

GenLCA: 3D Diffusion for Full-Body Avatars from In-the-Wild Videos

Yiqian Wu^{2,1} , Rawal Khirondkar¹, Egor Zakharov¹ ,
Timur Bagautdinov¹, Lei Xiao¹, Zhaoen Su¹,
Shunsuke Saito¹ , Xiaogang Jin² , and Junxuan Li¹ 

¹ Codec Avatars Lab, Meta, USA

² State Key Lab of CAD&CG, Zhejiang University, China

Project Page: <https://onethousandwu.com/GenLCA-Page>



Fig. 1: GenLCA is a diffusion-based generative model for generating and editing full-body 3D Gaussian avatars from text and image inputs. (A) **Generation**. GenLCA generates avatars that are visually realistic and consistent with both the identity in the input face image and the semantic descriptions in the input text, while supporting high-fidelity facial and full-body animations. We present zoomed-in and animated face results. (B) **Editing**. By leveraging text, RGB images, or scribbles as control signals, GenLCA enables seamless multi-modal editing of the generated avatars. (C) Diverse 3D avatars generated by GenLCA from text inputs.

Abstract. We present **GenLCA**, a diffusion-based generative model for generating and editing photorealistic full-body avatars from text and image inputs. The generated avatars are faithful to the inputs, while supporting high-fidelity facial and full-body animations. The core idea is a novel paradigm that enables training a full-body 3D diffusion model from partially observable 2D data, allowing the training dataset to scale to millions of real-world videos. This scalability contributes to the superior photorealism and generalizability of GenLCA. Specifically, we scale up the dataset by repurposing a pretrained feed-forward avatar reconstruction model as an animatable 3D tokenizer, which encodes unstructured video frames into structured 3D tokens. However, most real-world videos only provide partial observations of body parts, resulting in excessive blurring or transparency artifacts in the 3D tokens. To address this, we propose a novel visibility-aware diffusion training strategy that replaces invalid regions with learnable tokens and computes losses only over valid regions. We then train a flow-based diffusion model on the token dataset, inherently maintaining the photorealism and animatability provided by the pretrained avatar reconstruction model. Our approach effectively enables the use of large-scale real-world video data to train a diffusion model natively in 3D. We demonstrate the efficacy of our method through diverse and high-fidelity generation and editing results, outperforming existing solutions by a large margin.

Keywords: Digital human · 3D diffusion model · Generative 3D human

1 Introduction

As our world becomes increasingly digital, 3D photorealistic avatars hold the key to more natural and expressive virtual experiences. Yet their creation usually requires multi-view images or long monocular videos [41, 60, 66, 67], which remain inaccessible to most users. Recent advances in generative models, particularly diffusion models, have shown the ability to create high-quality 3D content from user-friendly inputs, such as text or incomplete images. In this paper, our goal is to investigate diffusion models as an efficient and scalable solution for 3D avatar generation.

Diffusion model training benefits from both the high quality and diversity of the training data. For digital humans, a common solution is to use synthesized data [68, 70, 83, 90], but its domain gap from real-world humans often compromises the quality and photorealism of the resulting models [68, 69, 83]. To achieve animatability and higher realism, calibrated and synchronized multi-view capture datasets are typically employed [9, 32, 80]. However, they cover only a few thousand subjects, which impairs model generalization and diversity [7, 78, 86]. As demonstrate in large-scale video diffusion models [8, 12, 20], monocular video data provides an ample resource for learning realistic appearance and motion at the 2D level. But training a 3D diffusion model generally requires accurate 3D assets, which remains challenging for monocular videos.

We propose **Generative Large 3D Codec Avatar Model (GenLCA)**, a multi-modal 3D diffusion model for generating and editing full-body avatars from text and image inputs, while enabling photorealistic appearance and animation. GenLCA trains a full-body 3D diffusion model from *partially observable, million-scale 2D data*, enabled by two key components: a feedforward avatar reconstruction network serving as a *tokenizer*, and a visibility-aware training strategy to mitigate artifacts in 3D tokens caused by imperfect video frames.

The avatar reconstruction network takes multiple body and face images of a single subject as input. These images are encoded into 3D tokens, which are then decoded by the reconstruction network into an animatable 3D Gaussian avatar. By applying this tokenizer to large-scale video collections, we construct a 3D token dataset for **~1.1 million identities**. These tokens are further compressed into compact latents using a compressor to facilitate efficient model training. However, due to the partial observability of monocular video frames and the reconstruction model’s inherent limitations in hallucinating unobserved regions, the 3D tokens for unobserved areas are often blurry or incomplete. *Directly training a generative model on such imperfect supervision leads to noticeable quality degradation.* To ensure data quality and fidelity, we introduce a novel visibility-aware training strategy. Specifically, we compute a mask for each identity based on the visibility of their tokens with respect to the input video frames. Tokens corresponding to unobserved areas are replaced with learnable placeholder features. Furthermore, we apply a masked loss function only to valid regions. This approach limits supervision to observable and reliable 3D regions, thereby mitigating the influence of corrupted information.

We then train a flow-based diffusion model on the compressed latent representations, inheriting the photorealism and animatability of the avatar reconstruction model. To support multi-modal generation and editing, we incorporate three types of modalities as conditional inputs: text, segmented body part images (e.g., hair, face, upper clothing, etc.), and scribble images.

To the best of our knowledge, GenLCA is the first method to train a 3D diffusion model at scale using real-world video data. Our method substantially relaxes the data requirements for generative 3D human modeling, **demonstrating the potential to scale 3D human datasets to a magnitude comparable to that of existing 2D datasets**. Despite being trained on incomplete observations, GenLCA effectively captures semantic relationships between visible tokens, enabling the generation of full-body, animatable avatars, and outperforming SOTAs by a significant margin.

In summary, we make the following contributions:

- We propose a 3D diffusion model to produce and edit high-quality, realistic, and animatable 3D full-body avatars from image and text.
- We employ a reconstruction model to tokenize unstructured images into 3D tokens, and introduce a novel training strategy that leverages partially observable inputs. Our framework facilitates constructing 3D realistic human datasets at a large scale and paves the way for generalized 3D human generative models.

2 Related Work

2.1 3D human reconstruction

Existing work has explored a variety of 3D representations to achieve high fidelity zero-shot avatar reconstruction and re-animation, including parametric meshes [17, 39, 52], neural radiance fields (NeRF) [3, 49], and 3D Gaussian splatting (3DGS) [33, 55, 61, 67, 91]. By conditioning these 3D representations on controllable parameters such as pose and lighting, dynamic details can be integrated into the avatar [16, 18, 41, 67, 73], enabling the creation of more expressive models. However, these approaches require extensive camera coverage [9, 32, 47, 80], disentangled attribute supervision, and sufficient capture of fine-grained details to achieve high-quality results. Feedforward reconstruction models focus on training 3D human priors [11, 37, 56, 57, 81, 90] to directly regress 3D human representations from 2D inputs in a single forward pass. **However, these reconstruction models fail to produce high-quality results for unobserved regions and lack editability.**

Unlike zero-shot or feed-forward reconstruction methods, GenLCA requires minimal input during inference and supports both generation and editing.

2.2 Zero-shot and one-shot 3D human creation

DreamFusion [53] introduces Score Distillation Sampling (SDS) for generating 3D content using guidance from 2D diffusion models [23, 59]. SDS-based 3D human creation utilizes 3D representations such as meshes [28, 29, 42], neural radiance fields [71, 85], and 3D Gaussian splatting (3DGS) [4, 31, 44, 89], followed by multi-step SDS optimization. However, SDS optimizes a single 3D content using a 2D diffusion prior, leading to ambiguities that cause over-saturation and unrealistic styles. Another line of research aims to directly reconstruct 2D avatars from multi-view data hallucinated by 2D diffusion models [5, 30, 38, 46, 54, 65, 75, 76] or video diffusion models [34, 45, 88]. To ensure geometric alignment and reduce view inconsistency, these methods either condition the diffusion model on 3D control signals [5, 34, 35, 54, 65] or incorporate reconstruction into the denoising process [30, 76, 77]. Nevertheless, inherent view inconsistencies still result in blurriness in the final outputs.

Compared to the aforementioned generation methods that rely on 2D diffusion models, **our GenLCA operates natively in 3D and is trained on real-world data, thereby inherently avoiding issues of blurriness and low realism.**

2.3 Generative 3D human model

Inspired by EG3D [6], which employs GANs [19] to generate implicit neural fields and uses 2D images for supervision, numerous works [1, 13, 25, 48, 72, 74, 79] have extended it to full-body human modeling. However, directly modeling 3D implicit distribution from single-view 2D collections introduces ambiguities and

Table 1: Comparison of training datasets for 3D human diffusion models. Our dataset contains the largest number of identities and the most realistic data among existing methods. “*” indicates that the corresponding model is trained on each sub-dataset individually, rather than on a mixture of them.

Dataset	# total IDs	# synthetic data	# captured data	# in-the-wild data
StructLDM [27]	1.8K	0.8K*	0.5K*	0.5K*
HumanLiff [26]	1.1K	1K*	0.1K*	0
Rodin [68]	100K	100K	0	0
RodinHD [83]	46K	46K	0	0
SimAvatar [40]	20K	20K	0	0
TeRA [69]	70K	70K	0	0
SIGMAN [78]	110K	100K	10K	0
GenLCA (Ours)	1,117K	0	4K	1,113K

leads to quality degradation. Diffusion models [23, 59] have been extended to 3D human modeling [7, 27, 40, 68, 69, 78, 83]. Since diffusion model training requires accurate 3D assets for each training sample, a typical training pipeline involves an encoding process. This process either trains an auto-encoder or performs zero-shot optimization to encode multi-view images into structured representations, such as feature planes [26, 68, 83], structured UV latents [14, 27, 63, 64, 69, 78, 86], or 3D primitives [7, 84]. However, since their encoding process is designed to accurately represent each identity in existing datasets, these methods require extensive camera coverage to achieve optimal performance and therefore cannot be generalized to in-the-wild data. Consequently, their training sources are limited to small-scale captured datasets [26, 27, 63, 64, 78, 86] or unrealistic synthetic datasets [7, 26, 40, 68, 69, 78, 83], as shown in Tab. 1.

In summary, there is no 3D avatar generator that performs native 3D generation while effectively utilizing in-the-wild data. To address this limitation, we propose to leverage a large-scale reconstruction model to extract training samples from in-the-wild videos. This approach substantially expands the available dataset for a more generalized 3D human diffusion model.

3 Methodology

In this section, we first introduce the 3D avatar tokenizer, which encodes images into 3D tokens (Sec. 3.1). Next, we present the overall architecture and training strategy (Sec. 3.2). GenLCA first employs a compressor to compress the 3D tokens into compact latents. To mitigate the influence of corrupted information in these 3D tokens, we propose a visibility-aware training strategy that utilizes a visibility mask to restrict training to observable and reliable 3D information. We then detail the generative model architecture and describe our conditional inputs. Finally, we explain the computation of the visibility mask (Sec. 3.3).

3.1 3D avatar tokenizer

To obtain a structured and unified 3D avatar representation from 2D images, we propose leveraging a pre-trained reconstruction model, **LCA** [2], as the *to-*

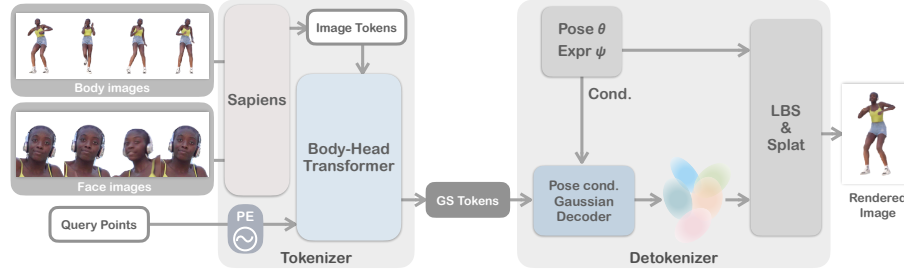


Fig. 2: The architecture of the reconstruction model. The transformer takes image tokens and query point embeddings as inputs, and outputs GS tokens. The GS tokens are decoded to get dynamic GS attributes. The resulting Gaussian splats are rendered using LBS to obtain the final renderings.

kenizer. LCA is a reconstruction-based model designed to produce high-fidelity 3D tokens for input frames. Further details are provided in Appendix B of the supplementary material.

We illustrate the high-level architecture of LCA in Fig. 2. The model can be divided into two components: the tokenizer and the detokenizer. The tokenizer encodes multiple input images into 3D tokens, while the detokenizer interprets these tokens as Gaussian splats. Specifically, the inputs to the tokenizer consist of multiple body and face images. Sapiens [36] then extracts image tokens from the input images. Subsequently, N query points are sampled on a template body mesh, denoted as $\mathbf{X} = \{x_i \in \mathbb{R}^3\}$, which is fixed and shared across all identities. Images tokens and point embeddings are fed into a transformer, which produces N GS tokens $\mathbf{T} = \{t_i \in \mathbb{R}^{D_T}\}$. Each GS token t_i is mapped to its corresponding query point x_i . During detokenization, each GS token is decoded by a lightweight MLP-based decoder into eight Gaussian splats, resulting in a total of $8N$ splats per identity. The decoder is further conditioned on pose and expression parameters to enable dynamic GS features. Finally, LBS and splatting are used to render the final image. During tokenization, we use four body images and four face images as input, which are extracted from video data as detailed in Sec. 4.1. The number of query points is set as $N = 8,192$, while the dimension of token is $D_T = 1,024$.

3.2 GenLCA

GenLCA is a flow-based diffusion model trained with the rectified flow objective [43]. In this section, we detail the training strategy and model architecture of GenLCA.

Token compressor The extracted GS tokens $\mathbf{T} \in \mathbb{R}^{N \times D_T}$ are high-dimensional representations that are scattered in a loosely structured space. To obtain a more compact space for generative model training, we use a compressor to encode the GS tokens into latents.

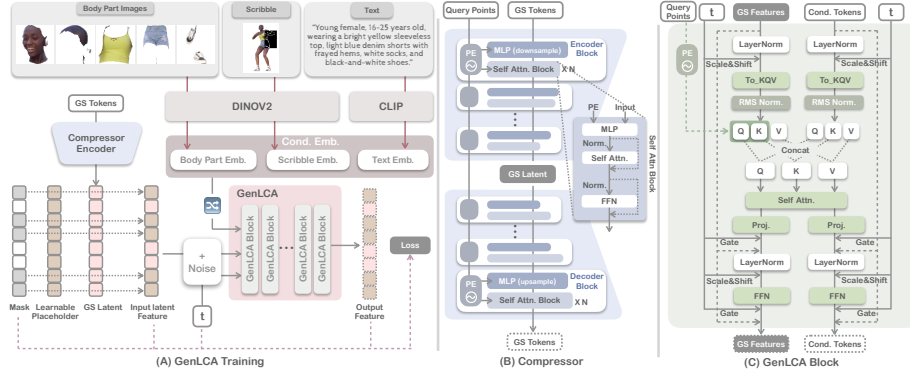


Fig. 3: (A) Training pipeline of GenLCA. During training, the high dimensional GS tokens are first encoded into compact GS latent by the compressor encoder. For conditional inputs, we use CLIP [58] to extract text embeddings and DINOv2 [50] to extract scribble and body part embeddings. To prevent the training process from being affected by corrupted information, we replace invalid regions (as indicated by the visibility mask) with learnable placeholder features and employ a masked loss. **(B) Detailed architecture of the compressor.** The compressor’s encoder and decoder consist of MLPs for downsampling or upsampling, combined with self-attention blocks for feature fusion. Positional encoding is applied within each block. **(C) Detailed architecture of the GenLCA block.** We adapt the MMDiT block as the basic block of GenLCA. Each GenLCA block takes the query points, time step, latent features, and conditional features as inputs. Separate branches are used to process latent and conditional features, and positional encoding is added to the latent features.

The detailed architecture of the compressor is presented in Fig. 3 (B). Both the encoder and decoder are composed of multiple blocks, each containing a MLP for downsampling or upsampling, as well as self-attention blocks for feature fusion. Additionally, the same set of query points $\mathbf{X} \in \mathbb{R}^{N \times 3}$ used during tokenization are encoded to obtain positional embeddings. The compressor is trained with the following loss function:

$$\mathcal{L}_{\text{compressor}} = \lambda_1 \mathcal{L}_1(\mathcal{D}(\mathcal{E}(\mathbf{T}, \mathbf{X}), \mathbf{X}), \mathbf{T}) + \lambda_2 \mathcal{L}_{\text{KL}}, \quad (1)$$

where $\mathcal{D}$ and $\mathcal{E}$ denote the encoder and decoder, respectively. The $\mathcal{L}_1$ reconstruction loss is computed between the reconstructed tokens $\mathcal{D}(\mathcal{E}(\mathbf{T}, \mathbf{X}), \mathbf{X})$ and the ground truth tokens $\mathbf{T}$. The KL divergence loss $\mathcal{L}_{\text{KL}}$ is also included. λ_1 and λ_2 are the corresponding weights for the losses. The compressed latents $\mathbf{Z} = \mathcal{E}(\mathbf{T}, \mathbf{X}) \in \mathbb{R}^{N \times D_{\mathbf{Z}}}$ have the same number of tokens as $\mathbf{T}$, but with a lower dimension $D_{\mathbf{Z}} = 8$.

Visibility-aware training After training the compressor, GenLCA is trained in the latent space. However, due to the partial observability of monocular video frames and the inherent limitations of the LCA reconstruction model in hallucinating unobserved regions, the information for these areas is often blurry or

incomplete, as discussed in Sec. 3.3. Directly training a generative model on such imperfect supervision leads to noticeable quality degradation (discussed in Sec. 5.3). To ensure data quality and fidelity, we propose a novel visibility-aware training strategy, which utilizes a visibility mask (Sec. 3.3) to apply different training strategies to latents corresponding to observable and unobservable regions. As shown in Fig. 3 (A), we introduce learnable placeholder features that are shared among all identities. To mitigate the influence of invalid information, we replace invalid latent components with these placeholder features. While mask tokens in masked image modeling [21] denote intentionally removed patches for self-supervised reconstruction, our placeholders represent invalid latent components caused by partial observability and prevent them from contaminating generative training. Additionally, during loss computation, we use masked weighting to ensure that the loss is computed only over valid regions.

Model architecture The detailed architecture of GenLCA is illustrated in Fig. 3 (C). We adopt the double-stream MMDiT block [15] from Hunyuan [87]. In this design, the latent features and conditional tokens are processed by separate network branches (distinct branches are used for different modalities, this is omitted from the figure for clarity) to obtain query, key, and value features. These features are concatenated and used to perform attention. These attention outputs are split into latent and conditional components, processed by separate branches, and fed into the next block.

For modulation, we only use the time step. Additionally, we add point embeddings to the query and key features of the latents, as our latent features maintain a one-to-one correspondence with the associated query points.

Conditional inputs For conditional inputs, as illustrated in Fig. 3 (A), we utilize text descriptions, scribble images, and body part images. For text, we use CLIP [58] to extract text embeddings $\mathbf{C}_{\text{text}}$. For scribble images, we use DINOv2 [50] to extract scribble embeddings $\mathbf{C}_{\text{scribble}}$. Similarly, for body part images, we input five body part images into DINOv2 and concatenate the resulting embeddings to obtain the body part embeddings $\mathbf{C}_{\text{body}}$.

To enable flexible controllability, we design three types of input modalities: *text-only*, *image-only*, and *text-plus-image*. During training, one of these modalities is uniformly sampled as the conditional input. If the selected modalities involve images (*image-only* or *text-plus-image*), we further uniformly sample between scribble images and body parts as the image input.

3.3 Visibility mask

As previously discussed, the tokenizer exhibits limitations in hallucinating unobserved regions. Consequently, artifacts tend to appear in areas with limited image information (Fig. 4 (B)). As described in Sec. 3.2, we employ a visibility-aware training strategy to address this issue, which requires a visibility mask to

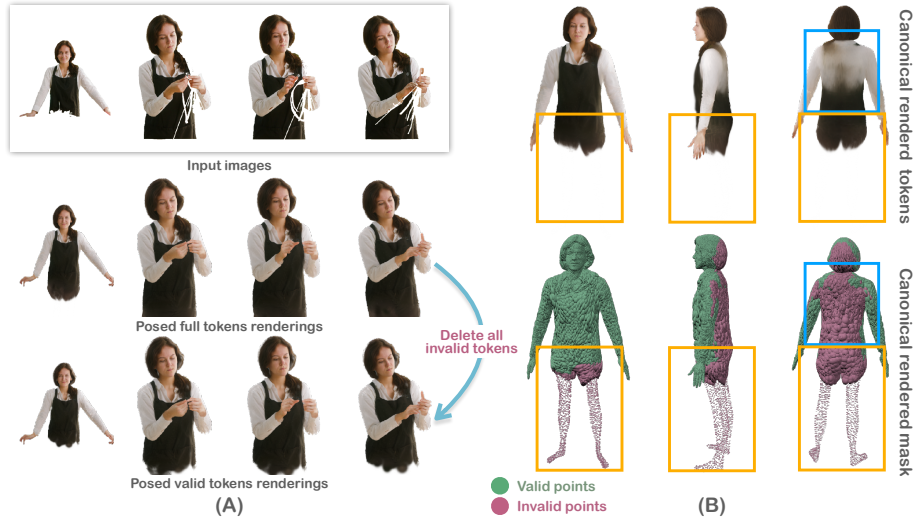


Fig. 4: (A) The tokens accurately reconstruct the visible regions. After filtering out all invalid tokens and retaining only the valid ones, the rendered results still achieve high-quality reconstruction. (B) We present the rendered tokens in canonical space (1st row), where blurry regions are highlighted with blue boxes and transparent regions with yellow boxes. The visibility mask (2nd row) separates the valid and invalid regions.

label invalid regions. In this section, we provide a detailed explanation of the visibility mask calculation.

As shown in Fig. 4 (B), when examining the complete set of tokens, we observe blurry back regions (highlighted by blue boxes) due to the input images being exclusively frontal views, and transparent lower body regions (highlighted by yellow boxes) resulting from the input images containing only upper body views. To obtain the visibility mask, we render the decoded Gaussian splats using the corresponding camera and body poses from the input body image and compute the gradients, which indicate each splat’s contribution to the rendered image. Splats with low contributions are considered to have low “visibility” relative to the input image. A token is defined as visible if at least two out of eight decoded splats are visible in at least one of the input views. This process produces the visibility mask, as illustrated in the second row of Fig. 4 (B). The corresponding Gaussian splats of filtered valid tokens are of high quality and appear realistic, whereas the invalid ones are typically blurry or even transparent.

4 Implementation Details

4.1 Training dataset

We construct the training dataset for GenLCA by encoding frames from monocular videos into structured 3D tokens using the tokenizer. Please refer to the supplementary material for visual examples.

Specifically, we reuse the video dataset employed for training LCA [2] to construct our training token dataset. The video dataset comprises:

- **In-the-wild data.** A total of 1,113,476 monocular, human-centric real-world videos are included to ensure diversity and broad generalization.
- **Captured data.** To provide cleaner data with comprehensive full-body coverage, the dataset additionally contains calibrated and synchronized multi-view videos of 2,737 identities recorded in a studio capture setup similar to [47]. Furthermore, 1,198 individuals are recorded using mobile phones, where participants perform a full-body rotation to ensure complete coverage from diverse viewpoints.

For each identity, we select frames with the largest differences in yaw angles as body input images, maximizing the coverage of observable body regions. Additionally, we randomly sample multiple frames from the video and crop the face region to serve as face input images. These images are processed by the tokenizer to obtain GS tokens $\mathbf{T}$. The final token dataset comprises **1,117,411** identities. For evaluation purposes, we sample 1,000 high-quality identities from the captured dataset to serve as our test set.

For each input image, we use Sapiens [36] for body segmentation and background removal. For multi-modal generation and editing, each image is annotated with three types of labels: text descriptions, scribble images, and body part images. Please refer to the supplementary material for further details.

4.2 Model architecture and training details

Compressor. Our compressor maps the token $\mathbf{T} \in \mathbb{R}^{8192 \times 1024}$ to a latent representation $\mathbf{Z} \in \mathbb{R}^{8192 \times 8}$ via an encoder, and reconstructs $\mathbf{T}' \in \mathbb{R}^{8192 \times 1024}$ using a decoder. The encoder consists of seven blocks with progressively reduced channel dimensions [512, 256, 128, 64, 32, 16, 8]. The decoder contains five blocks with channel dimensions [32, 64, 128, 512, 1024]. The number of tokens (8,192) remains constant throughout. We use SiLU activations and Layer Normalization in all MLP layers and at the input of each self-attention block. The compressor is trained on 32 NVIDIA A100 GPUs with a batch size of 256 for one day. The learning rate is linearly warmed up from 4×10^{-10} to 4×10^{-4} over the first 1K iterations. The reconstruction loss weight is set to 1.0, while the KL divergence weight is increased linearly from 1×10^{-3} to 1×10^{-2} over 10K iterations.

GenLCA. The denoising network of GenLCA consists of 28 blocks, each with 1,024 channels, 16 attention heads, and an FFN with an MLP ratio of 4.0. RMSNorm [82] is applied to the query and key features. The number of latent tokens (8,192) remains constant across all blocks. For conditional input tokenization, we use the huge version of MetaCLIP and the big version of DINOv2 with registers. GenLCA is trained with the rectified flow objective using the Conditional Flow Matching (CFM) loss [43] with $\sigma_{\min} = 1 \times 10^{-5}$. Training is performed on 64 NVIDIA A100 GPUs with a batch size of 128 for four days. The learning rate is linearly warmed up from 2×10^{-10} to 2×10^{-4} over the first 1K iterations. Classifier-free guidance [24] is employed by randomly replacing conditional to-



Fig. 5: 3D avatars generated by GenLCA from texts. All results are generated with CFG scale = 5.0, 50 sampling steps, and animated with random poses.

kens with zero tokens with a probability of 0.25. During inference, sampling is performed using an explicit Euler solver with 50 steps.

5 Results

5.1 Visual results

We present text-conditioned generations in Fig. 5. GenLCA is capable of generating animatable and realistic 3D humans that accurately align with the input text descriptions. Our method supports a wide range of variations, including gender, age, as well as diverse clothing styles and hairstyles. In Fig. 1, we demonstrate sequential editing using text, scribble images, and body part images as inputs. Please refer to the supplemental material for details about the editing implementation and additional visual results.

5.2 Comparison

We compare GenLCA with SOTA SDS-based 3D full-body avatar generation methods (TADA [42], HumanGaussian [44], and DreamWaltz-G [31]) and diffusion models that directly model the 3D human distribution (TeRA [69] and

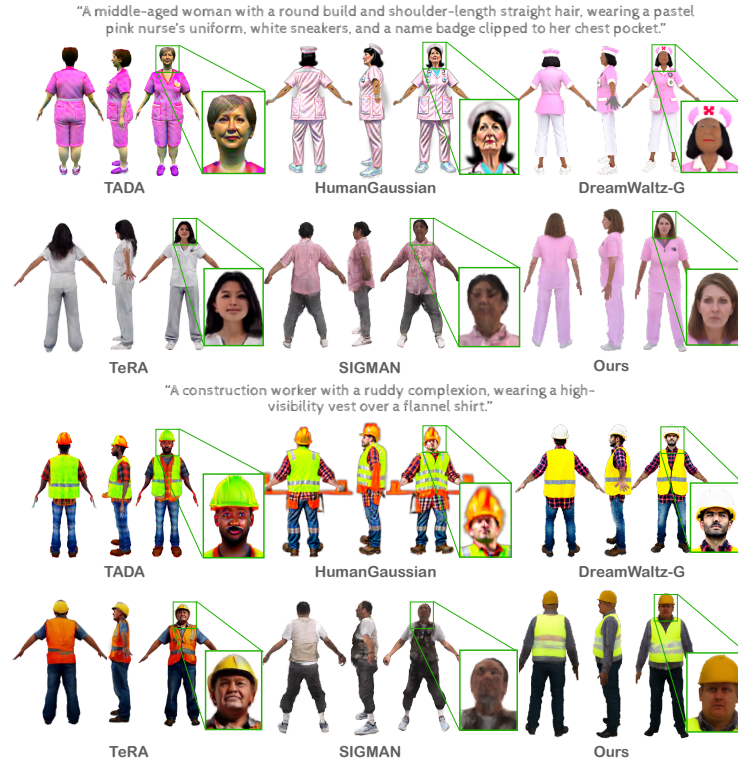


Fig. 6: We compare our GenLCA with SOTA methods, including SDS based approaches: TADA [42], HumanGaussian [44], and DreamWaltz-G [31], and text conditioned 3D human diffusion models, TeRA [69] and SIGMAN [78]. We use the same text prompt as input. In addition to full-body renderings, we also provide zoomed-in views for comparison of facial regions.

SIGMAN [78]). Additionally, we provide comparisons with 3D human reconstruction methods in the supplemental material.

Qualitative comparison Using the same text prompts, we show examples generated by SOTAs and GenLCA in Fig. 6. All SDS-based methods exhibit unrealistic visual styles. Both TeRA and SIGMAN demonstrate poor semantic alignment compared to other approaches. In the nurse case, although TeRA successfully generates a nurse avatar, it fails to produce the correct color of the uniform (“pastel pink nurse’s uniform”). SIGMAN fails to generate an aligned appearance in both cases. Additionally, TeRA suffers from a synthetic appearance, whereas SIGMAN demonstrates low visual quality. In contrast, GenLCA produces superior generation results in both semantic alignment and color, with realistic facial details and overall higher fidelity.

Table 2: Quantitative comparison results with SOTA methods. and denote the 1st and 2nd places.

Method	Semantic Align		Quality		FID ↓			Inference Time↓
	BLIP-VQA ↑	Text CLIP Score ↑	CLIB-FIQA ↑	HyperIQA ↑	2D Diffusion	THuman 2.0	HuGe100K	
TADA [42]	0.50	0.71	0.48	55.02	188.19	N/A	N/A	2.5h
HumanGaussian [44]	0.62	0.73	0.39	33.61	239.33	N/A	N/A	1.2h
DreamWaltz-G [31]	0.58	0.75	0.50	59.33	175.23	N/A	N/A	3.0h
TeRA [69]	0.42	0.67	0.44	44.01	151.80	N/A	N/A	12s
SIGMAN [78]	0.29	0.58	0.42	56.11	280.06	121.40	160.48	3s
GenLCA	0.64	0.76	0.55	63.05	160.91	96.03	76.50	12s

Table 3: User studies. and denote the 1st and 2nd places.

Method	User study ↑			
	Semantic align.	Consistency	Visual quality	Geometric quality
TADA [42]	2.89	3.18	2.29	2.15
HumanGaussian [44]	3.59	3.37	2.79	2.77
DreamWaltz-G [31]	3.68	3.93	3.41	3.37
TeRA [69]	2.63	3.74	3.30	3.28
SIGMAN [78]	1.65	1.86	1.40	1.52
GenLCA	4.56	4.68	4.65	4.63

Quantitative comparison We use 50 text prompts to generate 50 avatars for each method. Each avatar is rendered from multiple viewpoints (frontal, side, and back), resulting in three rendered images per avatar for evaluation.

Semantic alignment. We use BLIP-VQA from Progressive3D [10] to measure semantic alignment. Additionally, we estimate captions from the rendered images, and compute the CLIP feature distance between the estimated captions and the ground-truth text (Text CLIP score).

Visual quality. To evaluate the quality of the rendered avatar images, we employ CLIB-FIQA [51], a specialized method for assessing human facial image quality. For full-body avatar image quality assessment, we adopt HyperIQA [62]. We report FID [22] between the renderings of text-generated avatars and 2D diffusion-generated images (using the same text). Note that SDS methods and TeRA cannot be conditioned on images, making it infeasible to report metrics of them on large-scale image datasets. For methods that support image as inputs (SIGMAN and GenLCA), we generate avatars using 200 images from THuman 2.0 [80] and 200 from HuGe100K [90], and compute FID between the avatar renderings and the ground-truth images. Further details of evaluation metrics are provided in the supplemental material.

User study. We recruited 30 participants to evaluate rotating videos of 50 text-generated avatars produced by different methods. Each participant was presented with the results of 10 randomly selected avatars and asked to rate them on a 5-point scale across four criteria: text alignment, multi-view consistency, visual quality, and geometric quality. The questionnaire template used in the user study is provided in the supplementary material.

Tabs. 2 and 3 summarize our quantitative evaluations. Our GenLCA outperforms all state-of-the-art methods in semantic alignment, visual quality, and human preference by leveraging large-scale, real-world video data. In contrast,

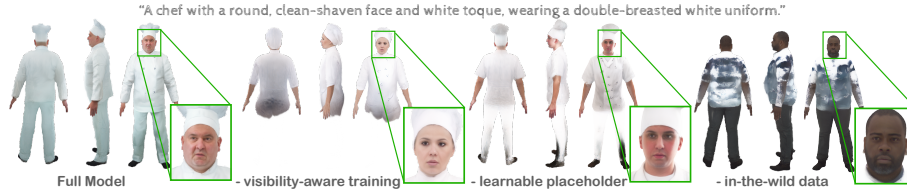


Fig. 7: We conduct ablation studies by individually removing the visibility-aware training, learnable placeholder, and in-the-wild training data components to demonstrate the effectiveness of each.

SDS-based approaches rely on 2D diffusion models to achieve strong semantic alignment, but this results in reduced visual quality. Meanwhile, 3D human diffusion model counterparts are trained on much smaller datasets, which negatively impacts both respects. Regarding FID, TeRA is trained on diffusion-generated images and therefore naturally aligns with the 2D diffusion distribution. GenLCA achieves improved FID on image datasets compared to SIGMAN.

5.3 Ablation studies

We conduct ablation studies to evaluate the effectiveness of the proposed training strategies. Fig. 7 shows the comparison results. Additional evaluations are provided in the supplemental material.

Visibility-aware training. We assess the impact of the visibility-aware training strategy by training the diffusion model directly on all tokens. We include both valid and invalid tokens as training data, and perform loss computation on all tokens. Without visibility-aware training, the model exhibits noticeable blurriness and transparency in the lower and back body, similar to the invalid regions present in the training data.

Learnable placeholder. To validate the effectiveness of the learnable placeholder, we replace all invalid region with fixed zero tokens during visibility-aware training. We observe that in the absence of a learnable placeholder, the generated avatars display unnatural color.

In-the-wild data. To evaluate the generalizability provided by in-the-wild data, we train GenLCA exclusively on indoor capture data, containing 3,000 identities. Without in-the-wild data incorporated into the training set, the model overfits to the captured data and fails to generate text-aligned results, showcasing poor generalization.

6 Conclusion

GenLCA achieves state-of-the-art quality through data scalability enabled by large-scale, imperfect real-world videos. This effective utilization of imperfect data is realized by (i) using a feed-forward reconstruction model to tokenize real-world videos, and (ii) introducing a visibility-aware training scheme that

handles partial observations. Experiments show that leveraging real-world videos significantly improves both the diversity and generalizability of the model, while the visibility-aware scheme filters unreliable signals to maximize use of high-quality data. The resulting 3D avatar diffusion model charts a path toward 2D-scale training for 3D digital humans.

The quality of GenLCA is constrained by its reliance on Linear Blend Skinning inherited from the reconstruction model for animation. Failures mainly occur under limited visibility or low coverage, e.g., extreme poses, rare text prompts, and over-smoothed or missing details in less observed regions such as the back (see examples in supplemental material). For future work, we aim to further strengthen the reconstruction model to boost fidelity and drivability, and to expand the data scale for continued gains.

Acknowledgements

Xiaogang Jin was supported by the National Natural Science Foundation of China (Grants: U25A20440, 62472373).

References

1. Abdal, R., Wang, Y., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D., Wetzstein, G.: Gaussian shell maps for efficient 3d human generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9441–9451. IEEE (2024) [4](#)
2. Anonymous: Large-scale codec avatars: The unreasonable effectiveness of large-scale avatar pretraining. In: Under review. See Supplementary Material Appendix B (2025) [5](#), [10](#)
3. Athar, S., Xu, Z., Sunkavalli, K., Shechtman, E., Shu, Z.: Rignerf: Fully controllable neural 3d portraits. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20332–20341. IEEE (2022) [4](#)
4. Cao, Y., Pan, L., Han, K., Wong, K.Y.K., Liu, Z.: AvatarGO: Zero-shot 4d human-object interaction generation and animation. In: The Thirteenth International Conference on Learning Representations (2025) [4](#)
5. Cha, H., Lee, I., Joo, H.: PERSE: personalized 3d generative avatars from A single portrait. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15953–15962 (2025) [4](#)
6. 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., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16123–16133 (2022) [4](#)
7. Chen, Z., Hong, F., Mei, H., Wang, G., Yang, L., Liu, Z.: Primdiffusion: Volumetric primitives diffusion for 3d human generation. In: Advances in Neural Information Processing Systems. vol. 36, pp. 13664–13677 (2023) [2](#), [5](#)
8. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., Sun, K., Tian, L., Wang, F., Wang, G., Wang, Q., Wang, Z., Xiao, J., Xu, S., Zhang, B., Zhang, P., Zhang, X., Zhang, Z., Zhou, J.,

- Zhuo, L.: Wan-animate: Unified character animation and replacement with holistic replication (2025) [2](#)
9. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., Ren, D., Yang, L., Liu, Z., Loy, C.C., Qian, C., Wu, W., Lin, D., Dai, B., Lin, K.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19925–19936. IEEE (2023) [2](#), [4](#)
 10. Cheng, X., Yang, T., Wang, J., Li, Y., Zhang, L., Zhang, J., Yuan, L.: Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. In: The Twelfth International Conference on Learning Representations, ICLR 2024 (2024) [13](#)
 11. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. In: Advances in Neural Information Processing Systems. vol. 37, pp. 57642–57670 (2024) [4](#)
 12. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 21086–21095 (2025) [2](#)
 13. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14916–14927 (2023) [4](#)
 14. Dong, Z., Duan, L., Song, J., Black, M.J., Geiger, A.: Moga: 3d generative avatar prior for monocular gaussian avatar reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13304–13314 (2025) [5](#)
 15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning, ICML 2024 (2024) [8](#)
 16. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (2021) [4](#)
 17. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Learning neural parametric head models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21003–21012. IEEE (2023) [4](#)
 18. Giebenhain, S., Kirschstein, T., Rünz, M., Agapito, L., Nießner, M.: Npga: Neural parametric gaussian avatars. In: SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24) (2024) [4](#)
 19. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27, pp. 2672–2680 (2014) [4](#)
 20. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion (2024) [2](#)
 21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022) [8](#)

22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems*. vol. 30, pp. 6626–6637 (2017) [13](#)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851 (2020) [4](#), [5](#)
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (2021) [10](#)
25. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: EVA3D: compositional 3d human generation from 2d image collections. In: *The Eleventh International Conference on Learning Representations, ICLR 2023* (2023) [4](#)
26. Hu, S., Hong, F., Hu, T., Pan, L., Mei, H., Xiao, W., Yang, L., Liu, Z.: Humanliff: Layer-wise 3d human diffusion model. *Int. J. Comput. Vis.* **133**(9), 5938–5957 (2025) [5](#)
27. Hu, T., Hong, F., Liu, Z.: Structldm: Structured latent diffusion for 3d human generation. In: *Computer Vision - ECCV 2024 - 18th European Conference*. vol. 15109, pp. 363–381. Springer (2024) [5](#)
28. Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., Wang, Q.: Human-norm: Learning normal diffusion model for high-quality and realistic 3d human generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4568–4577 (2024) [4](#)
29. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. In: *International Conference on 3D Vision, 3DV 2024*. pp. 1531–1542. IEEE (2024) [4](#)
30. Huang, Y., Yuan, Y., Li, X., Kautz, J., Iqbal, U.: Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 13533–13543 (2025) [4](#)
31. Huang, Y., Wang, J., Zeng, A., Zha, Z.J., Zhang, L., Liu, X.: Dreamwaltz-g: Expressive 3d gaussian avatars from skeleton-guided 2d diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–18 (2025) [4](#), [11](#), [12](#), [13](#)
32. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(7), 1325–1339 (2014) [2](#), [4](#)
33. Jiang, Y., Shen, Z., Guo, C., Hong, Y., Su, Z., Zhang, Y., Habermann, M., Xu, L.: Reperformer: Immersive human-centric volumetric videos from playback to photoreal reperformance. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11349–11360. Computer Vision Foundation / IEEE (2025) [4](#)
34. Jin, Y., Peng, S., Wang, X., Xie, T., Xu, Z., Yang, Y., Shen, Y., Bao, H., Zhou, X.: Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 11047–11057 (October 2025) [4](#)
35. Kant, Y., Weber, E., Kim, J.K., Khirodkar, R., Zhaoen, S., Martinez, J., Gilitschenski, I., Saito, S., Bagautdinov, T.M.: Pippo: High-resolution multi-view humans from a single image. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*. pp. 16418–16429. Computer Vision Foundation / IEEE (2025) [4](#)
36. Khirodkar, R., Bagautdinov, T.M., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: *Computer Vision - ECCV 2024 - 18th European Conference. Lecture Notes in Computer Science*, vol. 15062, pp. 206–228 (2024) [6](#), [10](#)

37. Li, J., Cao, C., Schwartz, G., Khirodkar, R., Richardt, C., Simon, T., Sheikh, Y., Saito, S.: Uravatar: Universal relightable gaussian codec avatars. In: SIGGRAPH Asia 2024 Conference Papers. SA '24 (2024) [4](#)
38. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., Luo, W., Guo, Y.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 16008–16018 (2025) [4](#)
39. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.* **36**(6), 194:1–194:17 (2017) [4](#)
40. Li, X., Yuan, Y., Mello, S.D., Daviet, G., Leaf, J., Macklin, M., Kautz, J., Iqbal, U.: Simavatar: Simulation-ready avatars with layered hair and clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26320–26330. Computer Vision Foundation / IEEE (2025) [5](#)
41. 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). pp. 19711–19722. IEEE (2024) [2](#), [4](#)
42. Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., Black, M.J.: Tada! text to animatable digital avatars. In: 2024 International Conference on 3D Vision (3DV). pp. 1508–1519 (2024) [4](#), [11](#), [12](#), [13](#)
43. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations, ICLR 2023 (2023) [6](#), [10](#)
44. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: Human-gaussian: Text-driven 3d human generation with gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6646–6657 (2024) [4](#), [11](#), [12](#), [13](#)
45. Lu, Y., Dong, J., Kwon, Y., Zhao, Q., Dai, B., De la Torre, F.: Gas: Generative avatar synthesis from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12883–12893 (2025) [4](#)
46. Lyu, W., Zhou, Y., Yang, M.H., Shu, Z.: Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12691–12701 (October 2025) [4](#)
47. Martinez, J., Kim, E., Romero, J., Bagautdinov, T., Saito, S., Yu, S.I., Anderson, S., Zollhöfer, M., Wang, T.L., Bai, S., Li, C., Wei, S.E., Joshi, R., Borsos, W., Simon, T., Saragih, J., Theodosis, P., Greene, A., Josyula, A., Maeta, S.M., Jewett, A.I., Venshtain, S., Heilman, C., Chen, Y.T., Fu, S., Elshaer, M.E.A., Du, T., Wu, L., Chen, S.C., Kang, K., Wu, M., Emad, Y., Longay, S., Brewer, A., Shah, H., Booth, J., Koska, T., Haidle, K., Andromalos, M., Hsu, J., Dauer, T., Selednik, P., Godisart, T., Ardisson, S., Cipperly, M., Humberston, B., Farr, L., Hansen, B., Guo, P., Braun, D., Krenn, S., Wen, H., Evans, L., Fadeeva, N., Stewart, M., Schwartz, G., Gupta, D., Moon, G., Guo, K., Dong, Y., Xu, Y., Shiratori, T., Prada, F., Pires, B.R., Peng, B., Buffalini, J., Trimble, A., McPhail, K., Schoeller, M., Sheikh, Y.: Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In: Advances in Neural Information Processing Systems. vol. 37, pp. 83008–83023 (2024) [4](#), [10](#)
48. Men, Y., Lei, B., Yao, Y., Cui, M., Lian, Z., Xie, X.: En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In: Proceedings of the

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9981–9991. IEEE (2024) [4](#)
49. Mihajlovic, M., Bansal, A., Zollhöfer, M., Tang, S., Saito, S.: Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In: *Computer Vision - ECCV 2022 - 17th European Conference. Lecture Notes in Computer Science*, vol. 13675, pp. 179–197 (2022) [4](#)
 50. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* **2024** (2024) [7](#), [8](#)
 51. Ou, F., Li, C., Wang, S., Kwong, S.: CLIB-FIQA: face image quality assessment with confidence calibration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1694–1704. IEEE (2024) [13](#)
 52. 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019) [4](#)
 53. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: *The 11th International Conference on Learning Representations, ICLR* (2023) [4](#)
 54. Prinzler, M., Zakharov, E., Sklyarova, V., Kabadayi, B., Thies, J.: Joker: Conditional 3d head synthesis with extreme facial expressions. In: *International Conference on 3D Vision, 3DV 2025, Singapore, March 25-28, 2025*. pp. 1583–1593. IEEE (2025) [4](#)
 55. 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)*. pp. 20299–20309. IEEE (2024) [4](#)
 56. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., Bo, L.: Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 14184–14194 (2025) [4](#)
 57. Qiu, L., Li, P., Zuo, Q., Gu, X., Dong, Y., Yuan, W., Zhu, S., Han, X., Chen, G., Dong, Z.: Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images (2025) [4](#)
 58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763. PMLR (2021) [7](#), [8](#)
 59. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10684–10695 (2022) [4](#), [5](#)
 60. Saito, S., Schwartz, G., Simon, T., Li, J., Nam, G.: Relightable gaussian codec avatars. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 130–141 (2024) [2](#)

61. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1606–1616. IEEE (2024) [4](#)
62. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3664–3673. Computer Vision Foundation / IEEE (2020) [13](#)
63. Tang, X., Zhang, B., Wonka, P.: Generative human geometry distribution. CoRR **abs/2503.01448** (2025) [5](#)
64. Tang, X., Zhang, B., Wonka, P.: Human geometry distribution for 3d animation generation. CoRR **abs/2512.07459** (2025) [5](#)
65. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: CAP4D: creating animatable 4d portrait avatars with morphable multi-view diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5318–5330 (2025) [4](#)
66. Wang, S., Antic, B., Geiger, A., Tang, S.: Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1877–1888 (2024) [2](#)
67. Wang, S., Simon, T., Santesteban, I., Bagautdinov, T., Li, J., Agrawal, V., Prada, F., Yu, S.I., Nalbone, P., Gramlich, M., Lubachersky, R., Wu, C., Romero, J., Saragih, J., Zollhoefer, M., Geiger, A., Tang, S., Saito, S.: Relightable full-body gaussian codec avatars. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (2025) [2](#), [4](#)
68. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., Guo, B.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4563–4573 (2023) [2](#), [5](#)
69. Wang, Y., Zhuang, Y., Zhang, J., Wang, L., Zeng, Y., Cao, X., Zuo, X., Zhu, H.: Tera: Rethinking text-guided realistic 3d avatar generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10686–10697 (2025) [2](#), [5](#), [11](#), [12](#), [13](#)
70. Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: Fake it till you make it: face analysis in the wild using synthetic data alone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3661–3671. IEEE (2021) [2](#)
71. Wu, Y., Xu, H., Tang, X., Chen, X., Tang, S., Zhang, Z., Li, C., Jin, X.: Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. ACM Trans. Graph. **43**(4) (2024) [4](#)
72. Wu, Y., Xu, S., Xiang, J., Wei, F., Chen, Q., Yang, J., Tong, X.: Aniportraitgan: Animatable 3d portrait generation from 2d image collections. In: SIGGRAPH Asia 2023 Conference Papers, SA 2023. pp. 51:1–51:9. ACM (2023) [4](#)
73. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1931–1941. IEEE (2024) [4](#)
74. XU, Z., Zhang, J., Liew, J.H., Feng, J., Shou, M.Z.: Xagen: 3d expressive human avatars generation. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34852–34865 (2023) [4](#)

75. Xue, Y., Xie, X., Kostyrko, M., Pons-Moll, G.: Infinihuman: Realistic 3d human creation with precise control. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. SA Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025) [4](#)
76. Xue, Y., Xie, X., Marin, R., Pons-Moll, G.: Human-3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In: Advances in Neural Information Processing Systems. vol. 37, pp. 99601–99645 (2024) [4](#)
77. Xue, Y., Xie, X., Marin, R., Pons-Moll, G.: Gen-3diffusion: Realistic image-to-3d generation via 2d & 3d diffusion synergy. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–17 (2025) [4](#)
78. Yang, Y., Liu, F., Lu, Y., Zhao, Q., Wu, P., Zhai, W., Yi, R., Cao, Y., Ma, L., Zha, Z.J., Dong, J.: Sigman: Scaling 3d human gaussian generation with millions of assets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5122–5133 (2025) [2](#), [5](#), [12](#), [13](#)
79. Yang, Z., Li, S., Wu, W., Dai, B.: 3dhumangan: 3d-aware human image generation with 3d pose mapping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23008–23019 (2023) [4](#)
80. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer RGBD sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5746–5756. Computer Vision Foundation / IEEE (2021) [2](#), [4](#), [13](#)
81. Yu, Z., Li, Z., Bao, H., Yang, C., Zhou, X.: Humanram: Feed-forward human reconstruction and animation model using transformers. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. SIGGRAPH Conference Papers '25 (2025) [4](#)
82. Zhang, B., Sennrich, R.: Root mean square layer normalization. In: Advances in Neural Information Processing Systems. vol. 32 (2019) [10](#)
83. Zhang, B., Cheng, Y., Wang, C., Zhang, T., Yang, J., Tang, Y., Zhao, F., Chen, D., Guo, B.: Rodinhd: High-fidelity 3d avatar generation with diffusion models. In: Computer Vision – ECCV 2024: 18th European Conference. p. 465–483. Springer-Verlag (2024) [2](#), [5](#)
84. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: A structured and explicit radiance representation for 3d generative modeling. In: Advances in Neural Information Processing Systems. vol. 37, pp. 97445–97475 (2024) [5](#)
85. Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, D.K., Zheng, M.: Avatarverse: High-quality & stable 3d avatar creation from text and pose. In: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024. pp. 7124–7132 (2024) [4](#)
86. Zhang, W., Yan, Y., Liu, Y., Sheng, X., Yang, X.: E3gen: Efficient, expressive and editable avatars generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 6860–6869. MM '24 (2024) [2](#), [5](#)
87. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., Shi, H., Liu, S., Wu, J., Lian, Y., Yang, F., Tang, R., He, Z., Wang, X., Liu, J., Zuo, X., Chen, Z., Lei, B., Weng, H., Xu, J., Zhu, Y., Liu, X., Xu, L., Hu, C., Yang, S., Zhang, S., Liu, Y., Huang, T., Wang, L., Zhang, J., Chen, M., Dong, L., Jia, Y., Cai, Y., Yu, J., Tang, Y., Zhang, H., Ye, Z., He, P., Wu, R., Zhang, C., Tan, Y., Xiao, J., Tao, Y., Zhu, J., Xue, J., Liu, K., Zhao, C., Wu, X., Hu, Z., Qin, L., Peng, J., Li, Z., Chen, M., Zhang, X., Niu, L., Wang, P., Wang, Y., Kuang, H., Fan, Z., Zheng, X., Zhuang, W., He, Y., Liu, T., Yang, Y., Wang,

- D., Liu, Y., Jiang, J., Huang, J., Guo, C.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025) [8](#)
88. Zhou, Z., Ma, F., Fan, H., Chua, T.: Zero-1-to-a: Zero-shot one image to animatable head avatars using video diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15941–15952 (2025) [4](#)
 89. Zhou, Z., Ma, F., Fan, H., Yang, Y.: Headstudio: Text to animatable head avatars with 3d gaussian splatting. In: Computer Vision - ECCV 2024 - 18th European Conference. Lecture Notes in Computer Science (2024) [4](#)
 90. Zhuang, Y., Lv, J., Wen, H., Shuai, Q., Zeng, A., Zhu, H., Chen, S., Yang, Y., Cao, X., Liu, W.: Idol: Instant photorealistic 3d human creation from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26308–26319 (2025) [2](#), [4](#), [13](#)
 91. Zielonka, W., Bagautdinov, T.M., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. In: International Conference on 3D Vision, 3DV 2025. pp. 979–990 (2025) [4](#)