DOI: 10.1111/cgf.15282
COMPUTER GRAPHICS *forum*
Volume 44 (2025), number 1, e15282

THGS: Lifelike Talking Human Avatar Synthesis From Monocular Video Via 3D Gaussian Splatting

Chuang Chen,[1,2] Lingyun Yu,[2,3] Quanwei Yang,[3] Aihua Zheng[1] and Hongtao Xie[3]

[1]School of Artificial Intelligence, Anhui University, Hefei, China
chenchuang010@gmail.com, ahzheng214@foxmail.com
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
yuly@ustc.edu.cn
[3]School of Information Science and Technology, University of Science and Technology of China, Hefei, China
yangquanwei@mail.ustc.edu.cn, htxie@ustc.edu.cn

Abstract

Despite the remarkable progress in 3D talking head generation, directly generating 3D talking human avatars still suffers from rigid facial expressions, distorted hand textures and out-of-sync lip movements. In this paper, we extend speaker-specific talking head generation task to ***talking human avatar synthesis*** *and propose a novel pipeline,* THGS*, that animates lifelike Talking Human avatars using 3D Gaussian Splatting (3DGS). Given speech audio, expression and body poses as input,* THGS *effectively overcomes the limitations of 3DGS human re-construction methods in capturing expressive dynamics, such as* ***mouth movements, facial expressions and hand gestures****, from a short monocular video. Firstly, we introduce a simple yet effective* ***Learnable Expression Blendshapes (LEB)*** *for facial dynamics re-construction, where subtle facial dynamics can be generated by linearly combining the static head model and expression blendshapes. Secondly, a* ***Spatial Audio Attention Module (SAAM)*** *is proposed for lip-synced mouth movement animation, building connections between speech audio and mouth Gaussian movements. Thirdly, we employ a* ***body pose, expression and skinning weights joint optimization strategy*** *to optimize these parameters on the fly, which aligns hand movements and expressions better with video input. Experimental results demonstrate that* THGS *can achieve high-fidelity 3D talking human avatar animation at 150+ fps on a web-based rendering system, improving the requirements of real-time applications. Our project page is at https://sora158.github.io/THGS.github.io/.*

Keywords: image and video processing, real-time rendering, monocular re-construction, 3D Gaussian Splatting

CCS Concepts: • Computing methodologies → Re-construction; Appearance and texture representations

1. Introduction

Existing 3D talking head generation approaches [GCL*21, TWZ*22, LZB*23, SLZ*22] have made significant progress in driving realistically lip-synced talking faces. However, they are confined to generating only the face area and typically model the head and shoulders separately. In domains such as the metaverse, live broadcasting and gaming, a more expressive full-body lifelike human avatar is needed. Thus, a challenging task called *Talking Human Avatar Synthesis* is introduced: Given speech audio and body poses, this task aims to generate a 3D human avatar incorporating realistic hand gestures, expressions and mouth movements exquisitely synchronized with the audio. However, existing techniques [LWZ*22, NRB*24], which either require large amounts of data or are not trained in an end-to-end manner, are still far from achieving the authenticity and liveliness of real-time talking human avatar generation. The challenge primarily encompasses two aspects: Firstly, learning the realistic appearance of a human body from a short monocular video including various expressions and poses, without 3D observations like 3D scans or RGBD images; secondly, expressively animating small-scale regions, such as the face and hand regions, involving lip-synced mouth movements, natural facial expressions and authentic hand gestures.

Previous data-driven methods, such as ACGVG and MakeYourAnchor [LWZ*22, HTZ*24], employed VQ-VAE or diffusion model to improve the realism of talking human avatars. However, huge labelled video data with aligned SMPL-X [PCG*19]

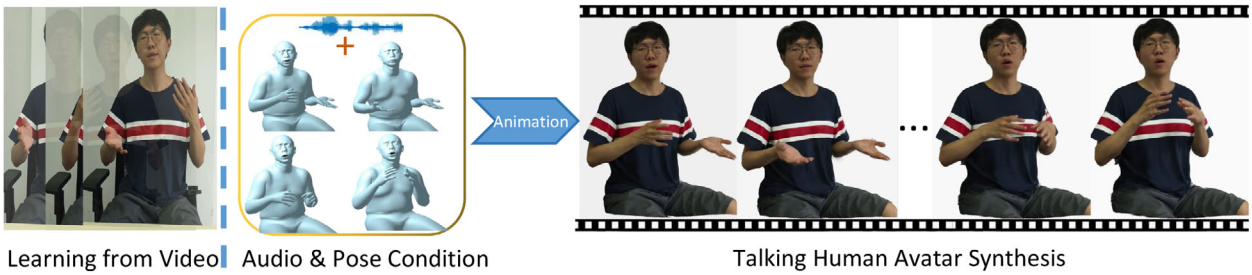

Figure 1: *Talking human avatar synthesis: **THGS** can re-construct expressive human avatar from a **1-min monocular video**. During inference, we can animate lip-synced 3D talking humans with authentic hand gestures, given audio, expression and body poses as input.*

annotations are needed for pre-training a foundation model, and such pre-training process can cost 7 days on one Nvidia A100 GPU in MakeYourAnchor, greatly increasing the computational burden. Besides, these methods only support generating 2D videos, lacking 3D consistency, which can lead to appearance distortion when different body orientations are presented during the inference stage. Moreover, like most GAN/diffusion-based methods, these methods are prone to overfitting the training poses. They struggle with generalization when generating a video sequence with pose variations due to the out-of-domain challenge.

Recently, powerful 3D Gaussian Splatting (3DGS) [KKLD23] has paved the way to re-construct animatable 3D human avatars from monocular videos efficiently. 3D Gaussians explicitly represent the static 3D scenes through discrete point cloud-like representation. Then we render the final image via differentiable point-based alpha-blending volumetric rendering techniques. These methods [HZZ*24, HHL24, LZWL24, LWP*24] typically model human shape and appearance in a pose-independent canonical space (e.g. T-pose) via 3D Gaussians and use Linear Blend Skinning (LBS) to animate these canonical 3D Gaussians, similar to how template meshes are controlled. Compared to data-driven methods, 3DGS avatar re-construction techniques can model humans from a short monocular video, effectively eliminating the need for large datasets and bypassing the model pre-training stage. These methods also ensure 3D consistency while capturing geometric and texture details. More importantly, unlike 3D talking head methods that separately model the head and torso [GCL*21, TWZ*22], 3DGS human avatar **inherently provides a full-body representation and has an explicit body driving process thanks to SMPL model**. However, existing 3DGS-based methods [LWP*24, HHL24, LWL*24] lack the capability for expressive modelling because they animate only 24 SMPL bones, which are insufficient for capturing complex deformations, and they do not incorporate 3D morphable face models [BV99, BV23, LBB*17] to capture facial dynamics. These result in significant gaps in talking human generation, **specifically in the expressive animation of hand gestures, facial expressions and lip-synced mouth movements**.

To address the aforementioned challenges, we introduce a novel end-to-end approach, THGS, which is designed to animate high-fidelity 3D talking human avatars, given speech audio and body poses as input. Different from the existing 3DGS-based human re-construction pipeline [LWP*24], THGS concentrates on capturing and animating expressive dynamics, as shown in Figure 1. Specifically, to enhance expressive animation capabilities, we introduce a simple yet effective Learnable Expression Blendshapes (LEB) for facial expression animation. LEB enables the re-construction of detailed facial dynamics, such as wrinkles and subtle expressions, through the linear combination of a static head model and expression blendshapes. More crucially, a Spatial Audio Attention Module (SAAM) is designed to improve lip synchronicity, given audio input. SAAM innovatively encodes the positions of 3D Gaussians into a spatial continuous tri-plane representation, where it integrates audio features through an attention mechanism to predict mouth Gaussians deformations. Additionally, we use a body pose, expression and skinning weights joint optimization strategy to optimize hand pose and expression during the training stage. Extensive experiments demonstrate that our method can generate high-fidelity talking human avatars, achieving a 27% improvement in LPIPS. Our contributions are summarized as follows:

- An innovative framework, THGS, is introduced to extend the field from 3D talking heads to 3D talking human avatar synthesis, enabling the learning of expressive avatars from monocular videos and achieving photorealistic, real-time rendering.
- To capture facial dynamics and achieve explicit control over expression, a LEB is proposed to effectively predict expression offsets by linearly combining the static head model with expression blendshapes.
- A novel SAAM is designed to integrate audio features through an attention mechanism to predict mouth 3D mouth Gaussians deformations, thus enhancing lip synchronicity.
- We present a joint optimization strategy for optimizing body poses, expressions and skinning weights, which aligns hand movements and expressions more accurately with video input.

2. Related Works

2.1. 2D avatar video generation

Recently, research has emerged in the field of 2D human avatar generation with gestures, under various conditional inputs like audio, text or motions. Liu *et al.* [LWZ*22] use VQ-VAE to encode the pose parameters into a codebook and generate co-speech gesture videos given the audio input only. DreamPose [KHWKS23] adapts Stable Diffusion, which is guided by input reference images and poses, allowing it to animate a static human image based on motion sequence. Similarly, DisCo [WLL*24], utilizing reference

images and poses as input, employs pose and background ControlNet [ZRA23] to blend various motions and scenes through a pre-training approach, yielding promising outcomes in human dance generation with enhanced generalization. To generate exact hand gestures, Make-Your-Anchor [HTZ*24] uses rendered 3D human mesh images as conditions and binds movements with specific human appearances via a pre-trained diffusion model. However, these methods require at least 27 h of aligned training video data, and the fine-tuning stage described in Huang *et al.* [HTZ*24] can take about a day on an Nvidia A100, with rendering speed at 0.45 fps. This indicates their extensive computational demands, high data requirements and slow inference speeds. **THGS** addresses the above issues using 3DGS, enabling the learning of talking human avatars from a 1-min training data, within an hour of training time and achieving real-time photorealistic avatar rendering at 150+ fps. Besides speed issues, these methods also suffer from insufficient controllability of hands and face [KHWKS23], as well as appearance distortion with different body orientations [QTZ*21, ZZ22]. These limitations underscore the need for 3D solutions.

2.2. Audio-driven talking head generation via NeRF/3DGS

Recently, talking head synthesis based on neural radiance field (NeRF) [GCL*21, LXW*22, YZY*22] has risen as an essential research area in computer vision. These methods are capable of controlling head pose and generating lip-synced mouth movements. RAD-NeRF [TWZ*22] first introduced multi-resolution hash grids to encode spatial and audio information separately, greatly accelerating training and inference compared to AD-NeRF [GCL*21]. ER-NeRF [LZB*23] explores attention-based audio-spatial correlations, enhancing the modelling of speech-lip movements.

While NeRF-based talking head methods have their advantages, their generated avatars suffer from visual jitter and rendering speed remains unsatisfactory at 20–30 fps. This has led researchers to explore 3DGS methods. GSTalker [CHC*24] and GaussianTalker [CLY*24] replace NeRF's implicit representation with explicit 3DGS for head modelling, resulting in improvements in training time, inference rates and synthesis quality. Similarly, TalkingGaussian [LZB*24] uses 3DGS for canonical head modelling and a dual-branch Grid-based model to separately model face-mouth deformations in talking head synthesis. Another system, GaussianTalker [YQY*24], completely decouples audio from 3DGS attributes prediction. It binds 3DGS to the FLAME [LBB*17] head mesh, first training a model to predict FLAME mesh movements from speech and then animating 3DGS via LBS. The advantage of this method lies in its precise lip control and high visual quality. Despite the impressive progress in talking faces, avatars limited to the facial region have restrictions in real-world applications. Therefore, these methods [GCL*21, TWZ*22] either model the movements of the head and shoulder separately in a two-stage training or utilize GAN-based techniques to roughly fuse head images with still shoulder [PHS*24]. **To extend the 3D talking head generation task to talking human avatars, we model the full-body human within a complete framework and train it end-to-end to reduce post-processing steps**. That is, we utilize human avatar re-construction pipeline as the backbone, which provides a full-body representation and has an explicit body-driving process.

2.3. Human avatar re-construction via 3D Gaussian Splatting

With real-time and photorealistic rendering capabilities demonstrated in 3DGS [KKLD23], researchers are now using it as a 3D representation to tackle human avatar re-construction challenges. 3DGS-based avatar re-construction [LWP*24, HHL24, HZZ*24, LZWL24, ZZS*24] has quickly transformed into a vibrant research area in a short time. These methods use a human-prior mesh to initialize 3D Gaussians in canonical space and warp them to a posed space through LBS given arbitrary body poses. The real images can supervise canonical 3D Gaussians, once the estimated human body mesh is aligned with the truth body pose sequences. Typically, GART [LWP*24] utilizes learnable forward skinning and model non-rigid transformations with latent bones. Similarly, Animatable Gaussians [LZWL24] learns an SDF-based human template from input videos and then parameterizes the template on two Gaussian maps which can be decoded into 3D Gaussians' properties. However, these methods can only model pure-body poses, while hands and face regions suffer from obvious misalignment and appearance distortion when training data have various hand gestures and expressions. Thus, **THGS** proposes a LEB and a Spatial Audio Attention Mechanism to offer finer-grained controls, like facial dynamics and lip-synced mouth movements. Moreover, our method proposes a joint pose and expression optimization pipeline, which can get better SMPL-X body pose alignment results in hands and face area and generates expressive talking human avatars with authentic hand gestures.

3. Preliminary

3.1. 3D Gaussian Splatting (3DGS)

3DGS [KKLD23] uses 3D Gaussian ellipsoids as its primary rendering units. Unlike NeRF, which employs implicit representations, 3DGS represents scenes with explicit Gaussian ellipsoids. A 3D Gaussian $G(\boldsymbol{x})$ is mathematically defined as follows:

$$G(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the ellipsoid's spatial mean position and covariance matrix, respectively. The covariance matrix $\boldsymbol{\Sigma} = r\mathbf{s}\mathbf{s}^\top r^\top$ characterizes how the 3D Gaussian is stretched and rotated. Here, the matrix $\mathbf{s}$ represents the scale of stretching in space, while the rotation matrix r represents the degree of rotation of the 3D Gaussian in $\mathbb{R}^3$ space. Each Gaussian also has an opacity η and a view-dependent colour c, which is represented by spherical harmonic coefficients f. During the rendering process from a specific viewpoint, 3D Gaussians are projected onto the view plane by splatting. To obtain the pixel colour, alpha-blending is performed on N sequentially layered 2D Gaussians, starting from the front and moving toward the back, as follows:

$$C = \sum_{i \in N} T_i \alpha_i \boldsymbol{c}_i, \quad \text{with} \quad T_i = \prod_{j=1}^{i} \left(1 - \alpha_j\right),$$

where α_i is the blending weight computed by opacity η and 2D Gaussian distribution [ZPBG01]. Finally, optimizable 3D Gaussian property is $\mathbf{P} = \{\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\eta}, \boldsymbol{f}\}$, where each property controls position, rotation, scale, opacity and colour of 3D Gaussians separately.

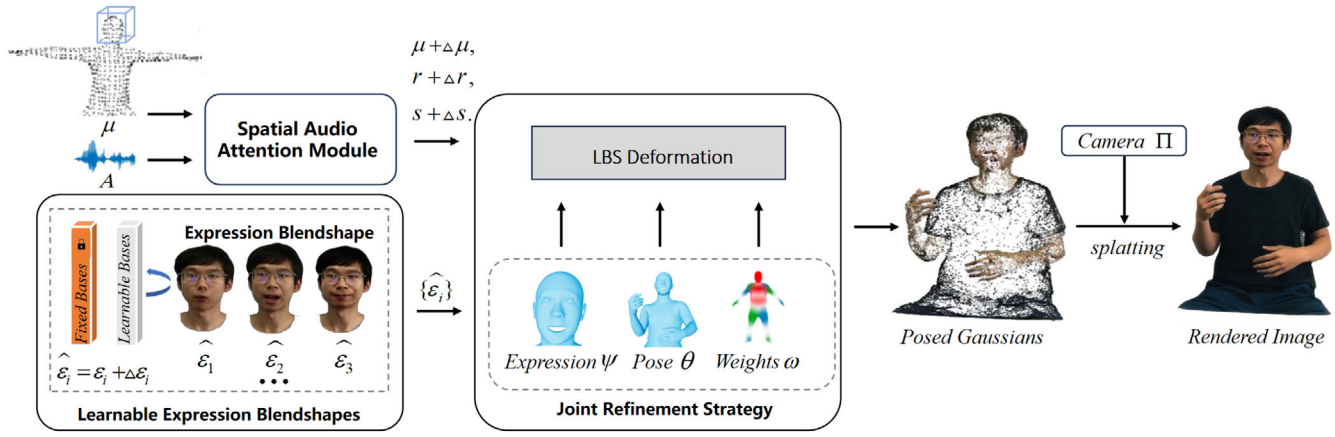

Figure 2: *Method overview. THGS learns a 3DGS representation of talking human avatars from a monocular video. 3D Gaussians* $P = \{\boldsymbol{\mu}, \mathbf{r}, \mathbf{s}, \boldsymbol{\eta}, \mathbf{f}\}$ *are initialized on the T-pose vertices in the canonical space. We introduce a Learnable Expression Blendshapes (LEB) for facial dynamics re-construction and explicitly control expression by linearly interpolating expression bases. Audio features* $\mathcal{A}$ *are processed through the Spatial Audio Attention Module (SAAM), predicting audio-driven mouth Gaussians deformation* $P = \{\boldsymbol{\mu} + \Delta\boldsymbol{\mu}, \mathbf{r} + \Delta\mathbf{r}, \mathbf{s} + \Delta\mathbf{s}, \boldsymbol{\eta}, \mathbf{f}\}$. *Then these Gaussians are deformed into posed Gaussians using LBS, given pose* θ *and expression* ψ *as inputs, incorporating expression offsets provided by LEB and mouth deformation from SAAM. We employ a joint optimization strategy to optimize* θ, ψ *and skinning weights* ω, *resulting in improved body alignment. Finally, a 3DGS rasterizer renders images based on camera poses* Π.

3.2. Parameterized SMPL-X model

SMPL-X [PCG*19] expands the joint and vertex set of the SMPL human mesh model [LMR*15] to include the face and fingers, providing enhanced control over facial expressions and hand movements. SMPL-X uses LBS method to drive the canonical human mesh template. This process can be formulated as: $M(\theta, \beta, \psi) : \mathbb{R}^{55\times3} \times \mathbb{R}^{300} \times \mathbb{R}^{100} \rightarrow \mathbb{R}^{3\times10595}$, where $\theta \in \mathbb{R}^{55\times3}$ denotes body poses including hand poses, $\beta \in \mathbb{R}^{300}$ denotes body shape, $\psi \in \mathbb{R}^{100}$ denotes expression coefficients, and whole body includes 10,595 vertices. LBS can be more formally formulated as:

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \omega), \tag{1}$$

$$T_P(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}), \tag{2}$$

where T_p represents human vertices in the canonical pose, adjusted for shape and facial expression. $\bar{T}$ denotes the average T-pose vertex positions of the SMPL-X template in canonical space. J represents the pre-trained regression matrix used for the joint location function, W denotes the standard LBS operation and ω refers to the pre-determined skinning weights. $B_S(\beta; \mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n \mathcal{S}_n$ is the body shape blendshape function, where β represents the shape coefficients. $B_E(\psi; \mathcal{E}) = \sum_{n=1}^{100} \psi_n \mathcal{E}_n$ is the expression blendshape function, where ψ represents the expression coefficients used for adjusting expressions. Lastly, B_P is used for pose adjustment.

4. Method

Given a monocular video of a talking person, our goal is to learn an expressive talking human avatar that can be driven by pose parameters and speech audio. In this section, we first describe problem settings (Section 4.1) and how we initialize and animate 3DGS properties based on Section 4.2. Then, a LEB (Section 4.3) is utilized to control facial expressions. Next, to achieve better lip-synced results, we design a SAAM to animate the mouth Gaussians movements in Section 4.4. Finally, we optimize pose, expression coefficients and skinning weights jointly during the training stage in Section 4.5 and give our training objectives (Section 4.6). The pipeline of THGS is illustrated in Figure 2.

4.1. Problem settings

Training data are usually a 1-min dynamic front-view talking person video with synchronized audio. For each frame, we get corresponding: (1) semantic parsing [KMR*23] of human for foreground extraction; (2) estimated SMPL-X [PCG*19] body poses θ and expression coefficients ψ; (3) audio features $\mathcal{A}$ processed by a pre-trained Automatic Speech Recognition model Wav2vec [BZMA20].

Therefore, our problem settings can be formulated as

$$G = F(\theta, \psi, \mathcal{A}; \Pi), \tag{3}$$

where G is the rendered image and Π is camera parameters. θ is body poses for control body movements including hand gestures. Expressions coefficient ψ is utilized to control facial movements, and $\mathcal{A}$ are viewed as audio feature inputs that control mouth Gaussian movements.

4.2. Dynamic human 3DGS representation

We combine 3DGS (Section 3.1) and SMPL-X human template (Section 3.2) together to create expressive human 3DGS representation:

$$G(\boldsymbol{\beta}, \theta, \psi, \mathbf{P}) = \text{Splatting}\,(W(\boldsymbol{\mu}^e, J(\boldsymbol{\beta}), \boldsymbol{\theta}, \omega), \mathbf{P} \setminus \boldsymbol{\mu}^e), \tag{4}$$

$$\boldsymbol{\mu}^e = \boldsymbol{\mu} + B_S(\beta; \mathcal{S}) + B_E(\psi; \widehat{\mathcal{E}}) + B_P(\theta; \mathcal{P}) + SAAM(\mu, \mathcal{A}), \tag{5}$$

where $G(\cdot)$ represents a rendered image, and Splatting $(\cdot)$ denotes the rendering process of 3D Gaussians from any viewpoint. $W(\cdot)$ is LBS function employed for reposing 3D Gaussians. $\boldsymbol{\mu}$ denotes the positions in canonical space under the T-pose, while the expressive positions $\boldsymbol{\mu}^e$ incorporate expression offsets and lip movements from SAAM (Section 4.4) as defined in Equation (5). $\mathbf{P} \setminus \boldsymbol{\mu}^e$ denotes the remaining properties of 3DGS except for positions $\boldsymbol{\mu}^e$.

Relationship with prior works: The LBS function in Equation (4) is consistent with previous GS-based avatar methods [LWP*24, LZWL24, HZZ*24], which is mainly for rigid body transformation. The main difference lies in our treatment of Equation (5): we utilize learnable expression bases described in Section 4.3 for re-constructing expression dynamics and the SAAM described in Section 4.4 to predict audio-driven lip movements, given audio features $\mathcal{A}$ and the positions of Gaussians $\boldsymbol{\mu}$ as inputs.

The key component of our method is how to build connections between the SMPL-X vertices and 3D Gaussian splats properties $\mathbf{P}$. Firstly, position $\boldsymbol{\mu}$ is initialized on T-pose vertices positions $\bar{T}$ of SMPL-X template in canonical space. To obtain expressive $\boldsymbol{\mu}^e$, we add offsets according to Equation (5). Secondly, given each vertex normal **uz** defined in SMPL-X, we select a random direction **uy** and compute $\mathbf{ux} = \mathbf{uy} \times \mathbf{uz}$ using a cross product. These vectors [**ux**; **uy**; **uz**] are then stacked to form the initial rotation matrix of 3D Gaussian splats $\boldsymbol{r}$. Thirdly, initial scale of each 3D Gaussian splats $\boldsymbol{s}$ is estimated by computing the radius for each vertex based on the area of its neighbouring faces. Given the average area $\mathrm{A}_{\mathrm{avg}}$ of the neighbouring faces, the radius is computed by $\text{radius} = \sqrt{\frac{\mathrm{A}_{\mathrm{avg}}}{\pi}}$, ensuring that the values remain within a defined range. This process ensures a more accurate initialization of the Gaussian splats.

Note that skinning weight ω of each 3D Gaussians is obtained from interpolating the voxel grid, which stores the skinning weight of the nearest SMPL-X vertices. Once we bind 3D Gaussians to the human template, we can directly repose these canonical 3D Gaussians to the motion space for free-view rendering by LBS according to Equation (4). Specifically, we rotate and translate the 3D position and rotation matrix of each 3D Gaussian with the LBS transformation matrix.

The joint hierarchy of SMPL-X is used to achieve LBS transformations and the LBS transformation matrix $\mathcal{T}(\theta)$ is calculated for each joint, given a pose θ. For each Gaussian point, we calculate the LBS transformation $\mathcal{M}$ based on the $P = 55$ joints:

$$\mathcal{M}(\theta) = \sum_{p=1}^{P} \omega_p(\boldsymbol{\mu}^e)\mathcal{T}(\theta), \tag{6}$$

where $\omega_p(\boldsymbol{\mu})$ is the bilinear interpolation skinning weights of the Gaussian point $\boldsymbol{\mu}^e$. The deformation of the Gaussian point from the canonical pose to the target pose θ can be formulated as:

$$\boldsymbol{\mu}_\theta = \mathcal{M}_{\mathrm{rot}}(\theta)\boldsymbol{\mu}^e + \mathcal{M}_t, \quad \boldsymbol{r}_\theta = \mathcal{M}_{\mathrm{rot}}(\theta)\boldsymbol{r}, \tag{7}$$

where $\mathcal{M}_{\mathrm{rot}}$ represents the rotation component, and $\mathcal{M}_t$ represents the translation component of the Gaussian point transformation. $\boldsymbol{r}$ is the rotation of the Gaussian point.

4.3. Learnable Expression Blendshapes

Blendshape model [LBB*17] is a classic representation for avatar animation. It consists of a set of 3D meshes, each corresponding to a basis of expression. In SMPL-X model, expression offset is calculated by expression blendshapes $B_E(\psi; \mathcal{E})$, where expression offsets are viewed as a linear combination of expression bases $\mathcal{E}$. The shape of $\mathcal{E}$ is $\mathbb{R}^{100\times 10595\times 3}$, representing vertex position matrices for $|\psi|$ different expressions. The magnitude of these expressions is adjusted by the expression coefficients ψ. However, **person-agnostic human models, such as SMPL-X, have a pre-defined expression basis that models facial expressions but cannot capture the dedicated facial dynamics**. Hence, our goal is to find a simple yet efficient way to model diverse expressions. To be specific, we make the original expression bases $\mathcal{E}$ fixed and optimize learnable expression bases during training.

$$\widehat{\mathcal{E}} = \mathcal{E} + \Delta\mathcal{E} \circ M_H, \tag{8}$$

$$\widehat{B_E}(\psi; \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \widehat{\mathcal{E}_n}, \tag{9}$$

where $\Delta\mathcal{E}$ is learnable expression bases, M_H denotes the mask corresponding to the FLAME head model within the SMPL-X model, and $\widehat{B_E}(\psi; \mathcal{E})$ denotes LEB. In **THGS**, learnable $\widehat{B_E}(\psi; \mathcal{E})$ replaces the $B_E(\psi; \mathcal{E})$ in Equation (2). We also add the constraint $\|\Delta\mathcal{E}\| < \epsilon$ to prevent blendshapes from deviating too far from the initial SMPL-X expression blendshapes.

4.4. Learning audio-driven deformation of mouth Gaussians

Previous NeRF-based talking head methods [GCL*21, TWZ*22, LXW*22] enhance a conditional NeRF representation by maintaining fixed 3D coordinates of sampling points along each ray while adjusting only the colour and density based on input audio. Unlike these methods, we propose to model the spatial movements of the mouth in physical space instead of adjusting its appearance. Because when real humans speak, there is no change in the colour or density of the mouth; instead, there are only the muscle movements of the mouth. In order to **model the spatial movements of the mouth and fully benefit from the explicit representation of 3DGS**, we deform the 3DGS based on audio signals $\mathcal{A}$. Namely, we do not manipulate the appearance attributes $(\boldsymbol{\eta}, \boldsymbol{f})$ of mouth Gaussians, but deform their spatial attributes $(\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{s})$. Besides, considering the absence of teeth details in FLAME [LBB*17] model, we first add teeth mesh inside the mouth, bind upper teeth with the back of the head, and let lower teeth move with the jaw joint, improving the perception of realism. The predicted mouth deformation will only modify the Gaussian properties in the mouth region. Therefore, we select mouth Gaussians based on their proximity to the teeth mesh. We calculate the bounding box of the teeth mesh and extend it outward by 0.02 units; any 3D Gaussians that fall within this bounding box are considered as mouth Gaussians.

SAAM: SAAM aims to predict mouth Gaussians movements given audio features $\mathcal{A}$ and position of Gaussians $\boldsymbol{\mu}$ as input, as shown in Figure 3. Among them, audio features $\mathcal{A}$ are predicted by a pre-trained ASR model [BZMA20]. When we refer to using audio to predict lip movements, it means to build the mappings between $\mathcal{A}$ and physical space $\mathbb{R}^3$. Because 3D Gaussians are

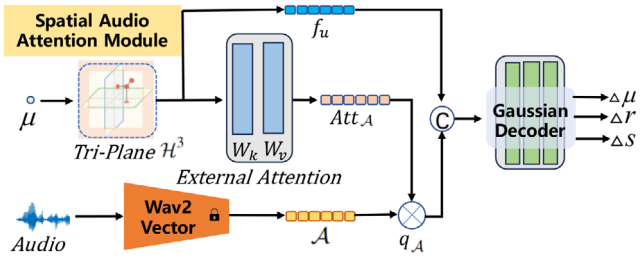

Figure 3: *Architecture of Spatial Audio Attention Module (SAAM). SAAM takes audio and position of 3D Gaussians as input, and predicts the deformation of mouth 3D Gaussians.*

discrete in physical space, we introduce an efficient and expressive continuous representation, the Tri-Plane Hash Representation [LZB*23], to model the head part in continuous space:

$$\mathcal{H}^3(\mathbf{x}) = \mathcal{H}^{XY}(x, y)\copyright\mathcal{H}^{YZ}(y, z)\copyright\mathcal{H}^{XZ}(x, z), \tag{10}$$

where © means the concatenation over 2D planes in channel. The Tri-Plane Hash Representation can effectively minimize hash collisions by projecting 3D features onto orthogonal 2D planes, thus streamlining spatial complexity, and is suitable for modelling continuous space. The positions of the Gaussians, $\boldsymbol{\mu}$, are input into the Tri-Plane $\mathcal{H}^3$ to obtain continuous spatial features f_u.

Audio has an uneven influence on the head area. Hence, the **External Attention** mechanism [GLMH21] is used to learn how audio affects different regions of head area:

$$Att_{\mathcal{A}} = (\mathrm{ReLU}(f_{\mu}W_k)W_v), \tag{11}$$

$$q_{\mathcal{A}} = Att_{\mathcal{A}} \circ \mathcal{A}, \tag{12}$$

where $Att_{\mathcal{A}}$ is regional-aware attention [LZB*23] of audio signals, f_{μ} denotes spatial features and $\circ$ means Hadamard product. W_k and W_v are two external memory units for individual levels connection and self-condition query for each region.

Then our audio-driven SAAM can be formulated as:

$$\mu' = \mu \circ M_H, \tag{13}$$

$$f_{\mu} = \mathcal{H}^3(\mu'), \tag{14}$$

$$\Delta\boldsymbol{\mu}, \Delta\mathbf{r}, \Delta\mathbf{s} = \mathrm{MLP}(f_{\mu}\copyright q_{\mathcal{A}}), \tag{15}$$

where M_H denotes the head mask to filter positions near the head area, $q_{\mathcal{A}}$ denotes the re-weighted audio features in Equation (12) and our Gaussian decoder (four-layer MLPs) transforms concatenated features into deformed Gaussian attributes ($\Delta\boldsymbol{\mu}$, $\Delta\mathbf{r}$, $\Delta\mathbf{s}$). Final audio-driven deformations of 3D Gaussians are $P = \{\boldsymbol{\mu} + \Delta\boldsymbol{\mu}, \mathbf{r} + \Delta\mathbf{r}, \mathbf{s} + \Delta\mathbf{s}, \boldsymbol{\eta}, \mathbf{f}\}$.

4.5. Joint optimization strategy

Creating a high-quality 3D Gaussian avatar relies heavily on the pose estimation accuracy derived from input videos. Accurate pose estimation speeds up the training process and significantly impacts the final rendering results. However, current full-body pose estimation methods like SHOW [YLL*23, ZTZ*21, ZTZ*23] struggle to accurately align hand and facial expressions when dealing with high-dynamic movements in monocular videos. While we use SHOW for initial body poses and expressions pre-processing, it often results in unstable body poses, leading to motion jitter. To address this, we implement our joint optimization strategy to achieve better body poses and expression alignment.

Camera and pose alignment before training: We adopt the data processing method called **SHOW**(https://github.com/yhw-yhw/SHOW) from the Talk-SHOW framework [YLL*23]. **SHOW** enables the re-construction of holistic, whole-body mesh results using only RGB images or videos. Specifically, **SHOW** employs SMPL-X 2D keypoint projections along with OpenPose pseudo ground truth to compute regression loss, which is used to iteratively optimize the SMPL-X body pose and expression parameters. Initially, **SHOW** provides a rough estimation of body poses θ, expressions ψ and camera translation T, assuming a pre-determined focal length f and camera rotation R. Our objective here is to minimize the projection error between the projected landmarks $\Pi_{f,R,T}(\theta, \psi)$ from SMPL-X and the actual landmarks L_{2D} under **the 3DGS rasterizer**. Formally, the optimal camera translation T is optimized by:

$$T_{\mathrm{opt}} = \arg\min_{R,T} E(L_{2D}, \Pi_{f,R,T}(\theta, \psi)). \tag{16}$$

E denotes the MSE metric, $\Pi_{f,R,T}$ denotes perspective projection of SMPL-X 3D joints given camera parameters (f, R, T), body poses θ and expression ψ. After obtaining the optimal camera parameters (f, R, T), we optimize body poses θ and expression ψ with given loss $\mathcal{L}_{\mathrm{motion}}$:

$$\mathcal{L}_{\mathrm{motion}} = \|(L_{2D} - \Pi_{f,R,T}(\theta, \psi))\| \cdot M(L_{2D}, tr) + \lambda_T \cdot \left(\frac{1}{N}\sum_{i=0}^{N-1}(\mathbf{V}_{i+1} - \mathbf{V}_i)^2\right), \tag{17}$$

where $M(L_{2D}, tr)$ represents a filtering operation that selects ground truth 2D keypoints with confidence greater than the threshold tr. $\mathbf{V}_i$ denotes the SMPL-X vertices at the ith time step. Part 2 of $\mathcal{L}_{\mathrm{motion}}$ corresponds to the temporal smoothness loss between adjacent time steps for SMPL-X vertices.

Skinning weight optimization during training: While optimizing the Gaussians' attributes, we finetune the SMPL-X body poses and expressions to ensure them more accurately align with the video input. In this stage, body poses and expressions can be optimized jointly with supervision from image loss, rather than calculating landmark loss as in the previous stage.

As for skinning weights ω, multi-layer perceptrons (MLPs) can be used to predict the LBS weight coefficients ω_k for each 3D Gaussian in the canonical space. However, it would be time-consuming and generate low-quality rendering results. To address this, we start with the LBS weights from SMPL-X and initialize a voxel grid $\mathcal{V}_{\omega}$ to approximate and store skinning weights ω [JHBZ22]. Since the pre-defined skinning weights are used for human template mesh and do not reflect the actual deformation such as clothes, we optimize

$\mathcal{V}_\omega$ on the fly:

$$\widehat{\mathcal{V}_\omega} = \frac{(\mathcal{V}_\omega^k + \Delta\mathcal{V}_\omega^k)}{\sum_{k=1}^{55}(\mathcal{V}_\omega^k + \Delta\mathcal{V}_\omega^k)}, \tag{18}$$

where k denotes the number of joints, and $\widehat{\mathcal{V}_\omega}$ are the optimized skinning weights in voxel grid. A 3D Gaussian queries its corresponding skinning weights from $\mathcal{V}_\omega^k$ using bilinear interpolation.

4.6. Rendering loss with differentiable rasterizer

SMPL-X LBS transformation is utilized to drive our avatar from canonical space to posed space. Our rendered avatar in canonical space can be optimized through a differentiable 3DGS rasterizer. Given the rendered and actual images, our losses include the reconstruction loss L_{recon}, the vertex loss L_{reg}, and the normal loss. Although the LPIPS loss can reduce the final L_{LPIPS}, the use of the VGG network for perceptual loss significantly slows down the optimization process. Therefore, the LPIPS loss is not used during training. We advance to a refinement stage after 7000 steps of optimization, incorporating the normal loss L_{recon} for view-consistency regularization, similar to 2DGS [HYC*24]. We designed a multi-stage training strategy, which first optimizes the static 3D Gaussians, and then sequentially initiates the optimization of body poses, skinning weights, expression, expression blendshapes and the SAAM module.

$$L_{\text{vertex}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\mu}_\text{i}) - \text{parent}(\boldsymbol{\mu}_\text{i})))^2, \tag{19}$$

$$L_{\text{recon}} = (1-\lambda_1)L_1 + \lambda_1 L_{D-SSIM}, \tag{20}$$

$$L_{\text{total}} = L_{\text{recon}} + L_{\text{normal}} + \lambda_2 L_{\text{vertex}}, \tag{21}$$

where $\lambda_1 = 0.2$ and $\lambda_2 = 1 \times 10^{-2}$. Besides, we also introduce vertex loss L_{vertex}, aiming to maintain a soft binding relationship between the 3D Gaussians and the SMPL-X vertices position. During initialization, we initialize the 3D Gaussians based on the SMPL-X vertices. At the same time, we assign each Gaussian a parent attribute, which points to the indices of the SMPL-X vertices. parent($\boldsymbol{\mu}_\text{i}$) refers to the vertex to which the 3D Gaussian is bound. When optimizing the positions of the 3D Gaussians, each Gaussian should not stray far from its corresponding parent vertices. We also maintain a parent array during the densification and pruning processes.

5. Experiments

5.1. Dataset preparation

To address the limitations of existing datasets, which either lack audio input and expression data [ZSZ*21, ZHY*22] or contain only a limited number of identities [YLL*23], we developed the TalkingAvatar dataset [CYY*24]. Our dataset includes 14 subjects, each with 1–3 min of front-facing RGB video recordings. For comparison purposes, we selected four subjects named 'Trading', 'PromoENG', 'Fah' and 'Law', while the remaining subjects are used to demonstrate visual effects only. Each subject has 3000–5000 frames, and each frame is accompanied by a corresponding foreground mask, audio feature, 2D joint keypoints, SMPL-X body poses, expression coefficients and camera parameters. For a subject's front-view monocular video (30fps), we use the first **1500–3000 frames sampled at intervals of 4 frames for training**, and the **remaining 600–1200 frames for testing**.

Table 1: *Learning rate for each module. Start indicates the optimization start step for each module, LR_Start indicates the learning rate at the start of optimization and LR_Final indicates the learning rate at the end of optimization. SAAM stands for parameters of Spatial Audio Attention Model. POSE and EXPRESSION correspond to SMPL-X parameters, and EXPR_SHAPEBLENDS refers to our Learnable Expression Blendshapes. Our final optimization step is 150,000, which costs about an hour on a Nvidia 4090.*

Parameters	Start	LR_Start	LR_Final
SKINNING_WEIGHTS	1500	1×10^{-4}	1×10^{-5}
POSE	1500	3×10^{-5}	1×10^{-5}
EXPRESSION	15,000	3×10^{-5}	1×10^{-5}
EXPR_BLENDSHAPES	25,000	3×10^{-7}	1×10^{-8}
SAAM	50,000	1×10^{-3}	1×10^{-4}

Here is a detailed description of data pre-processing. For each target subject, we require 1 min of talking portrait video with a corresponding speech audio for training. Given a talking video, we first use human pose estimation method SHOW [YLL*23] to estimate corresponding camera translation, body poses θ and expression ψ under a fixed camera focal length and rotation angle. Then, Openpose [CMS*19, SJMS17] is utilized to obtain 2D human keypoints L_{2D} as pesudo ground truth. As described in Section 4.5, we will optimize better body poses and hand pose alignment results under 3DGS rasterizer. Then we use Segment-Anything [KMR*23] with 2D keypoints as prompts to obtain masks for human segmentation. However, when subjects are holding something, hand segmentation often fails. We use Rembg [HCH*19] to get the human foreground, which performs better in such cases. Finally, wav2vec [BZMA20] is utilized to extract audio features from speech audio. So we can use audio signals to drive mouth Gaussian movements.

5.2. Implementation details

Our method is implemented via Pytorch. For each subject, we train a separate, speaker-specific Gaussian model independently. 2DGS, a variant of the 3DGS rasterizer [HYC*24], is utilized in our training process for better view-consistent geometry. For the Gaussian attributes, the learning rate is 1.6×10^{-4} for position $\boldsymbol{\mu}$, 5×10^{-2} for opacity $\boldsymbol{\eta}$, and 5×10^{-3} for the remaining attributes $\{\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{f}\}$. For more config settings, please refer to Table 1. We initially sampled Gaussians on the T-pose vertices of SMPL-X model following GART [LWP*24], and the SMPL-X model is modified by adding additional teeth mesh following GaussianAvatar Head Model [QKS*24]. We add an Expression Blendshapes constraints in Section 4.3, based on $\|\Delta\mathcal{E}\| < \epsilon$. This constraints help LEB from deviating from original bases. We realize it by adding an L2 norm of $\Delta\mathcal{E}$ to the final loss function and multiply it by a factor of 0.1. During the Full-Body Motion Alignment (FMA) stage, our keypoints threshold is set to 0.2.

Table 2: *Multi-task capabilities. We evaluate our model against other methods to demonstrate the various tasks our approach can handle, including upper body synthesis, novel view synthesis and novel body pose synthesis. Additionally, our method can explicitly animate expressions and hand movements.*

Method	Body synthesis	Novel view	Novel pose	Audio-driven	Hand control	Expression control	Training time
InstantAvatar	✓	✓	✓	×	×	×	10 minutes
GART	✓	✓	✓	×	×	×	5 minute
RAD-NeRF	×	✓	×	✓	×	×	6 hours
ER-NeRF	×	✓	×	✓	×	×	3 hours
TalkingGaussian	×	✓	×	✓	×	×	2.5 hours
GaussianTalker	×	✓	×	✓	×	×	2 hours
Ours	✓	✓	✓	✓	✓	✓	1 hour

As we know, LBS is designed for the SMPL-X human template, which has fixed vertex sequences and a pre-defined skinning weights matrix ω. This ω matrix determines each joint's influence on the SMPL-X model's vertices. To compute forward LBS efficiently, we use a voxel-based method to compute the skinning weights of each 3D Gaussian. Given ω, we use a $64 \times 64 \times 64$ grid to pre-compute the skinning weights for each grid vertex based on the linear interpolation of the skinning weights of the nearest three vertices. The skinning weights of each grid point are stored accordingly. The values for an arbitrary 3D Gaussians can be effectively computed as the bi-linear interpolation of the values of the eight grid points nearest to the Gaussian centre.

For our THGS, training a subject costs about an hour on a single Nvidia 4090. We also built a web-based 3D interactive viewer following Viser (https://viser.studio/latest/) for runtime frame rates testing. Our avatar can be animated at 150+ fps on the Viser server.

5.3. Evaluation metrics

The evaluation during testing involves self-reenactment synthesis using PSNR, SSIM and LPIPS metrics. These metrics measure image quality and are commonly used in human re-construction tasks. We use testing frames from our dataset as ground truth and use corresponding audio and body pose from the testing set for self-reenactment synthesis. SyncNet [CZ16] is utilized for comparing lip synchronization. We assess confidence score (Conf), Average frame Offsets between audio and video (AO) and error minimal distance (MD) using SyncNet. AO should be closer to GT, MD values (↓) should be as low as possible and Conf (↑) should be as high as possible.

5.4. Comparison with SoTA

As no existing work re-constructs **expressive** 3D human avatars from **monocular** videos, we structure the experiments into the following two parts: comparisons with human re-construction methods and with 3D talking head methods. For a fair comparison, we use silent audio input and neutral expressions, as current **monocular human re-construction methods** focus on rigid bodies and do not incorporate audio or expressions. On the other hand, the comparison with **3D talking head methods** centres on the quality of self-driven facial dynamics re-construction and lip synchronization. We detail various tasks that our method can handle in Table 2.

Human re-construction comparison includes the recent efficient monocular human re-construction method InstantAvatar [JCSH23] and GART [LWP*24]. InstantAvatar, a state-of-the-art (SoTA) NeRF-based method, uses Instant-NGP as its backbone and employs the efficient Fast-SNARF for root finding from posed space to canonical space. This approach can be viewed as a GPU-efficient reverse LBS method. On the other hand, GART, a SoTA 3DGS-based human rendering method, uses 3D Gaussians as an explicit representation, allowing it to directly use forward LBS to drive human models, significantly increasing rendering speed. Neither methods support expressive control of hand and mouth movements. This highlights the advantages of our approach: precise control over expressions and enhancing the expressive animation capabilities of existing pipeline.

In the comparison with 3D talking head methods, we include NeRF-based approaches RAD-NeRF [TWZ*22] and ER-NeRF [LZB*23], as well as GS-based SoTA methods GaussianTalker [CLY*24] and TalkingGaussian [LZB*24]. RAD-NeRF employs Instant-NGP as its backbone and introduces a Decomposed Audio-spatial Encoding Module to effectively capture facial dynamics. ER-NeRF further explores attention-based audio-spatial correlations, enhancing the modelling of audio-driven lip movements. GaussianTalker replaces NeRF's implicit representation with explicit 3DGS for head modelling, leading to improvements in training time, inference rates and synthesis quality. Similarly, TalkingGaussian utilizes 3DGS for continuous head modelling and incorporates a dual-branch grid-based model to independently address face and mouth deformations in talking head synthesis. Since 3D talking head methods generally generate only the facial region, our comparison focuses on lip-synced mouth movements and facial dynamics synthesis.

5.5. Qualitative experiments

Our method can generate photorealistic hand gestures based on SMPL-X hand poses and create audio-driven mouth movements. Our high-fidelity 3D talking human avatar generation results can be found in Figure 4. THGS produces high-fidelity facial dynamics for different subjects, including detailed facial expressions like wrinkles and smiling.

Comparison with human re-construction methods: Different from previous methods, our approach generates accurate facial dynamics with a wide range of expressions, as shown in Figures 5

Figure 4: *Self-reenactment talking avatar synthesis from testing frames, driven by body poses and speech audio. Our method can generate lifelike 3D talking human avatars with various hand gestures, lip-synced mouth movements and realistic facial dynamics.*

and 6. This improvement is due to our method's ability to process expression coefficients through the LEB and audio input via SAAM, which distinguishes our method from monocular human re-construction approaches like InstantAvatar and GART. The qualitative comparison in **self-reenactment synthesis task** with these methods is illustrated in Figure 5. Figure 5 highlights how our approach overcomes limitations in facial dynamics modelling that earlier monocular human re-construction methods neglected. Unlike previous works, which did not incorporate audio and expression coefficients as inputs, our method uses these inputs to enhance facial dynamics synthesis, introducing novel features in our THGS. Additionally, our method can produce more realistic appearances for common outputs such as hands and arms. We can also achieve better SMPL-X alignment results due to our joint optimization strategy, which can benefit small-region re-construction like hands and arms. In the novel view synthesis task, our method also maintains good 3D consistency thanks to explicit pointcloud-like 3DGS representation, as shown in Figure 7.

Comparison with 3D talking head methods: A visual comparison of facial synthesis quality with 3D talking head methods is shown in Figure 8. We crop the head areas from our generated results and highlight the lip and neck artifacts presented in previous methods, which tend to occur when avatars have large head rotation angles. We will discuss the above issues in Section 5.6.

5.6. Quantitative results

Our main evaluation is concerned with the self-reenactment task. For this purpose, all human avatars are trained on a set of monocular training videos alongside their respective tracking results. We animate the avatars with body poses from a held-out test sequence.

Comparison with human re-construction methods: To ensure fairness among InstantAvatar and GART, we conduct quantitative evaluations using silent audio input and neutral expressions, consistent with previous approaches. The quantitative results of the self-reenactment synthesis are presented in Table 3, which focuses on the comparison of common outputs like hands and body. Our method shows a 27.07% improvement in LPIPS over the SoTA monocular human re-construction methods.

Comparison with 3D talking head methods: Moreover, we compare lip synchronization with 3D talking head methods. For a fair comparison, expression inputs are excluded from the pipeline and highlighted in the 'w/o expr.' row. We adopt SyncNet metrics: average offsets (AO), confidence (Conf) and MD to measure audio-lips synchronization, as shown in Table 4. The results demonstrate that our method effectively enables the **audio-driven mouth movements**, thanks to our predicted mouth Gaussian movements from SAAM. As presented in Table 5, the improvement in facial dynamics synthesis is not solely due to the

Figure 5: *Comparison of generated key frame results with SoTA of monocular human re-construction methods. We compared the pronunciation: 'easily', 'thousand', 'ten', 'how' and 'always'.* ***The highlighted red words indicate the corresponding pronunciation moments; please focus on the mouth area****. Our generated results exhibit a photorealistic hand appearance,* ***lip-synced mouth movements*** *and dynamic facial dynamics. The facial dynamic synthesis is a new aspect of our method. In common outputs, such as the synthesis of hands and the torso, our method also demonstrates superiority over previous approaches.*

SAAM module's handling of audio-driven mouth movements. Compared to other 3D talking head methods, our approach benefits from using the SMPL-X model as a full-body human prior. Previous methods often rely on the BFM model [PKA*09] either for vertex initialization [YQY*24, LZB*24] or for head pose initialization [TWZ*22, LZB*23]. In contrast, our approach adopts the FLAME model [LBB*17], which offers superior accuracy in fitting facial shapes and head poses [EST*20] in dynamic sequences. Furthermore, prior methods that **model the head and torso separately** often encounter artifacts in the lip and neck

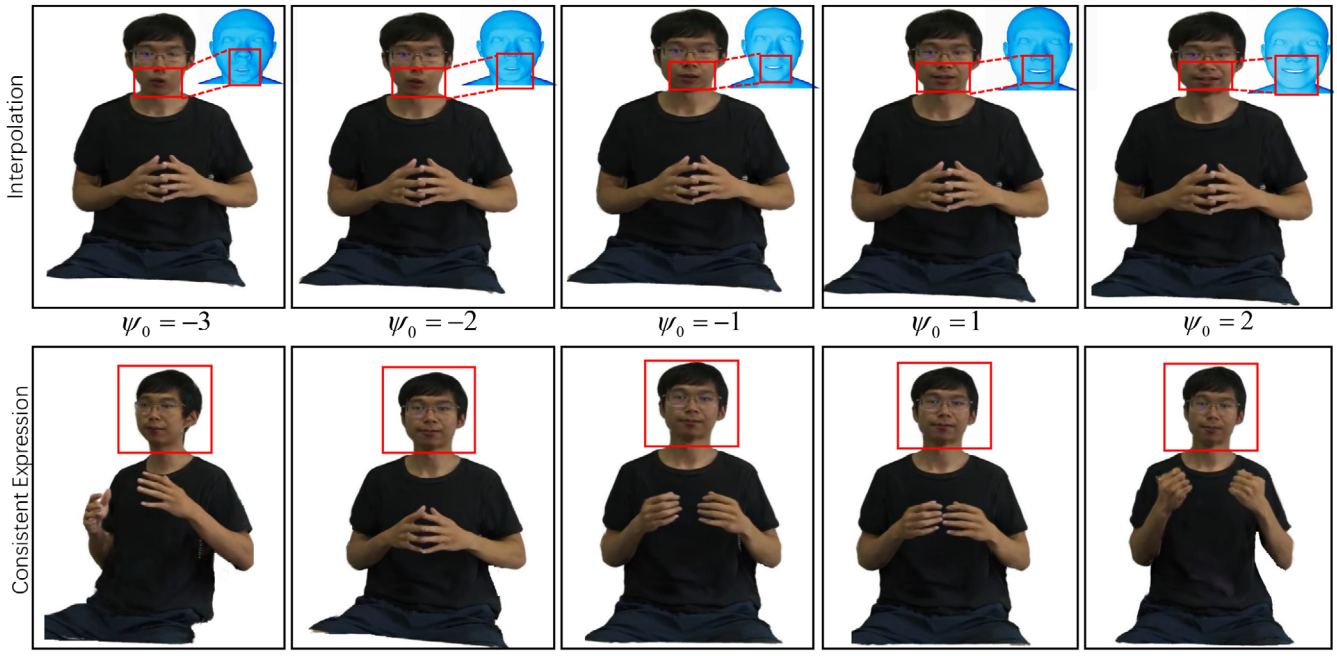

Figure 6: *Explicit expression control with Learnable Expression Blendshapes. In the first row, we interpolate the expression coefficients ψ_0 to demonstrate the* ***continuous transition from a 'puzzled' expression to a 'laugh' expression*** *controlled by LEB. In the second row, we maintain* ***a constant expression*** *ψ while varying the camera positions and hand gestures to demonstrate the consistency of the 'smile' expression. Our Learnable Expression Blendshapes (LEB) enable explicit 3D-consistent control of facial expressions.*

Figure 7: ***Novel view synthesis*** *by rotating the camera along the (x, y) axis, demonstrating that our re-constructed Gaussian avatar maintains* ***3D consistency****. Note that the training data only consists of front-view monocular video, which means that tilting the camera too much along certain angles may lead to artifacts in the novel view.*

regions at extreme angles, resulting in decreased image quality metrics.

5.7. Ablation study

LEB: We study the effect of LEB in self-reenactment synthesis (Table 6). Without LEB, the generated avatars experience a decline in image quality across all subjects. With LEB, we can explicitly control facial expression by adjusting expression coefficients ψ_0 from -3 to 2, which is shown in the first row of Figure 6. In the second row, when we keep a consistent expression coefficient, we can maintain the facial expression of the human avatar across various camera perspectives and body poses. We also present a visual ablation study featuring our LEB module, as shown in Figure 9. More results can be found in video materials. These results demonstrate that our LEB module not only significantly enhances facial dynamics but also explicitly generates diverse expressions more accurately, given expression coefficients as input. Unlike previous methods, our approach leverages additional expression inputs to generate more expressive human avatars, with only three FPS drops compared with GART.

SAAM: Our SAAM can achieve better audio-driven mouth movements, greatly enhancing the lip synchronization ability, as shown in Table 4. We use sync quality metrics used in SyncNet to compare audio and lip movement synchronization. Without SAAM, image quality suffers a slight decline, as shown in Table 6.

Joint optimization strategy: Our joint body pose, expression and skinning weights optimzation strategy uses Landmark Distances (LMD) between projected SMPL-X joints and estimated OpenPose pseudo Ground Truth for comparison. With FMA, we achieve better alignment results before training, as shown in Table 7.

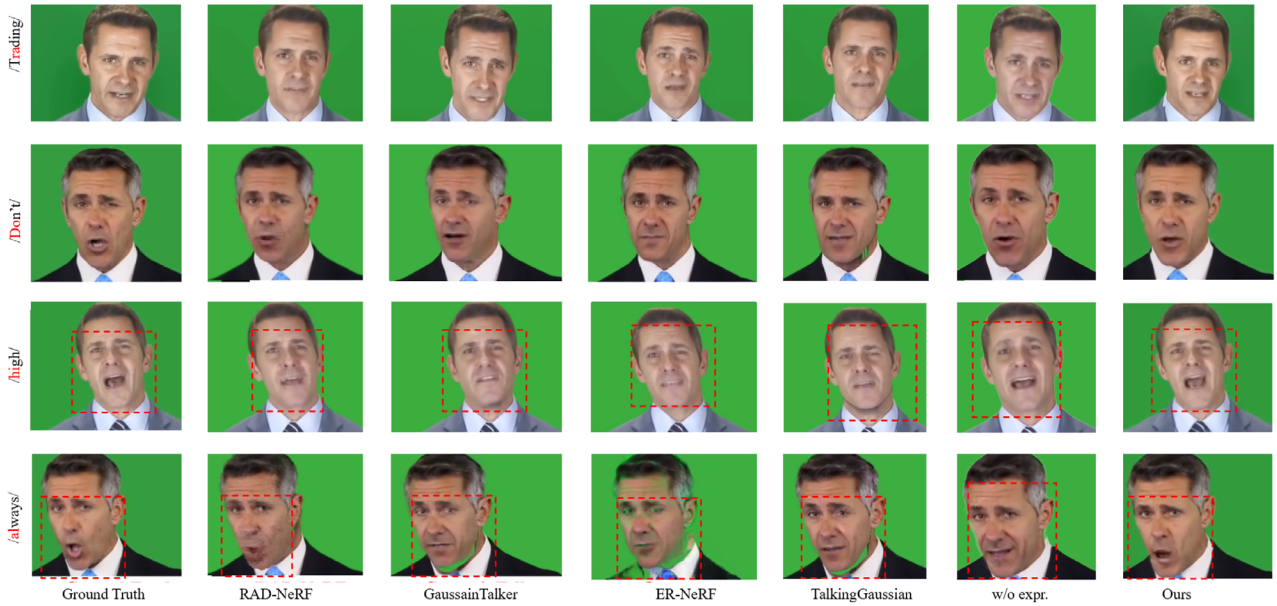

Figure 8: *Comparison of generated key frame results with SoTA 3D talking head methods. We compared the pronunciation: 'Trading', 'Don't', 'high' and 'always'. We highlight the lip and neck artifacts in the previous methods, which can occur when avatars have large head rotation angles. This is mainly caused by separate modelling of the head and shoulder in these methods.* ***w/o expr****. means that expression inputs are excluded from pipeline for a fair comparison.*

Table 3: *Quantitative results on self-reenactment synthesis compared to SoTA, using silent audio input and neutral expressions. We present self-reenactment synthesis results for four subjects on testing frames, demonstrating that our method achieves the highest image quality for novel pose synthesis. Note: LPIPS values have been scaled by a factor of* 10^2*. IA refers to InstantAvatar. The 'smplx' suffix refers to using the SMPL-X model in place of the original SMPL model in these methods.*

	Trading			Fah			Law			PromoENG			
Method	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	FPS
IA	25.099	0.9318	10.742	28.320	0.9613	4.633	23.275	0.9317	9.245	23.369	0.9342	11.276	15
IA_{smplx}	25.113	0.9328	10.173	28.868	0.9636	4.155	23.640	0.9318	8.735	23.120	0.9306	10.751	15
GART	26.123	0.9389	6.036	29.558	0.9633	2.576	29.212	0.9550	5.431	27.572	0.9389	7.611	**158**
$GART_{smplx}$	26.198	0.9392	6.223	29.497	0.9631	2.774	29.059	0.9539	5.765	27.645	0.9396	7.825	**158**
Ours	**26.215**	**0.9401**	**3.756**	**30.112**	**0.9664**	**1.944**	**29.412**	**0.9621**	**3.836**	**27.834**	**0.9414**	**5.642**	155

Bold signifies the best in each column.

Table 4: *Comparison on our lip-synced mouth movements generation with 3D Talking head methods across different subjects.* ***AO*** *stands for AV offset: the offset between audio and video,* ***MD*** *stands for Min Dist: the minimal distance of extracted features between audio and video and* ***Conf*** *stands for confidence: how much confidence we should trust that audio and video are synchronized. The best results are in* ***bold****, and the second-best results are underlined.* ***w/o expr****. means expression inputs are excluded from pipeline for a fair comparison.*

	Trading			Fah			Law			PromoENG		
Method	AO	MD ↓	Conf ↑	AO	MD ↓	Conf ↑	AO	MD ↓	Conf ↑	AO	MD ↓	Conf ↑
Ground Truth	−2	7.49	7.58	−3	7.10	8.79	0	6.88	8.64	−3	7.80	7.36
RAD-NeRF	−4	11.51	2.62	−5	12.35	2.25	−7	11.53	3.50	−4	12.40	2.14
ER-NeRF	−5	12.80	2.48	−6	13.02	1.98	−8	12.71	3.41	−5	13.08	2.00
TalkingGaussian	−5	12.74	2.31	−5	13.17	2.55	−8	12.54	3.24	−5	12.79	2.23
GaussianTalker	−5	12.21	2.85	−2	12.71	2.45	−9	12.32	3.54	−4	11.21	2.39
w/o SAAM	−3	10.05	4.39	−2	12.38	3.49	−4	10.06	4.82	−1	12.89	2.62
w/o expr.	−3	**9.85**	4.64	−2	11.62	4.28	−2	9.80	5.17	−2	10.99	4.11
Full	**−2**	9.99	**5.10**	**−4**	**9.79**	**5.17**	**−2**	**9.62**	**5.27**	**−3**	**9.85**	**4.63**

Table 5: *The quantitative results of the self-reconstruction setting with 3D talking head methods. The comparison is limited to the head and neck region, excluding the shoulder area. w/o expr. means that expression inputs are excluded from pipeline for a fair comparison.*

Method	PSNR ↑	LPIPS ↓	FPS ↑
RAD-NeRF	29.881	4.459	32
ER-NeRF	29.338	4.227	35
TalkingGaussian	29.795	3.730	108
GaussianTalker	30.289	3.825	120
w/o expr.	31.086	3.271	**155**
Ours	**31.257**	**2.923**	**155**

Bold signifies the best in each column.

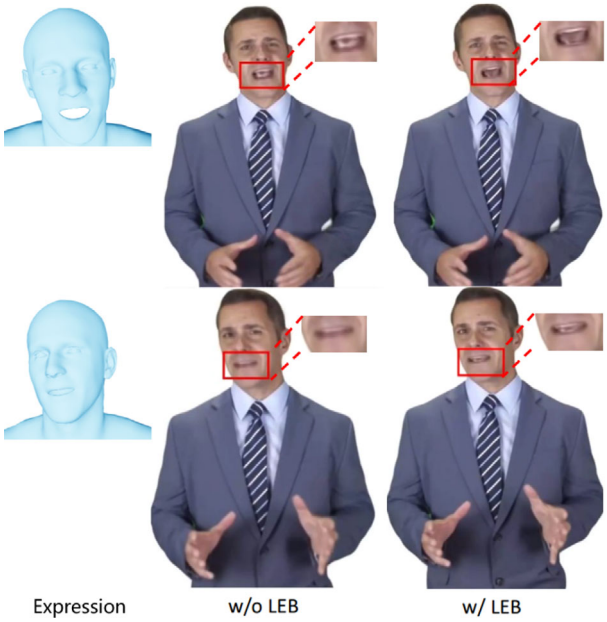

Figure 9: *Ablation study on Learnable Expression Blendshapes. This improvement in the realism of facial dynamics demonstrates that our LEB module can explicitly animate our avatars with various expressions more vividly and accurately.*

We compare LMD for the face, body and hands separately, with the most significant drops in the hand and face regions. Visual alignment result is showcased in Figure 10.

Disentanglement of expression and audio inputs: To evaluate the contributions of expression coefficients and audio inputs, we remove each input separately. This allows us to assess their respective impacts on lip synchronization and image fidelity. Specifically, we observe a larger drop in lip synchronization quality when audio inputs are removed as shown in Table 8 and Figure 11. In contrast, removing only the expression inputs has a minimal impact on lip-synced ability, but results in a slight decline in image fidelity. Expression coefficients provide explicit control over facial expressions, as demonstrated in Figure 6 and the accompanying video materials, offering an advantage over comparison methods. In summary, both audio and expression inputs are essential for reproducing realistic facial dynamics. Expression inputs primarily control facial expressions and can also manage rough lip movements, while audio serves to further enhance lip synchronization.

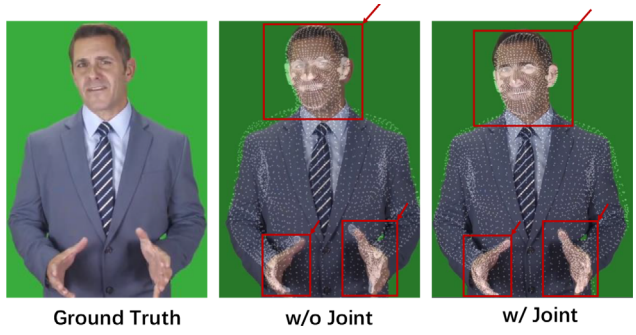

Figure 10: *Joint optimization strategy provides more accurate body pose alignment, particularly for the alignment of hands and faces. Please zoom in for better results. w/o Joint denotes the initial pose provided by the talk show, while w/ Joint shows the complete alignment of hand and facial keypoints after Joint optimization strategy.*

Figure 11: *Ablation study on different inputs. Using only audio inputs, the SAAM module is capable of generating realistic lip-synced mouth movements. When expression inputs are incorporated, subtle facial expressions can be effectively re-enacted.*

6. Limitations and Future Work

In our experiments, we demonstrate that THGS can create controllable and high-fidelity talking human avatars from short monocular video data. However, due to the limitations of monocular video reconstruction, when the human torso undergoes significant rotation, body parts not captured by the training video may have artifacts. Training with multi-view data should mitigate this issue. Additionally, our method faces challenges in re-constructing and driving the tongue, which is fundamentally restricted to the SMPL-X model animation ability. In the future, we need a physics-integrated [XZQ*24] human model, beyond a learned human mesh model. Furthermore, as a learning-based approach to avatar generation, despite the short training time, our method is somewhat limited by training separate model for each individual. We believe that the recent large-scale multi-view video dataset of full-body human [ZSZ*21, ZHY*22, ZHW*23] open up gates to learn a more

Table 6: *Ablation study on Spatial Audio Attention Module (SAAM) and Learnable Expression Blendshapes (LEB). Comparison of models with and without SAAM and Learnable Expression Blendshapes across different subjects. Note: LPIPS values have been scaled by a factor of* 10^2.

	Trading			Fah			Law			PromoENG		
Method	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Vanilla	26.123	0.9389	6.036	29.558	0.9633	2.576	29.212	0.9550	5.431	27.572	0.9389	7.611
w/o SAAM	25.672	0.9358	3.603	29.917	0.9653	1.793	29.322	0.9561	**3.949**	27.131	0.9329	5.804
w/o LEB	25.661	0.9354	3.776	29.742	0.9644	1.919	28.919	0.9543	4.185	27.124	0.9326	6.010
Full	**26.157**	**0.9390**	**3.454**	**30.021**	**0.9656**	**1.728**	**29.342**	**0.9562**	4.043	**27.754**	**0.9405**	**5.790**

Bold signifies the best in each column.

Table 7: *Comparison of our pose alignment. Comparison of averaged landmark distance (LMD) between Vanilla model, Full-Body Motion Alignment (+FMA) results and our Full model.*

Method	LMD (Hand)↓	LMD (Body)↓	LMD (Face)↓
Vanilla	9.4595	6.8904	10.3728
+FMA	5.9740	4.4772	4.6682
Full	**4.4407**	**4.3587**	**3.2619**

Bold signifies the best in each column.

Table 8: *Ablation study on different input conditions. The comparison is limited to the head and neck region.*

Method	PSNR ↑	LPIPS ↓	MD ↓	Conf ↑
Ground Truth	N/A	0	7.32	8.09
w/o audio	31.115	3.148	11.35	3.83
w/o expr.	31.086	3.271	10.57	4.55
Full	**31.255**	**2.925**	**9.81**	**5.04**

Bold signifies the best in each column.

generalized human model with Large Gaussian Model [TCC*24], through the use of photometric optimization and efficient rendering, like 3DGS.

7. Conclusion

This paper extends the 3D talking head generation task to talking human avatar synthesis by introducing a novel THGS pipeline. Unlike existing 3DGS-based human re-construction methods, THGS enhances expressive animation capabilities and achieves high-fidelity, real-time 3D talking human avatar generation. We design LEB to model facial dynamics and explicitly control the facial expressions. A SAAM is proposed to generate lip-synced mouth movements. A body pose, expression, and skinning weights joint optimization strategy is proposed for better alignment results with monocular video inputs. Our method can render lifelike talking human avatars in real time, with various hand gestures, subtle facial expressions, and audio-driven lip synchronization, on a web-based visualizer (see Supporting Information).

Acknowledgements

This work is supported by the National Nature Science Foundation of China (62472395), the Fundamental Research Funds for the Central Universities (WK2100000047) and University Synergy Innovation Program of Anhui Province under Grant (GXXT-2022-036).

Conflicts of Interest

None of the authors have a conflict of interest to disclose. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[BV99] Blanz V., Vetter T.: A morphable model for the synthesis of 3D faces. In *SIGGRAPH'99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (USA, 1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194. https://doi.org/10.1145/311535.311556.

[BV23] Blanz V., Vetter T.: *A Morphable Model for the Synthesis of 3D Faces* (1st edition). Association for Computing Machinery, New York, NY, USA, 2023. https://doi.org/10.1145/3596711.3596730.

[BZMA20] Baevski A., Zhou Y., Mohamed A., Auli M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (Vancouver, Online, Canada, 2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), vol. *33*, Curran Associates, Inc., pp. 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[CHC*24] Chen B., Hu S., Chen Q., Du C., Yi R., Qian Y., Chen X.: Gstalker: Real-time audio-driven talking face generation via deformable Gaussian splatting. http://arxiv.org/abs/2404.19040 (2024).

[CLY*24] Cho K., Lee J., Yoon H., Hong Y., Ko J., Ahn S., Kim S.: GaussianTalker: Real-time talking head synthesis with 3D Gaussian splatting. In *MM'24: Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC,

Australia, 2024), Association for Computing Machinery, pp. 10985–10994. https://doi.org/10.1145/3664647.3681627.

[CMS*19] Cao Z., Martinez G. H., Simon T., Wei S., Sheikh Y. A.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 1 (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257.

[CYY*24] Chen C., Yu L., Yang Q., Zheng A., Xie H.: Talkingavatar [dataset]. Github. https://github.com/sora158/THGS (2024).

[CZ16] Chung J. S., Zisserman A.: Out of time: Automated lip sync in the wild. In *Workshop on Multi-View Lip-Reading, ACCV* (2016).

[EST*20] Egger B., Smith W. A. P., Tewari A., Wuhrer S., Zollhoefer M., Beeler T., Bernard F., Bolkart T., Kortylewski A., Romdhani S., Theobalt C., Blanz V., Vetter T.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics 39*, 5 (June 2020). https://doi.org/10.1145/3395208.

[GCL*21] Guo Y., Chen K., Liang S., Liu Y., Bao H., Zhang J.: AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).

[GLMH21] Guo M.-H., Liu Z.-N., Mu T.-J., Hu S.-M.: Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 5 (2023), 5436–5447. https://doi.org/10.1109/TPAMI.2022.3211006.

[HCH*19] Hou Q., Cheng M.-M., Hu X., Borji A., Tu Z., Torr P. H. S.: Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*, 4 (Apr. 2019), 815–828. https://doi.org/10.1109/TPAMI.2018.2815688.

[HHL24] Hu S., Hu T., Liu Z.: GauHuman: Articulated Gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 20418–20431.

[HTZ*24] Huang Z., Tang F., Zhang Y., Cun X., Cao J., Li J., Lee T.-Y.: Make-your-anchor: A diffusion-based 2D avatar generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 6997–7006.

[HYC*24] Huang B., Yu Z., Chen A., Geiger A., Gao S.: 2D Gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH'24: ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA, 2024), Association for Computing Machinery. https://doi.org/10.1145/3641519.3657428.

[HZZ*24] Hu L., Zhang H., Zhang Y., Zhou B., Liu B., Zhang S., Nie L.: GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D Gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024).

[JCSH23] Jiang T., Chen X., Song J., Hilliges O.: InstantAvatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada, 2023), pp. 16922–16932.

[JHBZ22] Jiang B., Hong Y., Bao H., Zhang J.: SelfRecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[KHWKS23] Karras J., Holynski A., Wang T.-C., Kemelmacher-Shlizerman I.: DreamPose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, October 2023), pp. 22680–22690.

[KKLD23] Kerbl B., Kopanas G., Leimkühler T., Drettakis G.: 3d Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics 42*, 4 (July 2023), 1–14. https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[KMR*23] Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R.: Segment anything. In IEEE/CVF International Conference on Computer Vision (ICCV) (Paris, France, 2023), pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371.

[LBB*17] 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 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813.

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics 34* (2015), 1–16.

[LWL*24] Liu X., Wu C., Liu J., Liu X., Wu J., Zhao C., Feng H., Ding E., Wang J.: GVA: Reconstructing vivid 3D gaussian avatars from monocular videos. CoRR abs/2402.16607 (2024). https://doi.org/10.48550/arXiv.2402.16607.

[LWP*24] Lei J., Wang Y., Pavlakos G., Liu L., Daniilidis K.: GART: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 19876–19887.

[LWZ*22] Liu X., Wu Q., Zhou H., Du Y., Wu W., Lin D., Liu Z.: Audio-driven co-speech gesture video generation. In *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA, 2022), Curran Associates Inc.

[LXW*22] Liu X., Xu Y., Wu Q., Zhou H., Wu W., Zhou B.: Semantic-aware implicit neural audio-driven video portrait generation. In *Computer Vision – ECCV 2022: 17th European Conference* (Tel Aviv, Israel, 2022), Springer-Verlag, pp. 106–125. https://doi.org/10.1007/978-3-031-19836-6_7.

[LZB*23] Li J., Zhang J., Bai X., Zhou J., Gu L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, Oct.r 2023), pp. 7568–7578.

[LZB*24] Li J., Zhang J., Bai X., Zheng J., Ning X., Zhou J., Gu L.: TalkingGaussian: Structure-persistent 3D talking head synthesis via Gaussian splatting. In *Computer Vision – ECCV 2024: 18th European Conference* (Milan, Italy, 2024), Springer-Verlag, pp. 127–145. https://doi.org/10.1007/978-3-031-72684-2_8.

[LZWL24] Li Z., Zheng Z., Wang L., Liu Y.: Animatable Gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024).

[NRB*24] Ng E., Romero J., Bagautdinov T., Bai S., Darrell T., Kanazawa A., Richard A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024).

[PCG*19] Pavlakos G., Choutas V., Ghorbani N., Bolkart T., Osman A. A. A., Tzionas D., Black M. J.: Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA, USA, 2019), pp. 10975–10985.

[PHS*24] Peng Z., Hu W., Shi Y., Zhu X., Zhang X., He J., Liu H., Fan Z.: SyncTalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024).

[PKA*09] Paysan P., Knothe R., Amberg B., Romdhani S., Vetter T.: A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (Genoa, Italy, 2009), pp. 296–301. https://doi.org/10.1109/AVSS.2009.58.

[QKS*24] Qian S., Kirschstein T., Schoneveld L., Davoli D., Giebenhain S., Nießner M.: GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 20299–20309.

[QTZ*21] Qian S., Tu Z., Zhi Y., Liu W., Gao S.: Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 11077–11086.

[SJMS17] Simon T., Joo H., Matthews I., Sheikh Y.: Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii, 2017).

[SLZ*22] Shen S., Li W., Zhu Z., Duan Y., Zhou J., Lu J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Tel Aviv, Israel, 2022).

[TCC*24] Tang J., Chen Z., Chen X., Wang T., Zeng G., Liu Z.: LGM: Large multi-view Gaussian model for high-resolution 3D content creation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV* (Milan, Italy, 2024), Springer-Verlag, pp. 1–18. https://doi.org/10.1007/978-3-031-73235-5_1.

[TWZ*22] Tang J., Wang K., Zhou H., Chen X., He D., Hu T., Liu J., Zeng G., Wang J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. https://arxiv.org/abs/2211.12368 (2022).

[WLL*24] Wang T., Li L., Lin K., Zhai Y., Lin C.-C., Yang Z., Zhang H., Liu Z., Wang L.: DisCo: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 9326–9336.

[XZQ*24] Xie T., Zong Z., Qiu Y., Li X., Feng Y., Yang Y., Jiang C.: PhysGaussian: Physics-integrated 3D Gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 2024), pp. 4389–4398.

[YLL*23] Yi H., Liang H., Liu Y., Cao Q., Wen Y., Bolkart T., Tao D., Black M. J.: Generating holistic 3D human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, Canada, 2023), pp. 469–480.

[YQY*24] Yu H., Qu Z., Yu Q., Chen J., Jiang Z., Chen Z., Zhang S., Xu J., Wu F., Lv C., Yu G.: GaussianTalker: Speaker-specific talking head synthesis via 3D Gaussian splatting. In *MM'24: Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia, 2024), Association for Computing Machinery, pp. 3548–3557. https://doi.org/10.1145/3664647.3681675.

[YZY*22] Yao S., Zhong R., Yan Y., Zhai G., Yang X.: DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering. https://arxiv.org/abs/2201.00791 (2022).

[ZHW*23] Zhou T., He K., Wu D., Xu T., Zhang Q., Shao K., Chen W., Xu L., Yu J.: Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada, 2023), pp. 4315–4327.

[ZHY*22] Zheng Z., Huang H., Yu T., Zhang H., Guo Y., Liu Y.: Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, June 2022).

[ZPBG01] Zwicker M., Pfister H., Baar J. V., Gross M.: Surface splatting. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (New York, NY, USA, 2001), ACM, pp. 371–378. https://doi.org/10.1145/383259.383300.

[ZRA23] Zhang L., Rao A., Agrawala M.: Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, 2023).

[ZSZ*21] Zheng Y., Shao R., Zhang Y., Yu T., Zheng Z., Dai Q., Liu Y.: DeepMultiCap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV)* (2021).

[ZTZ*21] Zhang H., Tian Y., Zhou X., Ouyang W., Liu Y., Wang L., Sun Z.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021).

[ZTZ*23] Zhang H., Tian Y., Zhang Y., Li M., An L., Sun Z., Liu Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45* (2023), 12287–12303.

[ZZ22] 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 (CVPR)* (New Orleans, LA, USA, 2022), pp. 3657–3666.

[ZZS*24] Zheng S., Zhou B., Shao R., Liu B., Zhang S., Nie L., Liu Y.: GPS-Gaussian: Generalizable pixel-wise 3D Gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024).

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information