

Instant Expressive Gaussian Head Avatars at Over 100 FPS

Kaiwen Jiang^{1*}, Xueting Li², Seonwook Park², Ravi Ramamoorthi¹, Shalini De Mello², and Koki Nagano²

¹ University of California, San Diego, United States

² NVIDIA, United States

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 feedforward facial animation methods – built upon 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 address this *portrait animation trilemma* (speed, 3D consistency, and expressiveness) and propose a pipeline that instantly converts an in-the-wild single image into a 3D-consistent, fast yet expressive animatable representation via a feed-forward encoder. 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. Furthermore, 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. Our method runs at 107.31 FPS for animation and pose control, representing a 3-4 order of magnitude speedup versus the state of the art while achieving comparable animation quality, thus surpassing alternative designs that trade speed for quality or vice versa.

1 Introduction

Creating a 4D digital twin from a single facial image that supports both 3D viewpoint control and animation is a long-standing goal in computer vision and graphics. Real-time animatable 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) [50] and 3D Gaussians [31], previous 3D-aware face animation works [11, 12, 42, 68, 108]

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

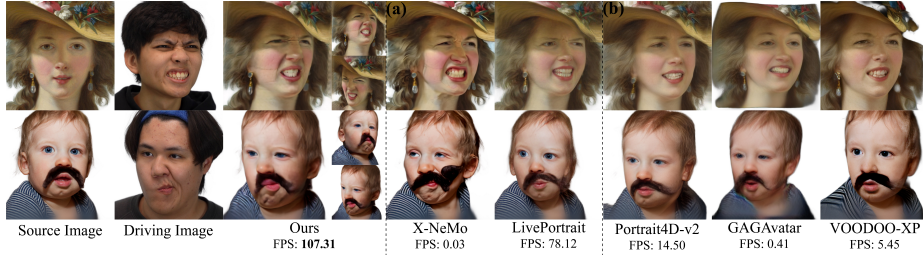


Fig. 1: We present an instant feedforward encoder that transforms an in-the-wild source image into an animatable 3D avatar. Our method introduces a fast, consistent yet expressive 3D animation representation. Given a driving image, we evaluate both the expression transfer quality and the animation speed (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 VODOO-XP [75] 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. 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. *FPS is measured in a fully end-to-end streaming setting throughout the paper, where incoming frames are processed sequentially and all modules, including preprocessing and fitting steps, are accounted for across all baselines.*

have achieved real-time and consistent animation with photo-realistic view synthesis by using parametric models [4, 20, 39]. However, parametric models inherently limit the animation quality. Follow-up methods [15, 16, 35, 75, 76] therefore resort to learning the animation purely from data. Nevertheless, their representations entangle 3D structure and animation (e.g., via 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, 37, 48, 58, 59, 62, 89, 95, 109] into portrait animation has brought the expressivity of achieved facial animation to a whole new level [95, 109]. 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 expressivity. We call this new challenge *the portrait animation trilemma*.

There are existing methods (e.g., [1, 24, 69, 71, 72]) proposing to optimize a person-specific animatable 3D avatar from captured videos or generated images that achieve these three goals. However, we seek to design a 3D-aware animation framework that *instantly* encodes a facial image into an animatable 4D avatar, while supporting fast, consistent and detailed animation. Our key insights for solving this problem are twofold. First, we observe that large-scale 2D diffusion models already encode rich facial dynamics. We reinterpret them as motion priors for 3D-aware animation and introduce a principled mechanism to leverage

their guidance while avoiding their inherent 2D inconsistencies and hallucinations. Second, we show that achieving both efficiency and 3D consistency without sacrificing expressiveness requires a novel animation representation, which we propose.

Specifically, we build upon 3D Gaussians [31] 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 [77] 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 animation bases in traditional blendshapes animation. This auxiliary vector is then combined with the driving signal to deform each Gaussian associated with 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 [44–46, 87] 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 and 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 via a state-of-the-art facial animation diffusion model [109] for transferring its expressive motion priors in a 3D consistent manner. To minimize the 3D inconsistency and hallucination issues of diffusion models, we first frontalize in-the-wild portraits and then synthesize expressions by the diffusion model only on frontal views, thereby avoiding hallucinations caused

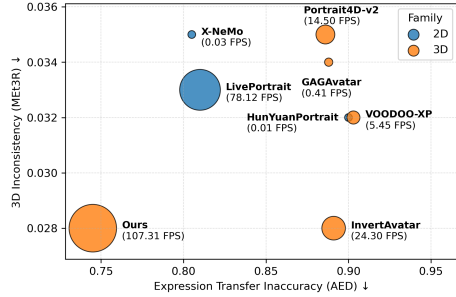


Fig. 2: A portrait animation trilemma: existing methods do not simultaneously achieve 3D consistency, speed, and expressive expression transfer. We provide a visualization of the quantitative comparison in terms of 3D inconsistency (measured by MET3R ↓), expression transfer inaccuracy (measured by AED ↓) and animation speed (measured by FPS ↑, visualized as the size of the circle) with other 2D- or 3D-based baselines, including [11, 16, 26, 75, 95, 108, 109], 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 [95, 109] while simultaneously achieving better 3D consistency and expression transfer accuracy, appearing on the lower left.

by viewpoint changes. Multi-view results are then separately synthesized during training using a pre-trained expert 2D-to-3D network, which enforces cross-view geometric consistency via multiview supervision. Therefore, our method is *free of* inconsistency or hallucination issues. In summary, our main contributions include:

- 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 this animation representation by leveraging guidance from existing diffusion-based methods (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 competing methods at the same level of expressivity. (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, 66]), which synthesizes driven faces given control signals. GAN-based methods [5, 17–19, 22, 26, 43, 49, 52, 60, 63, 64, 80–83, 94, 101, 107, 110] 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, 47, 48, 55, 58, 73, 78, 79, 86, 89, 90, 93, 95, 99, 109], 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.

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, 40] or facial motion representations (e.g., rasterized coordinates [111] or facial keypoints) as priors [11, 24, 32, 41, 42, 65, 68, 70, 88, 91, 100, 108] 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, 75, 76] implicitly learns the motion as residual features to the triplanes [8] through data. Notably, Portrait4D [15] transfers knowledge 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.

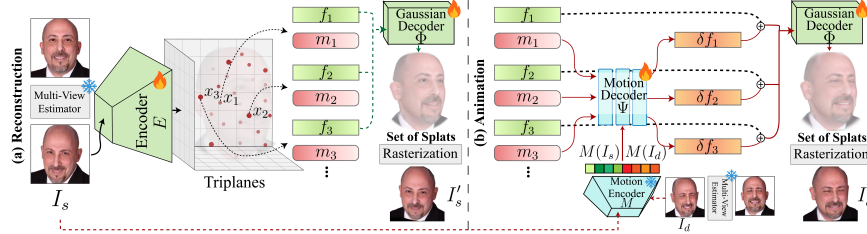


Fig. 3: Overview of our **training** pipeline with the two-part self-reenactment task. **(a) Reconstruction:** Given a frontalized source frame with an expression synthesized by a pre-trained diffusion model [109], we first use a pre-trained multi-view estimator [77] to generate its another viewpoint I_s' . The encoder E converts I_s into triplanes, from which we sample feature vectors $f_1, f_2, \dots$ and paired motion basis vectors $m_1, m_2, \dots$. A Gaussian decoder Φ 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 Ψ to predict residual features $\delta f_1, \delta f_2, \dots$ from paired motion basis vectors. Adding these residuals to the original features and decoding them with Φ 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.

4D representation and rendering. Aiming at a native 4D representation, many works [21, 53, 54, 57] build upon NeRFs to construct a spatial deformation field, which, however, is computationally expensive. With the advent of Gaussian splatting [31], another family of methods identify the potential of using the spatial deformation of 3D Gaussians to represent motion. One group of methods [24, 45, 46, 87] learn an implicit representation, such as HexPlanes [6] or a coordinate-based MLP, to deform the Gaussians. Another group of methods [24, 27, 69, 105, 112] utilizes an explicit mesh to drive the 3D Gaussians. However, we propose to deform Gaussians in a high-dimensional feature space, thereby preserving greater expressivity (Sec. 5.2).

Notably, Avat3r [35] and ScaffoldAvatar [1] also use 3D Gaussians to build animatable avatars. However, they are not a feed-forward method from a single image and require either 3D GAN inversion or mesh tracking. Prior work [71, 72] uses diffusion-generated multi-view images for animated 3D head synthesis, but relies on hours-long optimization to fit a single avatar. In contrast, we 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 [77] 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 $\in \mathbb{R}^{256 \times 256 \times 32}$. These planes

can then be used for rendering arbitrary viewpoints using volumetric rendering [50]. 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 across the three planes 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 (Sec. 4.1).

3D Gaussian splatting. Kerbl et al. [31] provides a differentiable and efficient solution to encoding a 3D scene as a set of anisotropic 3D Gaussian and rendering them efficiently 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^{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 [77], to encode I_s into a set of 3D Gaussians [31] 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 at viewpoint p .

For training, we construct a *frontalized* synthetic extreme facial expression dataset (Sec. 4.2), in which each identity is represented by multiple images with different expressions. We first frontalize the FFHQ [30] in-the-wild portraits and leverage a pre-trained diffusion-based portrait animation model [109] to synthesize facial expressions on the frontalized views. Multi-view images are synthesized on the fly during the training by a pre-trained expert 2D-to-3D network [77] for multi-view supervision. 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 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, 35, 75], leading to faster inference. The rendering of the Gaussians from arbitrary viewpoints is realized through rasterization [31].

4.1 Encoder Design and Animation Representation

3D Gaussians decoder. LP3D [77] lifts single 2D image to a 3D avatar represented by a neural triplane representation. However, its implicit radiance field representation is less ideal for representing dynamics as opposed to 3D Gaussians. 3D Gaussians offer more independent control over each primitive and fast rendering speed while maintaining 3D consistency [29].

Therefore, we make a minimal change to adapt the architecture of [77] into using 3D Gaussians while preserving its strong capability of faithfully lifting 2D images into 3D. We sample 3D Gaussians from the encoded triplanes, as explored in [3]. We will first explain how we decode 3D Gaussians from features sampled from neural triplane and then clarify how we decide the sampling locations.

We first use 96 channels for the encoded triplanes. Given a sampled 3D location $\mathbf{x}_i$, we project it onto each plane and sum 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 [77] with an MLP Φ 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 [31] from the set of 3D Gaussians and the specified camera parameters as shown in Fig. 3. Notably, unlike previous works [11, 16, 75, 108], 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 pre-trained radiance field model to propose the sampling locations. Instead, we find that simply adapting the original ray shooting operation and two-pass importance sampling strategy in [8, 50] 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 Gaussians, which is sufficient to render at 512×512 resolution [34, 104].

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 Gaussians. The

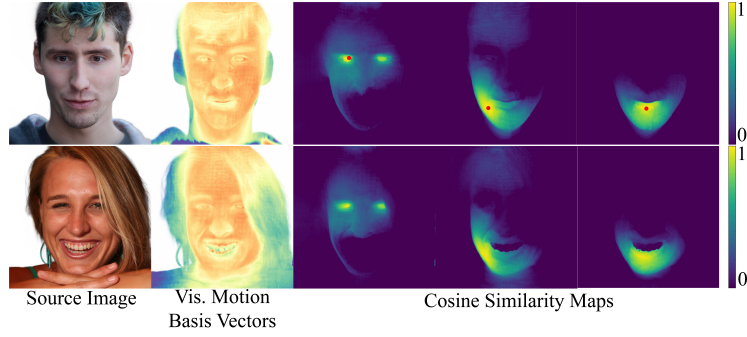


Fig. 4: Demonstration of the similarity among motion basis vectors within and across subjects. Given the source images, we render the motion basis vectors of their Gaussian kernels via splatting. For the first-row subject, we select three specific points (red points) and compute the cosine similarity 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.

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. We explain this later.

Feature-space deformation for animation. Typically, dynamics with Gaussians are modeled by deforming the Gaussian attributes directly, i.e., updating their position, scaling and rotation vectors and optionally their color [24, 45, 46, 87]. However, we find that such a design has limited capacity to model expressions and struggles with learning 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 [109] 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 localized bases in blendshape animation [38], but it is uniquely adapted for each individual and each facial part to capture personalized differences such as wrinkles. We show that the learned motion basis vector is coherent, semantically-meaningful and localized in Fig. 4. Altogether, a residual feature $\delta f_i(I_s \rightarrow I_d) \in \mathbb{R}^{48}$ due to the motion from I_s to I_d for each sampled Gaussian is *individually* predicted by a motion decoder Ψ using both the motion basis vector $\mathbf{m}_i$ and motion coefficients

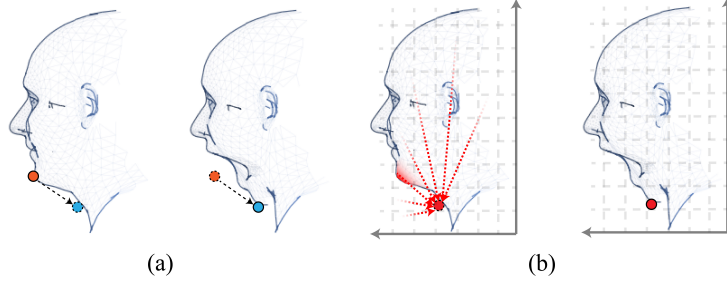


Fig. 5: Conceptual comparison between predicting residual features per Gaussian versus per grid point on the triplanes [16, 75] in the case of realizing the expression of opening the mouth. (a) In our framework, the 3D Gaussian can be transformed independently from the red point to the blue point because 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.

$M(I_s)$ and $M(I_d)$. Formally, the i^{th} 3D Gaussian is updated as:

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

Details of motion decoder Ψ . We adopt a single-layer AdaLN [92] to modulate the motion basis vector $\mathbf{m}_i$ 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.

Discussion. We highlight the difference between our design of predicting residual features and the design of existing works (i.e., [16, 76]) in Fig. 5 with visual explanations. Our design is more efficient by locally deforming each 3D Gaussian individually without aggregating dense global context, while also being expressive to capture motion details (Table 2).

4.2 Training with Guidance from a Diffusion Model

We adopt a state-of-the-art 2D facial animation diffusion model X-NeMo [109] as our data synthesizer. Specifically, we construct a synthetic dataset for self-reenactment-based training, containing over 60,000 real identities from the FFHQ dataset [30], each of which has 8 synthesized expressions with the driving expressions sampled from both the FFHQ and FEED datasets [18]. 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 [67, 77] to frontalize the source 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.



Fig. 6: Qualitative comparison between our method and other 2D methods and 3D-aware methods in terms of expression and pose transfer. We denote [95] 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.

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 [77] and estimate their other viewpoint images using a frozen pre-trained LP3D [67, 77] for multi-view supervision, as shown in Fig. 3. When I_d has the same expression as I_s , we intend to learn zero 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 and Gaussians are therefore instantiated from $p(I_d)$, we design the loss to compare them along with an adversarial objective to enhance the image quality as:

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

where $\lambda_{L1} = 1$; $\mathcal{L}_{LPIPS}$ denotes the perceptual loss [106] with $\lambda_{LPIPS} = 1$; $\mathcal{L}_{ID}$ denotes the identity loss [14] with $\lambda_{ID} = 0.1$; λ_{Detail} computes an additional separate L1 loss over the eye and mouth regions [102] with $\lambda_{Detail} = 0.1$; λ_{Norm} regularizes the averaged L2 norm for each predicted residual feature to enforce sparsity as in [84, 88] with $\lambda_{Norm} = 0.001$; $\mathcal{L}_{adv}$ is the adversarial loss as in [77], which is further conditioned on the motion coefficients extracted by the motion encoder M , with $\lambda_{adv} = 0.025$. Similarly, we also compute the loss for I'_s and I_s and sum it with loss for I'_d and I_d as our final loss.

Implementation details. We initialize all our networks including our instant lifting encoder and the adversarial discriminator used in $\mathcal{L}_{adv}$ from random weights, except the motion encoder M from [109] that is pre-trained and frozen, and optimize with the Adam optimizer [33] and batch size 32 and learning rate

	Method	Met3R↓	PSNR↑	SSIM↑	LPIPS↓	IoU↑	Mem.↓	FPS↑
2D	LivePortrait [26]	0.032	22.18	0.821	0.174	0.79	0.5 GB	78.12
	HYPportrait [95]	0.028	22.31	0.834	0.169	0.83	4.2 GB	0.01
	X-NeMo [109]	0.032	21.92	0.823	0.174	0.83	6.0 GB	0.03
3D	GAGAvatar [11]	0.032	21.45	0.821	0.191	0.80	1.4 GB	0.41
	InvertAvatar [108]	0.026	23.43	0.846	0.159	0.79	3.0 GB	0.07
	VOODOO-XP [75]	0.030	21.14	0.809	0.197	0.77	1.2 GB	5.45
	Portrait4D-v2 [16]	0.030	22.17	0.833	0.171	0.82	1.5 GB	14.50
	GAGAvatar [†] [11]	0.027	21.33	0.826	0.199	0.77	1.4 GB	0.41
	Portrait4D-v2 [†] [16]	0.030	21.25	0.822	0.195	0.82	1.5 GB	14.50
	Ours	0.025	22.34	0.835	0.181	0.85	0.4 GB	107.31

Table 1: Quantitative comparison on **self-reenactment**. Red indicates best and orange indicates second best throughout the paper. [†] indicates training with our synthetic dataset.

	Method	Met3R↓	ID↑	EMO↑	AED↓	APD↓	FPS↑
2D	LivePortrait [26]	0.033	0.74	0.716	0.810	0.026	78.12
	HYPportrait [95]	0.032	0.74	0.752	0.900	0.077	0.01
	X-NeMo [109]	0.035	0.72	0.760	0.805	0.032	0.03
3D	GAGAvatar [11]	0.034	0.77	0.654	0.888	0.025	0.41
	InvertAvatar [108]	0.028	0.80	0.565	0.891	0.049	0.07
	VOODOO-XP [75]	0.032	0.77	0.699	0.903	0.028	5.45
	Portrait4D-v2 [16]	0.035	0.79	0.589	0.886	0.029	14.50
	GAGAvatar [†] [11]	0.030	0.67	0.578	0.944	0.038	0.41
	Portrait4D-v2 [†] [16]	0.034	0.76	0.448	0.972	0.026	14.50
	Ours	0.028	0.76	0.785	0.731	0.028	107.31

Table 2: Quantitative comparison on **cross-reenactment**.

$5e-5$. We gradually increase the rendering resolution from 64 to 512 with fixed number of sampled Gaussians 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, including a real-time demo, please refer to the supplementary material and the accompanying video.

Metrics. We conduct experiments for facial animation using the VOODOO-XP test set as in [75], 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 the 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 memory consumption for static model storage required during inference and the rendering speed measured in FPS on an NVIDIA 6000 Ada GPU, we evaluate performance using the following five aspects: (a) Met3R [2] for dense 3D inconsistency; (b) PSNR, SSIM [85] and LPIPS [106] to evaluate the quality of image reconstruction; (c) IoU to evaluate 3D reconstruction quality by computing the intersection-over-union between the binary silhouettes of the full head region in the driving frame and the resulting frame of the same subject; (d) face ID consistency [14]; (e) the accuracy of expression and pose transfer where we use SMIRK [61] to extract the FLAME [39] 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”. Besides, we also use EmoNet [74] as in [109] to measure the emotion similarity between the driving and resulting frames as “EMO” that is more sensitive to extreme motions. Please note that “APD” reveals *coarse* 3D shape quality while Met3R focuses more on dense photometric 3D consistency across views.

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 [95] and X-NeMo [109], 3DMM-based 3D facial animation methods InvertAvatar [108] and GAGAvatar [11], and 3D facial animation methods with learned motion space Portrait4D-v2 [16] and VOODOO-XP [75]. 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 GAGAvatar 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 (third and forth rows). 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

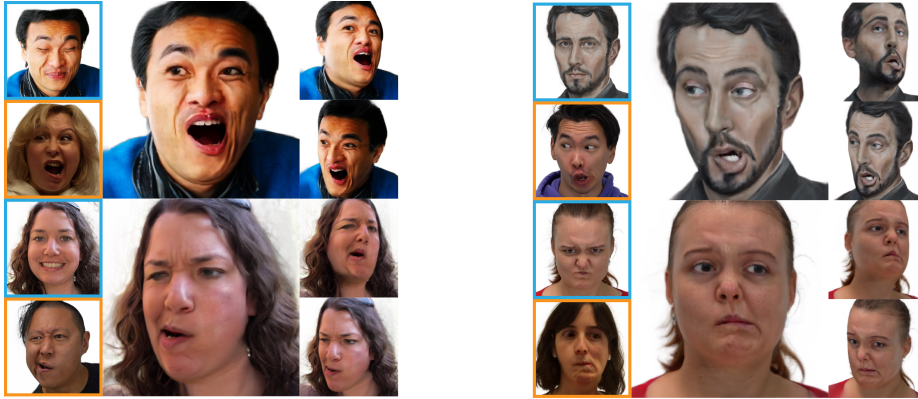


Fig. 7: More qualitative results with our method from a **source** image and **driving** image. We provide the multi-view rendered results next to the driven results.

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 $3500\times$ faster (see Tab. 2). Please find more comparison in the supplementary. We further provide multi-view rendering results in Fig. 7. Despite extreme expressions being present in the source images, our method can infer the occluded regions, such as the closed eyes, produces outputs with identity consistent with that of the source image and motions consistent with the driving image.

Quantitative results. As shown in Table 1, 2, our method achieves the best 3D consistency (MET3R), 3D shape (IOU), and compute efficiencies (Mem and FPS) and second best image reconstruction (in terms of PSNR and SSIM) across all methods, for the task of self-reenactment. 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. Even though InvertAvatar achieves the best image quality, its 3D quality is significantly limited, as captured by the worse IoU for the 3D head shape compared to us. 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. We note that ID scores conflate identity preservation with expression transfer accuracy, since simply keeping the source image without modifying the expression can maximize the ID score (please see more analysis in the supplement). For example, InvertAvatar achieves strong identity similarity (high ID) but fails to faithfully reproduce the target expression in Fig. 6 (fourth row), resulting in worse EMO and AED scores in Table 2. InvertAvatar’s collapsed 3D shape is also reflected in poor APD due to inaccurate pose transfer.

While our training data is synthesized using X-NeMo and LP3D, we restrict X-NeMo to operate on frontal images only and use LP3D to estimate multi-view images to resolve identity drift. This improves consistency beyond that of the teacher model (X-NeMo) as evidenced by our method’s better ID metric and consistency metric versus X-NeMo (Table 2).

Method	MEt3R ↓	PSNR ↑	SSIM ↑	LPIPS ↓	AED ↓	IoU ↑	Mem. ↓	FPS ↑
Ours (128 × 128)	0.073	23.18	0.785	0.104	0.507	0.88	0.32 GB	132.87
w/ DINO-v2 Encoding	0.074	22.88	0.776	0.110	0.597	0.87	4.32 GB	22.47
w/ Real Dataset	0.073	23.10	0.782	0.109	0.543	0.88	0.32 GB	132.87
w/ Spatial Deformation	0.075	23.27	0.788	0.109	0.634	0.89	0.27 GB	141.61

Table 3: Ablation studies with self-reenactment.

Our method animates faces at 107.31 FPS, surpassing all other baselines, including ones using attention mechanisms [16, 75]. 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.

5.2 Ablation Studies

We validate each component of our model using the self-reenactment task on the same VOODOO-XP test set with a rendering resolution of 128×128 without an adversarial loss for more stable comparison. We use the exact same metrics for quantitatively evaluating self-reenactment as before along with AED for measuring expression transfer accuracy, reported in Table 3 and Fig. 8.

Choice of motion encoder. We study the importance of the selected motion encoder. Instead of the motion encoder in X-Nemo [109] we use DINO-v2 [51] as in [75]. Notably, this encoder is computationally more expensive than our selected motion encoder [109]. It increases the memory consumption and reduces the FPS. It fails to capture expression details in Fig. 8 and leads to a worse AED.

Usage of diffusion model. We study the necessity of using a synthetic dataset generated with a pre-trained diffusion model by using a real dataset instead. We use CelebVText [103] 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 Ψ , similarly to the previous dynamic modeling methods (e.g., [24, 45, 46, 87]). Since we do not need to go through the non-linear decoder in this setup, the memory consumption and FPS are slightly improved. However, the expression transfer accuracy is significantly compromised, leading to worsened AED. This is also reflected in Fig. 8.

6 Discussion

Conclusion. In this work, we propose an instant 4D method that encodes a single image into 3D-consistent, efficient yet expressive avatar. We propose



Fig. 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.

an animation representation that deforms both the Gaussian appearance and geometry based on the encoded motion basis vectors, inspired by bases representations in traditional facial animation. We believe our method paves the way for a real-time expressive representation, enabling real-world applications such as digital twins, where real-time performance, and controllability are critical.

Limitation & Future work. Our method finds it hard to extend to high-quality eyewear reconstruction due to thin structures, and complex reflection and refraction. We plan to extend our method into disentangling appearance properties (lighting). Besides, even though we demonstrate our method using only image-driven 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 [36, 96–98] or works that authenticate the authorized driving subject of the avatar [56].

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. Aneja, S., Weiss, S., Baeza, I., Chandran, P., Zoss, G., Nießner, M., Bradley, D.: Scaffoldavatar: High-fidelity gaussian avatars with patch expressions. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: Computer Vision and Pattern Recognition (CVPR) (2024)
3. Barthel, F., Beckmann, A., Morgenstern, W., Hilsmann, A., Eisert, P.: Gaussian splatting decoder for 3d-aware generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7963–7972 (2024)

4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of SIGGRAPH (1999)
5. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13786–13795 (2020)
6. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
7. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. arXiv preprint arXiv:2504.14899 (2025)
8. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16123–16133 (2022)
9. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2403–2410 (2025)
10. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025)
11. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems* **37**, 57642–57670 (2024)
12. Chu, X., Li, Y., Zeng, A., Yang, T., Lin, L., Liu, Y., Harada, T.: Gpavatar: Generalizable and precise head avatar from image(s) (2024)
13. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
15. Deng, Y., Wang, D., Ren, X., Chen, X., Wang, B.: Learning one-shot 4d head avatar synthesis using synthetic data. arXiv preprint arXiv:2311.18729 (2023)
16. Deng, Y., Wang, D., Wang, B.: Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. arXiv (2024)
17. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF International conference on Computer Vision. pp. 14398–14407 (2021)
18. Drobyshev, N., Casademunt, A.B., Vougioukas, K., Landgraf, Z., Petridis, S., Pantic, M.: Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8498–8507 (2024)
19. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: Megaportraits: One-shot megapixel neural head avatars. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2663–2671 (2022)
20. Egger, B., Smith, W.A., Tewari, A., Wuhler, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)* **39**(5), 1–38 (2020)

21. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
22. Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., Lu, Y.: High-fidelity and freely controllable talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5609–5619 (2023)
23. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Monophm: Dynamic head reconstruction from monocular videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10747–10758 (2024)
24. Giebenhain, S., Kirschstein, T., Rünz, M., Agapito, L., Nießner, M.: Npga: Neural parametric gaussian avatars. arXiv preprint arXiv:2405.19331 (2024)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
26. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168 (2024)
27. He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., Bo, L.: Lam: Large avatar model for one-shot animatable gaussian head. arXiv preprint arXiv:2502.17796 (2025)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
29. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4220–4230 (2024)
30. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
31. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
32. Ki, T., Min, D., Chae, G.: Learning to generate conditional tri-plane for 3d-aware expression controllable portrait animation. arXiv preprint arXiv:2404.00636 (2024)
33. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
34. Kirschstein, T., Giebenhain, S., Tang, J., Georgopoulos, M., Nießner, M.: Gghead: Fast and generalizable 3d gaussian heads. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
35. Kirschstein, T., Romero, J., Sevastopolsky, A., Nießner, M., Saito, S.: Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars (2025), <https://arxiv.org/abs/2502.20220>
36. Koki, N.: StyleGAN3 Synthetic Image Detection (10 2021), <https://github.com/NVlabs/stylegan3-detector>
37. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. In: arXiv (2024)
38. Lewis, J.P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F., Deng, Z.: Practice and Theory of Blendshape Facial Models. In: Eurographics 2014 - State of the Art Reports (2014)

39. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: 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), <https://doi.org/10.1145/3130800.3130813>
40. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.* **36**(6) (Nov 2017). <https://doi.org/10.1145/3130800.3130813>, <https://doi.org/10.1145/3130800.3130813>
41. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: 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*. pp. 17969–17978 (2023)
42. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems* **36** (2024)
43. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3377–3386 (2022). <https://doi.org/10.1109/CVPR52688.2022.00338>
44. Lin, C., Lin, Y., Pan, P., Yu, Y., Yan, H., Fragkiadaki, K., Mu, Y.: Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065* (2025)
45. Ling, H., Kim, S.W., Torralba, A., Fidler, S., Kreis, K.: 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*. pp. 8576–8588 (2024)
46. Liu, I., Su, H., Wang, X.: Dynamic gaussians mesh: Consistent mesh reconstruction from monocular videos. *arXiv preprint arXiv:2404.12379* (2024)
47. Liu, R., Ma, B., Zhang, W., Hu, Z., Fan, C., Lv, T., Ding, Y., Cheng, X.: 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*. pp. 2114–2123 (2024)
48. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900* (2024)
49. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems* **35**, 22438–22450 (2022)
50. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: *ECCV* (2020)
51. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
52. Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: Dpe: Disentanglement of pose and expression for general video portrait editing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 427–436 (2023)

53. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
54. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
55. Paskaleva, R., Holubakha, M., Ilic, A., Motamed, S., Van Gool, L., Paudel, D.: A unified and interpretable emotion representation and expression generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2447–2456 (2024)
56. Prashnani, E., Nagano, K., Mello, S.D., Luebke, D., Gallo, O.: Avatar fingerprinting for authorized use of synthetic talking-head videos. ECCV (2024)
57. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. <https://arxiv.org/abs/2011.13961> (2020)
58. Qiu, D., Fei, Z., Wang, R., Bai, J., Yu, C., Fan, M., Chen, G., Wen, X.: Skyreels-a1: Expressive portrait animation in video diffusion transformers. arXiv preprint arXiv:2502.10841 (2025)
59. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: 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)
60. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13759–13768 (2021)
61. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: 3d facial expressions through analysis-by-neural-synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2490–2501 (2024)
62. Sang, S., Zhi, T., Gu, T., Liu, J., Luo, L.: Lynx: Towards high-fidelity personalized video generation. arXiv preprint arXiv:2509.15496 (2025)
63. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2377–2386 (2019)
64. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. *Advances in neural information processing systems* **32** (2019)
65. Siarohin, A., Menapace, W., Skorokhodov, I., Olszewski, K., Lee, H.Y., Ren, J., Chai, M., Tulyakov, S.: Unsupervised volumetric animation. arXiv preprint arXiv:2301.11326 (2023)
66. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019)
67. Stengel, M., Nagano, K., Liu, C., Chan, M., Trevithick, A., Mello, S.D., Kim, J., Luebke, D.: Ai-mediated 3d video conferencing. In: ACM SIGGRAPH Emerging Technologies (2023)
68. Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20991–21002 (2023)

69. Tang, J., Davoli, D., Kirschstein, T., Schoneveld, L., Niessner, M.: Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5546–5558 (2025)
70. Tang, J., Zhang, B., Yang, B., Zhang, T., Chen, D., Ma, L., Wen, F.: 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *IEEE Transactions on Visualization and Computer Graphics* (2023)
71. Taubner, F., Zhang, R., Tuli, M., Bahmani, S., Lindell, D.B.: MVP4D: Multi-view portrait video diffusion for animatable 4D avatars (2025), <https://arxiv.org/abs/2510.12785>
72. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5318–5330. IEEE Computer Society (2025)
73. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision. pp. 244–260. Springer (2024)
74. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* (2021), <https://www.nature.com/articles/s42256-020-00280-0>
75. Tran, P., Zakharov, E., Ho, L.N., Hu, L., Karmanov, A., Agarwal, A., Goldwhite, M., Venegas, A.B., Tran, A.T., Li, H.: Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *arXiv preprint arXiv:2405.16204* (2024)
76. Tran, P., Zakharov, E., Ho, L.N., Tran, A.T., Hu, L., Li, H.: Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10336–10348 (2024)
77. Trevithick, A., Chan, M., Stengel, M., Chan, E.R., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. In: *ACM Transactions on Graphics (SIGGRAPH)* (2023)
78. Varanka, T., Khor, H.Q., Li, Y., Wei, M., Kung, H., Sebe, N., Zhao, G.: Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175* (2024)
79. Wang, C., Tian, K., Zhang, J., Guan, Y., Luo, F., Shen, F., Jiang, Z., Gu, Q., Han, X., Yang, W.: V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511* (2024)
80. Wang, D., Deng, Y., Yin, Z., Shum, H.Y., Wang, B.: Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17979–17989 (2023)
81. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. *arXiv preprint arXiv:2011.15126* (2020)
82. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. In: *International Conference on Learning Representations* (2022)
83. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Lia: Latent image animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–16 (2024)
84. Wang, Y., Yang, D., Chen, X., Bremond, F., Qiao, Y., Dantcheva, A.: Lia-x: Interpretable latent portrait animator. *arXiv preprint arXiv:2508.09959* (2025)

85. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
86. Wei, H., Yang, Z., Wang, Z.: Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694* (2024)
87. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20310–20320 (2024)
88. Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., Tong, X.: Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems* **35**, 36188–36201 (2022)
89. Xie, Y., Xu, H., Song, G., Wang, C., Shi, Y., Luo, L.: X-portrait: Expressive portrait animation with hierarchical motion attention. In: *ACM SIGGRAPH 2024 Conference Papers*. pp. 1–11 (2024)
90. Xu, C., Liu, Y., Xing, J., Wang, W., Sun, M., Dan, J., Huang, T., Li, S., Cheng, Z.Q., Tai, Y., 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*. pp. 1292–1302 (2024)
91. Xu, H., Song, G., Jiang, Z., Zhang, J., Shi, Y., Liu, J., Ma, W., Feng, J., Luo, L.: Omniavatar: Geometry-guided controllable 3d head synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12814–12824 (2023)
92. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. *Advances in neural information processing systems* **32** (2019)
93. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation (2024)
94. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems* **37**, 660–684 (2024)
95. Xu, Z., Yu, Z., Zhou, Z., Zhou, J., Jin, X., Hong, F.T., Ji, X., Zhu, J., Cai, C., Tang, S., et al.: Hunyuanportrait: Implicit condition control for enhanced portrait animation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 15909–15919 (2025)
96. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: 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)
97. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 22412–22423 (2023)
98. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: Deepfakebench: A comprehensive benchmark of deepfake detection. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 4534–4565. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf
99. Yang, S., Li, H., Wu, J., Jing, M., Li, L., Ji, R., Liang, J., Fan, H.: Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851* (2024)

100. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
101. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022)
102. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision* **129**, 3051–3068 (2021)
103. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A large-scale facial text-video dataset. In: CVPR (2023)
104. Yu, Z., Li, T., Sun, J., Shapira, O., Park, S., Stengel, M., Chan, M., Li, X., Wang, W., Nagano, K., De Mello, S.: Gaia: Generative animatable interactive avatars with expression-conditioned gaussians. *SIGGRAPH Conference Papers '25* (2025)
105. Yu, Z., Li, T., Sun, J., Shapira, O., Park, S., Stengel, M., Chan, M., Li, X., Wang, W., Nagano, K., Mello, S.D.: GAIA: Generative animatable interactive avatars with expression-conditioned gaussians. In: *ACM SIGGRAPH* (2025)
106. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
107. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3657–3666 (2022)
108. Zhao, X., Sun, J., Wang, L., Suo, J., Liu, Y.: Invertavatar: Incremental gan inversion for generalized head avatars. In: *ACM SIGGRAPH 2024 Conference Papers*. pp. 1–10 (2024)
109. Zhao, X., Xu, H., Song, G., Xie, Y., Zhang, C., Li, X., Luo, L., Suo, J., Liu, Y.: X-nemo: Expressive neural motion reenactment via disentangled latent attention. arXiv preprint arXiv:2507.23143 (2025)
110. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4176–4186 (2021)
111. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 146–155 (2016)
112. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. arXiv preprint arXiv:2311.08581 (2023)