EMOSH: Expressive Motion and Shape Disentanglement for Human Animation

Dongbin Zhang[1,2*], Hao Liu[2], Binquan Dai[1], Kangjie Chen[1], Chuming Wang[1], Chen Li[2†], Jing LYU[2], and Haoqian Wang[1†]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]WeChat Vision, Tencent Inc.

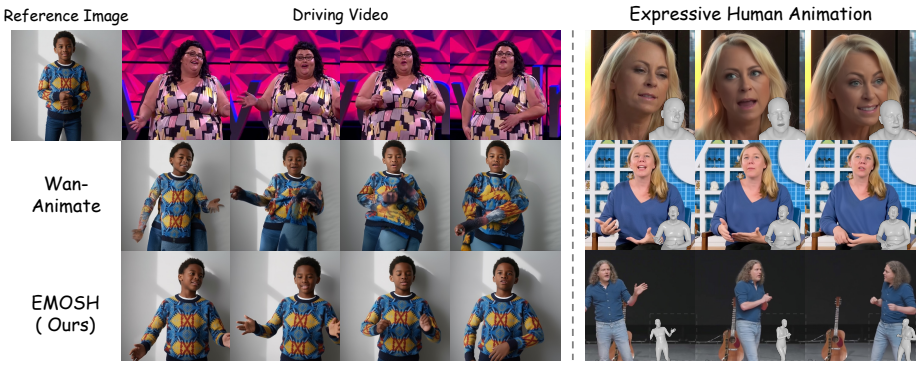


Fig. 1: Given a reference image and a driving video, EMOSH achieves high-fidelity, mesh-guided expressive human animation while disentangling expressive motion from body shape to prevent shape leakage.

Abstract. High-fidelity and expressive controllable human animation is essential for content creation and digital avatar applications. However, existing methods face a dilemma between expressiveness and disentanglement. Mainstream 2D pose-conditioned approaches suffer from "motion-shape entanglement", leading to the leakage of the driving subject's body shape. Conversely, methods relying on 3D priors (e.g., SMPL) achieve geometric disentanglement but struggle to capture facial expressions and complex gestures, resulting in rigid animations. To this end, we propose EMOSH, a novel framework for high-fidelity controllable human video generation. First, an Expressive Human Model (EHM) is introduced as the core control representation. By explicitly disentangling shape and pose parameters, we fundamentally resolve the body shape leakage issue. Alongside this, a robust motion tracker is designed to accurately estimate EHM parameters from video. Second, we propose a Coarse-to-Fine Hybrid Motion Injection strategy, enabling more fine-grained control over expressions and gestures. Furthermore, we introduce a Spatially-Aligned Conditioning mechanism to bridge the domain gap between training and

* Intern at WeChat Vision. † Corresponding authors.

inference, improving identity consistency. Extensive experiments demonstrate that EMOSH outperforms previous methods in both self-driven and cross-driven scenarios, producing high-fidelity videos with vivid expressions while maintaining shape disentanglement. Video demos and additional results are available at our Project page.

1 Introduction

Driven by rapid advancements in generative AI, high-fidelity and expressive human animation has become a research hotspot in computer vision, with broad applications in digital avatars, film, and social media. The core objective is to animate a reference image using a driving video. This requires faithfully replicating the driver's motions, expressions, and spatial layout, while strictly preserving the reference subject's identity and appearance.

Extensive research has explored various methodologies in this domain. Early works [56, 57, 71] primarily relied on GANs [12] and warping fields, but these methods often fail with large motions or unseen regions. Recently, diffusion models [14, 61] have become the mainstream choice due to their superior generation quality. Pioneering explorations like ControlNet [80] established the paradigm of injecting 2D skeletal poses as spatial conditions into diffusion models. Building on this, works like Animate Anyone [17] introduced temporal layers for more fluid motion transfer, while MagicPose [4] incorporated fine-grained facial landmarks to enhance expression control. Most recently, driven by Video Foundation models, approaches [21,33,59,73,83] have scaled human animation to large DiT-based backbones, further elevating generation quality and dynamic fidelity.

Despite these advancements, existing methods still face a dilemma between **expressiveness** and **disentanglement**. Mainstream approaches relying on 2D poses suffer from severe **"Motion-Shape Entanglement"**. Because 2D skeletons implicitly encode the driver's body proportions—an issue difficult to resolve even with pose retargeting—these models inevitably leak the driver's physique into the generated output. To address this, some works [88, 89] adopt 3D parametric models (e.g., SMPL [30]) to geometrically disentangle shape and motion. However, such priors lack the capacity to capture and represent facial expressions and complex gestures, resulting in **rigid animations**. Furthermore, a pervasive **domain gap** exists between training and inference. Models are typically trained on spatially aligned self-driven data but must handle spatially misaligned cross-driven tasks during inference. This discrepancy frequently causes identity loss and visual artifacts, particularly in long video generation.

To address these challenges, we propose EMOSH, a novel framework for high-fidelity human video generation that achieves both expressive motion control and shape disentanglement, as shown in Fig. 1. First, we adopt a 3D parametric model, the **Expressive Human Model (EHM)** [79], as our core representation to resolve the shape leakage of 2D pose-based methods. By explicitly disentangling shape and pose parameters, EHM allows us to precisely transfer driving motions while preserving the reference subject's body shape. Alongside

this, we design a **Confidence-Aware Motion Tracker** to robustly capture full-body motions—including facial expressions and hand gestures—from monocular videos. Second, to overcome the lack of high-frequency details (e.g., eye blinking) in naive mesh guidance, we propose a **Coarse-to-Fine Hybrid Motion Injection strategy**. This combines global geometric priors of 3D meshes with the local precision of 2D keypoints, enhancing control precision while maintaining structural integrity. Finally, to bridge the domain gap between training and inference, we introduce a **Spatially-Aligned Conditioning mechanism**. By injecting spatially-aligned visual anchors, it mitigates error accumulation, reduces visual artifacts, and ensures superior identity consistency in video generation.

Extensive experiments demonstrate that EMOSH achieves state-of-the-art performance in both self-driven and cross-driven scenarios, surpassing 2D pose-based methods plagued by shape coupling and previous mesh-based methods limited by expressiveness. It successfully achieves high-fidelity motion-shape disentanglement while preserving vivid expressions, gestures, and identity. In summary, our main contributions are as follows:

- We propose EMOSH, a novel framework that utilizes the Expressive Human Model for controllable human video generation. It delivers highly expressive human animation while strictly maintaining motion-shape disentanglement.
- We design a Confidence-Aware Motion Tracker to accurately capture expressive motions from videos, coupled with a Coarse-to-Fine Hybrid Motion Injection strategy that facilitates highly precise control.
- We propose a Spatially-Aligned Conditioning strategy to mitigate the training-inference domain gap. This mechanism reduces visual artifacts and improves identity preservation, setting a new SOTA over existing methods.

2 Related Work

2.1 Video Generation

Driven by the success of diffusion models [14, 60] in image synthesis [49, 50, 52], the video generation paradigm has rapidly shifted away from traditional GANs [6,65] and autoregressive models [66,75]. Early video diffusion models [15, 58] typically adapted pre-trained image models by inflating 2D U-Nets into 3D U-Nets and adding temporal layers. To further enhance visual quality, works like Imagen-Video [13] introduced cascaded frameworks, initially generating low-resolution videos and progressively upscaling them using spatial super-resolution and temporal interpolation models.

Recently, Diffusion Transformers (DiT) [42] have emerged as a powerful alternative to U-Net backbones, overcoming the scalability bottlenecks of convolutional networks and exhibiting strong scaling laws. Leveraging this architecture and massive video datasets, subsequent works [1, 11, 51, 86] have significantly scaled up video generation. These DiT-based models achieve high-quality, high-resolution long video synthesis with remarkable physical realism.

Within the DiT paradigm, various architectural designs have emerged. For instance, HunyuanVideo [24] trains a causal 3D VAE for efficient compression and employs a dual-stream DiT to process text and video tokens independently. Notably, it replaces standard factorized spatiotemporal attention with full 3D attention, while extending RoPE [62] to 3D to support dynamic resolutions and durations. In contrast, Wan2.1 [67] introduces Wan-VAE with caching mechanisms to alleviate memory bottlenecks during long video encoding. Structurally, it adopts a single-stream DiT that injects text via cross-attention and utilizes a shared MLP for timestep modulation to improve parameter efficiency.

Concurrently, numerous studies are dedicated to advancing specific dimensions of video generation, including consistent generation [7, 18, 25], acceleration [45, 82], duration extension [19, 77], and joint audio-video generation [8, 31].

2.2 Human Animation

Beyond general text-to-video, researchers have actively explored controllable human animation. Early works [34, 90] treated pose transfer as an image-to-image translation task, encoding the reference image into latent features and decoding them after injecting the target pose. However, these methods often struggled with large motions. Subsequently, GAN-based warping techniques became prominent for both facial animation [16, 44, 69] and body pose transfer [10, 28, 63, 71]. By predicting flow-based warping fields, these approaches spatially transform reference features to align with target poses. While effective at preserving texture details, they are prone to artifacts in occluded regions.

Leveraging the generative power of diffusion models, numerous works [35, 68, 74] have adopted ControlNet-like frameworks [80] to achieve controllable animation driven by 2D poses [3, 76]. Animate Anyone [17] pioneered the use of ReferenceNet to extract spatial details, injecting them into the main U-Net via attention to improve appearance preservation. In contrast, UniAnimate [70] simplified this by stacking the reference latent with noisy latents and utilizing Temporal Mamba for efficient long-sequence generation. To mitigate identity loss under extreme motions, StableAnimator [64] introduced a distribution-aware ID adapter, while MimicMotion [85] incorporated confidence-aware guidance to handle pose estimation errors. Similarly, HyperMotion [73] leverages a DiT-based framework with enhanced RoPE to further improve generation quality. Most recently, Wan-Animate [5] unified character animation on the Wan-14B [67] foundation, employing implicit encoding to replicate subtle expressions.

Despite significant progress, 2D Pose-based methods remain limited by motion-shape entanglement. To resolve this, several approaches [2, 53, 87, 89] employ SMPL-rendered [30] normal or depth maps as conditions. However, these explicit priors are often confined to coarse limb movements and fail to match the fine-grained control over expressions and gestures offered by 2D Pose. In contrast, our EMOSH extends the representation and controllability of 3D mesh, achieving both shape disentanglement and superior motion expressiveness.

Alternatively, methods based on 3D radiance fields (e.g., NeRF [37], 3DGS [22]) inherently maintain 3D consistency by driving reconstructed avatars. While of-

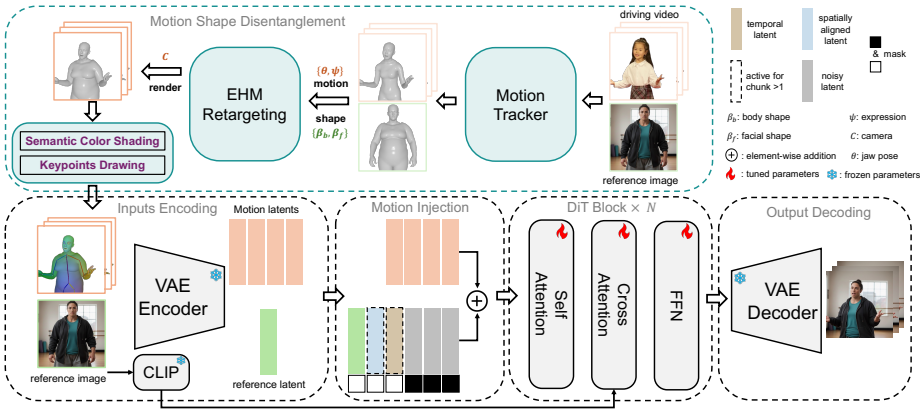


Fig. 2: First, the motion tracker extracts motion (θ^d, ψ^d) and camera (C^d) parameters from the driving video and shape parameters (β_b^r, β_f^r) from the reference image, achieving motion-shape disentanglement via EHM retargeting. The retargeted model is then rendered into hybrid control signals through semantic color shading and keypoint drawing, and encoded into motion latents. During the generation phase, the reference latent and noisy latent are concatenated; for subsequent video chunks (chunk > 1), the spatially-aligned and temporal latents are additionally activated and concatenated. The motion latents are injected into the main sequence via addition and fed into the DiT network for denoising, and are finally decoded into the output video.

fering superior efficiency, approaches based on per-subject optimization [43, 78] or feed-forward reconstruction [46, 79] often struggle to capture complex backgrounds and fine-grained dynamics, such as hair or clothing wrinkles. Consequently, they fall short of the visual realism provided by diffusion-based models.

3 Method

As illustrated in Fig. 2, we present the overview of EMOSH. We first review the preliminaries in Sec. 3.1. In Sec. 3.2, we introduce the Expressive Human Model (EHM) as our motion representation, alongside the design of a robust motion tracker. Sec. 3.3 elaborates on the Coarse-to-Fine Hybrid Motion Injection strategy, as well as our approach to disentangling motion from body shape. Finally, Sec. 3.4 describes the Spatially-Aligned Conditioning mechanism designed to enhance identity consistency, and details our training strategy.

3.1 Preliminaries

Video Generation. We adopt advanced Wan2.1-I2V [67] as our baseline framework. It employs a Diffusion Transformer (DiT) [42] to process spatiotemporal latents compressed by a 3D Causal VAE. Text and image semantics are extracted via T5 [48] and CLIP [47], respectively, and injected into the backbone

through cross-attention. To achieve long video generation, Wan2.1-I2V relies on an explicit latent conditioning strategy: the encoded reference image is concatenated with the noisy latent along the temporal dimension, while a binary mask is concatenated channel-wise to distinguish reference frames (value 1) from target generation regions (value 0).

Expressive Human Model (EHM). To address the limited facial expressiveness and head-body shape coupling in SMPL-X [40], we adopt EHM [79], a mesh-based parametric model integrating SMPL-X with FLAME [26]. It represents variations in human shape and expression within a linear space, explicitly decoupling identity into body shape $\beta_b \in \mathbb{R}^{|\beta_b|}$ and head shape $\beta_f \in \mathbb{R}^{|\beta_f|}$, while capturing facial dynamics via expression parameters $\psi \in \mathbb{R}^{|\psi|}$. Skeletal poses—encompassing body, hands, and facial joints—are defined by axis-angle rotations $\theta \in \mathbb{R}^{|\theta|}$. Using Linear Blend Skinning (LBS) for vertex deformation, EHM maps these parameters to a 3D mesh: $V = M_{ehm}(\beta_b, \beta_f, \psi, \theta)$, where $V \in \mathbb{R}^{N\times 3}$ denotes the deformed vertex coordinates.

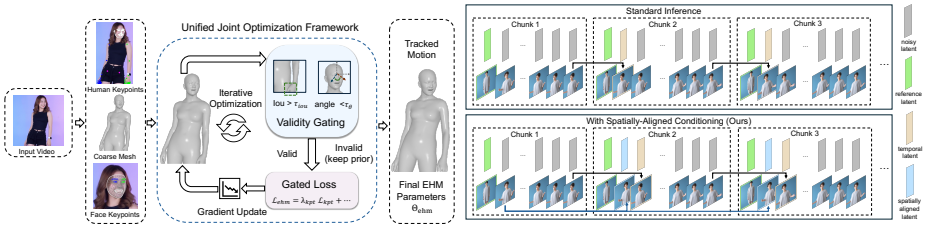
3.2 Expressive Motion Representation

Unlike previous methods [87, 89] restricted to coarse SMPL-based limb control [30], we introduce the highly expressive EHM alongside a robust motion tracker. This enables the accurate capture of facial expressions and hand poses, pushing the upper limits of mesh-based control to achieve both rich details and high-fidelity video generation.

Given a driving video, our goal is to estimate the per-frame EHM parameters $\Theta_{ehm} = \{\beta_b, \beta_f, \theta, \psi\}$ and camera parameter C. While GUAVA [79] offers a preliminary tracking workflow, its cascaded strategy—optimizing facial parameters before body pose—suffers from inefficiency and limitations in challenging scenarios, such as side or back views.

As illustrated in Fig. 3a, we propose a unified joint optimization framework that simultaneously recovers body, hand, facial, and camera parameters, simplifying the pipeline and improving tracking efficiency. The process follows a coarse-to-fine strategy: we first initialize the shape, pose, and camera parameters for each frame using off-the-shelf estimators. Then, leveraging differentiable rendering, we jointly fine-tune these parameters by minimizing the reprojection error $\mathcal{L}_{kpt}$ between the projected EHM vertices and detected 2D keypoints.

To address the instability of keypoint estimation and enhance robustness under complex motions (e.g., large-angle head turns, limb occlusions), we introduce a Confidence-Aware Validity Gating mechanism. Specifically, for hand tracking, we employ a dual-criteria check based on keypoint confidence and the projection IoU of the initial 3D mesh. The optimizer utilizes keypoint guidance only when hands are strictly visible; otherwise, it falls back to initial priors. For facial tracking, we compute the angle between the head orientation and camera view, dynamically masking the facial keypoint loss during large-angle or back-view scenarios to prevent structural collapse from forced alignment. Finally, incorporating the inter-frame smoothness $\mathcal{L}_{smooth}$, parameter regularization $\mathcal{L}_{reg}$, and



(a) Optimization pipeline of Motion Tracker. **(b)** Spatially-Aligned Conditioning strategy.

Fig. 3: (a) It extracts 2D/3D priors from the input video and dynamically filters unreliable guidance signals via Validity Gating, obtaining EHM parameters through joint iterative optimization. (b) For subsequent chunks in long video inference, the initial latent from the first chunk is injected as an additional spatially-aligned latent.

Fig. 4: Visual results of our motion tracker. Our tracker accurately extracts expressive motion parameters across diverse scenarios.

3D hand alignment $\mathcal{L}_{hand}^{3D}$, the overall joint optimization objective is defined as:

$$\mathcal{L}_{ehm} = \lambda_{kpt}\mathcal{L}_{kpt} + \lambda_{3d}\mathcal{L}_{hand}^{3D} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{smooth}\mathcal{L}_{smooth}. \tag{1}$$

As demonstrated in Fig. 4, our motion tracker enables accurate and stable motion capture across diverse and challenging scenarios.

3.3 Motion-Appearance Control

Hybrid Motion Injection. To achieve precise motion control, we design a Coarse-to-Fine Hybrid Motion Injection strategy based on the per-frame tracked EHM mesh. First, to enable the network to spatially distinguish different body regions, we employ a geometric semantic coloring scheme. Specifically, the normalized 3D coordinates of each vertex in a canonical T-pose are mapped to fixed RGB values V_{color}. Using a differentiable renderer $\mathcal{R}$, we then render these semantically colored meshes to generate dense 2D condition map for each frame.

However, while this rendering effectively represents macroscopic body structures, it provides predominantly low-frequency signals. In distant views or at low resolutions, the rendered maps often miss subtle high-frequency dynamics, such as mouth closures or blinking. To address this, we introduce sparse 2D keypoints to provide complementary high-frequency details. Specifically, we select a set of key semantic vertices V_{kpt} from the mesh and explicitly draw them onto the rendered map using distinct colors, allowing the model to perceive these detailed

changes. The final condition map I_{mesh} is thus formulated as:

$$I_{mesh} = \mathcal{P}\Big(\mathcal{R}(M_{ehm}, V_{color}, C), V_{kpt}\Big) \ , \tag{2}$$

where $\mathcal{P}$ denotes the keypoint drawing operation. Finally, I_{mesh} is processed by a VAE encoder and a lightweight embedding layer. The extracted features are directly added to the noisy video latent, providing explicit motion guidance during the denoising process.

Identity Injection. To effectively inject appearance information from the reference image, we follow the explicit latent conditioning of Wan2.1-I2V [67]. Specifically, the VAE-encoded reference latent is temporally concatenated with the noisy video latents, with its corresponding mask set to 1 to explicitly designate it as the visual context.

Motion-Shape Disentanglement. During inference, the reference image and the driving video often originate from distinct subjects. Directly applying the tracked mesh of the driving video would inevitably leak the driving subject's body shape into the generated video (e.g., imposing a robust physique onto a slender reference). To address this, we implement an explicit motion-shape disentanglement strategy. Specifically, we retarget the driving mesh by retaining the pose and expression from the driving video, while injecting the shape parameters extracted from the reference image. The retargeted mesh M_{ehm}^{target} is formalized as:

$$M_{ehm}^{target} = M_{ehm}(\theta^d, \psi^d, \beta_b^r, \beta_f^r) \ , \tag{3}$$

where θ^d and ψ^d denote the pose and expression parameters from the driving video, respectively, while β_b^r and β_f^r represent the body and facial shape parameters estimated from the reference image. Additionally, the camera parameter C^d from the driving video is applied during rendering. Through this parameter recombination, we ensure the generated motion remains faithful to the driving video, while the physical shape is strictly anchored to the reference subject.

3.4 Inference and Training

Long Video Inference. Due to GPU memory constraints, generating long videos in a single pass is intractable. We thus adopt an autoregressive strategy based on overlapping frames. Specifically, the last five frames of the preceding chunk are encoded into temporal latents and serve as the temporal context for the subsequent chunk. These latents are concatenated after the reference latent with their mask values set to 1. To ensure character consistency across the entire video, following [45, 54], we fix the original reference latent as a "visual anchor" at the beginning of the latent sequence for every generated chunk.

Spatially-Aligned Conditioning. We identify the spatial misalignment between training and inference as a critical factor contributing to identity degradation in long video generation. During training, the reference image and target frames originate from the same video, exhibiting high spatial alignment. Conversely, cross-identity inference introduces spatial discrepancies between the reference image and driving motion. This training-inference gap biases the model

to attend more to the spatially aligned temporal latents, gradually neglecting the information in the reference latent. Consequently, as autoregressive errors accumulate, the generated video suffers from visual artifacts and identity drift.

To mitigate this issue, we propose a Spatially-Aligned Conditioning strategy, as illustrated in Fig. 3b. After generating the first video chunk, we retain the latent of its initial frame. For all subsequent chunks, this spatially-aligned latent is inserted between the original reference latent and the temporal latents, with its mask set to 1. Acting as a structural bridge, it preserves both the original identity and the spatial layout of the target scene. Furthermore, since this latent accumulates fewer errors than the recursively generated temporal latents, it acts as a reliable visual anchor to redistribute the model's attention. This prevents over-reliance on noisy temporal cues, thereby significantly enhancing appearance consistency and reducing artifacts in long video generation.

Training. To adapt the model for both cold-start and autoregressive generation, we probabilistically toggle the use of temporal latents. Additionally, to simulate spatial misalignment, we randomly sample reference frames from the video, apply simple geometric augmentations (e.g., rotation, translation), and randomly inject an auxiliary reference latent, mirroring the Spatially-Aligned Conditioning setup used during inference. Finally, we train our model using the standard Flow Matching [27] objective with an Optimal Transport path.

4 Experiment

4.1 Setup

Implementation Details. Our model is implemented in PyTorch [39] and trained using Adam [23]. Training is conducted on 32 NVIDIA H20 GPUs with a global batch size of 32. The entire training process spans approximately 100 hours, covering roughly 6,500 iterations to reach convergence. During training, we randomly sample 77-frame video clips at a resolution of 512×512, paired with a reference image as the conditioning input. Our backbone adopts a DiT [42] architecture consistent with Wan2.1-I2V [67]. Since Wan-Animate [5] shares this structure and has been pre-trained on large-scale human videos, we initialize our DiT module with its weights to leverage this strong prior. For the Motion Tracker, we utilize GVHMR [55], TEASER [29], and HaMeR [41] for the coarse estimation of body pose, facial expressions, and hand parameters, respectively. DWPose [76] and MediaPipe [32] are employed to extract body and facial keypoints for tracking refinement.

Training Dataset. We construct a composite dataset comprising approximately 900k video clips. The core data originates from human motion videos (37.8%) crawled from the Internet and the Speaker-Vid [84] dataset (60.6%). To ensure high visual quality and motion dynamics, we rigorously filter these sources by discarding clips with undetected faces, insufficient face scales, low hand visibility, or static subjects. Finally, to enhance facial animation, we incorporate the VFHQ [72] dataset (1.6%) into our final training corpus.

Table 1: Quantitative results on three datasets under the self-driven setting. Our method achieves the best performance across all evaluation metrics.

Method	EchoMimicV2 dataset				TikTok dataset				Self-collected dataset				condition
	PSNR↑	L1↓	SSIM↑	LPIPS↓	PSNR↑	L1↓	SSIM↑	LPIPS↓	PSNR↑	L1↓	SSIM↑	LPIPS↓	
Moore-AA	20.39	0.0513	0.7388	0.2059	16.23	0.1022	0.6742	0.3263	18.11	0.0762	0.6464	0.2788	2D Pose
StableAnimator	16.76	0.0796	0.6724	0.2836	11.76	0.1845	0.5732	0.4437	15.97	0.1024	0.6179	0.3285	2D Pose
MimicMotion	19.52	0.0622	0.7211	0.2450	14.17	0.1335	0.3876	0.6316	16.00	0.1057	0.6179	0.3552	2D Pose
UniAnimate	16.39	0.1045	0.6933	0.2678	11.73	0.1939	0.6096	0.4320	14.69	0.1286	0.6048	0.3543	2D Pose
UniAnimate-DiT	17.90	0.0678	0.7266	0.2533	12.79	0.1571	0.6151	0.4213	14.76	0.1302	0.5778	0.3688	2D Pose
HyperMotion	22.32	0.0391	0.7975	0.1800	15.48	0.1154	0.6529	0.3536	19.48	0.0660	0.7082	0.2512	2D Pose
Wan-Animate	22.86	0.0379	0.7750	0.1802	17.18	0.0932	0.6808	0.3209	20.72	0.0555	0.7281	0.2229	2D Pose & Face
Champ	17.40	0.0756	0.6502	0.2911	13.34	0.1482	0.5902	0.4156	14.85	0.1159	0.5602	0.3727	Mesh
EMOSH (Ours)	**23.66**	**0.0308**	**0.8240**	**0.1428**	**17.80**	**0.0831**	**0.740**	**0.2933**	**21.18**	**0.0522**	**0.751**	**0.2066**	Mesh

Table 2: Comparison of Identity Preservation Score (IPS) in the cross-driven setting. Our method better maintains ID consistency.

	Moore-AA	Champ	StableAnimator	MimicMotion	UniAnimate	UniAnimate-DiT	HyperMotion	Wan-Animate	EMOSH (Ours)
IPS↑	0.2174	0.2619	0.0935	0.0744	0.1618	0.3130	0.2872	0.3802	**0.4445**

Baselines. We compare EMOSH against several state-of-the-art (SOTA) human animation methods, including Wan-Animate [5], HyperMotion [73], MimicMotion [85], Moore-AA (AnimateAnyone by Moore) [38], StableAnimator [64], UniAnimate-DiT and UniAnimate [70], and Champ [89]. To ensure a fair comparison, we conduct inference for all baseline models at their native resolutions and subsequently evaluate metrics on a unified spatial scale.

Evaluation Protocols. We evaluate our model under two distinct settings: Self-driven and Cross-driven. *Self-driven:* The driving video and the reference image originate from the same video. We utilize the public benchmarks EchoMimicV2 [36] and TikTok [20] datasets, alongside a self-collected test set containing 150 videos (∼55k frames). For evaluation, the first frame of the video serves as the reference image, and the model reconstructs the full video sequence based on the driving signals. To quantitatively evaluate the reconstruction quality, we employ PSNR, L1, SSIM, and LPIPS [81] to compare the generated videos against the ground truth. *Cross-driven:* We collect 12 reference images covering diverse body shapes (e.g., varying heights and weights) and 10 driving videos featuring distinct motions. These are cross-paired to generate a total of 120 video samples (∼51k frames). Due to the absence of Ground Truth, we primarily rely on Human Evaluation. Additionally, we use the Identity Preservation Score (IPS) as a supplementary metric. Specifically, we utilize ArcFace [9] to extract facial features from both the generated video and the reference image, calculating the average cosine similarity between them.

4.2 Quantitative Results

Self-driven. Tab. 1 summarizes the quantitative comparisons between our method and the baseline models across three datasets. Our method achieves the best performance across all metrics. This demonstrates that our approach can generate

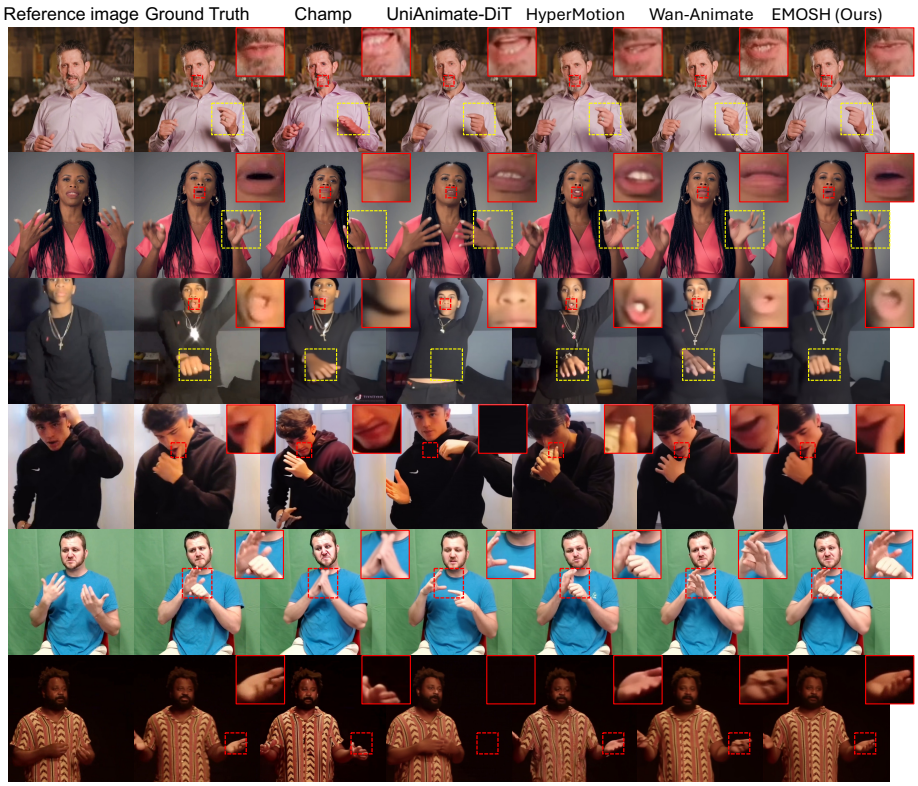


Fig. 5: Qualitative results on three datasets under the self-driven setting. EMOSH generates higher-fidelity videos and achieves more accurate control of facial expressions and hand gestures compared to the baselines.

higher-fidelity videos conditioned on the driving video and reference image while maintaining precise motion control.

Cross-driven. Tab. 2 presents the Identity Preservation Score (IPS) under the Cross-ID setting. The results show that our approach achieves the highest IPS, verifying its superior robustness in preserving the identity features of the reference image under diverse driving poses.

4.3 Qualitative Results

Self-driven. As illustrated in Fig. 5, we present the qualitative comparison of various methods across three datasets under the self-driven setting. While most approaches can reasonably preserve the identity of the reference image, they struggle with fine-grained details. Specifically, Champ's reliance on a human mesh control mechanism makes it difficult to achieve fine-grained control over expressions and gestures. UniAnimate-DiT, although utilizing 2D poses, lacks facial keypoints, leading to inaccurate expression control. Moreover, it may strug-

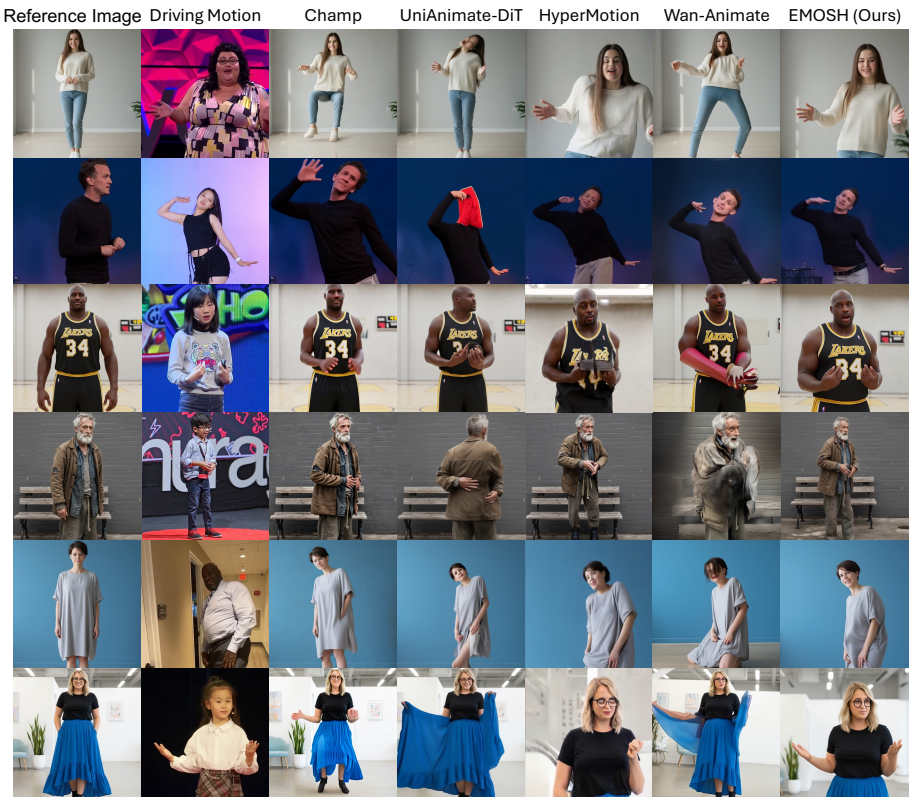


Fig. 6: Visual quality comparison under the cross-driven setting. While achieving precise motion control, our method better preserves the identity and body shape characteristics of the reference subject, effectively avoiding the shape distortion and limb artifacts seen in other methods.

gle with complex body poses and gestures, resulting in failure on some instances. HyperMotion and Wan-Animate produce videos of higher overall quality and can generally control expressions and gestures. However, they still suffer from incorrect mouth shapes, and blurry or merged finger artifacts. In contrast, although our method is mesh-guided, it achieves control performance that is comparable to, or even surpasses, methods relying on fine-grained 2D poses. This demonstrates our method generates videos with superior accuracy and higher fidelity.

Cross-driven. As shown in Fig. 6, we present qualitative results under the cross-driven setting. Some baselines struggle to preserve the reference identity, causing shape distortions like unnatural thinning or widening. Some also fail to follow the driving video's camera framing, rigidly keeping the original body scale. Specifically, Champ fails to accurately control expressions and gestures. UniAnimate-DiT struggles with expression driving and frequently generates merged finger artifacts. HyperMotion improves expression control but lacks ID preservation,

Table 3: Human evaluation results are based on the standard GSB (Good/Same/Bad) protocol for pairwise comparisons between our approach and three baseline methods. The values and colored bars indicate the percentage of user preferences for **Ours**, **Both (Same)**, and **Competitor**.

Ours vs Wan-Animate			**Ours vs HyperMotion**			**Ours vs UniAnimate-DiT**		
89.21%	3.96%	**6.83%**	**98.42%**	0.75%	**0.83%**	**95.62%**	1.80%	**2.58%**

Table 4: Quantitative results of our ablation study. (Left) Under the self-driven setting, removing the Tracker or Hybrid Motion leads to a significant drop in generation metrics. (Right) Under the cross-driven setting, removing Disentanglement or Spatially-Aligned Conditioning (SAC) degrades the identity preservation score.

Method	EchoMimicV2 dataset				Self-collected dataset			
	PSNR↑	L1↓	SSIM↑	LPIPS↓	PSNR↑	L1↓	SSIM↑	LPIPS↓
w/o Tracker	20.51	0.0434	0.7789	0.1909	16.48	0.0959	0.7095	0.3300
w/o Hybrid Motion	22.90	0.0345	0.8127	0.1492	17.55	0.0850	0.7330	0.2973
Full (Ours)	**23.66**	**0.0308**	**0.8240**	**0.1428**	**17.80**	**0.0831**	**0.740**	**0.2933**

Method	IPS↑
w/o Disentanglement	0.4258
w/o SAC	0.4237
Full (Ours)	**0.4445**

leading to appearance deformations. Although Wan-Animate uses 2D pose retargeting to mitigate shape leakage, it still generates disproportionate bodies, or even abnormal limb lengths with severe artifacts in some cases. In contrast, our EMOSH demonstrates remarkable superiority. It perfectly disentangles and preserves the reference's authentic body shape and identity while achieving highly accurate control over poses, expressions, and camera framing.

4.4 Human Evaluation

To further subjectively assess visual performance in the cross-driven setting, we conducted a user study comparing EMOSH against recent DiT-based state-of-the-art open-source baselines: Wan-Animate, HyperMotion, and UniAnimate-DiT. We designed a pairwise comparison test with 20 participants. In each trial, participants viewed a reference image, a driving video, and two anonymized, randomly ordered generated videos. They were instructed to select their preferred result based on three criteria: video generation quality, motion and expression accuracy, and identity preservation. As reported in Tab. 3, EMOSH consistently obtains the highest user preference, further validating that our approach produces superior video results that better align with human visual perception.

4.5 Ablation Studies

To verify the contribution of each core component, we perform ablation experiments under the self-driven and cross-driven settings, with quantitative and qualitative results detailed in Tab. 4 and Fig. 7.

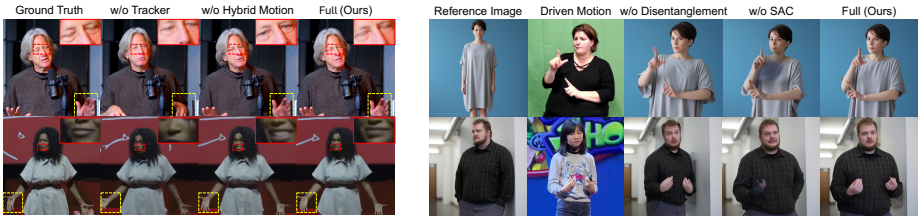


Fig. 7: Qualitative ablation results. (Left) Under the self-driven setting, full method enables more accurate control over expressions and gestures. (Right) Under the cross-driven setting, full method disentangles shape and motion, whereas removing Spatially-Aligned Conditioning (SAC) leads to artifacts and reduces identity preservation.

w/o Tracker. By removing the proposed Motion Tracker, this baseline relies on off-the-shelf models for motion estimation, akin to the strategy used in Champ. As illustrated in Fig. 7, it struggles to accurately control facial expressions and intricate hand gestures. Furthermore, it suffers a significant performance drop across all metrics in the self-driven setting, as shown in Tab. 4.
w/o Hybrid Motion. This baseline discards 2D keypoints drawing operation and depends on naive mesh renderings for motion conditioning. Visually, it struggles to animate facial expressions like blinking, and its gesture control granularity falls short of the full method. Additionally, its quantitative metrics under the self-driven setting drop significantly.
w/o Disentanglement. By removing the EHM Retargeting, this baseline discards the motion-shape disentanglement. Without this, the model suffers from shape leakage and fails to preserve the reference body shape. This directly translates to a noticeable drop in identity consistency, reflected by the lower IPS score.
w/o SAC. This variant removes the Spatially-Aligned Conditioning (SAC) mechanism. Consequently, the model produces artifacts during video generation. Furthermore, it weakens the model's capability to preserve identity features, as evidenced by the degraded IPS metric.

5 Conclusion

In this paper, we propose EMOSH, a high-fidelity controllable human video generation framework that achieves expressive motion and shape disentanglement. By introducing the Expressive Human Model as the core motion control representation, we explicitly disentangle the motion from the body shape in the driving signal. Coupled with our proposed Confidence-Aware Motion Tracker, EMOSH enables fine-grained control over human poses, facial expressions, and complex hand gestures in video generation. This successfully overcomes the inherent bottleneck of traditional mesh-guided generation methods, which typically struggle to achieve precise control over facial and hand movements. To further enhance control precision, we propose a Coarse-to-Fine Hybrid Motion Injection strategy. Finally, our Spatially-Aligned Conditioning mechanism effectively bridges

the domain gap between training and inference, significantly reducing visual artifacts and enhancing identity preservation in cross-driven scenarios. Extensive experiments demonstrate the superiority of EMOSH in control precision, identity preservation, and overall visual quality. Further discussions about the limitations of our method are provided in the supplementary material.

Acknowledgements

The authors are grateful to the user study participants for their valuable feedback and subjective assessments. This research was funded by the National Key Research and Development Program of China (Project No.2022YFB36066) and partially funded by the Shenzhen Science and Technology under Grant (KJZD20240903103210014, JCYJ20220818101001004).

References

1. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
2. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
4. Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. arXiv preprint arXiv:2311.12052 (2023)
5. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025)
6. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571 (2019)
7. DeepMind, G.: Genie 3: A new frontier for world models (2025), `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`, accessed: 2026-02-15
8. DeepMind, G.: Veo 3 (2025), `https://deepmind.google/technologies/veo/`, accessed: 2026-02-15
9. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
10. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. Advances in neural information processing systems **31** (2018)
11. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)

12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
16. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
17. Hu, L.: 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. pp. 8153–8163 (2024)
18. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
19. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
20. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12753–12762 (2021)
21. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
25. Li, J., Tang, J., Xu, Z., Wu, L., Zhou, Y., Shao, S., Yu, T., Cao, Z., Lu, Q.: Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. arXiv preprint arXiv:2506.17201 **2**(3), 6 (2025)
26. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813
27. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
28. Liu, W., Piao, Z., Tu, Z., Luo, W., Ma, L., Gao, S.: Liquid warping gan with attention: A unified framework for human image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5115–5133 (2021)
29. Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., Li, Y.: Teaser: Token enhanced spatial modeling for expressions reconstruction. arXiv preprint arXiv:2502.10982 (2025)

30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
31. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284 (2025)
32. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
33. Luo, Y., Rong, Z., Wang, L., Zhang, L., Hu, T.: Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11036–11046 (2025)
34. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. Advances in neural information processing systems **30** (2017)
35. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
36. Meng, R., Zhang, X., Li, Y., Ma, C.: Echomimicv2: Towards striking, simplified, and semi-body human animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5489–5498 (2025)
37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
38. Moore Threads: Moore-animateanyone. `https://github.com/MooreThreads/Moore-AnimateAnyone` (2024), accessed: 2026-02-15
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
40. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
41. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024)
42. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
43. 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. pp. 20299–20309 (2024)
44. Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., Wang, H.: Geometry-contrastive gan for facial expression transfer. arXiv preprint arXiv:1802.01822 (2018)
45. Qiu, H., Liu, S., Zhou, Z., An, Z., Ren, W., Liu, Z., Schult, J., He, S., Chen, S., Cong, Y., et al.: Histream: Efficient high-resolution video generation via redundancy-eliminated streaming. arXiv preprint arXiv:2512.21338 (2025)
46. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model for single

image to 3d in seconds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14184–14194 (2025)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
48. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
49. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
50. 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. pp. 10684–10695 (2022)
51. Runway: Introducing runway gen-4.5: A new frontier for video generation. `https://runwayml.com/research/introducing-runway-gen-4.5` (2025), accessed: 2026-02-15
52. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
53. Shao, R., Pang, Y., Zheng, Z., Sun, J., Liu, Y.: 360-degree human video generation with 4d diffusion transformer. ACM Transactions on Graphics (TOG) **43**(6), 1–13 (2024)
54. Shen, L., Qiao, Q., Yu, T., Zhou, K., Yu, T., Zhan, Y., Wang, Z., Tao, M., Yin, S., Liu, S.: Soulx-flashtalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation (2025), `https://arxiv.org/abs/2512.23379`
55. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
56. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
57. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3408–3416 (2018)
58. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
59. Song, G., Xu, H., Zhao, X., Xie, Y., Gu, T., Li, Z., Zhang, C., Luo, L.: X-unimotion: Animating human images with expressive, unified and identity-agnostic motion latents. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
60. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
61. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

62. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
63. Sun, Y.T., Fu, Q.C., Jiang, Y.R., Liu, Z., Lai, Y.K., Fu, H., Gao, L.: Human motion transfer with 3d constraints and detail enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4682–4693 (2022)
64. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z.: Stableanimator: High-quality identity-preserving human image animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21096–21106 (2025)
65. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
66. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
67. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
68. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for realistic human dance generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9326–9336 (2024)
69. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10039–10049 (2021)
70. Wang, X., Zhang, S., Gao, C., Wang, J., Zhou, X., Zhang, Y., Yan, L., Sang, N.: Unianimate: Taming unified video diffusion models for consistent human image animation. Science China Information Sciences **68**(10), 200103 (2025)
71. Wei, D., Xu, X., Shen, H., Huang, K.: Gac-gan: A general method for appearance-controllable human video motion transfer. IEEE Transactions on Multimedia **23**, 2457–2470 (2020)
72. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 657–666 (2022)
73. Xu, S., Zheng, S., Wang, Z., Yu, H., Chen, J., Zhang, H., Li, B., Jiang, P.T.: Hypermotion: Dit-based pose-guided human image animation of complex motions. arXiv preprint arXiv:2505.22977 (2025)
74. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
75. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
76. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
77. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
78. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Chen, K., Qin, M., Li, Y., Wang, H.: Hravatar: High-quality and relightable gaussian head avatar. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26285–26296 (2025)

79. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14205–14217 (2025)
80. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
81. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
82. Zhang, S., Li, W., Chen, S., Ge, C., Sun, P., Zhang, Y., Jiang, Y., Yuan, Z., Peng, B., Luo, P.: Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. arXiv preprint arXiv:2502.05179 (2025)
83. Zhang, S., Zhuang, J., Zhang, Z., Shan, Y., Tang, Y.: Flexiact: Towards flexible action control in heterogeneous scenarios. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
84. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025)
85. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv preprint arXiv:2406.19680 (2024)
86. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
87. Zhou, J., Wang, B., Chen, W., Bai, J., Li, D., Zhang, A., Xu, H., Yang, M., Wang, F.: Realisdance: Equip controllable character animation with realistic hands. arXiv preprint arXiv:2409.06202 (2024)
88. Zhou, J., Wu, Y., Li, S., Wei, M., Fan, C., Chen, W., Jiang, W., Wang, F.: Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. arXiv preprint arXiv:2504.14977 (2025)
89. Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: European Conference on Computer Vision. pp. 145–162. Springer (2024)
90. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2347–2356 (2019)