UIKA: Fast Universal Head Avatar from Pose-Free Images

Zijian Wu[1,2,*] Boyao Zhou[2,#] Liangxiao Hu[2] Hongyu Liu[2,3] Yuan Sun[2,4]
Xuan Wang[2,4] Xun Cao[1] Yujun Shen[2] Hao Zhu[1,†]

[1]Nanjing University [2]Ant Group [3]HKUST [4]Xi'an Jiaotong University

Figure 1. We present UIKA, a novel **feed-forward** approach for high-fidelity 3D Gaussian head avatar reconstruction from **an arbitrary number** of input images (e.g., *a single portrait image* or *multi-view captures*) **without requiring** extra camera or expression annotations.

Abstract

We present UIKA*, a feed-forward animatable Gaussian head model from an arbitrary number of unposed inputs, including a single image, multi-view captures, and smartphone-captured videos. Unlike the traditional avatar method, which requires a studio-level multi-view capture system and reconstructs a human-specific model through a long-time optimization process, we rethink the task through the lenses of model representation, network design, and data preparation. First, we introduce a UV-guided avatar modeling strategy, in which each input image is associated with a pixel-wise facial correspondence estimation. Such correspondence estimation allows us to reproject each valid pixel color from screen space to UV space, which is independent of camera pose and character expression. Furthermore, we design learnable UV tokens on which the attention mechanism can be applied at both the screen and UV levels. The learned UV tokens can be decoded into canonical Gaussian attributes using aggregated UV information from all input views. To train our large avatar model, we additionally prepare a large-scale, identity-rich synthetic training dataset. Our method significantly outperforms existing approaches in both monocular and multi-view settings. Project page: https://zijian-wu.github.io/uika-page/*

1. Introduction

Creating a 3D-aware human portrait avatar is a crucial research area for downstream applications such as telepresence systems, the filmmaking industry, and virtual reality. This area remains challenging in two ways: lifelike avatar quality and a flexible capture setup. Our goal is to reconstruct a photo-realistic avatar model from an arbitrary number of unposed images, eliminating the requirement for estimating camera and expression parameters.

Early 2D approaches [21, 41, 43, 68, 75, 86] leverage the powerful generative capabilities of GANs [22, 29, 30] to drive source images by integrating facial landmarks [63, 85] or latent codes [3] as control signals. Recent methods [19, 44, 45, 91] leverage advances in diffusion models to improve animation performance further. Although these 2D approaches achieve promising results, they often exhibit long inference times and are not robust to extreme camera poses due to the lack of explicit 3D representation.

In terms of 3D avatar modeling, classic methods [35, 56, 72, 88] typically require long-term optimization for a specific identity using studio-level videos. In particular, a sophisticated multi-view camera system is always necessary for comprehensive 3D reconstruction with the representation of either NeRF [49] or Gaussian-Splatting [4, 56, 70, 77]. Such methods demand accurate camera calibration, while some approaches [40, 77] rely on computationally intensive post-processing networks, thereby hindering their practical deployment in downstream applica-

* Work done during an internship at Ant Group.
# Project lead. † Corresponding author.

Method	Inputs	FF	PF	RTA
GAGAvatar [7]	1	✓	✓	✓
Portrait4D-v2 [13]	1	✓	✗	✗
AvatarArtist [42]	1	✓	✓	✗
LAM [25]	1	✓	✓	✓
Avat3r [34]	4	✓	✗	✗
CAP4D [66]	≥ 1	✗	✗	✓
GPAvatar [8]	≥ 1	✓	✓	✗
InvertAvatar [90]	≥ 1	✗	✗	✗
DiffusionRig [14]	≥ 1	✗	✗	✗
Ours	≥ 1	✓	✓	✓

Table 1. Comparison with state-of-the-art 3D head reconstruction methods. **FF** denotes a feed-forward pipeline that requires no test-time optimization or fine-tuning. **PF** indicates pose-free input, i.e., camera and expression parameters are not required. **RTA** denotes real-time animatability ($\geq$ 30 FPS).

tions. Even if some approaches [6] take monocular videos as input, they typically rely on high-precision expression capture data to ensure accurate avatar modeling. Regardless of whether monocular or multi-view data are used as input, such optimization-based methods remain fundamentally constrained in generalizing to novel portrait expressions and camera poses.

Recently, an increasing number of approaches have shifted towards feed-forward avatar modeling by leveraging a large reconstruction model [28, 78] from a single image [7, 25] or limited images [34, 90]. LAM [25] and GAGAvatar [7] reconstruct head avatars from a single input image and are typically trained on monocular portrait videos, which often limits their ability to generalize to novel view synthesis under large camera poses. Avat3r [34], in contrast, requires a fixed set of four calibrated input images, a restrictive setting that reduces practical applicability and also confines training to existing identity-scarce multi-view datasets, thereby limiting generalization. More recent methods, such as GPAvatar [8] and PF-LHM [58], extend the input setting to an arbitrary number of images, but they lack explicit correspondence across input frames, making multi-view information aggregation less reliable. Tab. 1 summarizes the flexibility and efficiency of our method relative to the baselines.

In this work, we present UIKA, a novel feed-forward framework for animatable Gaussian head modeling from an arbitrary number of unposed input images. To establish explicit correspondences across unposed input images, we design a facial correspondence estimator that supports an arbitrary number of inputs, inspired by Pixel3DMM [20]. Given a set of unposed input images, our facial correspondence estimator first estimates UV coordinates in the pixel level, and the corresponding colors are reprojected onto the shared UV space. The reprojected images and original images are embedded with a frozen DINOv3 [65] encoder followed by a trainable lightweight fusion module, producing multi-scale features from both screen and UV spaces. Typically, prior works [25, 57, 58, 96] build a connection between learnable tokens and screen space features by using Transformer blocks, which lack a structural correspondence. Other than conventional screen attention, we introduce a UV attention branch that enables our learnable UV tokens to interact with UV-space features. This design allows our model to simultaneously leverage local details from the screen space and structured global context from the reprojected UV space in a complementary manner. Furthermore, the processed learnable tokens can be decoded into canonical Gaussian primitives, including appearance and geometry attributes. Although the predicted appearance is globally coherent to input images, it typically lacks realistic details. Thus, we propose a self-adaptive fusion strategy per Gaussian primitive that blends these two color sources via learned weights. This design dynamically balances accurate but potentially incomplete local cues against globally coherent yet sometimes imprecise predictions, leading to high reconstruction quality. In addition, the resulting canonical Gaussian head avatar is immediately animatable using standard linear blend skinning and supports real-time rendering at 220 FPS, in contrast to approaches [7, 8, 51] that rely on an additional neural renderer at inference time to produce the final outputs. To strengthen multi-view learning, we construct a synthetic dataset with diverse identities and rich expression variations. Training on our collected and synthetic datasets, our method outperforms the prior state of the art in both monocular and multi-view settings. The contributions of our work can be summarized as:

- We present UIKA, a feed-forward framework that reconstructs animatable 3D Gaussian head avatars from an arbitrary number of unposed input images.
- We design a novel UV attention branch that leverages predicted facial correspondence to efficiently align multi-observation within a unified canonical space, facilitating robust cross-image information matching.
- We propose a self-adaptive fusion strategy to dynamically balance predicted global appearance and reprojected local details from input images to improve overall quality.
- To mitigate the limited identity diversity, view coverage, and motion in existing datasets, we curate a large-scale, multi-view synthetic head dataset for head avatar reconstruction and generation.

2. Related Work

2.1. Generative 2D Head Avatar

Recent years have witnessed significant advances in image-based talking head synthesis, with most state-of-the-art approaches operating within a 2D generative framework [3, 15, 21, 41, 43, 62–64, 68, 75, 86]. Two families of generative models have been widely adopted in recent years. Early work leverages Generative Adversarial Networks (GANs)

to learn motion-driven warping of a static reference image. To enable faithful transfer of facial expressions and dynamic motion, diverse motion descriptors are used, including 2D facial keypoints [63, 85], depth maps [26], and latent embeddings [3]. Diffusion models, renowned for their generative capability, have been adopted in several recent approaches, especially for one-shot talking-face generation. Benefiting from strong priors learned from large-scale image datasets, diffusion-based methods [14, 44, 71, 74] can drive stylized or out-of-domain faces. However, their inherent 2D modeling assumptions and the lack of 3D awareness cause them to struggle with large pose variations and consistent head-movement synthesis.

2.2. Optimize-based 3D Head Avatar

Optimization-based methods for 3D head avatars commonly employ explicit or implicit 3D representations, including meshes [23], NeRFs [18, 27, 49, 76, 84, 97], SDFs [93], point-based representations [94], and 3D Gaussians [6, 72, 79, 92], to reconstruct per-subject avatars. Among these, methods that rely on monocular video input [6, 18, 72, 76, 88, 97] can reconstruct faithful 3D head avatars for the observed viewpoints but struggle to generalize to novel views. Other approaches [27, 49, 79, 84, 92] exploit large-scale multi-view datasets [1, 33, 47, 52, 80] to learn expressive, generalizable priors over both geometry and appearance, achieving state-of-the-art visual realism. Despite their impressive reconstruction quality, the need for per-subject optimization limits their practicality and broader applicability. More recently, diffusion-based frameworks such as CAP4D [66] synthesize multi-view images from a single portrait to guide avatar reconstruction. While this strategy improves identity preservation across views, it still requires time-consuming optimization, hindering deployment in real-time or one-shot settings.

2.3. Feed Forward 3D Head Avatar

Feed-forward avatar reconstruction methods create realistic, animatable 3D head avatars from as few as one or a handful of input images by learning expressive priors from large-scale datasets, whether monocular [73] or multi-view [33, 47]. Prior work [5, 8, 38, 39, 46, 49, 67, 81–83, 95] introduces NeRFs [48] for feed-forward reconstruction of 3D head avatars, delivering high-fidelity results and precise control over novel camera viewpoints. To enhance performance, several methods incorporate 3D supervision either from monocular data [9, 11, 17] or synthetic multi-view data [12, 13, 42, 87]. However, NeRF-based architectures struggle to support real-time avatar animation. Recent methods [7, 25, 34] have demonstrated the utility of 3D Gaussian Splatting (3DGS) [31] for avatar modeling, achieving rapid rendering without sacrificing visual quality. Nevertheless, GAGAvatar [7] relies on a 2D-to-3D lifting scheme that fails to reconstruct regions not visible in the input plausibly. LAM [25] suffers reduced reconstruction accuracy under novel viewpoints due to its one-shot setting. Avat3R [34] departs from the one-shot setting, requiring multiple input views and rerunning the entire network inference to generate the 3D head Gaussian attributes for each new expression. In contrast, the proposed UIKA reconstructs realistic, animatable 3D head avatars from multiple unposed input images in a feed-forward manner.

3. Method

Given an arbitrary number of unposed images $\{\mathrm{I}_{\mathrm{s}}^{i}\}_{i=1}^{N}$, without additional camera or expression parameters, our goal is to reconstruct a high-fidelity and animatable Gaussian head avatar represented by a set of Gaussians [31] $\mathcal{G} = \{\mathbf{c}_k, o_k, \boldsymbol{\mu}_k, \boldsymbol{s}_k, \boldsymbol{r}_k\}_{k=1}^{M}$. The overall pipeline is illustrated in Fig. 2. Firstly, we present a facial correspondence estimator and introduce the color reprojection and aggregation in Sec. 3.1. Then, Sec. 3.2 introduces a novel UV attention branch into Transformer architecture. In the Sec. 3.3, we present the proposed self-adaptive fusion strategy in the UV decoder. We furthermore introduce our synthetic multi-view head dataset in Sec. 3.4. Finally, Sec. 3.5 outlines the training objectives.

3.1. Facial Correspondence Prediction

Facial correspondence estimator. Inspired by prior work [20, 69], we develop a facial correspondence estimator that accepts an arbitrary number of unposed images $\{\mathrm{I}_{\mathrm{s}}^{i}\}_{i=1}^{N}$ as input and predicts facial correspondence, in the format of pixel-aligned UV coordinates $\{\mathrm{U}^{i}\}_{i=1}^{N}$, $\mathrm{U} = (u, v) \in [0, 1]^2$, for corresponding input images as follow:

$$\mathrm{U}^{i} = \mathcal{U}\left(\mathrm{I}_{\mathrm{s}}^{i}\right), \quad i \in [1, N]; \tag{1}$$

where $\mathcal{U}(\cdot)$ denotes our facial correspondence estimator network. Specifically, the input images are processed with a frozen pre-trained VGGT [69] encoder, which extract robust feature representations. These features are subsequently decoded into dense UV coordinate maps through a trainable DPT head [59, 60]. Further architectural details are provided in the supplementary material.

Color reprojection & Aggregation. As shown in Fig. 2 (a), we reproject input images $\{\mathrm{I}_{\mathrm{s}}^{i}\}_{i=1}^{N}$ from screen-space into a shared UV space by leveraging the predicted facial correspondence $\{\mathrm{U}^{i}\}_{i=1}^{N}$, alleviating the ambiguity of camera pose and facial expression from different frames. Thus, we can obtain reprojected images $\{\mathrm{I}_{\mathrm{uv}}^{i}\}_{i=1}^{N}$ by pixel-to-pixel matching. We then aggregate all reprojected images into an averaged UV observation $\mathrm{I}_{\mathrm{aggr}}$ and a confidence map γ_{aggr}:

$$\mathrm{I}_{\mathrm{uv}}^{i} = \mathrm{Reproj}\left(\mathrm{I}_{\mathrm{s}}^{i}, \mathrm{U}^{i}\right), \quad i \in [1, N]; \tag{2}$$

$$\mathrm{I}_{\mathrm{aggr}},\ \gamma_{\mathrm{aggr}} \leftarrow \mathrm{Aggr}\left(\mathrm{I}_{\mathrm{uv}}^{1}, \mathrm{I}_{\mathrm{uv}}^{2}, \ldots, \mathrm{I}_{\mathrm{uv}}^{N}\right); \tag{3}$$

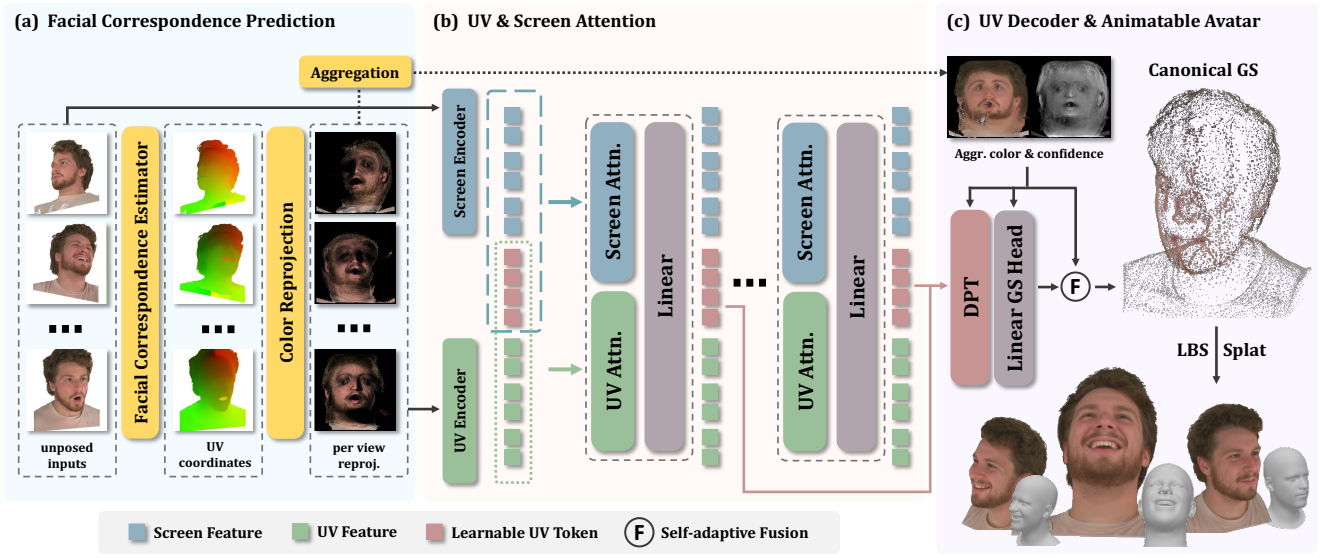

Figure 2. **Pipeline Overview**. Given a set of unposed input images, our pipeline begins with a facial correspondence estimator that predicts UV coordinates for valid facial pixels, and the corresponding colors are reprojected onto the shared UV space. The source images (screen space) and reprojected images (UV space) are encoded through two dedicated encoders, producing multi-scale features from both screen space and UV space. We then apply screen attention and UV attention to inject these into learnable UV tokens, which are then decoded into UV Gaussian attribute maps while incorporating the aggregated color and confidence map. The resulting canonical Gaussian head avatar supports animation via standard linear blend skinning and achieves real-time rendering at 220 FPS.

In practice, I_{aggr} is computed by pixel-wise averaging over the reprojected images. For each UV pixel, we count the number of valid projections n_{hit} and define the aggregated confidence as $\gamma_{\text{aggr}} \coloneqq \text{Norm}\left(\log(1+n_{\text{hit}})\right)$, here $\text{Norm}\left(\cdot\right)$ denotes *min–max normalization*.

3.2. UV & Screen Attention

As shown in Fig. 2 (b), given source input images $\{\mathrm{I}_{\text{s}}^{i}\}_{i=1}^{N}$ and their corresponding reprojected images $\{\mathrm{I}_{\text{uv}}^{i}\}_{i=1}^{N}$, we extract screen features $\mathcal{F}_{\text{s}}$ and UV features $\mathcal{F}_{\text{uv}}$ via:

$$\mathcal{F}_{\text{j}} = \mathcal{E}_{\text{j}}\left(\mathrm{I}_{\text{j}}^{1}\right) \oplus \mathcal{E}_{\text{j}}\left(\mathrm{I}_{\text{j}}^{2}\right) \oplus \cdots \oplus \mathcal{E}_{\text{j}}\left(\mathrm{I}_{\text{j}}^{N}\right); \tag{4}$$

where $\text{j} \in [\text{s}, \text{uv}]$ for either screen or UV space and $\oplus$ denotes the concatenation in the length dimension. Encoder $\mathcal{E}_{\text{j}}$ is composed of a frozen pretrained DINOv3 [65] backbone and a trainable lightweight CNN fusing features derived from both shallow and deep layers of the backbones.

To exploit semantic features from both screen space and UV space into our learnable UV tokens $\mathcal{Z} \in \mathbb{R}^{\text{L}_z \times \text{D}}$, we perform attention mechanism [16] Attn in both spaces:

$$\Delta\mathcal{Z}_{\text{j}},\ \Delta\mathcal{F}_{\text{j}} = \text{Attn}_{\text{j}}\left(\mathcal{Z}, \mathcal{F}_{\text{j}}\right); \tag{5}$$

$$\mathcal{Z}' = \mathcal{Z} + \text{MLP}\left(\mathcal{Z} + \Delta\mathcal{Z}_{\text{s}} + \Delta\mathcal{Z}_{\text{uv}}\right); \tag{6}$$

$$\mathcal{F}'_{\text{j}} = \mathcal{F}_{\text{j}} + \text{MLP}\left(\mathcal{F}_{\text{j}} + \Delta\mathcal{F}_{\text{j}}\right); \tag{7}$$

where $\text{j} \in [\text{s}, \text{uv}]$ for either screen or UV space and $\mathcal{Z}', \mathcal{F}'_{\text{j}}$ denotes the updated $\mathcal{Z}, \mathcal{F}_{\text{j}}$ in a Transformer block.

3.3. UV Decoder

UV Gaussian prediction. As shown in Fig. 2 (c), starting from the UV aggregation map $\{\mathrm{I}_{\text{aggr}}, \gamma_{\text{aggr}}\}$ produced in Sec. 3.1 and the multi-depth learned UV tokens $\mathcal{Z}^{l}$ obtained from our Transformer in Sec. 3.2, we feed them into our UV decoder $\mathcal{D}\left(\cdot\right)$ to obtain the canonical Gaussian attributes:

$$\{\hat{\boldsymbol{c}}_k, w_k, o_k, \boldsymbol{\Delta\mu}_k, \boldsymbol{s}_k, \boldsymbol{r}_k\}_{k=1}^{M} = \mathcal{D}\left(\mathcal{Z}^{l}; \mathrm{I}_{\text{aggr}}, \gamma_{\text{aggr}}\right); \tag{8}$$

where $l = 3, 6, 9, 12$ denotes different depth of our Transformer blocks and $\hat{\boldsymbol{c}}_k, w_k, o_k, \boldsymbol{\Delta\mu}_k, \boldsymbol{s}_k, \boldsymbol{r}_k$ represent the predicted color, color fuse weight, opacity, position offset, scaling, and rotation of the canonical Gaussian attributes, respectively. And we define a **self-adaptive fusion strategy** to balance the impact between the predicted $\boldsymbol{c}_k$ and real-captured aggregation $\boldsymbol{c}_k^{aggr}$ via:

$$\boldsymbol{c}_k = w_k * \hat{\boldsymbol{c}}_k + (1 - w_k) * \boldsymbol{c}_k^{aggr}, \quad \boldsymbol{c}_k^{aggr} \subset \mathrm{I}_{aggr}; \tag{9}$$

The final canonical Gaussians $\mathcal{G}$ is updated with color $\boldsymbol{c}_k$ and position $\boldsymbol{\mu}_k^{m} + \boldsymbol{\Delta\mu}_k$, where $\boldsymbol{\mu}_k^{m}$ represents the initial position on the template FLAME [37] mesh surface.

Novel expression animation. Given the reconstructed canonical Gaussian head avatar, we reenact it under novel FLAME poses and expressions. Each Gaussian originates from a valid UV pixel; FLAME UV rasterization [72, 88] provides its associated triangle assignments and corresponding barycentric coordinates. Using barycentric interpolation over the corresponding mesh triangle, we obtain per-Gaussian quantities, e.g., LBS weights, posedirs, and

shapedirs. Conditioned on target FLAME pose and expression, we then apply standard vertex-based linear blend skinning (LBS) to deform the Gaussians from the canonical space to the posed space, yielding the animated head avatar. Finally, we obtain the rendered images I_{pred} through differentiable Gaussian splatting $\mathcal{R}(\cdot)$ as follows:

$$\mathrm{I}_{\text{pred}} = \mathcal{R}\left(\mathrm{LBS}\left(\mathcal{G},\ \Theta\right),\ \Pi\right); \tag{10}$$

where Θ denotes the target FLAME pose and expression parameters and Π denotes the target camera parameters.

3.4. Synthetic Multi-view Head Dataset Curation

Prior work predominantly trains on head datasets that are monocular, leading to restricted camera viewpoints and limited expression variability dominated by speech-related motions. Although recent multi-view datasets such as NeRSemble [33], Ava-256 [47], and RenderMe-360 [52] alleviate the view limitation, they suffer from small identity counts due to costly capture setups and are typically recorded under studio lighting, hindering generalization to in-the-wild conditions. To address these limitations, we introduce a scalable data curation pipeline that combines a 3D head generation model with an efficient 2D portrait animation model to produce identity-diverse, multi-view sequences with extreme expressions. Concretely, we leverage SphereHead [36], a 3D head generator trained on in-the-wild images spanning wide camera poses to synthesize multi-view and 3D consistent head renderings. For each identity, we sample 9 fixed viewpoints and render the corresponding views. We then employ LivePortrait [24], an efficient 2D portrait animation model that drives a source head image using a driver video. For each view, we select the same driver sequence from a curated motion library to animate the rendered view, producing temporally synchronized multi-view head sequences. In total, we curate over 7,500 identities, each with 9 views and more than 13,000 frames per identity, covering complex and exaggerated facial expressions while avoiding expensive studio capture and improving robustness to in-the-wild scenarios. Further dataset details are provided in the supplementary material.

3.5. Training Objectives

During training, we randomly sample 1 to N_{ref} frames from the same video as source inputs to reconstruct the canonical Gaussian representation, and additionally sample N_{d} frames as driving and target views for reenactment supervision. We supervise the rendered images against the corresponding ground truth frames using a photometric objective that combines L1, SSIM, and VGG-based perceptual losses:

$$\mathcal{L}_{\text{l1}} = ||\mathrm{I}_{\text{pred}} - \mathrm{I}_{\text{gt}}||_1; \tag{11}$$

$$\mathcal{L}_{\text{lpips}} = \mathrm{LPIPS}\left(\mathrm{I}_{\text{pred}},\ \mathrm{I}_{\text{gt}}\right); \tag{12}$$

$$\mathcal{L}_{\text{ssim}} = \mathrm{SSIM}\left(\mathrm{I}_{\text{pred}},\ \mathrm{I}_{\text{gt}}\right); \tag{13}$$

We also add geometry regularization to prevent Gaussians from drifting too far from its initialized position via:

$$\mathcal{L}_{\text{reg}} = ||\max\left(\boldsymbol{\Delta\mu},\ \epsilon\right)||_2; \tag{14}$$

where $\boldsymbol{\Delta\mu}$ is the predicted Gaussian offsets, ϵ is an hyper parameters set close to 0. The overall training objectives are the weighted sum of the image supervision over all supervised frames and the offset regularization:

$$\mathcal{L} = \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}; \tag{15}$$

where λ_{l1} and λ_{lpips} are 1.0, λ_{ssim} and λ_{reg} are 0.1.

4. Experiments

4.1. Experiments settings

Implementation Details. We implement our framework using PyTorch [53]. Our Transformer architecture consists of $\mathrm{L} = 12$ MM-Transformer [16] blocks, each equipped with $\mathrm{h} = 16$ attention heads and a hidden feature dimension $\mathrm{D} = 1024$. For per-view reprojected images, the resolution is kept identical to the source inputs in 512×512. We introduce learnable UV tokens of length $\mathrm{L}_{\text{z}} = 9216$, which are then reshaped into a 96×96 grid before being fed into the UV decoder. These UV tokens are jointly processed with the UV aggregate map via a DPT-based [59, 60] decoder, producing a UV space feature map of size $384 \times 384 \times 256$. Subsequently, we rasterize the UV representation of the FLAME [37] mesh using PyTorch3D [61] to obtain a valid UV mask, from which we sample approximately 130K feature points via bilinear interpolation. These features are then passed through two fully connected layers, followed by separate MLP heads for each Gaussian attribute, decoding the corresponding property values. In practice, we set $\mathrm{N}_{\text{ref}} = 16$ and $\mathrm{N}_{\text{d}} = 8$. We train the model for 150K steps using the Adam [32] optimizer and a cosine warm-up learning rate scheduler. Our training is conducted on 32 NVIDIA H20 GPUs, taking approximately two weeks to complete.
Datasets. We train our model on four datasets: VFHQ [73], HDTF [89], NeRSemble-v2 [33] and our synthetic dataset. For all datasets, we obtain pose and expression parameters of FLAME 2023 w/ jaw version and camera parameters using the VHAP [54] tracker, following the preprocessing protocol in GaussianAvatars [55]. To prepare model inputs, we first detect the facial region using the method from GAGAvatar [7]. We then enlarge the bounding box, crop the region of interest, and resize it to 512×512. We also perform background removal on each input image and randomly replace the background with one of three solid colors: black, white, or gray. For evaluation, we use 50 test clips from VFHQ, together with 25 identity clips split from NeRSemble-v2.
Evaluation Metrics. We evaluate our model under two input configurations: monocular and multi-view settings.

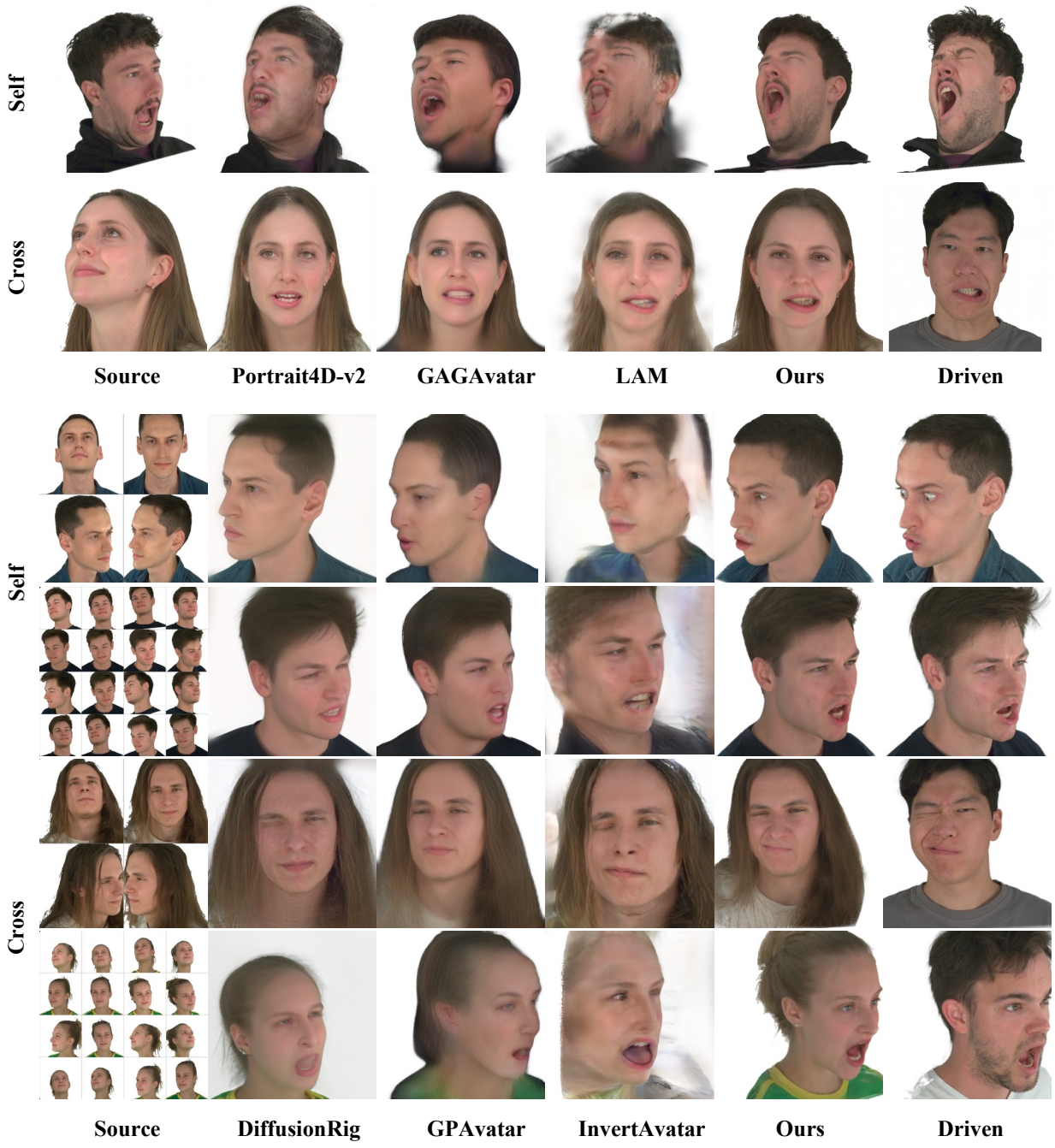

Figure 3. **Qualitative results for comparison to baselines in both monocular and multi-view settings in NeRSemble-v2 datasets.**

In both cases, we focus on two reenactment scenarios: self reenactment and cross reenactment, and report performance across multiple quantitative metrics. For self reenactment, where ground-truth images are available, we measure image reconstruction quality using PSNR, SSIM, and LPIPS. Identity similarity (CSIM) is computed as the cosine distance between facial feature vectors extracted by ArcFace [10]. Expression and head pose fidelity are assessed via the Average Expression Distance (AED) and Average Pose Distance (APD), respectively, estimated by the 3DMM-based facial parameter regressor from Deep3DFaceRecon [11]. Additionally, we measure facial geometry consistency using Average Keypoint Distance (AKD), obtained from a facial landmark detector [2].

Method	Self Reenactment							Cross Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
Portrait4D-v2 [13]	21.03	0.859	0.134	0.688	0.094	0.113	3.718	0.654	0.132	0.149
GAGAvatar [7]	20.34	0.850	0.160	0.693	0.071	0.075	4.372	**0.678**	0.151	0.142
LAM [25]	18.29	0.810	0.206	0.602	0.104	0.112	4.631	0.612	0.126	0.130
Ours	**21.69**	**0.867**	**0.105**	**0.738**	**0.055**	**0.056**	**3.066**	0.649	**0.114**	**0.123**

Table 2. Quantitative results on the monocular setting in VFHQ and NeRSemble-v2 datasets.

Method	Self Reenactment							Cross Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
DiffusionRig [14]	16.97	0.768	0.395	0.598	0.209	0.138	9.585	0.616	0.263	0.218
GPAvatar [8]	17.11	0.783	0.313	0.553	0.129	0.108	6.423	0.492	0.210	0.168
InvertAvatar [90]	16.35	0.776	0.394	0.449	0.084	0.069	7.402	0.491	0.198	0.177
Ours	**22.50**	**0.855**	**0.120**	**0.740**	**0.064**	**0.063**	**3.437**	**0.666**	**0.145**	**0.153**

Table 3. Quantitative results on the multi-view setting in NeRSemble-v2 datasets.

For cross reenactment, where ground-truth images are unavailable, we evaluate performance using CSIM, AED, and APD metrics.

4.2. Main Results

Baselines. In the monocular input setting, we compare our approach against state-of-the-art methods, including LAM [25], GAGAvatar [7], and Portrait4D-v2 [13]. In the multi-view input setting, the SOTA baselines are listed as follows: InvertAvatar [90], GPAvatar [8], and DiffusionRig [14]. Among them, LAM, GAGAvatar, Portrait4D-v2 and GPAvatar are feed-forward methods. However, DiffusionRig requires a fine-tuning phase for each identity, which relies on iterative denoising steps, resulting in slow inference for approximately 30 minutes. Avat3r [34] is restricted to a fixed four-view input configuration; due to the absence of an open-source implementation, it is excluded from our baseline comparisons.

Monocular Setting. Our method achieves superior results in Tab. 2, especially for self reenactment. As a conditional generative model, Portrait4D-v2 can not handle extreme expressions in Fig. 3, because implicit control signal is not efficiently encoded. Similar to our method, GAGAvatar and LAM utilize canonical Gaussian representations. When target views are very different from the input source view, GAGAvatar and LAM could degrade the rendering results as shown in Fig. 3. In contrast, our method produces plausible and photo-realistic rendering results on the monocular setting. Further monocular comparison experiments are provided in the supplementary material.

Multi-view Setting. It is not trivial to tackle several images in different frames and from different views as inputs. DiffusionRig and InvertAvatar aggregate latent codes of all input images as conditions to guide a generative model, e.g., a 2D diffusion model or a 3D GAN, to generate the final results. However, such methods can not efficiently encode the expression from driven images, as shown in Fig. 3. Due to the lack of explicit correspondence modeling, GPAvatar and InvertAvatar could even degrade the rendering results when increasing input images. Thanks to our UV guided modeling, our method outperforms all these baselines in both self and cross reenactments in Tab. 3. Further multi-view comparison experiments are provided in the supplementary material.

Thanks to our UV attention branch and self-adaptive fusion strategy, our method is able to aggregate more and more observed information against initial occlusion in Fig. 4 (a) and to improve 3D consistency in Fig. 4 (b) as well as rendering details in Fig. 4 (c, d), when progressively increase number of input images. Our method also generalizes well to out-of-domain data, including samples from the Ava-256 dataset [47] (Fig. 6 (a)) and in-the-wild internet images (Fig. 6 (b)). Additional in-the-wild results are provided in the supplementary material.

Method	PSNR↑	LPIPS↓	AED↓	AKD↓
w/o synth	21.86	0.093	0.060	3.078
w/o uv_attn	22.21	0.091	0.056	3.086
w/o aggr	22.39	0.088	0.059	3.120
Ours	**22.61**	**0.082**	**0.055**	**3.037**

Table 4. Quantitative results for ablation study on the monocular setting in NeRSemble-v2 datasets for self reenactment.

4.3. Abalation Study

In the following, we study the efficacy of our designed choices for UIKA. The ablations are performed on the monocular setting in the NeRSemble-v2 dataset. Quantitative results are shown in Tab. 4.

UV attention branch. When removing the UV attention branch from our Transformer network, the learnable UV tokens perform attention only with screen tokens. Due to the lack of structural information, the ablated version suffers from a significant detail loss in Fig. 5 (c).

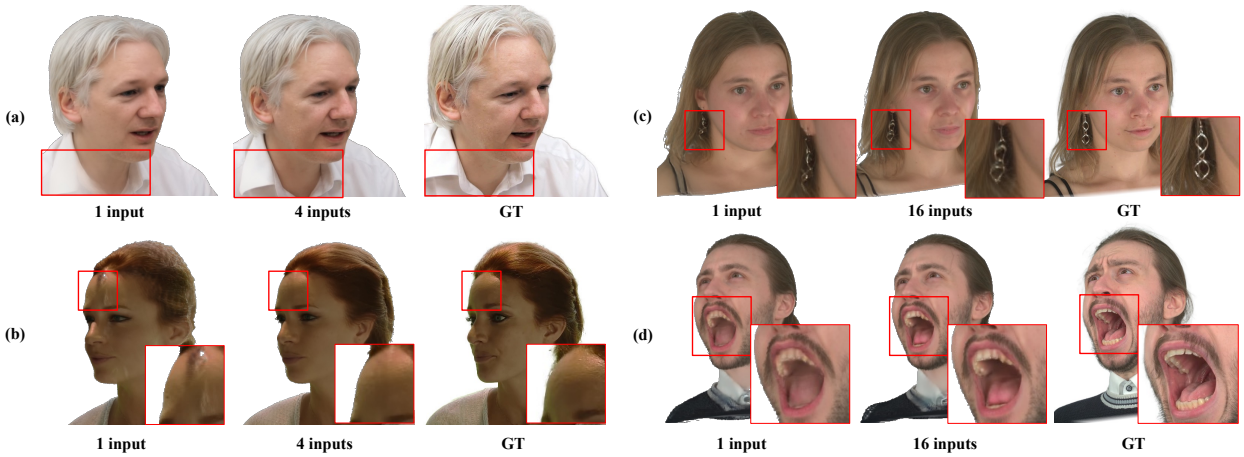

Figure 4. **Qualitative results of different numbers of input views in VFHQ and NeRSemble-v2 dataset.**

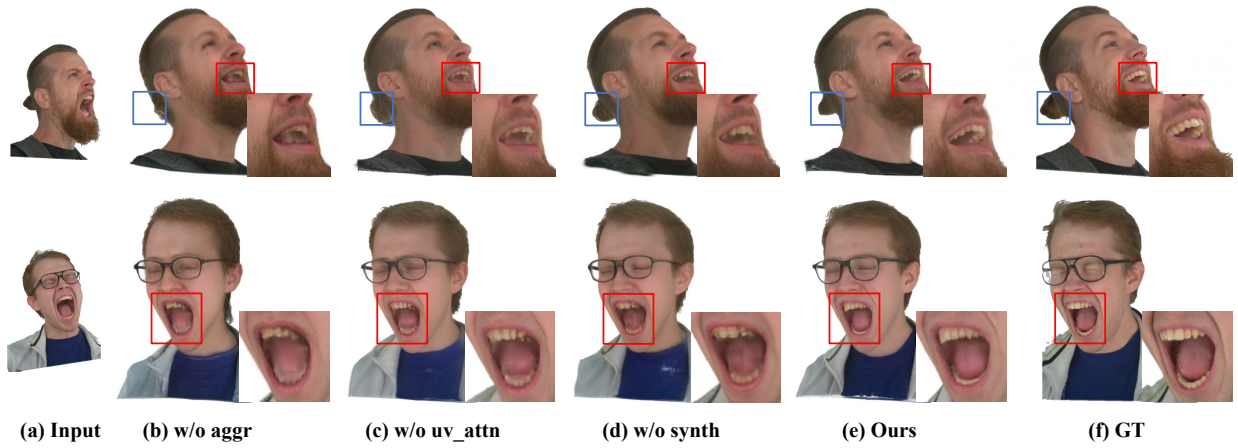

Figure 5. **Qualitative results for ablation study in the monocular settings in NeRSemble-v2 dataset.**

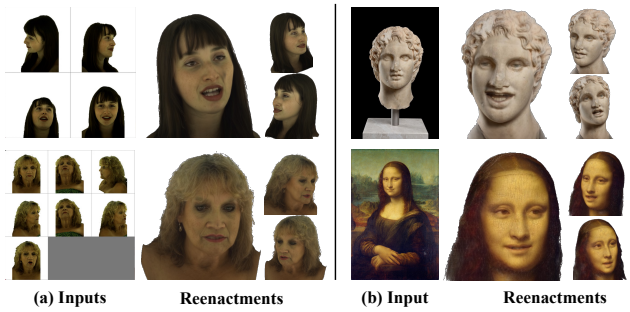

Figure 6. **Qualitative results for in-the-wild cases.**

Self-adaptive fusion strategy. In the ablated version, we do not add the aggregated UV map into our decoding stage. As shown in Fig. 5 (b), without injection from the observed image, it's hard to yield correct and coherent details.

Importance of our synthetic dataset. In this version, we trained our model only on VFHQ and NeRSemble-v2 dataset. Comparing to the results in Fig. 5 (d), our full model, Fig. 5 (e), preserves view consistency and reconstructs more high-frequency details when using multi-view synthetic data. Further ablation study experiments are provided in the supplementary material.

5. Conclusion

In this work, we present UIKA, a feed-forward framework for animatable Gaussian head avatar modeling from an arbitrary number of unposed inputs. By leveraging pixel-wise facial correspondence estimation, we introduce a UV-guided avatar modeling pipeline. We further design a novel UV attention branch to facilitate robust cross-image information matching. Finally, a self-adaptive fusion strategy is applied to guarantee plausible and complete avatar modeling. Our method achieves superior results in both monocular and multi-view settings, with a fast run time.

References

[1] Marcel C Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, et al. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 6

[3] 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. 1, 2, 3

[4] Hongrui Cai, Yuting Xiao, Xuan Wang, Jiafei Li, Yudong Guo, Yanbo Fan, Shenghua Gao, and Juyong Zhang. Hera: Hybrid explicit representation for ultra-realistic head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 260–270, 2025. 1

[5] 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. 3

[6] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 2, 3

[7] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3, 5, 7

[8] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 2, 3, 7

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 3

[10] 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

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3, 6

[12] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 3

[13] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 2, 3, 7

[14] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12736–12746, 2023. 2, 3, 7

[15] 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. 2

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 4, 5

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 3

[18] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[19] Xuan Gao, Jingtao Zhou, Dongyu Liu, Yuqi Zhou, and Juyong Zhang. Constructing diffusion avatar with learnable embeddings. In *ACM SIGGRAPH Asia Conference Proceedings*, 2025. 1

[20] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 2, 3, 14

[21] Yuan Gong, Yong Zhang, Xiaodong Cun, Fei Yin, Yanbo Fan, Xuan Wang, Baoyuan Wu, and Yujiu Yang. Toontalker: Cross-domain face reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7690–7700, 2023. 1, 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[23] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18664, 2022. 3

[24] 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. 5

[25] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 2, 3, 7

[26] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 3

[27] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[28] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[29] 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. 1

[30] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 1

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

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[33] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 3, 5, 15

[34] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025. 2, 3, 7, 15

[35] Jaeseong Lee, Taewoong Kang, Marcel Buehler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. Surfhead: Affine rig blending for geometrically accurate 2d gaussian surfel head avatars. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[36] Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation, 2024. 5

[37] 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. 4, 5, 16

[38] 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

[39] 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. 3

[40] Zhanfeng Liao, Yuelang Xu, Zhe Li, Qijing Li, Boyao Zhou, Ruifeng Bai, Di Xu, Hongwen Zhang, and Yebin Liu. Hhavatar: Gaussian head avatar with dynamic hairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[41] Hongyu Liu, Xintong Han, Chengbin Jin, Lihui Qian, Huawei Wei, Zhe Lin, Faqiang Wang, Haoye Dong, Yibing Song, Jia Xu, et al. Human motionformer: Transferring human motions with vision transformers. *arXiv preprint arXiv:2302.11306*, 2023. 1, 2

[42] Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. Avatarartist: Open-domain 4d avatarization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10758–10769, 2025. 2, 3

[43] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1, 2

[44] 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. 1, 3

[45] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Zhifeng Li, Wei Liu, Zhang linfeng, and Qifeng Chen. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *International Journal of Computer Vision (IJCV)*, 2025. 1

[46] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16910, 2023. 3

[47] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu,

Thomas Dauer, Peter Seleznik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 3, 5, 7, 15

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

[49] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 14

[51] Antonio Oroz, Matthias Nießner, and Tobias Kirschstein. Perchead: Perceptual head model for single-image 3d head reconstruction & editing, 2025. 2

[52] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. Renderme-360: a large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5, 15

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. Curran Associates, Inc., 2019. 5

[54] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 5

[55] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 5

[56] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1

[57] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14184–14194, 2025. 2

[58] Lingteng Qiu, Peihao Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. *arXiv preprint arXiv:2506.13766*, 2025. 2

[59] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 3, 5, 14

[60] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 3, 5, 14

[61] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 5

[62] 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

[63] 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. 1, 3

[64] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 2

[65] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2, 4

[66] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330, 2025. 2, 3

[67] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. 2023. 3

[68] 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. 1, 2

[69] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 14, 15

[70] Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, and Lizhuang Ma. 3d gaussian head avatars with expressive dynamic appearances by compact tensorial representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21117–21126, 2025. 1

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

[72] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 1, 3, 4

[73] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 3, 5, 15

[74] 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. 3

[75] 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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[76] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 3

[77] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2024. 1

[78] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 2

[79] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3d gaussian parametric head model. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 3

[80] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 3

[81] 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

[82] Houteng Yu, Hao Zhu, and Xun Cao. RealityAvatar: Comprehensive Head Avatar Generation with 360° Rendering . In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Los Alamitos, CA, USA, 2025. IEEE Computer Society.

[83] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 3

[84] Zhixuan Yu, Ziqian Bai, Abhimitra Meka, Feitong Tan, Qiangeng Xu, Rohit Pandey, Sean Fanello, Hyun Soo Park, and Yinda Zhang. One2avatar: Generative implicit head avatar for few-shot user adaptation. *arXiv preprint arXiv:2402.11909*, 2024. 3

[85] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 1, 3

[86] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 1, 2

[87] Jiawei Zhang, Lei Chu, Jiahao Li, Zhenyu Zang, Chong Li, Xiao Li, Xun Cao, Hao Zhu, and Yan Lu. Bringing your portrait to 3d presence. *arXiv preprint arXiv:2511.22553*, 2025. 3

[88] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5535–5545, 2025. 1, 3, 4

[89] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 5

[90] 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, 7

[91] 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, 2025. 1

[92] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan

Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 3

[93] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13545–13555, 2022. 3

[94] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21057–21067, 2023. 3

[95] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, pages 268–285. Springer, 2022. 3

[96] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. *arXiv preprint arXiv:2412.14963*, 2024. 2

[97] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 3

UIKA: Fast Universal Head Avatar from Pose-Free Images

Supplementary Material

In this supplementary material, we first provide additional implementation details for our model, along with further visualization results (Sec. A). We then present our synthetic dataset (Sec. B). Next, we report additional comparative experiments, covering both self and cross reenactment on monocular and multi-view settings (Sec. C). We also include extended ablation studies, examining the impact of training data size as well as ablations of our method itself (Sec. D). We then provide more in-the-wild cases and applications, e.g., text-to-head-avatar generation (Sec. E). Finally, we discuss the limitations of our method (Sec. F) and its associated ethical implications (Sec. G). Additional dynamic results are provided in our supplementary video.

A. Additional Implementation Details

A.1. Facial Correspondence Estimator

Our facial correspondence estimator architecture is illustrated in Fig. S3. We adopt the pretrained backbone of VGGT [69] and keep it frozen, as this backbone encodes rich multiview image priors. On top of it, we initialize a trainable DPT [59, 60] head to predict two-channel UV coordinates within the range $[0, 1]$. The predicted UV coordinates map is further multiplied by the input image mask to extract the valid foreground region of the human head. Since VGGT is built upon DINOv2 [50] with a patch size of 14×14, the input resolution must be divisible by 14. Consequently, both the input image and the predicted UV coordinates map are resized to 518×518, and the prediction is finally resized to 512×512.

A.2. UV coordinates map

We visualize the predicted UV coordinates map and compare them with Pixel3DMM [20]. As shown in Fig. S1, our approach produces significantly smoother results in boundary regions, particularly around hair. This smoothness is crucial for our subsequent operation of reprojecting screen space color back into the UV space, enabling more coherent and reliable reprojection outcomes.

A.3. Hyperparameters

In Tab. S1, we provide additional detailed hyperparameters used in our model configuration.

B. Synthetic Dataset

In Sec. 3.4 of the main paper, we explain the curation process of our synthetic multi-view head dataset. In this section, we provide visualization results of this dataset, as shown in Fig. S4, which illustrates the results of each identity under different camera viewpoints and expressions. Our synthetic data achieves a well-balanced combination of identity diversity and expression richness, while maintaining multi-view and 3D consistency. Such a dataset contributes to training a more robust model. Please refer to our supplementary video for additional dynamic results.

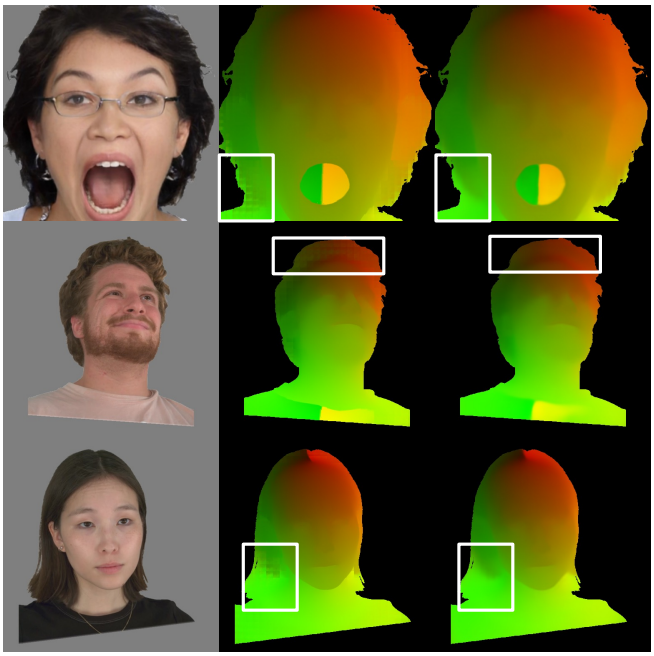

Figure S1. **Visualization and Comparison of UV coordinates map.**

Figure S2. **Visualization of ablation study results of training data.**

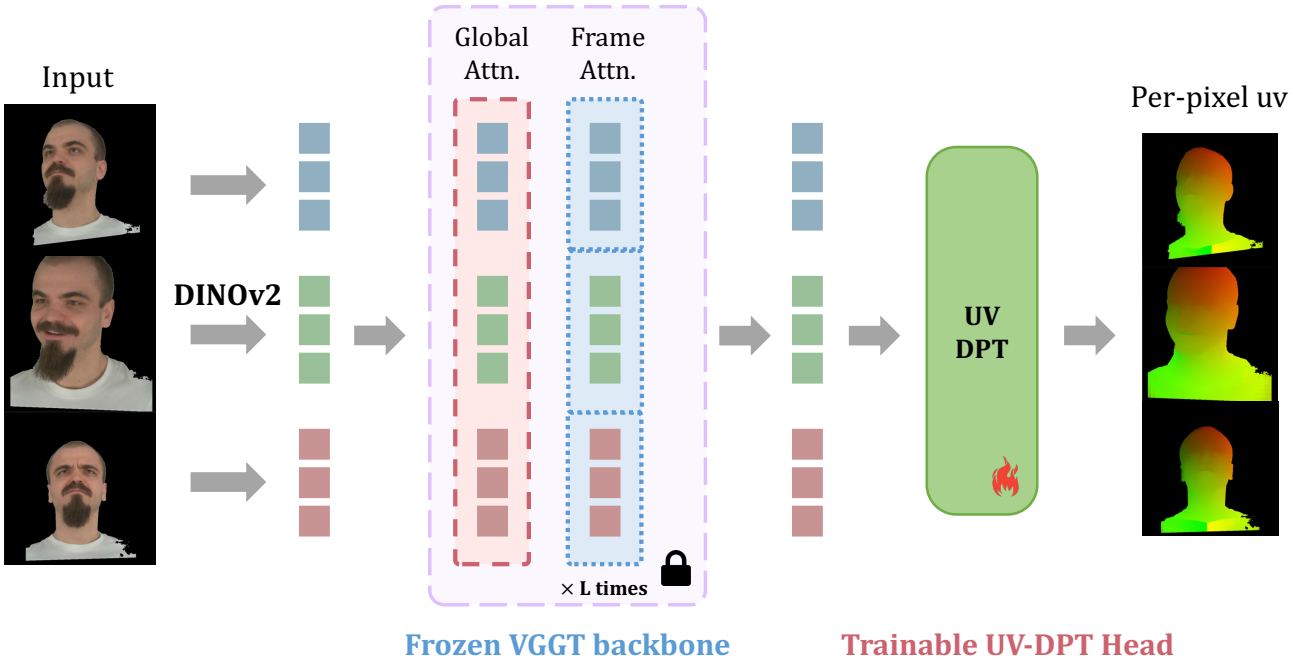

Figure S3. **Architecture of our facial correspondence estimator network.**

	Hyperparameter	Value
Input & Output	Input image resolution	512 × 512
	Train render resolution	512 × 512
Feature Extractor	DINOv3 version	vitl16
	DINOv3 patch size	16 × 16
	DINOv3 feature size	$\mathcal{N}$ × 1024 × 1024
	DINOv3 intermediate layer	4, 11, 17, 23
MultiModal Transformer	Hidden dimension	1024
	Head numbers	16
	Self attention layers	12
	Learnable UV token size	96 × 96 × 1024
UV Gaussian Decoder	Gaussian attribute map size	384 × 384
	Aggregated UV map size	384 × 384
	UV DPT inner dimension	256
	MLP inner dimension	512
	MLP layers	3
	MLP activation	SiLU
Gaussian Settings	Offset max range	0.2
	Scaling clip range	0.01
	Init scaling	exp(-5.0)
	Init density	0.1

Table S1. **Hyperparameters used in our method.** $\mathcal{N}$ represents the number of input views.

C. Additional Comparison Results

Monocular Setting. In Fig. S5, we show more self and cross reenactment results on the VFHQ dataset and NeRSemble-v2 dataset.

Multi-view Setting. In Fig. S6, we show more results of self and cross reenactment on the NeRSemble-v2 dataset. Please refer to our supplementary video for additional dynamic results.

D. Additional Ablation Results

D.1. Ablation on training data

Thanks to the paradigm of our framework, the model can accept an arbitrary number of input images. Although the number of input views during training is limited to $1 \sim 16$ due to VRAM constraints, similar to VGGT [69], our model can take more than 16 input images during inference. This flexibility enables us to train on monocular video datasets, unlike methods such as Avat3r [34] that require a fixed set of four input views and therefore rely exclusively on multi-view datasets. The monocular video dataset VFHQ [73] contains approximately 7k identities, which is an order of magnitude larger than existing multi-view datasets such as NeRSemble-v2 [33], Ava-256 [47], and RenderMe-360 [52], each of which typically includes only a few hundred identities.

To evaluate the importance of high-quality training data, we prepare two ablated versions. One model is only trained on the NeRSemble-v2 dataset, as shown in Fig. S2 (b), which can hardly preserve the identity of input images. When using both NeRSemble-v2 and a rich-identity dataset VFHQ, the model generalizes better to novel identities, but would collapse in some extreme viewpoint in Fig. S2 (c). When including our multi-view synthetic data, our model demonstrates superior generalization capability of identity and preserves 3D consistency, as shown in Fig. S2 (d).

D.2. Ablation on our method

In this section, we provide additional visualizations of ablation studies on our method. Other than the ablated versions in the main paper, we further include an extra ablation on our self-adaptive fusion strategy, as shown in Fig. S7 (d). In our full model, the fusion weight for each Gaussian is predicted by the network as a per-Gaussian value in the range $[0, 1]$. In contrast, this ablated variant replaces the learned weight with a fixed value computed as 0.5 times the UV-domain confidence map described in Sec. 3.1. The results demonstrate that our proposed full model effectively leverages information from the input views, leading to higher-fidelity head avatar reconstruction. Please refer to our supplementary video for additional dynamic results.

E. Applications

In-the-wild Image Reenactment. We also demonstrate the reenactment results of our method on in-the-wild Internet cases, as shown in Fig. S8.

Text-to-Head-Avatar Generation. In addition, we visualize the pipeline for generating controllable head avatars from text prompts. Given a textual description, we employ advanced multimodal large models such as ChatGPT or Gemini to synthesize corresponding images, which are then fed into our model to produce a animatable head avatar. Detailed visualizations are provided in Fig. S9.

Such results show that our method generalizes well to a wide variety of visual styles, benefiting from both our proposed approach and the synthetic dataset. Please refer to our supplementary video for additional dynamic results.

F. Limitations

Despite its effectiveness, our approach has several limitations. First, the expressiveness of our reconstructed head avatars is inherently constrained by the FLAME [37] model used for both data tracking and avatar driving. As a result, fine-grained facial dynamics such as subtle wrinkles, micro-expressions, and tongue motions cannot be reliably captured or reproduced. Second, although our training includes both real and synthetic data, the combined dataset still exhibits certain demographic biases, which may lead to degraded performance or failure cases for underrepresented groups. Third, while our framework supports an arbitrary number of input images, the computational cost and memory consumption grow with the number of views, whereas the performance improvement saturates beyond a certain point. These limitations highlight important directions for future work, such as integrating more expressive parametric models, reducing data bias, and improving scalability for large-view inference.

G. Ethics

Our work focuses on feed-forward reconstruction of animatable head avatars from arbitrary numbers of input facial images. While the proposed method advances the efficiency and accessibility of personalized head avatar creation, it also raises several potential ethical concerns. First, the ability to reconstruct high-fidelity 3D human heads from sparse or casually captured images introduces risks of misuse, such as generating unauthorized digital replicas of individuals or producing manipulated content that may compromise privacy, consent, or identity integrity. Second, reconstructed avatars could be misappropriated for malicious applications, including impersonation, deepfake-style synthesis, or other forms of deceptive media generation.

To mitigate these risks, our research uses only publicly available datasets with established licenses and synthetic data generated in-house. We emphasize that our method is intended for legitimate applications such as virtual telepresence, animation, and human computer interaction. We strongly discourage any use of this technology for surveillance, non-consensual persona reproduction, or deceptive content creation. Future deployment of systems built upon our approach should incorporate suitable safeguards, such as perceptual watermarking, provenance tracking, or identity verification mechanisms, to ensure responsible and ethical use.

Figure S4. **Visualization of examples from our synthetic dataset.**

Figure S5. **Visualization of self and cross reenacted results on the VFHQ and NeRSemble-v2 datasets for the monocular input setting.**

Figure S6. **Visualization of self and cross reenacted results on the NeRSemble-v2 dataset for the multi-view setting.**

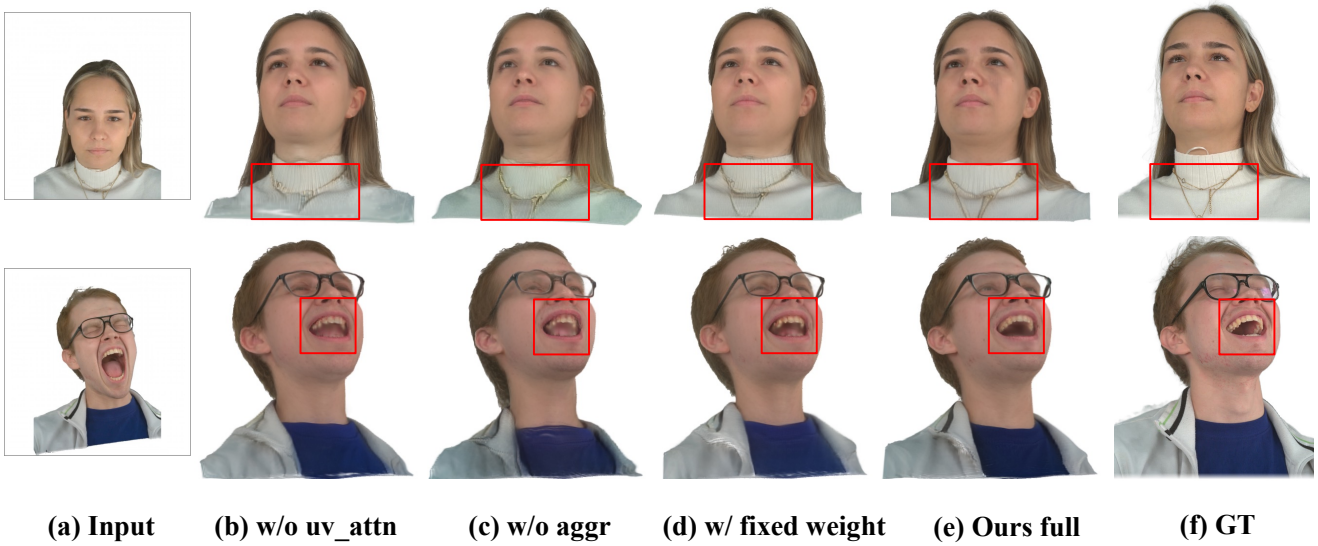

Figure S7. **Visualization of ablation study results of our method.**

Figure S8. **Visualization of in-the-wild cases.**

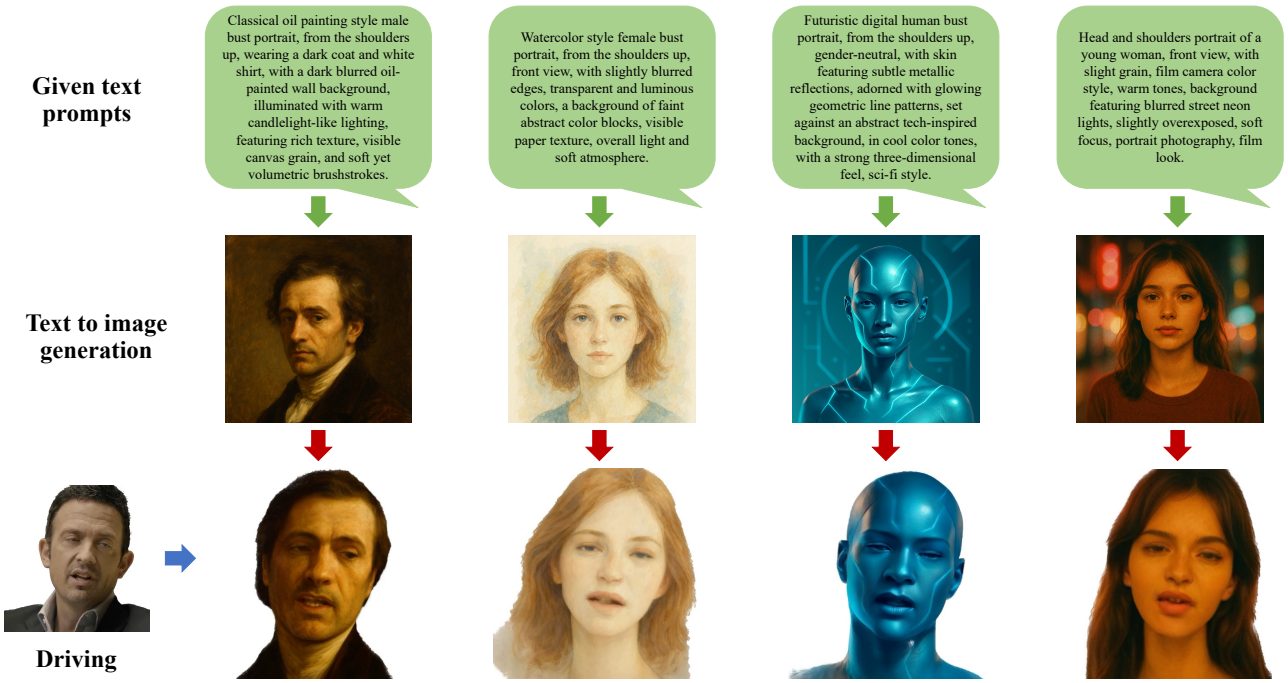

Figure S9. **Visualization of text-to-head-avatar generation.**