TimeWalker: Personalized Neural Space for Lifelong Head Avatars

Dongwei Pan[1,2], Yang Li[1,3], Hongsheng Li[4] ✉, and Kwan-Yee Lin[5] ✉ *

[1] Shanghai AI Laboratory [2] Baidu [3] Institute of Automation, CAS
[4] CUHK MMLab [5] University of Michigan

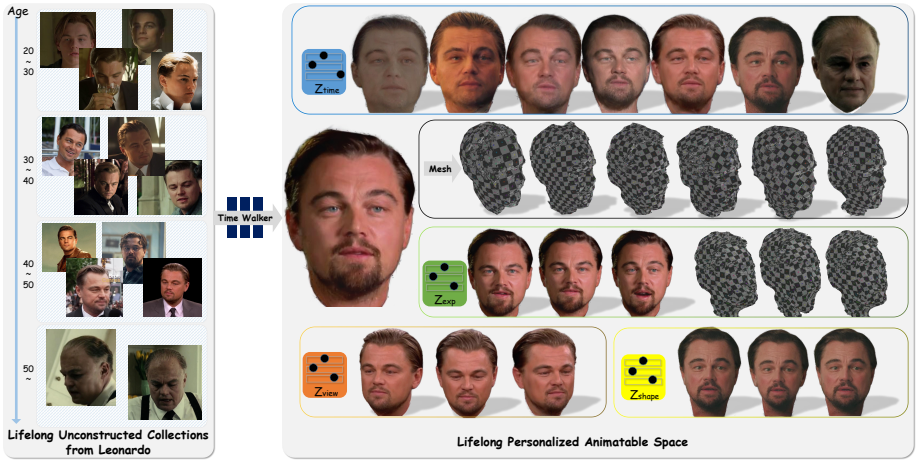

Fig. 1: TimeWalker. Given a set of unstructured data from the Internet or photo collection across years, we build a personalized neural parametric morphable model, *TimeWalker*, towards replicating a life-long 3D head avatar of a person. With the TimeWalker, we can control and animate one's avatar in terms of shape, expression, viewpoint, and appearance across his/her different age periods. In this Figure, we show Leonardo DiCaprio's lifelong avatar reconstructed and animated by our proposed model.

Abstract. We present **TimeWalker**, a novel framework that models realistic, full-scale 3D head avatars of a person on *lifelong scale*. Unlike current human head avatar pipelines that capture a person's identity only at the momentary level (*i.e.,* instant photography, or short videos), TimeWalker constructs a person's comprehensive identity from unstructured data collection over his/her various life stages, offering a paradigm to achieve full reconstruction and animation of that person at different moments of life. At the heart of TimeWalker's success is a novel neural parametric model that learns personalized representation with the disentanglement of shape, expression, and appearance across ages. Central to our methodology are the concepts of two aspects: 1) We track back to the principle of modeling a person's identity in an additive combination of his/her average head representation in the canonical space, and moment-specific head attribute representations driven from a set of neural head basis. To learn the set of head

*✉ H. Li and K. Lin are co-corresponding authors.

basis that could represent the comprehensive head variations of the target person in a compact manner, we propose a **Dy**namic **N**eural B**a**sis-Blending **Mo**dule (**Dynamo**). It dynamically adjusts the number and blend weights of neural head bases, according to both shared and specific traits of the target person over ages. 2) We introduce **Dyna**mic **2D G**aussian **S**platting (**DNA-2DGS**), an extension of Gaussian splatting representation, to model head motion deformations like facial expressions without losing the realism of rendering and reconstruction of the full head. DNA-2DGS includes a set of controllable 2D oriented planar Gaussian disks that utilize the priors from a parametric morphable face model, and move/rotate with the change of expression. Through extensive experimental evaluations, we show TimeWalker's ability to reconstruct and animate avatars across decoupled dimensions with realistic rendering effects, demonstrating a way to achieve personalized "time traveling" in a breeze. Project page: https://TimeWalker2026.github.io/.

Keywords: Head Reconstruction · Head Modeling · Animation · Lifelong Avatar Modeling

1 Introduction

"*As we grow older, our life stories become our identity. This ongoing narrative of the self is constructed from the past and anticipated future.*"

What forms a person's identity? In the realm of Computer Graphics and Vision, researchers have traditionally considered an individual's shape as invariant, assuming it to represent the person's identity. This assumption has driven significant advancements in human faces and head modeling over decades. From classic 3DMMs (3D morphable models) [2] and its follow-up line of work [23, 29], to current advanced neural representations for head modeling [13, 15, 30, 41], 3D head avatars become increasingly lifelike, serving as *momentary* replicas of humans.

However, sociology and psychology provide a different perspective to answer the question of identity formation – the fields emphasis the idea that a person's identity is shaped by a continuous process influenced by various lifestages and experiences of self over time, rather than a single moment. Since the $1960s$, sociologists and psychologists (*e.g.,* [22, 35]) have recognized the intrinsic value of *lifelong* construction of self-identity, and have developed several cornerstone theories upon this basis.

Motivated by the critical identity formation gap, in this work, we aim to explore a paradigm of **modeling 3D head avatars towards a person's lifelong scale**. This new problem definition introduces three major challenges to head avatar modeling: 1)**Long-horizon Identity Consistency Preserving.** The life-long modeling breaks the long-lasting assumption of the shape invariant of a person. As the musculoskeletal structure changes with age growth, a person's facial/head shape could change significantly over his/her different lifestages. Besides, aesthetic transformation and unique life experiences also stimulate significant physical changes like facial texture features, hairstyle, accessories, and motion behavior of a person, compounding the difficulties of learning effective representation to preserve identity. 2) **Limited Data Quantity and Quality for Each Lifestage.** In prior research, specific momentary data is required with either "standard inputs" (*i.e.,* high-quality front-view images for 3D-aware

head generation/editing [31]), or sufficient geometric cues (*i.e.,*multi-view capture systems [21, 28, 39, 43], depth sensors [8, 25], or short video sequences [12, 48] for 3D/4D head avatar reconstruction). In contrast, it is infeasible to capture lifelong data of a person with sufficient geometric cues, or a unified frontal camera view for each lifestage. Oftentimes, one's lifetime is recorded through unstructured image collection, with extremely uneven data volume and viewpoints over different moments. Such data poses a significant challenge to high-fidelity 3D head avatar modeling in both appearance-realistic and geometry-plausible aspects. 3) **Explicit-controlled Animation in Full Scales.** Aside from lifelike reconstruction, a lifelong avatar is also expected to be controllable in terms of expression, shape, viewpoint, headpose, and appearance across a person's different age periods. Fueled by the disentangled space of parametric morphable head models like FLAME [23] and the expressiveness of neural field representation [26, 27], previous arts of 3D head avatar animation [12, 17, 30, 45, 46, 48] could support head animation at different scales with vivid details. However, none of these methods could manipulate *moments of life*. How to learn a personalized disentangled space from highly unstructured data, that enables full-scale control without losing realism is yet unknown.

We present **TimeWalker** – a baseline solution that tackles the above challenges for lifelong head avatar modeling from two aspects: **1)** To get invariant identity representation, and overcome the limited data issues, we track back to the principle of the classic 3DMMs, where a specific 3D mesh can be approximated by a mean template with shape/expression's main modes of variation in an additive combination manner. Analogously, TimeWalker models a person's each lifestage by the additive combination of a shared representation in canonical space, and a set of neural head basis. The former serves as the invariant, *i.e.,*average head representation, of the person's identity across different ages. The latter encodes the main modes of variation of moment-specific head attributes via a *Dynamic Neural Basis-Blending Module (Dynamo)*. The personalized space parameterized by these two components could provide both geometric and appearance priors to life moments with rare data. **2)** To achieve full-scale control while ensuring realism, we introduce *Dynamic 2D Gaussian Splatting (DNA-2DGS)* upon the additive combination framework. It extends static Gaussian splatting representation to a set of animatable Gaussian Surfels for human heads via two deformation guidelines - the deformation fields driven from the neural head basis, and inverse warping operation rooted from FLAME expression coefficients. These ensure the surfels could be animated properly without losing geometric realism caused by FLAME's fixed topology, or explicit control problems caused by implicit neural head basis representation. With above designs, TimeWalker can learn a disentangled personalized neural space (Fig. 1) only from one's lifelong *unstructured* image collection.

To enable modeling heads on lifelong scale, a substantial dataset of the same individual at different time points is necessary. However, this requirement is hard to fulfill with current open-source datasets since they neglect the lifelong human concept. Thus, we construct TimeWalker-1.0, a large-scale head dataset comprising 40 celebrities' lifelong image collection sourced from various Internet data. It contains over 2.7 million individual frames, with each identity consisting of $15K - 260K$ frames and diverse age and head pose distributions. In experiments, we show TimeWalker's ability to reconstruct and animate avatars across decoupled dimensions with realistic rendering effects. Then, we

demonstrate the superior rendering and geometry reconstruction quality by comparing our models to SOTA momentary 3D head models. We also demonstrate our designs' effectiveness with ablation studies. Finally, we show our model's potential benefits to downstream applications like 3D Editing.

2 Related Works

Personalized Head Avatars. 3DMM [2], as the foundational work in 3D human head modeling, constructs a generic *head* space via linear combination of a mean template mesh and low-dimensional linear subspaces of shape and expression from PCA. Subsequent research extends it to personalized head mesh modeling. For instance, [5] predicts personalized corrections on a 3DMM prior to obtain user-specific expression blendshapes and dynamic albedo maps. [47] learns personalized face details via multi-view image fusion from virtually rendered multi-view input images. Recent advancements have extended focus to creating animatable personal *head* avatars with realistic rendering. Methods like NerFace [12] and IM Avatar [45] leverage FLAME [23] expression coefficients to drive neural scene representation networks, implicitly representing head avatars from monocular video inputs. INSTA [48] enhances training speed and enables avatar control through a mesh-based warping field. Efforts like GaussianAvatars [30], FlashAvatar [36] and other recent studies [34, 37] leverage 3DGS's representational capabilities. They associate Gaussian points with a 3D parametric model or head mesh, generating personalized head avatars by using expression parameters as conditions for Gaussian offsets. PAV [3] shares a conceptual similarity with our approach in modeling variations of a person. Building upon INSTA, it extends to support multiple appearances of an individual from unstructured video data through the use of appearance embeddings. In contrast, TimeWalker diverges by emphasizing an interpretable, scalable, and steerable framework for constructing lifelong personalized spaces, offering explicit control over multiple-scale animations and preserving identity consistency.

By harnessing generative prior, DreamBooth [32] and Lora [16] customize diffusion models from multiple images of a particular subject to produce personalized outcomes. However, these approaches are limited to static 2D results, lacking 3D consistency and animation capability. Other works like DiffusionAvatars [20] and GANAvatar [17] utilize generative models to create personalized head avatars. The former fine-tunes ControlNet [44] with NPHM [13] features. The later distill EG3D [4] to single appearance. Focusing on one-shot talking head generation, methods like [7, 11, 24, 40] could achieve effective head reconstruction and animation, benefiting from the generative priors/pretrained vision models to extract fundamental human head features and 3D representation for geometric reasonability. While these pipelines excel at face reenactment, their momentary representations limit their ability to address the challenge of modeling personalized spaces over a lifelong scale. Our pipeline takes a step in this direction with a baseline solution, enabling the construction of a lifelong replica. A comparison between TimeWalker and representative methods is shown in Tab. 1.

Neural Representation for Static Reconstruction. In contrast to traditional explicit reconstruction methods like meshes, point cloud and voxel grids, neural representations like NeRF [26], show promise with high-fidelity rendering [1], efficient training [27, 42],

Table 1: TimeWalker enables preserving identity consistency in the long-horizon time spectrum (*Lifelong Replica*), with explicit-controlled animation at full scale (*Animation-Expression/Shape*). It also supports surface reconstruction and produces dynamic mesh efficiently under sparse view observations for each life stage (*Mesh Reconstruction*), without losing rendering realism (*High-Fidelity Rendering*). **GS**: Gaussian Splatting, **N**: Nerf-based, **Tri**: Tri-plane. **All the mentioned methods are compared in our experiments.**

	Lifelong Replicas	Animation Expression	Animation Shape	Mesh Reconstruction	High-Fidelity Rendering	Open Source	Base Representation
Gaussian Surfels [9]	×	×	×	✓	✓	✓	**GS**
INSTA [48]	×	✓	✓	✓	✓	✓	**N**
FlashAvatar [36]	×	✓	×	×	✓	✓	**GS**
GANAvatar [17]	×	✓	×	×	✓	✓	**Tri**
GAGAvatar [7]	×	✓	×	×	✓	✓	**GS**
LAM [14]	×	✓	×	×	✓	✓	**GS**
GPHM [38]	×	✓	×	×	✓	×	**GS**
PAV [3]	✓	✓	×	×	✓	×	**N**
TimeWalker (ours)	✓	✓	✓	✓	✓	✓	**GS**

and mobile deployment [6]. These models leverage differentiable rendering to refine parameters and minimize overfitting. Recent enhancements introduce explicit structures to boost rendering performance and training efficiency: InstantNGP [27] adopts a multi-resolution hashgrid to streamline scene feature storage and expedite training. 3DGS [19] uses explicit Gaussian Splatting for rendering, achieving fast inference rates without network reliance. Gaussian Surfels [9], refines Gaussian kernels for depth inconsistency problems rooted in 3DGS, results in high-quality static mesh reconstruction and realistic rendering under sparse view conditions. Our work extends this representation to dynamic human head modeling, enabling effective dynamic avatar animation.

3 TimeWalker

Our work targets to construct a 3D head avatar of a person towards a lifelong scale, as opposed to current trends that reconstruct and animate a person at the momentary level.

To address the challenges this new task brings, ***we introduce a novel neural parametric model (Fig. 2) that captures an average representation of a person's identity in canonical space, and extends to moment-specific head attributes through a set of Neural Head Bases.*** In Sec. 3.1, we briefly overview 2D Gaussian representation [9], which forms the foundational representation of our pipeline, ensuring high-fidelity rendering and dense meshing for the base frame. In Sec. 3.2, we detail how our Personalized Neural Parametric Model constructs a personalized neural feature space via a Linear Combination Space formulation. The model leverages a Linear Combination Space formulation to handle the complexities introduced by different lifestages through Neural Head Basis, while maintaining a shared canonical representation. Then, we explore the geometric representation behind Personalized Neural Parametric Model in Sec. 3.3– how our DNA-2DGS, a dynamic extension of 2DGS, fosters rendering and drives the dense mesh with different motion signals. In Sec. 3.4, we introduce training process,

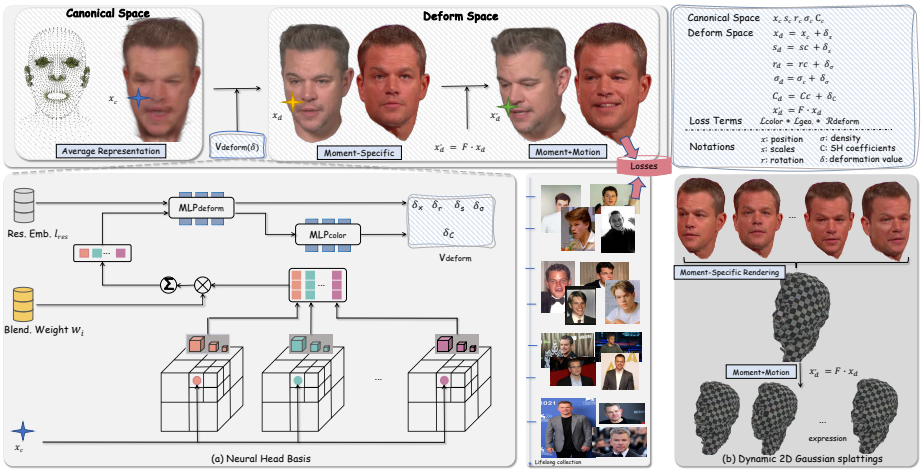

Fig. 2: Method Overview. TimeWalker constructs a lifelong scale 3D avatar from unstructured photo collections spanning years, maintaining realism and animation fidelity. The model is rooted in the principle of linear combination space, and innovated into an interpretable, scalable, and steerable neural personalized feature space with two key components: Dynamic Neural Basis-Blending Module (Dynamo) and Dynamic 2D Gaussian Splatting (DNA-2DGS). Concretely, Gaussian Surfels initialized in canonical space using the FLAME template represent an individual's average head. Neural Head Basis deformations model moment-specific representations, while DNA-2DGS applies inverse warping operation driven by FLAME parameters to capture expressions and movements, generating multi-dimensional head avatars (*i.e., moment-motion* in deform space).

and discuss how we build a lifelong personalized space with disentangled control along different dimensions in Sec. 3.5.

3.1 Preliminaries

Gaussian Surfels (2DGS) extends Gaussian Splatting [19] via reducing one dimension and transforms Gaussian ellipsoids into Gaussian ellipses. Specifically, a scene is depicted by a set of unconstructed Gaussian kernels with attribute $\{\mathbf{x}_i, \mathbf{r}_i, \mathbf{s}_i, \sigma_i, \mathcal{C}_i\}_{i\in\mathcal{P}}$, where i is the index of each Gaussian kernel, and $\mathbf{x}_i \in \mathbb{R}^3$, $\mathbf{r}_i \in \mathbb{R}^4$, $\mathbf{s}_i \in \mathbb{R}^3$, $\sigma_i \in \mathbb{R}$ and $\mathcal{C}_i \in \mathbb{R}^k$ respectively denotes the center position/rotation/opacity/spherical harmonic coefficients of each Gaussian's kernel. 2D Gaussian kernel could be obtained by flatten the 3D Gaussian's rotation on z-axis. By blending all Gaussian kernels via depth-ordered rasterization, a pixel's color $\tilde{C}$/normal value $\tilde{N}$/depth value $\tilde{D}$ can be obtained. Unlike 3DGS uses volumetric Gaussians that can blur depth boundaries, especially in areas with complex or discontinuous surfaces, 2DGS are surface-conforming primitives that project directly onto image plane, ensuring better local depth and normal blending processes. We represent human head upon 2DGS. Notably, 2DGS, built for static reconstruction, would encounter a significant challenge in generating dynamic mesh sequences, due to time-consuming nature of post Poisson Meshing [18]. Thus, we surmount this limitation within framework designs.

3.2 Personalized Neural Parametric Model.

Neural Linear Combination Space. In the quest to pioneer lifelong personalized animation spaces, our primary objective is to enable precise and controlled animation of individuals on a comprehensive scale. Leveraging linear combination suits for this context, as it offers flexibility and universality in constructing compact representations through the combination of base vectors that encapsulate core variations. It provides theoretical foundation for constructing a wide range of shapes, textures, and expressions even from limited data. Besides, it simplifies optimization and computation, ensuring consistent and predictable results. This concept traces its origins to traditional 3D Head Morphable Models, where a head is modeled by linearly combining expression and shape parameters with their corresponding basis functions[1]:

$$\mathbf{M}(\alpha,\beta,\gamma) = \left(\bar{S} + \sum_{i=1}^{N_s} \alpha_i \mathbf{s}_i,\ \bar{T} + \sum_{i=1}^{N_t} \beta_i \mathbf{t}_i,\ \bar{E} + \sum_{i=1}^{N_e} \gamma_i \mathbf{e}_i\right) \tag{1}$$

Building upon this foundation, we extend the concept to a neural linear combination space, integrating multiple animation dimensions in an additive, learnable, and scalable manner, as formulated in Eq 5. Particularly, as illustrated in Fig. 2, to characterize the average head representation of individuals, we initialize a set of 2DGS in canonical space based on FLAME template. With linear addition of the deformation value produced from Neural Head Basis (detailed in Sec. 3.2), the canonical 2DGS could be deformed from the average representation to a specific lifestage (*i.e.,* moment-specific representation in the Figure). This process facilitates the modeling of a neutral head at any life moment, and moreover, within our framework, this is achieved without the need for direct supervision. Sequentially, to enable motion-based modeling (*e.g.,* expression changes) and dynamic avatar animation, we further utilize inverse warping operation rooted from FLAME parameters (Sec. 3.3), to drive the moment-motion modeling. Along with analysis-by-synthesis training (Sec. 3.4), these designs allow creating multi-dimensional realistic head avatars in a breeze.

Neural Head Basis. With the assumption that any lifestage of one person can be linearly blended from his/her several key characteristic variations across lifestages, we now introduce how to capture and learn these variations within the linear neural feature space from the design of the Neural Head Basis, and how to obtain a *compact* set of the bases from our *Dynamo* module. Suppose the number of head basis is N, for a point $\mathbf{x}_c$ picked from the canonical space, its feature at a specific lifestage could be formed as a linear combination of N neural head basis $\mathcal{H}$: $\mathbf{f}(\mathbf{x}_c) = \sum_{i=1}^{N} \omega_i \mathcal{H}_i(\mathbf{x}_c)$, where ω_i is the learnable blending weight for each neural head basis. To store the features compactly, we utilize multi-resolution Hashgrid [21, 27], a hashmap-based cubic structure that enables the learnable features stored in a condensed form. When querying features for $\mathbf{x}_c$, the hashgrid looks up nearby features at various scales J, and cubic linear interpolation is

[1]Where $M(\alpha,\beta,\gamma)$ is the full model (shape, texture, expression), $\bar{S},\bar{T},\bar{E}$ are the mean shape, texture, and expression, $\mathbf{s}_i,\mathbf{t}_i,\mathbf{e}_i$ are the basis vectors, $\alpha_i,\beta_i,\gamma_i$ are the corresponding coefficients, and N_s,N_t,N_e are the number of components for shape, texture, and expression, respectively. Refer to the original paper for more details [2, 23].

applied to determine the final feature corresponding to the location. Thus, for each H_i, it is constructed by $\mathcal{H}_i = \text{C-LinearInterp}\ (\{\ h_j^i\}_{j=1}^J)$.

Dynamic Neural Basis-Blending Module (Dynamo). *How to set the number of head basis?* One intuitive way would be – assigning a fixed amount that is equivalent to the number of appearances in the character's data, with each basis independently learning the character's features for a specific lifestage. However, this idea is inefficient and redundant in the feature space, as a person's appearance evolves over their lifetime, their core characteristics that define their identity remain fundamentally consistent. Even with temporary alterations like heavy makeup or accessories, we can still recognize individuals by their underlying features. Besides, we expect the feature space could be interpretable, scalable, and controllable. Thus, *our goal for the neural basis is to capture these deeper characteristics rather than solely memorizing superficial appearances.* To this end, we introduce *Dynamo* to dynamically adjust the number of bases during the learning process of blending weight and hashgrids. We begin by initializing a set of learnable blending weights $\{\omega\}_{i=1}^N$ and grids $\mathcal{H}$, which align with the number of lifestages of the target person. During learning, if $\{Indicator(\omega_i^k < \kappa) = 1\}$ consistently reveals that a hashgrid's weight falls below lower a preset threshold κ across modeling multiple lifestages k at an iteration Q with sampled iteration interval q, this signals that this grid is not effectively learning meaningful features. In response, we deactivate that hashgrid. By the end of training process, this pruning strategy ensures that each remaining Neural Head Basis is efficiently capturing core, identity-defining features across lifestages. Additionally, it reduces memory requirements for storing these features, leading to a more efficient and scalable representation.

Residual Embedding. Using *Dynamo* outlined above, we obtain a collection of feature embeddings that compactly capture the subject's moment-specific attribute. As global compensation for each lifestage, we introduce a set of residual embedding $\mathbf{l_{res}}$, which are concatenated with the blended features $\mathbf{f}(\mathbf{x}_c)$. The concatenated features are then forwarded into a $\text{MLP}_{\text{deform}}$ to derive position $\mathbf{x}$, rotation $\mathbf{r}$, scale $\mathbf{s}$, and opacity σ deformations of the Gaussian kernel, as well as another feature vector which is subsequently passed through a $\text{MLP}_{\text{color}}$ to generate the SH coefficients $\mathcal{C}$ deformation:

$$\delta_{\mathbf{x}}, \delta_{\mathbf{r}}, \delta_{\mathbf{s}}, \delta_{\sigma}, \mathbf{f}_{\text{deform}}(\mathbf{x}_c) = \text{MLP}_{\text{deform}}(\mathbf{f}(\mathbf{x}_c), \mathbf{l_{res}}) \tag{2}$$

$$\delta_{\mathbf{C}} = \text{MLP}_{\text{color}}(\mathbf{f}_{\text{deform}}(\mathbf{x}_c)), \tag{3}$$

The network learns Gaussian attributes' deformations δ, which are then additively combined with the Gaussian average in canonical space. It yields the character's moment-specific features:

$$[\mathbf{x}_d, \mathbf{r}_d, \mathbf{s}_d, \sigma_d, \mathbf{C}_d] = [\mathbf{x}_c, \mathbf{r}_c, \mathbf{s}_c, \sigma_c, \mathbf{C}_c] + [\delta_{\mathbf{x}}, \delta_{\mathbf{r}}, \delta_{\mathbf{s}}, \delta_{\sigma}, \delta_{\mathbf{C}}]. \tag{4}$$

The whole combination process can be summarized below:

$$\begin{aligned} (\mathbf{x}, \mathbf{r}, \mathbf{s}, \sigma)_d &= (\mathbf{x}, \mathbf{r}, \mathbf{s}, \sigma)_c + \text{MLP}_{\text{d}}(\sum_{i=1}^{N} \omega_i \mathcal{H}_i(\mathbf{x}_c), \mathbf{l_r}) \\ \mathbf{C}_d &= \mathbf{C}_c + \text{MLP}_{\text{c}}(\text{MLP}_{\text{d}}(\sum_{i=1}^{N} \omega_i \mathcal{H}_i(\mathbf{x}_c), \mathbf{l_r})) \end{aligned} \tag{5}$$

3.3 Dynamic 2D Gaussian Splattings (DNA-2DGS)

The question now is – given a motion target/reference, how can we effectively drive the moment-specific representation to capture motion dynamics with precision, controllability and realism? This requires further deformation of the Gaussian Surfels to reflect various motion-related changes like expressions or shapes. Thus, we introduce *DNA-2DGS*, which tackles the problem from both rendering (upper area of Fig. 3) and meshing (lower area of Fig. 3) aspects.

Dynamic Gaussian Rendering. To realize animation, one intuitive strategy is to incorporate an additional MLP-based warping field. [2] While this method shows proficiency in handling deformations induced by various expressions, it falls short in capturing the entire range of human head motions, particularly around eyes. Moreover, disentangling the animation of appearance and expression remains challenging when both components of the warping field are designed similarly. In contrast, we integrate inverse warping operation inspired by INSTA [48]. During preprocessing, we acquire a tracked mesh M^{def} through FLAME fitting and a mean template M^{canon} defined within a canonical space, both sharing identical topology. Unlike INSTA that utilizes the warping field to inversely project points from deformed space back to canonical space, we define a deformation gradient $\boldsymbol{F} \in \mathbb{R}^{4\times4}$ in the form of a transformation matrix. It projects Gaussian Surfels $\mathbf{x}_d$ from moment-specific static space to dynamic motion space. Concretely, for each Gaussian Surfel $\mathbf{x}_d$, we employ a nearest triangle search algorithm to compute $\boldsymbol{F} = \boldsymbol{L}_{def} \cdot \Lambda^{-1} \cdot \boldsymbol{L}_{canon}^{-1}$, where $\boldsymbol{L}_{canon}$ and $\boldsymbol{L}_{def}$ is Frenet coordinate system frames, and Λ is diagonal matrix that takes scaling factor into account. Gaussian Surfels are further deformed $\mathbf{x}_d^{'} = \boldsymbol{F} \cdot \mathbf{x}_d$, guided towards specific movements. This process enables translation of expression/shape signals into animated outputs. The double-deformed 2DGS are then rasterized to produce final renderings.

Dynamic Gaussian Meshing. Although vanilla 2DGS [9] enables high-quality surface reconstructions, it is primarily suited for *static* settings, and the time-intensive nature of Poisson reconstruction process makes it impractical for mesh sequence reconstruction. These limit its applicability to dynamic head avatars. To address these challenges, we introduce *Defer-Warping*, an adapted Gaussian surface reconstruction strategy tailored specifically for dynamic head reconstruction. Specifically, unlike standard rendering & meshing processes that apply deformations before meshing, we delay the motion animation after the Poisson meshing process of moment-specific representation. This allows us to render results under moment-specific conditions customized to every single lifestage, effectively eliminating motion-induced artifacts (such as artifacts on mouth regions caused by talking) and generating appearance-specific static meshes. After Poisson meshing of the static result, the delayed motion animation generates dynamic mesh sequences by directly manipulating the vertices of reconstructed mesh. This process faithfully captures motion-driven deformations (as demonstrated in Fig. 7).

[2] In our experimental setup, we implemented this component as a naive dynamic version of Gaussian Surfels. Please refer to the results of Gaussian-Surfel++ in Appendix Sec. 7.1.

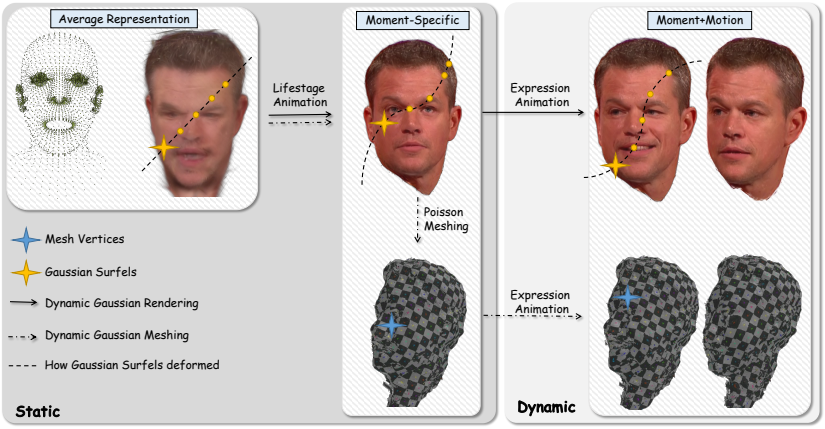

Fig. 3: Dynamic Gaussian Rendering & Meshing. For rendering, GS are first animated with lifestage and expression, respectively, followed by rasterization to produce high-fidelity results. For meshing, after deformation with lifestage, Poisson meshing is used to obtain a dense mesh. Then, expression animation is performed on the vertices.

3.4 Training

We apply end-to-end training manner that enables the simultaneous optimization of the explicit Gaussian surfels, multiple hashgrid, feature latent, and implicit deform MLP & color MLP. Before formal training, we introduce a warm-up phase where we suspend the optimization of the Neural Head Basis. This step aims to guide the Gaussian surfels in canonical space towards a mean representation. We use densify and pruning strategies to adaptively adjust the number of Gaussians. Our total loss for whole system $\mathcal{L}_{\text{total}}$ consists of three parts: (1) image-level supervision; (2)geometry level supervision; and (3) regulation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{geometry}} + \mathcal{L}_{\text{regulation}}, \tag{6}$$

Refer to Appendix Sec. 4 for more details.

3.5 Building a Life Long Personalized Space

How does TimeWalker construct a lifelong personalized space? The Gaussian Surfels in canonical space that characterize individual average representation, are additively combined with moment-specific head attribute representations driven from a set of Neural Head Basis, spanning a moment-specific head avatar. For each moment-specific avatar, we further warp Gaussian Surfels via expression or shape signals to get motion-specific head performing. By separating moment-specific deformation from motion-specific warping, we can decouple the driving of the head in multiple dimensions - lifestage, expression, shape, novel view, and *etc*, - constructing a comprehensive, steerable personalized space. Refer to Appendix Sec. 3 for details.

What are the core benefits of learned space? As life-long modeling inherently suffers from rare/long-tail data issues that certain stages have missing observations (like views,

expression), the personalized space can effectively mitigate the task's intrinsic data sparsity by learning a basic, common prior to represent a person's identity across stages. Thus, it can help solve several challenges posed by sparse data: **1**) multi-scale animation (Fig. 1), **2**) steerable reenactment, **3**) detailed mesh reconstruction (Fig 7), and **4**) ID-preserving rendering across ages and camera views (Appendix Sec. 6). With these, our method could improve downstream tasks' quality. See Sec. 4 and Appendix for details.

4 Experiments

Due to space limits, please refer to Appendix Sec. 6 for more internal experiments like additional ablations, personalized space, large-viewpoint rendering, lighting control and lifestage animation, and Appendix Sec. 7 for comparison experiments with sota multi-view reconstruction methods, generative methods, and one-shot head avatar methods. In main paper, we unfold our method with key comparisons with sota, reenactments, mesh comparison and key ablations.

Table 2: Quantitative Evaluation on TimeWalker-1.0. We evaluate our method using two protocols. The left half of columns shows *#Protocol-1* (1 vs 1), and the right half shows *#Protocol-2* (1 vs N). Pink is best, orange is second best.

	1 vs 1 (Protocol-1)				**1 vs N (Protocol-2)**				
Metric	*INSTA* [48]	*INSTA++*	*Flash Avatar* [36]	*Ours*	*GS* [9]	*GS++*	*INSTA* [48]	*Flash Avatar* [36]	*Ours*
PSNR↑	20.68	26.39	22.14	27.28	26.98	27.61	25.47	24.90	27.28
SSIM↑	0.697	0.879	0.771	0.949	0.950	0.948	0.860	0.848	0.949
LPIPS↓	0.299	0.139	0.267	0.071	0.141	0.134	0.170	0.165	0.071
ID Score↑	51.59	85.37	66.97	93.71	82.73	76.98	87.96	84.17	93.71

4.1 Datasets and Evaluation Protocol

In our experiments, if not specified, we select 5 representative identities from our TimeWalker-1.0 dataset, each comprising 8-13 lifestages with a total of 8000-20000 frames, with resized to 512x512. It is important to note that the number of frames for each lifestage ranges from 500 to 3000, generally much less than the frames in other SOTA personalized head avatar methods, which increases the challenge of this task. Refer to Appendix Sec. 4 for implementation details and Sec. 5 for dataset details.

We address the uneven frame distribution among appearances by allocating the last 10% of frames from each appearance as the test set and the remaining frames as the training set. We conduct experiments with two protocals – In 1) ***#Protocol-1 (1 vs 1 Comparison)*** We train one model for one identity with multiple lifestages. In 2) ***#Protocol-2 (1 vs N Comparison)*** Our model maintains the same pipeline setup as *#Protocol-1*, but for the baseline models, we train one model for each lifestage, allowing multiple models for one identity. This allows us to compare the results of our pipeline using a single model against the baseline using multiple models. We utilize three metrics, PSNR, SSIM, LPIPS, to assess the quality of individual generated frames.

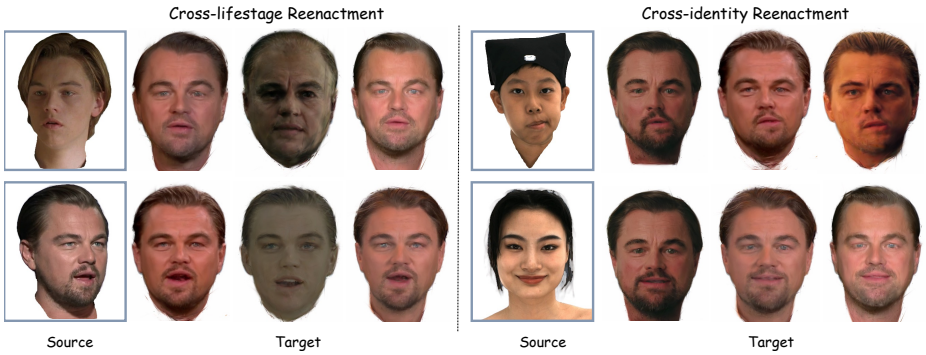

Fig. 4: Reenactment. We include cross-lifestage reenactment (self-reenacment) with TimeWalker-1.0 and cross-identity reenactment with RenderMe-360 [28], in Leonardo personalized space.

4.2 Comparisons with State-of-the-Arts

Baseline. We compare TimeWalker with **8** representative baselines: **1) NeRF-based methods** (INSTA [48]); **2) 3DGS-based methods** (FlashAvatar [36], Gaussian-Surfels [9], and its extension Gaussian-Surfels++); **3) generative-based methods** (GANAvatar [17]), and large-scale one-shot head avatar methods (GAGAvatar [7], GPHM [38]), and **4) personalized avatar over time scale** (a re-implementation of *PAV* [3], denoted as **INSTA++**), ***Refer to Appendix Sec.*** 7 ***for the detailed description of these methods and comparisons on GANAvatar [17], GAGAvatar [7] and GPHM [38].***
Results. Quantitatively, as shown in Tab. 2, our method achieves the best results in the *#Protocol-1*, and best or second best results in the *#Protocol-2* on our TimeWalker-1.0 dataset. Our method significantly outperforms other baselines in both protocols when assessed using TimeWalker-1.0. Although INSTA++ shows considerable enhancements post multi-lifestage adaptation, there remains a noticeable gap compared to our approach, underscoring the necessity of meticulously designing our pipeline for creating high-fidelity lifestage replicas and highlighting the effectiveness of our design in handling multi-stage data, rather than simply relying on dataset memorization. In the 1 vs N Comparison, our results remain competitive, with lower LPIPS values. Our method consistently achieves stronger identity preservation in both 1 vs 1 and 1 vs N Comparison settings. This demonstrates that the edited attributes are effectively disentangled from identity information. The results therefore provide quantitative support for both our disentanglement design and our ID consistency claim.

4.3 Animation

Reenactment. Fig. 4 shows expression animation with two types of reenactments: 1) The *cross-lifestage reenactment* on the left showcases how head avatars from different lifestages can consistently perform the same expression, animated from a source avatar belonging to a different time period. 2) The *cross-identity reenactment* on the right shows Leonardo in different lifestages are driven by unseen novel expression from RenderMe-360 [28]. The rendering result shows that multiple head avatars generated by the same personal space from TimeWalker are able to extrapolate novel expressions.

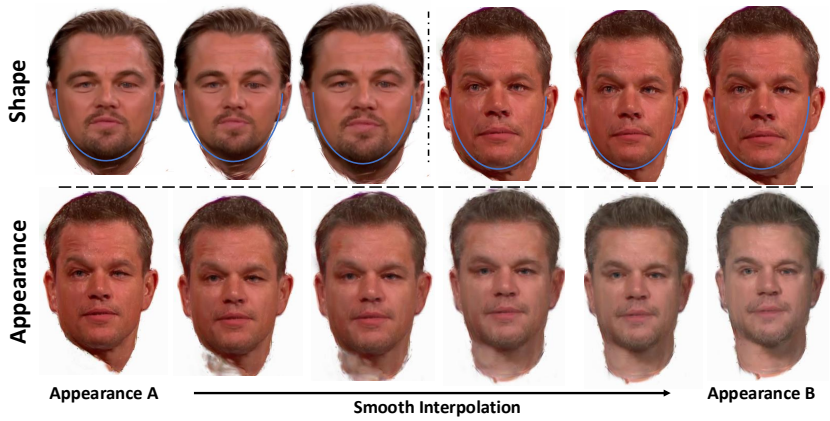

Fig. 5: Disentangled Shape/Appearance Transition. Blue contour lines as standard reference to show the changes of facial shape.

Shape and Appearance Transitions. As illustrated in Fig. 5, our framework achieves disentangled control over shape and appearance. The upper panel demonstrates shape-only morphing, where identity and expression remain invariant while morphology evolves. The bottom panel showcases age-dependent appearance interpolation, modeling the temporal evolution of a single identity across life stages.

Despite the sparsity of longitudinal observations, our personalized latent space learns a shared representation that maintains identity consistency across large temporal gaps. By linearly interpolating latent coefficients, we synthesize high-fidelity intermediate appearances, representing the aging process as a continuous trajectory rather than discrete jumps. This is further validated in Fig. 6, where we outperform sota baseline GAGAvatar in identity consistency across ages. These results demonstrate that our framework effectively models robust identity representations across ages from limited monocular data, without the need for large-scale pretraining.

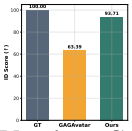

Fig. 6: Identity Consistency Cross Lifestages. We compared the id score calculated through ArcFace [10] between our method and GAGAvatar.

4.4 Mesh Comparison

To evaluate the effectiveness of our mesh reconstruction pipeline, we compare our geometric outcomes with meshes extracted from various pipelines: Colmap [33], INSTA++, and Gaussian Surfels [9]. Data from a single lifestage is provided to Colmap and Gaussian Surfels, tailored for static scene reconstruction. Conversely, our pipeline and INSTA++, designed for handling multiple lifestages, receive all data of an individual.

As demonstrated in Fig. 7, our pipeline successfully reconstructs the head mesh with precise shape, while the original Gaussian Surfel and Colmap fail to derive a meaningful mesh. This disparity is partly attributed to the restricted data from a single appearance and camera pose. Both pipelines struggle to recreate a plausible head mesh without prior knowledge of human head anatomy, given that monocular reconstruction

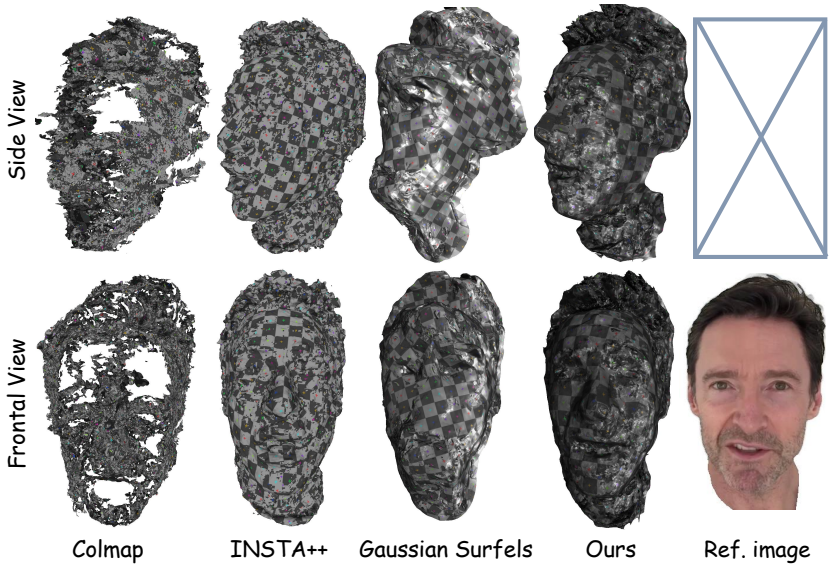

Fig. 7: Static Mesh Comparison. We visualize and compare static mesh reconstruction results from Gaussian Surfels [9] and Ours (adaptive version of GS with DNA-2DGS). We render the meshes in both frontal and side views using Blender software under identical rendering conditions. Note that to ensure a fair comparison, for INSTA++ and Gaussian Surfel, we also provide FLAME-guided initialization. Despite these consistent settings, rendering results exhibit significant differences due to the diverse topologies of the meshes (Zoom in for details).

Table 3: Ablation on Hashgrid.

	w/o Dynamo	1 Hashgrid	All Hashgrid	20 Hashgrid	Ours
PSNR↑	21.69	24.84	26.86	26.94	27.20
SSIM↑	0.767	0.890	0.938	0.935	0.941
LPIPS↓	0.197	0.119	0.078	0.077	0.077

Table 4: Multi- vs Single-lifestage.

Data	PSNR ↑	SSIM ↑	LPIPS ↓
One stage data	23.69	0.959	0.072
Full stages (Ours)	27.22	0.955	0.070

is inherently an ill-posed problem. In contrast, our mesh reconstruction, utilizing shared representations learned data from multiple lifestages, achieves superior quality with more accurate surface details. Although INSTA++ can also reconstruct a plausible head mesh, it exhibits more noises. This comparison underscores the efficacy of our DNA-2DGS component in reconstructing head meshes from unstructured photo collections.

4.5 Ablation Studies

To validate the effectiveness of TimeWalker's components, we conduct ablation experiments on 3 individuals with respective 9, 10 and 13 lifestages. We keep other settings unchanged except the ablation term. We show results about our Dynamo design here and ***Refer to Appendix Sec.*** 6.1 ***for more detailed ablations.***

Dynamo. We conduct ablation on Dynamo module in the Neural Head Basis, a key contribution that enables our pipeline to preserve the moment-specific attributes of individuals. Two protocols are used: 1) removing Dynamo and replacing it with the position

$\mathbf{x}_c$ of each Gaussian surfels; 2) predefining and fixing hashgrid numbers (either 1, 20 or matching the number of life stages). The quantitative result in Tab. 3 shows the necessity of Dynamo with multiple hashgrids, as reducing the hashgrid number or removing Dynamo leads to performance dropout. Interestingly, our pipeline with adaptive hashgrid number performs slightly better than the setting with all hashgrid and a large number of hashgrid(20) that exceeds the number of lifestages. This improvement demonstrates the effectiveness of our adaptation mechanism, which fosters the framework to learn compact representations of both shared and unique characteristics across different life stages, confirming that the dynamic setting achieves the best performance by balancing representational capacity and generalization.

How Temporal Diversity Helps Single Stage. To further evaluate the benefit of our multi-lifestage supervision, we conduct a controlled study comparing two training paradigms: Single- and Multi-Lifestage training. As shown in Tab 4, our multi-lifestage training yields a substantial improvement in rendering quality. The core benefit of this paradigm relies on the learned prior knowledge afforded by our Dynamo and Residual Embedding modules.

5 Limitations and Conclusion

Limitations. Our pipeline still has several limitations. 1) It anchors on FLAME priors; therefore, exaggerated expressions and unseen inner-mouth regions, such as the tongue, may lead to artifacts. 2) Even with large-scale data, collecting and annotating heads across ages in the wild inevitably introduces noise and imbalance, such as long-tailed age/illumination/camera view distributions, imperfect matting around head boundaries, and less accurate FLAME fitting compared with standardized captures. These factors make age interpolation, lighting consistency, and rendering from extreme views more challenging. 3) TimeWalker provides a preliminary exploration on lifelong modeling, while not fully spanning an individual's entire life from infancy to old age. This is mainly due to data constraints, and the methodological focus of our framework is reconstruction within a personalized space. As a result, it can interpolate across life stages based on the learned personalized space, but may not fully hallucinate unseen ages. On the other hand, generative methods also face challenges such as preserving identity consistency and geometric fidelity. As demonstrated and discussed in our experiments, neither reconstruction-based nor generative approaches alone can fully solve lifelong modeling. While we showcase preliminary ways to combine the two paradigms in Appendix, achieving a seamless integration remains an important direction for future work.

Conclusion. We propose to construct human head avatars at a life-long scale. For this new task, we present TimeWalker, a baseline solution for constructing personalized spaces that maintain long-term identity consistency while enabling explicit, multi-scale animation control. Rooted in the additive combination principles of classic 3DMM, our approach innovates through a neural feature-based design with two key components: a Dynamic Neural Basis-Blending model to represent head variations in a compact manner and a Dynamic 2D GS module to construct dynamic dense head meshes. Experiments show the effectiveness of TimeWalker and reveal the current status of lifelong head avatar development, opening opportunities for future research.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
3. Caliskan, A., Kicanaoglu, B., Kim, H.: Pav: Personalized head avatar from unstructured video collection. In: ECCV (2024)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: CVPR (2022)
5. Chaudhuri, B., Vesdapunt, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: ECCV (2020)
6. Chen, Z., Funkhouser, T., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In: CVPR (2023)
7. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. In: NeurIPS (2024)
8. Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: ICCV (2011)
9. Dai, P., Xu, J., Xie, W., Liu, X., Wang, H., Xu, W.: High-quality surface reconstruction using gaussian surfels. In: SIGGRAPH (2024)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
11. Deng, Y., Wang, D., Wang, B.: Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In: ECCV (2024)
12. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021)
13. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Learning neural parametric head models. In: CVPR (2023)
14. He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., Bo, L.: Lam: Large avatar model for one-shot animatable gaussian head. In: SIGGRAPH (2025)
15. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: CVPR (2022)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
17. Kabadayi, B., Zielonka, W., Bhatnagar, B.L., Pons-Moll, G., Thies, J.: Gan-avatar: Controllable personalized gan-based human head avatar. In: 3DV (2024)
18. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. TOG (2013)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023)
20. Kirschstein, T., Giebenhain, S., Nießner, M.: Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In: CVPR (2024)
21. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: Nersemble: Multi-view radiance field reconstruction of human heads. TOG (2023)
22. Kroger, J.: Identity development: Adolescence through adulthood. Sage publications (2006)
23. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. TOG (2017)
24. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot 3d neural head avatar. NeurIPS (2024)
25. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PLOS (2018)

26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
27. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. TOG (2022)
28. Pan, D., Zhuo, L., Piao, J., Luo, H., Cheng, W., Wang, Y., Fan, S., Liu, S., Yang, L., Dai, B., et al.: Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. NeurIPS (2024)
29. 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. Ieee (2009)
30. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: CVPR (2024)
31. Rai, A., Gupta, H., Pandey, A., Vicente-Carrasco, F., Takagi, S.J., Aubel, A., Kim, D., Prakash, A., la Torre, F.D.: Towards realistic generative 3d face models. In: WACV (2024)
32. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
33. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
34. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In: CVPR (2024)
35. Waterman, A.S.: Identity development from adolescence to adulthood: an extension of theory and a review of research. Developmental Psychology (1982)
36. Xiang, J., Gao, X., Guo, Y., Zhang, J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In: CVPR (2024)
37. Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., Zheng, Z., Liu, Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In: CVPR (2024)
38. Xu, Y., Su, Z., Wu, Q., Liu, Y.: Gphm: Gaussian parametric head model for monocular head avatar reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
39. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: CVPR (2020)
40. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., Zhang, C., Yin, X., Ma, Z., Zhao, Z.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. In: ICLR (2024)
41. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: CVPR (2021)
42. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: ICCV (2021)
43. Yu, Z., Yoon, J.S., Lee, I.K., Venkatesh, P., Park, J., Yu, J., Park, H.S.: Humbi: A large multiview dataset of human body expressions. In: CVPR (2020)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
45. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: CVPR (2022)
46. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: CVPR (2023)
47. Zhu, X., Yu, C., Huang, D., Lei, Z., Wang, H., Li, S.Z.: Beyond 3dmm: Learning to capture high-fidelity 3d face shape. TPAMI (2023)
48. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: CVPR (2023)