AudioAvatar: Personalized Audio-driven Whole-body Talking Avatars

Seungeun Lee[1], SeungJun Moon[2], Hah Min Lew[3], Ji-Su Kang[1], Gyeong-Moon Park[3*]
[1]Klleon AI Research, [2] RLWRLD, [3]Korea University

Abstract

Prior expressive whole-body conversational avatar systems map audio to parametric poses and then render, creating a lossy bottleneck where quantization, retargeting, and tracking errors accumulate. This degrades audio–motion synchronization and suppresses micro-articulations critical for realism—such as bilabial closures, cheek inflation, nasolabial motion, blinks, and fine hand gestures—especially under single-image personalization. We propose an end-to-end framework that builds a full-body, photorealistic conversational avatar from a single image and drives it directly from audio, bypassing intermediate pose prediction. The avatar is modeled as a particle-based deformation field of 3D Gaussian primitives in a canonical space, with an audio-conditioned dynamics module that outputs per-particle trajectories for face, hands, and body, enabling localized high-frequency control with globally coherent motion. A splat-based differentiable renderer preserves identity, texture, and photo realism, while feature-level distillation from a large audio-driven video diffusion model and weak supervision from synthetic audio-conditioned clips further improve synchronization and natural expressivity. Joint photometric and temporal objectives shape the audio-conditioned deformation and rendering. Experiments across diverse speakers show improved lip–audio sync, fine facial detail, and conversational gesture naturalness over pose-driven baselines.

1. Introduction

Building highly realistic and animatable human avatars has been a central ambition in computer vision and graphics for decades. Beyond static reconstruction, recent work increasingly targets controllable, identity-preserving avatars driven by external signals such as pose, audio, or driving video [1, 3, 49, 55, 99]. Two major paradigms have emerged: (a) large-scale audio-driven video diffusion models that synthesize talking human videos directly from a single reference image and audio [10, 14, 50], and (b) 3D avatar methods that combine parametric body models with neural renderers such as NeRF or 3D Gaussian Splatting for photorealistic, pose-driven animation [69, 74, 92]. We refer to an animatable avatar as a personalized model that encodes a subject's canonical shape and appearance, deforms coherently under a driving signal, and renders photorealistically. Despite rapid progress in both directions, two capabilities remain underexplored in combination: personalizing a full-body avatar from a single image, and expressing conversational talking motion directly from audio. Achieving both is challenging because it requires recovering identity and deformation readiness from minimal input, and aligning subtle audio-conditioned dynamics across face, hands, and body at high temporal precision.

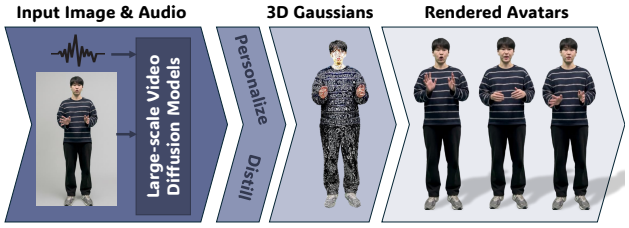

Figure 1. Overview of our personalized **Audio**-driven whole-body talking **Avatar**s (**AudioAvatar**): End-to-end framework that distills large video diffusion models to generate expressive, synchronized talking avatars from a single image.

Recent audio-driven video diffusion models [10, 14, 50] have demonstrated impressive capabilities in generating talking human videos from a single reference image and an audio signal, bypassing explicit audio-to-pose conversion. However, these approaches face several fundamental limitations: they are often restricted to head or upper-body generation, struggle with fine-grained hand and facial details, and exhibit identity drift over long sequences due to the lack of subject-specific personalization. Moreover, they operate without an explicit 3D representation, resulting in slow iterative denoising at inference time and limiting their applicability to scenarios that demand efficient, repeatable rendering of the same identity.

On the other hand, template-based 3D avatar pipelines fit SMPL/SMPL-X [47, 60], learn canonical geometry/appearance, and drive them with pose-dependent LBS [35], often coupled with NeRF or 3D Gaussians [31,

*Corresponding author

51], yielding photorealistic results across poses. However, they struggle with audio-synchronized conversational behavior—where millisecond lip closures, coarticulation, and fine hand gestures strongly affect naturalness since they require separate audio-to-motion module [11, 43, 54] that predicts parametric body/face/hand poses to drive a pose-conditioned renderer, where this introduces a lossy bottleneck, failing to capture tongue–lip contacts, cheek inflation, nasolabial detail, finger nuance and frame-by-frame deformation with weak temporal constraints, leading to sync errors, as illustrated in Fig. 2. These issues intensify under single-image personalization, where recovering a deformation-ready canonical avatar and learning an expressive, audio-aligned controller from one photo is especially ill-posed. Neither paradigm alone addresses the full challenge: video diffusion models lack efficient, identity-preserving rendering, while 3D avatar methods cannot directly leverage audio without a lossy pose intermediary.

We address these challenges with an end-to-end pipeline that builds, from a single user image, a full-body conversational avatar whose motion is driven directly by audio, where we modulate audio features to learn a dense deformation and fine appearance field inside differentiable avatar deformer and neural renderer that preserves identity and photorealism. It enhances temporal alignment by sequence-level rendering losses, allowing gradients to flow through time and synchronize deformations with speech prosody, rather than relying on per-frame pose tracking. Training on paired audio–video sequences enables the model to realize micro-articulations and coordinated face–hand–body dynamics without a lossy audio-to-pose bottleneck.

At the core of our approach is a particle-based deformation field embedded in a differentiable 3D Gaussian renderer. From a single user photo, we reconstruct a canonical, identity-preserving avatar and instantiate Gaussian particles that are dense over expressive regions and sparse elsewhere for efficiency. Audio features directly modulate per-particle trajectories—without an intermediate parametric pose—so that micro-articulations at the mouth, eyes, and hands can be controlled locally while the body motion remains globally coherent. Running this control at audio-synchronous rates expresses both rapid transients (e.g., plosive closures) and longer prosodic movements (e.g., head nods, beat gestures) with precise timing. Regularizers on locality and spectrum curb jitter yet preserve the high-frequency components essential for intelligible articulation.

To strengthen synchronization and realism under the single-image regime, we distill audio–motion priors from pretrained, large-scale, generative multi-modal video diffusion models. Diffusion features provide a sequence-level alignment signal that nudges our particle dynamics toward plausible coarticulation and conversational gesturing; in addition, synthetic audio-conditioned clips serve as weak supervision to diversify motion while keeping it synchronized to the same audio. Training is end-to-end: rendering losses propagate through time into the audio-conditioned deformation field, allowing the renderer and dynamics to co-adapt for tight audio–visual alignment while preserving identity and photorealistic appearance.

Contributions. (1) We propose an end-to-end, single-image pipeline that maps audio directly to a dense differentiable deformation field inside a Gaussian primitives, eliminating the lossy audio-to-pose and pose-to-render handoffs where quantization/retargeting/per-frame tracking errors accumulate, thereby reducing drift and improving temporal alignment. (2) We introduce a particle-based representation that affords localized, high-frequency facial/hand control with globally coherent full-body motion, yielding precise conversational expressivity. (3) We develop a diffusion-distillation scheme that transfers audio–motion priors via feature alignment and synthetic audio-conditioned clips, enabling realistic, well-synchronized behavior with minimal personalization data.

2. Related Work

2.1. Animatable Full-body Human Avatars

Early systems reconstructed actors from 3D capture or multi-view studios and animated the resulting meshes via hand-crafted pipelines—artist-designed rigging and skinning (e.g., LBS/DQS) or low-dimensional, PCA-based template models [2, 19, 29, 37, 47, 59, 64, 71, 75]. Pose-parameterized articulation enabled cross-subject transfer, but heavy expert intervention made these pipelines costly and time-consuming. The advent of continuous implicit representations ushered in neural renderers such as NeRF [51], powering photorealistic avatars [36, 61, 62, 72, 76, 85, 97] and free-view synthesis [17, 32, 33, 45, 87]. Yet NeRFs often train and infer slowly and need additional structure for reliable driving and retargeting. Acceleration via multi-resolution hash encodings and 3D Gaussian Splatting delivers real-time rendering with high-fidelity textures [28, 31], though many methods still rely on multi-view capture [39, 57] or monocular motion-capture signals [18, 24, 25, 34, 52, 53, 68] rather than commodity monocular inputs. Complementary lines leverage video diffusion to obtain animatable avatars from a single image, achieving view-consistent appearance even with limited data [74, 92].

Motivated by these observations, we pursue high-quality conversational full-body avatars that reduce dependence on pose-template intermediates. Our approach couples implicit motion-based deformation with a particle-based deformation layer designed to retain fine facial dynamics and finger gestures, while remaining compatible with efficient neural rendering. This hybrid control aims to preserve expressiveness and temporal coherence under realistic driving signals,

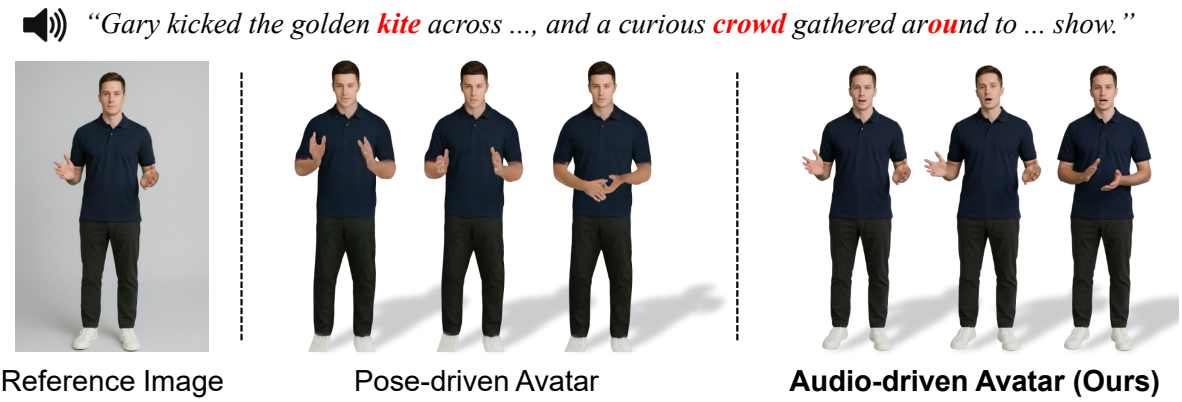

Figure 2. **Motivation**. When animating a talking avatar with conversational motion from audio, state-of-the-art pose-driven deformation approach degrades facial expressions, yields less natural motion, and exhibits poor audio–motion synchronization. In contrast, our method directly controls the avatar from the audio signal, yielding substantial improvements in visual quality, motion naturalness, and synchronization. For each method, we show the rendered frames aligned to the highlighted words in the driving audio.

closing the gap between head-only audio-driven animation and fully articulated, photorealistic human avatars.

2.2. Human Video Diffusion Models

Video diffusion models [7, 81] have become strong backbones for human video synthesis, enabling pose-guided animation from keypoints, dense or parametric poses [23, 78, 91, 94, 96, 98]. While these methods yield temporally consistent motion, they largely focus on coarse body animation and require audio-to-motion conversion. More recent large audio-driven diffusion models [10, 14, 50, 79, 82] generate talking videos directly from a single reference image and audio, producing realistic lip motion and gestures. However, they are typically limited to head/upper-body, rely on handcrafted or ground-truth guidance, operate at modest resolution, and struggle with fine-grained hand and facial details as well as identity preservation.

In contrast, our approach constructs a full-body audio-driven, and Gaussian Splatting-based avatar from a single image, overcoming the scope and fidelity limitations of prior work. By synthesizing diverse identity-specific talking videos from one image and varied audio, we enrich supervision for robust identity retention. Moreover, by distilling motion priors from large audio-driven diffusion models, our method achieves consistent coordination across body, hands, and face—capturing nuanced gestures and dynamic appearance beyond what existing approaches can deliver.

3. Method

Overview. Given a reference image I_0 and a driving audio sequence, our objective is to synthesize a full-body talking human video $V = \{I_t\}_{t=0}^{T}$ that faithfully preserves the identity specified in the reference. To maintain the visual quality characteristic of large-scale audio-driven video diffusion models while enabling computationally efficient inference, we represent the scene using a set of 3D Gaussians $\mathcal{G} = \{g_i\}_{i=1}^{N}$, which support fast and photorealistic rendering through Gaussian splatting [31].

To achieve generalized animation of Gaussians for arbitrary audio inputs, we map both the audio sequence and the temporal Gaussian deformations $\Delta\mathcal{G} = \{\Delta\mathcal{G}_0, \Delta\mathcal{G}_1, \ldots, \Delta\mathcal{G}_T\}$ into a unified latent manifold. Specifically, audio features $A = \{a_0, a_1, \ldots, a_T\}$ and latent Gaussian motion features, referred to as particle motions, $X = \{x_0, x_1, \ldots, x_T\}$ are embedded into a shared semantic space in which semantically corresponding audio signals and particle motions are placed close to one another.

The audio-driven particle motion generative model takes a sequence of embedded audio features as input and predicts the corresponding particle motions. For each time step t, the model first produces whole-body particle motions $x_t = \{x_t^{\text{body}}, x_t^{\text{face}}, x_t^{\text{hand}}\}$ and subsequently refines the face and hand regions, to capture fine-grained facial expressions and gesture dynamics. The generated particle motions are decoded into Gaussian attributes $\{\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T\}$ through a Gaussian decoder. These attributes are finally rendered via Gaussian splatting, producing the full-body talking human video V.

Problem Setting. We first synthesize identity-specific talking videos $\hat{V}_i$ using large-scale audio-driven video diffusion models (Sec. 3.3). Given these generated videos, we learn the corresponding Gaussian motions by training a deformable Gaussian splatting model [89] to reconstruct them, thereby producing Gaussian deformations $\Delta\mathcal{G}$ These pseudo ground-truth motions serve as supervision for learning the audio–particle motion embedding (Sec. 3.1) and the audio-driven particle motion generation model (Sec. 3.2).

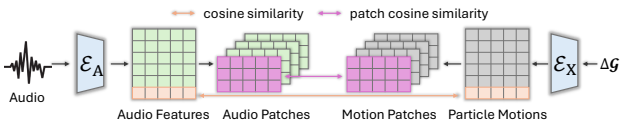

Figure 3. Overview of the audio–particle motion embedding pipeline, Sec. 3.1. The model learns a shared embedding space by aligning audio features with particle motion features through frame-level and patch-wise contrastive learning. This hierarchical alignment enables the network to capture modality-invariant motion cues that support accurate audio-driven particle motion generation.

3.1. Audio and Particle Motion Embedding

The goal of this module is to embed audio features $A = \{a_0, a_1, \ldots, a_T\}$ and particle motions $X = \{x_0, x_1, \ldots, x_T\}$ into a shared semantic manifold in which semantically corresponding elements are placed nearby. To achieve this cross-modal alignment, we follow a contrastive learning similar to CLIP [70, 77]. A particle motion encoder $\mathcal{E}_X$ is trained to map Gaussian deformations $\Delta\mathcal{G}$ into latent particle motion features X such that pairs of audio and particle motions representing the same underlying speech content exhibit high cosine similarity. Given an audio–particle pair at time t, we minimize the cosine distance $\mathcal{L}_{sim}$ between a_t and $x_t{=}\mathcal{E}_X(\Delta\mathcal{G}_t)$ while contrasting against mismatched pairs, encouraging modality-invariant motion representation.

To further incorporate local temporal structure and promote smooth motion trajectories, we introduce a patch-wise contrastive learning. Using a sliding temporal window, both audio features a_t and particle motions x_t are grouped into short-term patches, from which we compute average pooled patch features. Cosine similarity is then computed between corresponding audio and particle motion patches, allowing the model to learn temporally coherent alignment across short temporal contexts. This hierarchical alignment—frame-level and patch-level—produces a robust shared manifold that effectively supports audio-driven particle motion generation in later stages.

3.2. Audio-driven Particle Motion Generation

To this end, we design an audio-driven particle generative model that takes the aligned audio features $A = \{a_0, a_1, \ldots, a_T\}$ as input and predicts the corresponding particle motions $X = \{x_0, x_1, \ldots, x_T\}$ that determine the animation of Gaussian primitives. We adopt a hierarchical design; the model first generates particle motions for the whole-body region, and then applies an auxiliary refinement module to the face and hand subsets, since these regions contain fine-grained expressive motions such as facial expressions and finger gestures. We first generate the whole-body motions, including the body, face, and hands. Leveraging the fact that our representations are already embedded in a compact, low-dimensional manifold, we employ a high-capacity diffusion Transformer to synthesize high-fidelity co-speech gesture motion.

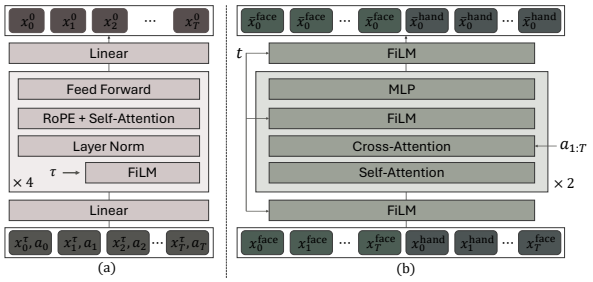

Figure 4. The pipeline of audio-driven particle motion generation, Sec. 3.2. Given aligned audio features, the model first predicts whole-body particle motions using a diffusion Transformer and then applies a dedicated refinement module to face and hand regions to capture fine-grained expressive motions. This hierarchical generation process enables high-fidelity, co-speech motion synthesis across both global body movement and detailed local articulations.

The model receives as input the noised particle motions X^τ and audio features A concatenated together through forward diffusion process. The diffusion timestep τ is conditioned on FiLM [63] layers, and the reverse process predicts progressively denoised particle motions, ultimately producing clean particle motions X^0 at timestep 0. In practice, following [56], we directly predict the clean particle motion X^0 at timestep 0 from the noisy one at timestep τ. Formally, this process can be expressed as:

$$X^0 = \mathcal{F}(X^\tau | \mathbf{A}, \tau), \tag{1}$$

where $\mathcal{F}$ denotes the diffusion Transformer to generate whole-body particle motions.

After generating the whole-body particle motions, the subsets corresponding to the face and hand regions are further refined using an additional Transformer-based refinement module. Unlike the diffusion model, this refinement module is conditioned not on diffusion timesteps τ, but rather on the actual motion time index t along the temporal axis of the talking sequence. This enables temporally consistent fine-grained motion refinement.

Finally, the complete set of particle motions, including the refined face and hand regions, is decoded into Gaussian attributes $\{\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T\}$ using a feedforward MLP. The resulting collection of Gaussians is rendered via Gaussian Splatting, producing the final full-body talking human video $\{\hat{I}_0, \hat{I}_1, \cdots, \hat{I}_T\}$. The architecture is illustrated in Fig. 4.

3.3. Video Diffusion Distillation

Video Data Synthesis. To inject the knowledge of large-scale video diffusion models into our particle-based rep-

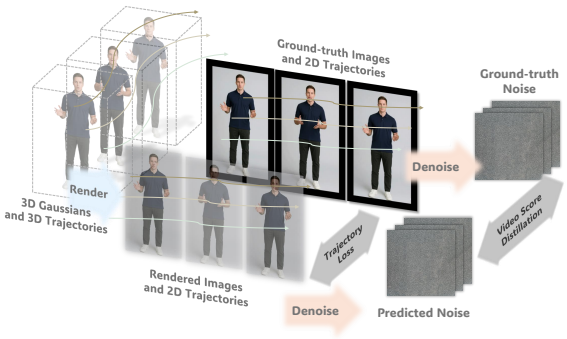

Figure 5. The schematic flow of the trajectory alignment loss and the video score distillation sampling loss. The deformed Gaussians are rendered to obtain images and their corresponding 3D and 2D trajectories. The 2D trajectories are aligned with the ground-truth trajectories to maximize correspondence. In parallel, the rendered images are passed through a denoising network to extract scores, which are optimized to match the ground-truth scores.

resentation, we synthesize identity-specific talking human video datasets and use them as pseudo–ground-truth supervision for learning time-varying Gaussian deformations. For this purpose, we propose a hybrid data construction strategy that leverages both a large-scale, multi-modal generative video foundation model [12] and audio-driven talking human video diffusion models [10, 14]. Thelarge-scale, multi-modal generative video foundation model is pretrained on massive open-domain video datasets and thus provides high-quality visual generation with rich appearance diversity. In contrast, the audio-driven human video diffusion models are capable of generating co-speech talking human videos that are temporally synchronized with input audio sequences. By combining these two complementary model categories, we obtain training videos that exhibit both high visual fidelity and accurate audio–motion synchronization, effectively improving the quality of supervision for Gaussian deformation learning. The full pipeline for automated data construction from audio-driven video diffusion models is illustrated in Fig. 6.

Video Score Distillation. To distill realistic appearance and motion knowledge for learning the dynamics of 3D Gaussians, we introduce a video score distillation sampling loss. Denote the teacher score network $s_\psi(\cdot, \tau, c)$ at noise level τ with variance schedule $\alpha(\tau)$ and $\sigma(\tau)$. We apply a video variant of score-distillation sampling to inject the teacher's generative prior:

$$\nabla_\Phi \mathcal{L}_{\text{vsd}} = \mathbb{E}_{t,\tau,\epsilon}\Big[w(\tau)\big(s_\psi(\tilde{I}_{t,\tau}, \tau, c) - \epsilon\big)\,\frac{\partial \tilde{I}_{t,\tau}}{\partial \Phi}\Big], \quad (2)$$

where $\tilde{I}_{t,\tau} = \alpha(\tau)\,\hat{I}_t + \sigma(\tau)\epsilon$, and Φ is learnable parameters of audio-driven particle generation model. It encourages $\hat{I}_{1:T}$ to lie on the teacher's audio-conditioned video manifold while inheriting its temporal coherence.

Trajectory Alignment Loss. To ensure temporally consistent deformation of Gaussians in our 4D representation, we introduce a trajectory alignment loss inspired by [84, 93]. Each Gaussian primitive produces a 2D pixel trajectory through rendering, while the diffusion-generated supervision provides the corresponding ground-truth pixel motion. We enforce alignment by penalizing deviations between predicted and reference trajectories across time. Concretely, for each Gaussian i, we compute a per-frame reprojection error between its rendered center $\hat{u}_t^i$ and the target pixel location u_t^i, and accumulate this over the full sequence:

$$\mathcal{L}_{\text{traj}} = \sum_i \sum_t \left\| \hat{u}_t^i - u_t^i \right\|_2^2. \quad (3)$$

This loss encourages Gaussians to follow coherent motion paths that remain consistent with observed visual trajectories, improving stability and reducing temporal drift during deformation learning. Finally, we supervise explicit video synthesis using an L1 loss between the Gaussian-rendered frames and the diffusion-generated images. To train the audio-driven particle motion generation module, we directly followed the $\mathcal{L}_{\text{simple}}$ loss function of [56]. To further regularize deformation, we impose an ARAP distance-preserving prior [83] between k-nearest neighbors across adjacent timesteps.

4. Experiments and Analysis

4.1. Experimental Setup

Dataset. Due to the limited availability of publicly accessible conversational human videos paired with audio, we aggregate data from multiple sources. We conducted test evaluations on motions that were not seen during training for a total of 30 subjects. The videos contain whole-body human figures and talking motions. The videos consist of publicly available datasets (causal conversational dataset [65], seamless-interaction dataset [1]) as well as data we collected and generated ourselves. Each sequence consists of a talking video with an average duration of 5–10 seconds. All videos depict a single human subject engaged in natural conversation, exhibiting both speaking activity and accompanying conversational motions, and each video is temporally aligned with its corresponding audio track.

Metrics. We evaluate our approach from multiple perspectives, using a variety of evaluation metrics. We evaluate the visual and aesthetic quality by evaluating IQA and ASE using Q-align [90]. We adopt SyncC and SyncD, introduced by [66], to quantify the synchronization accuracy between lip motion and the corresponding audio. To evaluate the preservation of facial identity, we compute the cosine similarity (CSIM) between facial features extracted from the reference image and those from the generated frames. We fur-

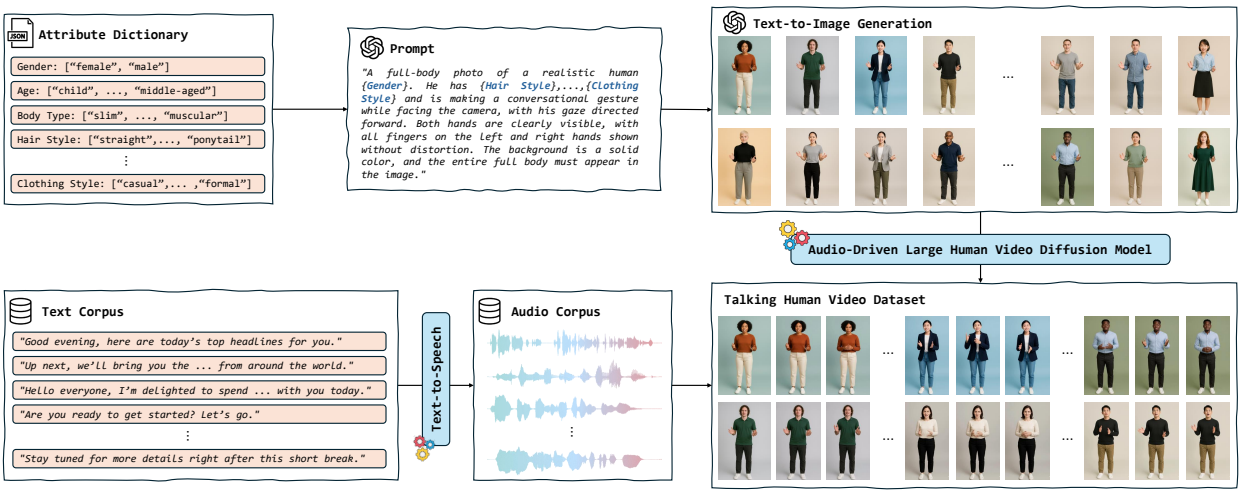

Figure 6. Overview of video data synthesis pipeline used for video diffusion distillation. We first generate diverse full-body human identities via a foundational text-conditioned image generative model and pair them with speech audio synthesized from a curated text corpus. Several Audio-driven video diffusion models is used to produce temporally synchronized talking human videos, yielding high-fidelity, audio-aligned training data for supervising Gaussian deformation learning.

ther assess gesture fidelity using the average keypoint distance, reporting the Hand Keypoint Confidence (HKC) and the Hand Keypoint Variance (HKV), defined as the average confidence score and standard deviation of detected hand keypoints. For low-level reconstruction fidelity, we report PSNR and SSIM [86] between the rendered images and the ground-truth video, given same audio driving signal. Lastly, we employ the Fréchet Inception Distance (FID) [21] and the Fréchet Video Distance (FVD) [80] to measure the generative diversity and overall coherence of rendered avatars.

Comparative Methods. We benchmarked our approach against the relevant state-of-the-art methods for creating animatable human avatars from a single image, publicly available PERSONA [74] and LHM [69], through both quantitative and qualitative evaluations. However, unlike our approach, they cannot directly drive an avatar from audio and therefore require a converter from audio to driving pose parameters. Therfore, we utilize a state-of-the-art whole-body motion converter [6] to generate motion, which is then used to control the avatars from baselines. We further extended our comparisons to include several state-of-the-art audio-driven human video diffusion models, OmniAvatar [14] and HunyuanAvatar [10], to provide a broader evaluation.

4.2. Results

Quantitative Comparisons. Table 1 shows that our approach outperforms all baselines across the metrics on the test set. Against single-image animatable Gaussian avatar methods, LHM [69] and PERSONA [74], our method achieves higher perceptual quality (IQA: +3.4%, ASE: +4.4%), better audio–lip synchronization (SyncC: +4.3%, SyncD: +20.3%), and stronger low-level fidelity (SSIM: +4.7%, PSNR: +4.3%). When compared with state-of-the-art audio-driven human video diffusion models, OmniAvatar [14] and HunyuanVideo-Avatar [10], our method delivers markedly improved video-level realism and temporal coherence, reducing FID by 27.9% and FVD by 25.0% relative to the strongest baseline. We also observe consistent gains in hand–gesture fidelity (HKC: +2.5%), reflecting more reliable control of fine-grained motions. Overall, these results substantiate the effectiveness of our proposed pipeline, yielding robust improvements across the metrics.

Qualitative Comparisons. Fig. 7 qualitatively compares the baselines that synthesize animatable human avatars from a single image on the test sets. Because prior approaches cannot directly control an avatar from audio, we evaluate rendering quality under the same motion for all methods to ensure a fair comparison. The results show that our approach produces sharper and more expressive facial expressions, improved lip synchronization, and finer hand–gesture details. Across time, our renderings also exhibit smoother, more natural motion transitions. Our rendering pipeline is built on a Gaussian rasterizer, enabling direct extraction of videos as sequences of images. We therefore also compare against state-of-the-art methods that generate human-animation videos from audio signals, as shown in Fig. 8. The visual comparisons indicate that our method achieves competitive image and motion quality even relative to generative video diffusion models. Notably, the second example highlights two consistent advantages of our approach: (i) motion-consistent preservation of fine hand details and (ii) stronger identity preservation throughout the sequence. Fig. 10 compares [92] in in-the-wild set-

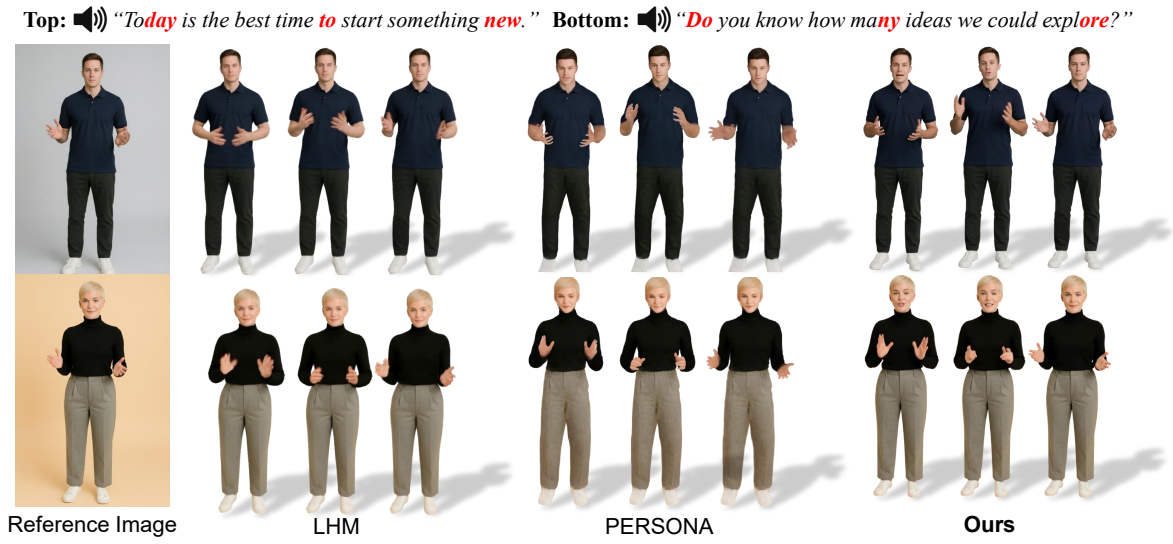

Figure 7. **Qualitative comparison with state-of-the-art animatable Gaussian Splatting-based avatar models.** For each method, we show the rendered frames aligned to the highlighted words in the driving audio. Our method outperforms state-of-the-art approaches in terms of visual quality, motion naturalness, and synchronization.

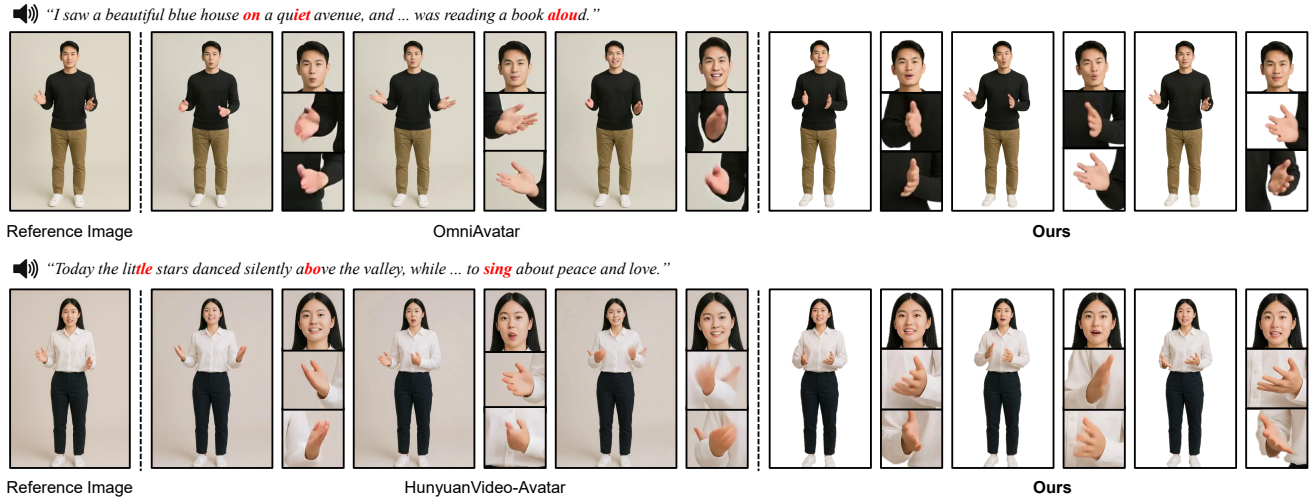

Figure 8. **Qualitative comparison with state-of-the-art audio-driven human video generation models.** For each method, we show the rendered frames aligned to the highlighted words in the driving audio, along with cropped views of the *face* and *hands* for finer inspection. Relative to diffusion-based baselines, our approach exhibits fewer motion artifacts (e.g., lip–audio desynchronization, hand jitter/warping) and stronger identity preservation across views and phonetic contexts.

tings. The proposed method preserves hand-gesture geometry and texture more sharply, while reducing blur and self-intersection artifacts seen in prior methods.

4.3. Ablation Study

We conduct a comprehensive ablation study to quantify the contribution of each component in our framework. As in Table 2 & Fig. 9, we evaluate visual quality (IQA, SSIM, PSNR), audiovisual coherence (ASE, Sync), identity and keypoint consistency (HKC, CSIM), and generative realism (FID, FVD) under different model configurations.

Audio–Particle Embedding. Removing the audio–particle embedding results in noticeable performance drops across synchronization metrics (e.g., SyncC decreases from 7.20 to 7.05, and SyncD increases from 5.42 to 5.60), demonstrating the module's importance in aligning temporal audio cues with facial dynamics. The patchified alignment strategy also proves essential for maintaining spatial consistency: without it, SSIM and PSNR decline while FID and FVD increase, indicating degraded reconstruction fidelity

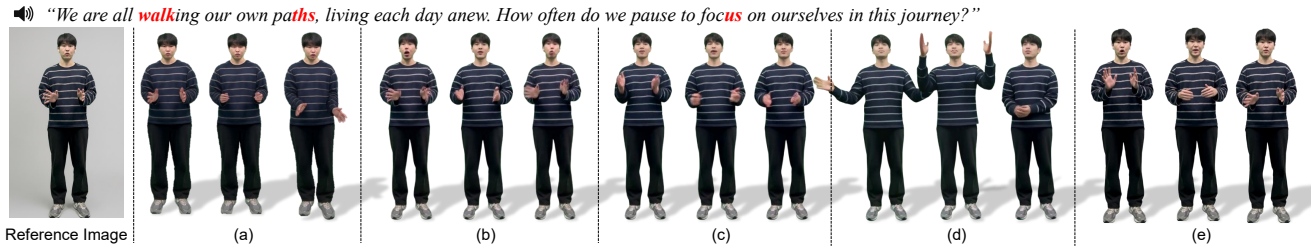

Figure 9. **Ablation of the proposed components.** We evaluate (a) w/o $\mathcal{L}_{\text{vsd}}$, (b) w/o $\mathcal{L}_{traj}$, (c) w/o hybrid talking video synthesis, (d) w/o patchfied strategy, and (e) full model. Removing any component harms visual fidelity, motion naturalness, and audio–motion sync, confirming each element's contribution to overall quality.

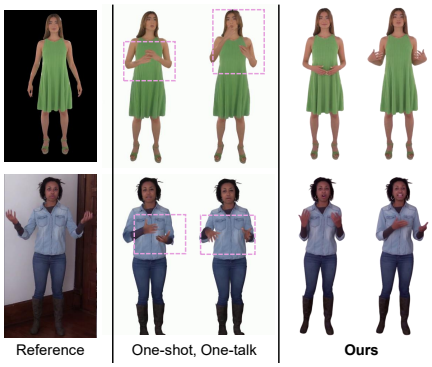

Figure 10. Visual comparison of talking animation in in-the-wild videos with a talking Gaussian avatar [92]. Our method captures hand-gesture details more accurately, whereas an existing method exhibits blurred textures and self-penetration artifacts.

Methods	IQA↑	ASE↑	SyncC↑	SyncD↓	HKC↑	CSIM↑	SSIM↑	PSNR↑	FID↓	FVD↓
EchoMimicV2 [50]	3.37	1.98	4.12	10.20	0.836	0.458	0.660	15.90	22.8	420
OmniAvatar [14]	3.99	2.64	6.40	7.60	0.858	0.525	0.705	17.20	18.6	350
HunyuanVideo-Avatar [10]	4.08	2.71	6.90	7.12	0.875	0.539	0.709	17.55	17.2	320
LHM [69]	3.80	2.50	6.10	7.00	0.860	0.500	0.700	16.90	19.5	365
PERSONA [74]	3.88	2.58	6.30	6.80	0.868	0.510	0.708	17.20	18.9	345
AudioAvatar (Ours)	**4.22**	**2.83**	**7.20**	**5.42**	**0.897**	**0.551**	**0.742**	**18.30**	**12.4**	**240**

Table 1. Quantitative comparisons on our defined test set. We compare our method with state-of-the-art Gaussian avatar and audio-driven human video generation models. Our approach shows consistently improved performance across all the metrics.

Methods	IQA↑	ASE↑	SyncC↑	SyncD↓	HKC↑	CSIM↑	SSIM↑	PSNR↑	FID↓	FVD↓
w/o audio–particle embedding	4.05	2.75	7.05	5.60	0.890	0.545	0.720	17.60	13.8	265
w/o patchfied alignment strategy	4.18	2.80	7.15	5.45	0.870	0.555	0.735	18.10	13.1	252
w/o hand and face refinement module	4.16	2.78	7.10	5.50	0.860	0.552	0.734	18.00	13.7	258
w/o hybrid talking video synthesis	4.00	2.74	6.85	5.80	0.888	0.520	0.730	17.90	14.2	272
w/o video score distillation	3.95	2.69	6.90	5.78	0.886	0.542	0.722	17.50	15.0	290
w/o trajectory alignment	4.02	2.73	6.70	6.20	0.872	0.544	0.725	17.60	15.6	310
AudioAvatar (Ours)	**4.22**	**2.83**	**7.20**	**5.42**	**0.897**	**0.551**	**0.742**	**18.30**	**12.4**	**240**

Table 2. Ablation study. We evaluate the impact of each proposed component by removing them in groups and comparing to our full model. Blocks correspond to (1) personalization modules, (2) audio-driven particle deformation, and (3) diffusion-based objective terms. Results show that every component contributes notably to overall performance.

and reduced video realism.

Audio-Driven Particle Deformation. The hand and face refinement module contributes significantly to the preservation of fine-grained motion details. Its removal yields declines in HKC (0.897 to 0.860) and SSIM/PSNR, confirming its role in enhancing structural coherence in highly articulated regions. Similarly, removing the hybrid talking-video synthesis module reduces audiovisual synchrony (SyncC drops from 7.20 to 6.85 and ASE from 2.83 to 2.74), highlighting its value in generating expressive and temporally aligned motion trajectories.

Diffusion-Based Objective Terms. Eliminating the video score distillation loss notably disrupts temporal smoothness, reflected in a substantial increase in FVD (240 to 290). The trajectory alignment loss is particularly critical: without it, SyncD increases sharply (5.42 to 6.20), representing the worst synchronization among all ablations. This demonstrates the necessity of trajectory-level alignment for stable and coherent motion generation. The full model consistently outperforms all ablated variants across every metric, confirming that the proposed personalization modules, audio-driven deformation mechanisms, and diffusion-based objectives are complementary and collectively essential for achieving high-quality, temporally consistent, and well-synchronized talking video generation.

5. Conclusion

We present a framework that builds a personalized full-body avatar from a single image and drives it directly with raw audio, bypassing intermediate pose representations to avoid error accumulation in conventional pipelines. The particle-based deformation model enables fine-grained facial and hand control with coherent body motion, and distilling motion priors from audio-driven video diffusion models further enhances synchronization and temporal consistency. Experiments confirm consistent improvements over baselines in lip-audio synchronization, perceptual quality, and gesture fidelity, demonstrating photorealistic talking avatars suitable for high-fidelity digital human applications.

Acknowledgements

This research was supported by Seoul Business Agency (SBA) (CC250007), and supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT), and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), RS-2024-00457882, AI Research Hub Project, and RS-2024-00456709).

References

[1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025. 1, 5, 4

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, 2018. 2

[3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 1

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 3

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[6] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1880–1888, 2025. 6, 1

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[8] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 3

[9] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024. 1

[10] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025. 1, 3, 5, 6, 8, 2, 4

[11] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1942–1953, 2024. 2, 1

[12] Google DeepMind. Flow: An ai filmmaking tool built with and for creatives. https://labs.google/flow/about, 2025. Accessed: 2025-11-14. 5

[13] ElevenLabs. Elevenlabs text-to-speech. https://elevenlabs.io, 2025. Online service. 2

[14] Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025. 1, 3, 5, 6, 8, 2, 4

[15] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3497–3506, 2019. 1

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 1

[17] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, 2023. 2

[18] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5559–5570, 2025. 2

[19] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 2

[20] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 1

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[22] Steven Hogue, Chenxu Zhang, Hamza Daruger, Yapeng Tian, and Xiaohu Guo. Diffted: One-shot audio-driven ted talk video generation with diffusion-based co-speech ges-

tures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1922–1931, 2024. 1

[23] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[24] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 634–644, 2024. 2

[25] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20418–20431, 2024. 2

[26] Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6997–7006, 2024. 1

[27] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1

[28] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2

[29] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 2

[30] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 3

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3

[32] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural Human Performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021. 2

[33] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. DELIFFAS: Deformable light fields for fast avatar synthesis. *NeurIPS*, 2024. 2

[34] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19876–19887, 2024. 2

[35] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023. 1

[36] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2

[37] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

[38] Xinjie Li, Ziyi Chen, Xinlu Yu, Iek-Heng Chu, Peng Chang, and Jing Xiao. Co-speech gesture video generation with implicit motion-audio entanglement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11384–11394, 2025. 1

[39] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable Gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 2

[40] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[41] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3764–3773, 2022. 1

[42] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. *arXiv preprint arXiv:2410.04221*, 2024. 1

[43] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 2, 1

[44] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024. 3

[45] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

[46] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*, 2025. 1

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1, 2

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[49] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 1

[50] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5489–5498, 2025. 1, 3, 8

[51] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[52] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2

[53] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 2

[54] M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16578–16588, 2025. 2, 1

[55] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 1

[56] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 4, 5

[57] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. ASH: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, 2024. 2

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 2

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[61] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2

[62] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[63] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[64] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 2

[65] Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10–17, 2023. 5

[66] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 5

[67] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11077–11086, 2021. 1

[68] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024. 2

[69] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 1, 6, 8

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[71] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[72] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-

avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 2

[73] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021. 1

[74] Geonhee Sim and Gyeongsik Moon. Persona: Personalized whole-body 3d avatar with pose-driven deformations from a single image. *arXiv preprint arXiv:2508.09973*, 2025. 1, 2, 6, 8, 4

[75] Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG)*, 29(6):1–10, 2010. 2

[76] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in neural information processing systems*, 34:12278–12291, 2021. 2

[77] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 4

[78] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024. 3

[79] Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025. 3

[80] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[81] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[82] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025. 3

[83] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 5

[84] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9660–9672, 2025. 5

[85] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European conference on computer vision*, pages 1–19. Springer, 2022. 2

[86] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[87] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2

[88] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 3

[89] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 3, 1

[90] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5

[91] Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. Musev: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising. *arxiv*, 2024. 3

[92] Jun Xiang, Yudong Guo, Leipeng Hu, Boyang Guo, Yancheng Yuan, and Juyong Zhang. One shot, one talk: Whole-body talking avatar from a single image. *arXiv preprint arXiv:2412.01106*, 2024. 1, 2, 6, 8

[93] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 5

[94] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[95] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 1

[96] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3

[97] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning

Liu, and Yebin Liu. AvatarRex: Real-time expressive full-body avatars. *ACM TOG*, 2023. 2

[98] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 3

[99] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *2025 International Conference on 3D Vision (3DV)*, pages 979–990. IEEE, 2025. 1