Baselines.** We compare our method against other existing open-source feed-forward methods, including the state-of-the-art GAN-based 2D facial animation method LivePortrait [26], diffusion-based 2D facial animation methods HunyuanPortrait [98] and X-NeMo [113], 3DMM-based 3D facial animation methods InvertAvatar [112] and GAGAvatar [11], and 3D facial animation methods with learned motion space Portrait4D-v2 [16] and VOODOO-XP [78]. Furthermore, for an additional fairer comparison and to illustrate the benefits of our proposed animation representation, we also train the best-performing prior 3D-based methods Portrait4D-v2 and GAGAavatar from scratch using our synthetic dataset with multi-view images estimated and denote them as Portrait4D-v2† and GAGAvatar†.

**Qualitative results.** We provide qualitative comparisons in Fig. 6. Generally, other 3D-aware methods create muted expressions even when retrained with our synthetic dataset and do not faithfully synthesize expressions in the driving image. Especially, in the first row, the 3D-aware methods cannot remove the extreme mouth motion in the source frame and the right wrinkles near the mouth are leaked into the driven results. InvertAvatar occasionally produces collapsed 3D head geometry. Among the 2D-based methods, X-NeMo mostly produces impressive results, but occasionally distorts head shape under a large pose change (third row) and hallucinates details (e.g., added cheek color in the fourth row). HunYuanPortrait cannot faithfully transfer the pose and expression in the second, third and fourth rows. LivePortrait creates dampened mouth expressions in the second to fourth rows. Even after retraining with our expressive synthetic dataset, Portrait4D-v2† and GAGAvatar† produce less accurate expression transfer results. In contrast, our method faithfully transfers the expression and pose and is on par with X-NeMo while maintaining the identity and being 3500× faster (see Tab. 1). Please find more comparison in the supplementary. We further provide multi-view rendering results in Fig. 7. Despite extreme expressions present in the source images, our method can infer the occluded regions, such as the closed eyes, and produces the output with consistent identity in the source image and motions in the driving image.

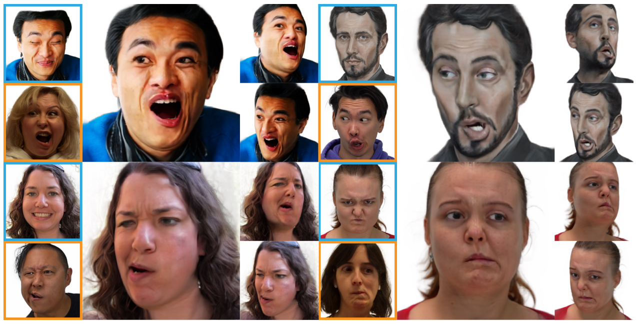

Figure 7. More qualitative results with our method from a **source** image and **driving** image. We provide the multi-view rendered results next to the driven results.

**Quantitative results.** As shown in Table 1, our method achieves the best 3D consistency (MEt3R) and best expression transfer quality (AED) among the 3D methods, for the task of self-reenactment. We also achieve state-of-the-art MEt3R, EMO, and AED scores across all methods for the task of cross-reenactment, even surpassing X-NeMo due to potentially more accurate pose transfer, which aids correct expression and emotion detection.

While LivePortrait performs comparably to X-NeMo. The latter typically produces more vivid and detailed facial expressions, such as wrinkles, which are not captured by AED but are captured by EMO and are evident in the

| Method | AED↓ | Mem.↓ | FPS↑ |
|---|---|---|---|
| Ours (128 × 128) | **0.507** | 0.32 GB | 132.87 |
| w/ DINO-v2 Encoding | 0.597 | 4.32 GB | 22.47 |
| w/ Real Dataset | 0.543 | 0.32 GB | 132.87 |
| w/ Spatial Deformation | 0.634 | **0.27 GB** | **141.61** |

Table 2. Ablation studies with self-reenactment.

qualitative comparisons (Fig. 6). Consequently, we select X-NeMo, rather than LivePortrait, as our teacher model.

All of the compared 3D baselines utilize camera-space 2D refinement to improve image quality, but this is known to degrade 3D consistency – reflected by their lower MEt3R score. Furthermore, InvertAvatar is able to copy texture from the source image, but its overall 3D shape is inaccurate as reflected in its lower "APD" and in the qualitative comparison (Fig. 6).

Our method is also theoretically bounded by the identity consistency of X-NeMo, and the image and 3D quality of LP3D, because of using a synthetic dataset generated from it. This may explain its less competitive image quality and ID metric versus X-NeMo. However, since we restrict X-NeMo to operate on frontal images only and use LP3D to estimate multi-view images, we mitigate the identity shift problem, evidenced by our method's better ID metric versus X-NeMo. We also find that the two 3D baselines [11, 16] that we retrained with our synthetic data still do not perform as well as ours as evidenced by worse metrics, while being slower. This again demonstrates the capability of our proposed animation representation compared to the existing methods. We provide a more detailed investigations to these re-trained 3D baselines in the supplementary.

Our method animates the face at 107.31 FPS, surpassing all other baselines, including ones using the attention mechanisms [16, 78]. In contrast, 2D diffusion methods could take up to almost a minute for driving a single image sequentially, or several seconds per frame when a sequence of images is used and the time cost is amortized across frames. For our method, encoding a facial image into 3D Gaussians is required only once per video sequence, and our method performs this step almost instantly in just 20ms. Besides, our method only requires 0.4 GB for static model storage during inference due to our decoupled motion representation. In contrast, X-NeMo requires over ten times more storage (∼ 6 GB).

### 5.2. Ablation Studies

We validate each component of our model using the self-reenactment on the same VOODOO-XP test set with a rendering resolution of 128 × 128 without an adversarial loss for comparison. We measure the quality of expression control accuracy, memory consumption for the static model storage and speed in Table 2 and Fig. 8.

**Choice of motion encoder.** We study the importance of the selected motion encoder. Instead of the motion encoder in X-Nemo [113] we use DINO-v2 [52] as in [78]. Notably, this encoder is computationally more expensive than our selected motion encoder [113]. It increases the memory consumption and reduces the FPS. It fails to capture expression details in Fig. 8 and leads to a worse AED.

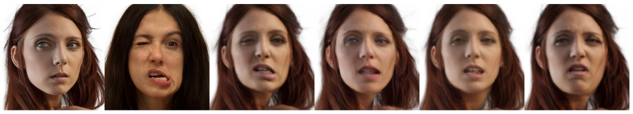

Figure 8. Comparison among different ablation models based on the expression transfer. "w/ S.D." denotes using the spatial deformation instead of feature-space deformation. "w/ R.D." denotes using a real dataset instead of the synthetic one.

**Usage of diffusion model distillation.** We study the necessity of distilling the knowledge from pre-trained diffusion-based models by using a real dataset instead. We use CelebVText [107] as the real dataset for its diversity of expressions and identities. Since a real dataset usually features common expressions such as smiling, opening the mouth, etc. and less extreme expressions, the AED is affected and the produced expression is muted (Fig. 8).

**Feature-space- vs. spatial-deformation.** We ablate the usage of the proposed feature-space deformation by replacing it with directly predicting the residual for the 3D Gaussians' position, scaling and quaternion vectors from the motion decoder $\Psi$, similarly to the previous dynamic modeling methods (e.g., [24, 46, 47, 90]). Since we do not need to go through the non-linear decoder in this setup, the memory consumption and FPS are slightly improved. However, all other metrics are significantly compromised. The expression is also not transferred accurately in Fig. 8.

## 6. Discussion

**Conclusion.** In this work, we investigate how to distill a 2D facial animation diffusion method into 3D-consistent, efficient yet expressive instant avatar encoder from a single image. We propose an animation representation that deforms both the Gaussian appearance and geometry based on the encoded motion basis vectors that resemble the muscle control mechanism or act as a learned version of "PCA bases" in traditional morphable models. We believe our method paves the way for an real-time and expressive representation distilled from a powerful diffusion model, enabling real-world applications such as digital twins, where real-time performance, and controllability are critical.

**Limitations and future work.** Our model may inherit potential errors in synthetic data generated by the diffusion model and the pre-trained 3D lifting algorithm. In the future, it will be interesting to extend our method into disentangling appearance properties such as lighting. Even though we demonstrate our method using only image-driving examples, it is possible to encode other conditions, such as audio or texts, into sequences of the 1D motion coefficients and drive our avatar.

**Ethics concern.** We propose an algorithm, which is capable of converting a single 2D facial image into a 3D-aware animatable avatar, which could be misused for generating malicious content. We do not condone such behavior and identify potential works that could be used for detecting fake information [37, 99–101] or works that authenticate the authorized driving subject of the avatar [57].

## Acknowledgements

We thank David Luebke, Michael Stengel, Yeongho Seol, Simon Yuen, Marcel Bühler, and Arash Vahdat for feedback on drafts and early discussions. We also thank Alex Trevithick and Tianye Li for proof-reading. This research was also funded in part by the Ronald L. Graham Chair and the UC San Diego Center for Visual Computing.

## References

[1] Shivangi Aneja, Sebastian Weiss, Irene Baeza, Prashanth Chandran, Gaspard Zoss, Matthias Nießner, and Derek Bradley. Scaffoldavatar: High-fidelity gaussian avatars with patch expressions. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2, 3

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[3] Florian Barthel, Arian Beckmann, Wieland Morgenstern, Anna Hilsmann, and Peter Eisert. Gaussian splatting decoder for 3d-aware generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7963–7972, 2024. 4

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH*, 1999. 2, 3

[5] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2

[6] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3

[7] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025. 2

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2, 3, 4

[9] 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. 3

[10] 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

[11] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024. 1, 2, 3, 4, 7, 8

[12] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image(s). 2024. 2

[13] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025. 2, 3

[14] 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. 6

[15] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Learning one-shot 4d head avatar synthesis using synthetic data. *arXiv preprint arXiv:2311.18729*, 2023. 2, 3

[16] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv*, 2024. 1, 2, 3, 4, 5, 7, 8

[17] 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

[18] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022.

[19] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2, 5

[20] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2, 3

[21] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[22] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023. 2

[23] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10747–10758, 2024. 3

[24] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. *arXiv preprint arXiv:2405.19331*, 2024. 2, 3, 5, 8

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[26] 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. 1, 2, 7

[27] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. *arXiv preprint arXiv:2502.17796*, 2025. 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[29] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4220–4230, 2024. 4

[30] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pages 428–447. Springer, 2024. 3

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 4

[33] Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Learning to generate conditional tri-plane for 3d-aware expression controllable portrait animation. *arXiv preprint arXiv:2404.00636*, 2024. 3

[34] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[35] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 4

[36] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars, 2025. 2, 3, 4

[37] Nagano Koki. StyleGAN3 Synthetic Image Detection, 2021. 9

[38] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon. Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *arXiv*, 2024. 2

[39] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports*, 2014. 5

[40] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 7

[41] 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), 2017. 3

[42] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 3

[43] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[44] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3377–3386, 2022. 2

[45] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065*, 2025. 2

[46] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 3, 5, 8

[47] Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from monocular videos. *arXiv preprint arXiv:2404.12379*, 2024. 2, 3, 5, 8

[48] Renshuai Liu, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2114–2123, 2024. 3

[49] 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. *arXiv preprint arXiv:2406.01900*, 2024. 2, 3

[50] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 2

[51] 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. 1, 3, 4

[52] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 8

[53] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 2

[54] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3

[55] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[56] Reni Paskaleva, Mykyta Holubakha, Andela Ilic, Saman Motamed, Luc Van Gool, and Danda Paudel. A unified and interpretable emotion representation and expression generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2447–2456, 2024. 3

[57] Ekta Prashnani, Koki Nagano, Shalini De Mello, David Luebke, and Orazio Gallo. Avatar fingerprinting for authorized use of synthetic talking-head videos. *ECCV*, 2024. 9

[58] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *https://arxiv.org/abs/2011.13961*, 2020. 3

[59] 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. 2, 3

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[61] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021. 2

[62] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024. 6

[63] Shen Sang, Tiancheng Zhi, Tianpei Gu, Jing Liu, and Linjie Luo. Lynx: Towards high-fidelity personalized video generation. *arXiv preprint arXiv:2509.15496*, 2025. 2

[64] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[65] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 3

[66] 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

[67] 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. 2

[68] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Hsin-Ying Lee, Jian Ren, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. *arXiv preprint arXiv:2301.11326*, 2023. 3

[69] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[70] Michael Stengel, Koki Nagano, Chao Liu, Matthew Chan, Alex Trevithick, Shalini De Mello, Jonghyun Kim, and David Luebke. Ai-mediated 3d video conferencing. In *ACM SIGGRAPH Emerging Technologies*, 2023. 5

[71] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20991–21002, 2023. 2, 3

[72] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 3

[73] Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5546–5558, 2025. 2, 3

[74] Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B. Lindell. MVP4D: Multi-view portrait video diffusion for animatable 4D avatars, 2025. 3

[75] 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)*, pages 5318–5330. IEEE Computer Society, 2025. 2, 3

[76] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024. 3

[77] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. 7

[78] 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*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[79] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10336–10348, 2024. 2, 3, 5

[80] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 2, 3, 4, 5, 6

[81] Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024. 3

[82] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 3

[83] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 2

[84] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *arXiv preprint arXiv:2011.15126*, 2020.

[85] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022.

[86] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Lia: Latent image animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2024. 2

[87] Yaohui Wang, Di Yang, Xinyuan Chen, Francois Bremond, Yu Qiao, and Antitza Dantcheva. Lia-x: Interpretable latent portrait animator. *arXiv preprint arXiv:2508.09959*, 2025. 6

[88] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

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

[90] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2, 3, 5, 8

[91] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 3, 6

[92] 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, 3

[93] Chao Xu, Yang Liu, Jiazheng Xing, Weida Wang, Mingze Sun, Jun Dan, Tianxin Huang, Siyuan Li, Zhi-Qi Cheng, Ying Tai, et al. Facechain-imagineid: Freely crafting high-fidelity diverse talking faces from disentangled audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1292–1302, 2024. 3

[94] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniavatar: Geometry-guided controllable 3d head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12814–12824, 2023. 3

[95] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019. 5

[96] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation, 2024. 3

[97] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024. 2

[98] 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, 6, 7

[99] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 9

[100] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, pages 4534–4565. Curran Associates, Inc., 2023.

[101] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent

space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 9

[102] 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. 3

[103] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 3

[104] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 2

[105] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 3

[106] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068, 2021. 6

[107] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023. 8

[108] Zhengming Yu, Tianye Li, Jingxiang Sun, Omer Shapira, Seonwook Park, Michael Stengel, Matthew Chan, Xin Li, Wenping Wang, Koki Nagano, and Shalini De Mello. Gaia: Generative animatable interactive avatars with expression-conditioned gaussians. 2025. 4

[109] Zhengming Yu, Tianye Li, Jingxiang Sun, Omer Shapira, Seonwook Park, Michael Stengel, Matthew Chan, Xin Li, Wenping Wang, Koki Nagano, and Shalini De Mello. GAIA: Generative animatable interactive avatars with expression-conditioned gaussians. In *ACM SIGGRAPH*, 2025. 3

[110] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[111] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3657–3666, 2022. 2

[112] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 2, 3, 4, 7

[113] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025. 2, 3, 4, 5, 6, 7, 8

[114] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 2

[115] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 3

[116] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 3