This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InfinityHuman: Towards Long-Term Audio-Driven Human Animation

Xiaodi Li[*1,2], Pan Xie[*1], Yi Ren[*2], Qijun Gan[*1,2],
Chen Zhang[2], Fangyuan Kong[1], Xiang Yin[1,†], Zehuan Yuan[1], Bingyue Peng[1]
[1]ByteDance [2]Zhejiang University
[*]Equal Contribution [†]Corresponding Author

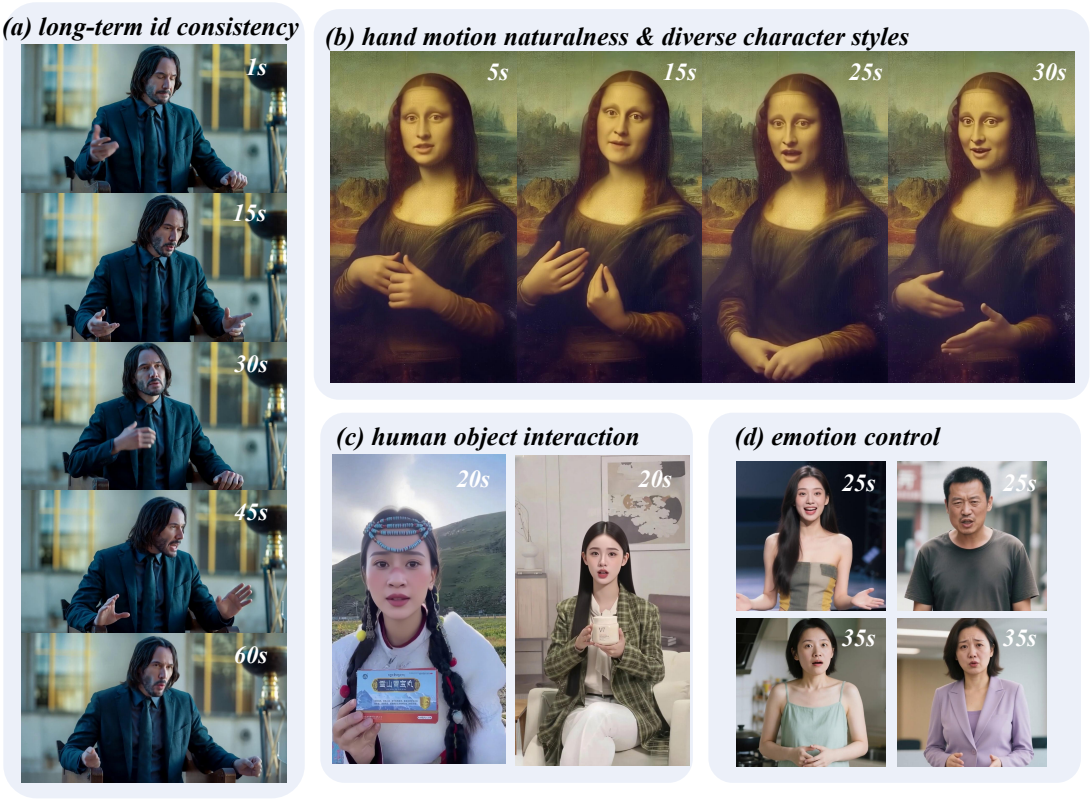

Figure 1. **InfinityHuman** is an audio-driven full-body animation framework that synthesizes long-duration videos with (a) temporally consistent visual appearance, (b) expressive and style-rich hand gestures, (c) dynamic human-object interactions, and (d) emotion-controllable, audio-aligned full-body motions.

Abstract

Audio-driven human animation has attracted wide attention thanks to its practical applications. However, critical challenges remain in generating high-resolution, long-duration videos with consistent appearance and natural hand motions. Existing methods extend videos using overlapping motion frames but suffer from error accumulation, leading to identity drift, color shifts, and scene instability. Additionally, hand movements are poorly modeled, resulting in noticeable distortions and misalignment with the audio. In this work, we propose InfinityHuman, a coarse-to-fine framework that first generates audio-synchronized representations, then progressively refines them into high-resolution, long-duration videos using a pose-guided refiner. Since pose sequences are decoupled from appearance and resist temporal degradation, our pose-guided refiner employs stable poses and the initial frame as a visual anchor to reduce drift and improve lip synchronization. Moreover, to enhance semantic accuracy and gesture realism, we introduce a hand-specific reward mechanism trained with high-quality hand motion data. Experiments on the EMTD and HDTF datasets show that InfinityHuman achieves state-of-the-art performance in video quality, identity preservation, hand accuracy, and lip-sync. Ablation studies further confirm the effectiveness of each module. And our project page is available at https://infinityhuman.github.io/.

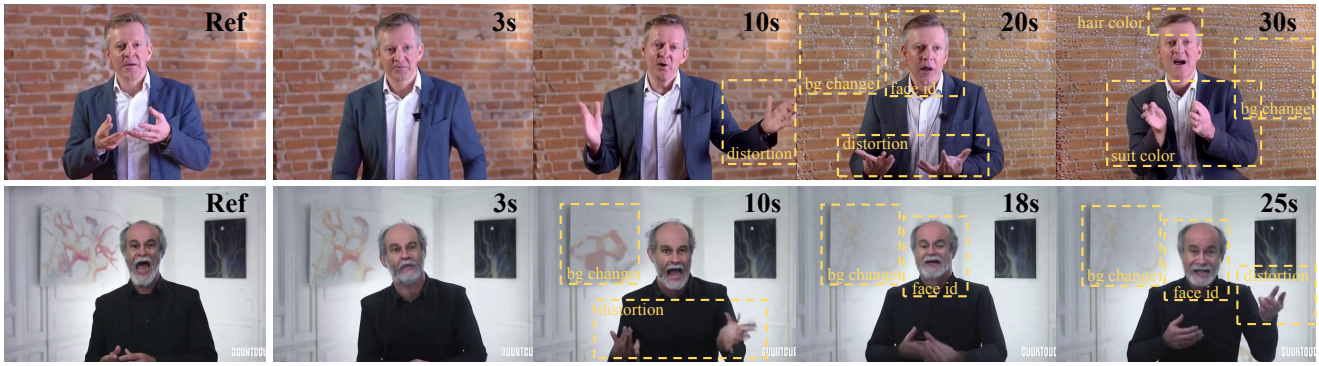

Figure 2. **Progressive Degradation in Long Video Animation by Previous Methods.** Existing methods suffer from cumulative errors leading to pronounced identity drift (facial inconsistencies), color shifts (hair, clothing), scene instability (background fluctuations), and hand motion artifacts. These challenges underscore the necessity of InfinityHuman's pose-guided refiner and hand-specific optimization for producing high-fidelity, temporally coherent animations over extended sequences.

1. Introduction

Audio-driven character animation aims to generate realistic human videos from a single image and audio input, transforming static portraits into speaking characters. This technology holds significant potential across various industries, including advertising, vlogging, and film production. With the rapid advancement of video generation models, recent research [8, 14, 22, 23, 30, 32, 41] has progressed from driving facial and head movements to full-body animation, greatly enhancing the expressiveness and richness of generated content.

Despite notable progress in full-body human animation, critical challenges remain in generating high-resolution, long-duration, and naturally coherent videos. These challenges can be grouped into two main areas: **i) Long-Term Visual Consistency**: Existing methods [5, 8, 10, 20, 23, 32] typically extend video sequences using overlapping motion frames. However, as sequence length increases, accumulated errors undermine visual coherence, resulting in progressive degradation. This degradation manifests in three key aspects: inconsistent character identity (e.g., variations in facial proportions or clothing); global color incoherence (e.g., erratic shifts in tone or brightness); and scene instability (e.g., shifting or disappearing background objects). **ii) Hand Motion Naturalness**: Prior work has predominantly focused on facial naturalness and coarse body movements, neglecting the nuanced handling of hand motions—small-range yet high-speed movements. Consequently, large hand gestures frequently lead to distortions or artifacts, and misalignment between hand movements and audio further diminishes the expressiveness and realism of generated videos.

To address the aforementioned limitations, we propose **InfinityHuman**, a novel coarse-to-fine generation framework. This framework first produces low-resolution motion frames synchronized with audio, and subsequently outputs high-resolution long-form videos via a dedicated Refiner.

Our method introduces innovations in two key aspects. First, we design a **pose-guided refiner** to address visual drift in long-duration sequences. Given that pose sequences are structurally decoupled from visual appearance, they inherently resist temporal degradation in appearance-related features. Consequently, we use them as reliable conditioning signals. In addition, during continuous continuation, we incorporate the initial frame as a visual anchor to further enhance temporal consistency. This combination offers both dynamic guidance for maintaining temporal coherence and a reference for visual fidelity. Furthermore, compared vanilla diffusion-based super-resolution, the pose signal provides strong anatomical structure and preserves fine-grained motion patterns. This enables more accurate lip-syncing while effectively reducing common artifacts such as motion distortions and finger overlap in diffusion-based super-resolution.

Secondly, considering that the human visual system is highly sensitive to hand distortions such as incorrect finger count, unnatural joint movements, we adopt a **hand-specific reward feedback mechanism** and incorporate high-quality hand motion data during training to guide hand generation. The mechanism encourages the model to produce temporally consistent and correct gestures, thereby enhancing character expressiveness and the realism of the video.

We evaluate InfinityHuman on the EMTD [25] and HDTF [42] datasets, covering long-duration upper-body and talking-head scenarios. Qualitative and quantitative results show it achieves SOTA performance in video quality, id preservation, hand accuracy, and lip-sync. Ablation studies further confirm the effectiveness of our proposed model. Our contributions are summarized as follows:

- We propose InfinityHuman, a coarse-to-fine generation framework specifically designed to address the chal-

lenges of visual realism and temporal consistency in long-duration audio-driven character animation.
- We develop a pose-guided refiner that leverages stable pose sequences and the initial frame as a visual anchor to correct accumulated errors, maintain lip-sync accuracy, and reduce artifacts in extended video sequences.
- To improve hand movement realism and expressiveness, we introduce a hand-specific reward feedback mechanism, integrated with high-quality hand motion data.
- Comprehensive experiments on EMTD and HDTF datasets demonstrate that InfinityHuman outperforms state-of-the-art methods across multiple metrics.

2. Related work

Long Video Generation Existing methods [1, 4, 11, 31] extend video diffusion models to longer durations by modifying objectives or architectures. Autoregressive pipelines and memory modules [1, 11] improve cross-segment consistency but require costly retraining on curated long-video datasets. In contrast, training-free extensions such as Gen-L-Video [31] and FreeNoise [27] improve efficiency via sliding-window attention and noise rescheduling. However, they offer limited temporal modeling, often causing temporal drift and less coherent transitions between segments. To balance quality and efficiency, recent works [21, 28, 36, 39] fine-tune short-video diffusion models with previous motion frames as conditions for autoregressive continuation. Despite their flexibility, these methods suffer from error accumulation at inference, leading to degraded fidelity and identity shifts. We adopt a similar strategy but address its limitations with a coarse-to-fine two-stage framework. A low-resolution long video is first generated, followed by a pose-guided refiner that corrects artifacts and restores spatial-temporal consistency, yielding high-resolution, identity-consistent long videos.

Audio-driven character animation. Recent advancements in audio-driven character animation have significantly improved lip-syncing and facial expression modulation using latent diffusion models. Works such as SadTalker [41] and Hallo [35] enhance audio-to-facial synchronization with 3D rendering and diffusion techniques, while V-Express [30] and EchoMimic [25] refine naturalness by integrating audio with facial landmarks and control signals. Loopy [17] and OmniHuman-1 [23] ensure identity consistency and mitigate data scarcity through multimodal training. Recent works have extended to full-body animation, with DiffTED [14] introducing a one-shot framework for synchronized talking head and gesture animations, and CyberHost [22] enhancing video quality using identity-independent features and human priors. Despite these advancements, generating high-resolution, long-duration, and natural videos remains a significant challenge, particularly in maintaining long-term identity consistency and ensuring the naturalness of hand motions. However, our Infinity Human leverages a pose-guided refiner and hand correction strategies to address these issues.

3. Methodology

Overview. As shown in Figure 3, InfinityHuman is a unified framework designed to generate long-duration, full-body talking high-resolution videos V_{hr} from a single reference image I_{ref}, audio $\mathbf{c}_{\text{audio}}$, and an optional text prompt ($\mathbf{c}_{\text{text}}$), ensuring visual consistency, precise lip-sync, and natural hand movements. The framework adopts a coarse-to-fine strategy, starting with **Low-Resolution Audio-to-Video**(§3.1) to produce coarse motion in V_{lr}, followed by **Pose-Guided Refiner**(§3.2) to generate high-resolution video V_{hr} conditioned on V_{lr} and I_{ref}. Additionally, **Hand Correction Strategies**(§3.3) are introduced to enhance the realism and structural integrity of hand movements.

3.1. Low-Resolution Audio-to-Video

Training Objective. We adopt Flow Matching [24] to train the low-resolution audio-to-video generation (LR-A2V). This approach enables efficient simulation of continuous-time dynamics by learning to predict the data's velocity field. The backbone of our method is a DiT [26], denoted as f_θ, which takes a noisy latent representation as input for all frames z^{lr}, along with conditioning information from multiple modalities: a reference image I_{ref}, text condition c_{text}, audio condition c_{audio}, and a continuous time step $t \in [0, 1]$. The low-resolution latent video $z^{\text{lr}} = \{z_i^{\text{lr}}\}_{i=0}^{f} \in \mathbb{R}^{(f+1)\times h\times w\times c}$ is produced by encoding the input video V_{lr} using a 3D VAE encoder.

To construct training samples, Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$ is sampled independently for each latent, and the noisy latent at diffusion time t for latent i is obtained by the diffusion process:

$$z_{i,t}^{\text{lr}} = \phi(z_i^{\text{lr}}, t) = (1-t)\cdot\epsilon_i + t\cdot z_{i,1}^{\text{lr}} \tag{1}$$

The target velocity is then defined as:

$$v_{i,t} = \frac{dz_{i,t}^{\text{lr}}}{dt} = z_{i,1}^{\text{lr}} - \epsilon_i \tag{2}$$

The DiT model is trained to predict these velocities for all frames jointly. The training objective minimizes the expected squared error:

$$\mathcal{L} = \mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,I),\, t\sim\mathcal{U}(0,1)} \Big\| f_\theta\big(\{z_{i,t}^{\text{lr}}\}_{i=0}^{f},\, I_{\text{ref}},\, c_{\text{text}},\, c_{\text{audio}},\, t\big) - \{v_{i,t}\}_{i=0}^{f} \Big\|_2^2 \tag{3}$$

Multimodal Condition Attention. To improve the incorporation and alignment of audio information, we decouple the audio condition from other modalities by introducing a

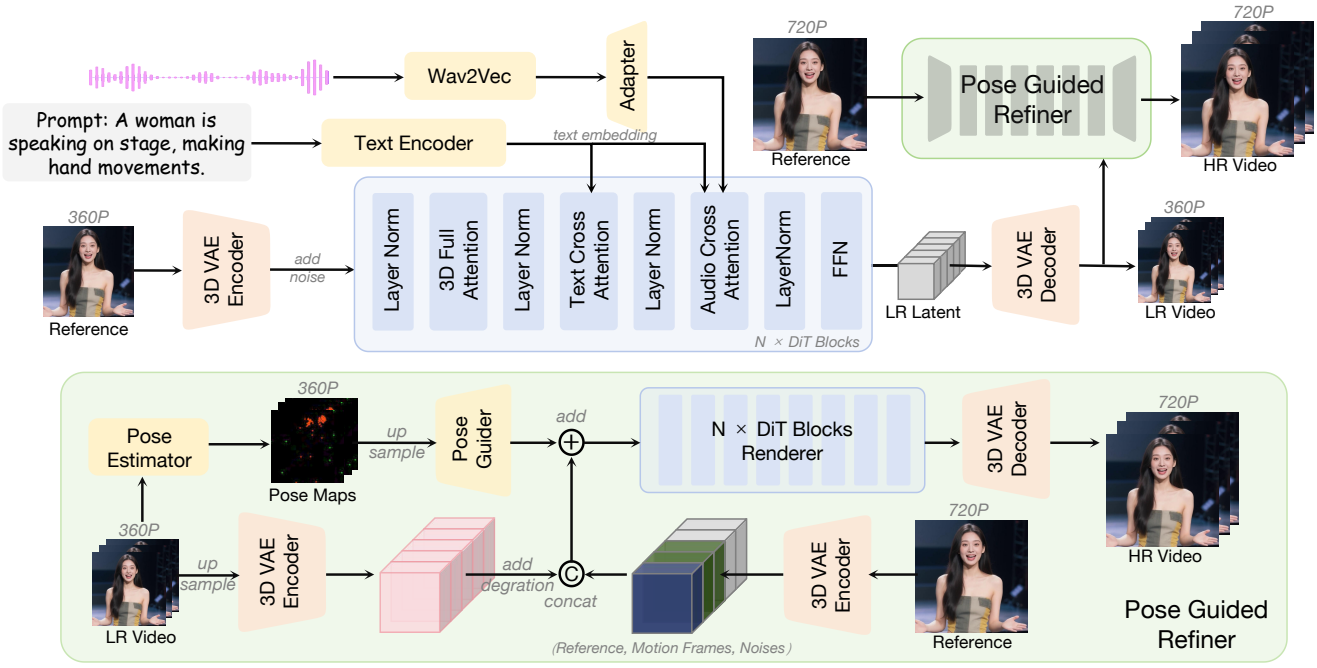

Figure 3. **InfinityHuman Pipeline.** The pipeline generates high-resolution (HR) audio-driven full-body videos through a two-stage coarse-to-fine process. First, a speech-aligned low-resolution (LR) video is generated using multimodal conditioning (text and audio) and DiT blocks. In the second stage, a pose-guided refiner utilizes pose guidance, LR latents, and reference images to restore degraded details, enhancing identity consistency, motion coherence, and hand realism.

separate cross-attention branch specifically for audio. Formally, the identity-aware cross-attention is extended as follows:

$$\mathrm{CA}_{\mathrm{mm}}\left(x^{\mathrm{lr}}, c_{\mathrm{text}}, c_{\mathrm{audio}}\right) = \mathrm{CA}\left(x^{\mathrm{lr}},\ c_{\mathrm{text}}\right) + \mathrm{CA}\left(x^{\mathrm{lr}},\ c_{\mathrm{audio}}\right) \tag{4}$$

In this way, we enable more precise control over multimodal interactions, allowing the model to better align audio cues with visual dynamics and enhance the generation quality.

3.2. Pose-Guided Refiner

In long-term generation tasks, low-resolution video V_{lr} tends to accumulate errors over time, resulting in visual drift where the appearance deviates from the reference image I_{ref}. To address this issue, the Pose-Guided Refiner (PG-Refiner) leverages the reference image I_{ref} as an identity prior and conditions on the low-resolution video V_{lr} along with its corresponding pose sequence $\mathcal{P} = \{p_i\}_{i=0}^{4f+1}$. This ensures both temporal coherence in motion and consistent appearance throughout the whole video.

Low-Resolution Video Latent Condition. To simulate the temporal degradation phenomenon, we filter out high-frequency signals from the low-resolution latent representation (z^{lr}) using a low-pass filter (LPF), and introduce noise augmentation to improve the model's ability to recover details and correct structural errors. Specifically, the degraded latent representation z^{deglr} is computed as:

$$z^{\mathrm{deglr}} = \mathrm{LPF}(z^{\mathrm{lr}}) + \alpha_{\mathrm{deg}} \cdot \epsilon \tag{5}$$

where $\mathrm{LPF}(z^{\mathrm{lr}})$ extracts the low-frequency components of the video latent, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is additive Gaussian noise, and α_{deg} controls the noise strength.

Pose Guidance Condition. Considering that pose sequence information possesses strong structural properties, preserves fine-grained motions such as lip movements, and remains highly stable with minimal error accumulation in long-duration generation tasks, we adopt it as the condition.

Based on this, we extract human and background keypoints from V_{lr}, forming a pose sequence $\{p_i\}_{i=0}^{4f+1}$. To avoid scale mismatch and keypoint overlap across different resolutions, we use an 8-channel pixel-level representation: the first 7 channels encode human keypoints, and the last channel encodes up to 20 background keypoints. The resulting pose tensor is denoted as $\mathcal{P} \in \mathbb{R}^{(4f+1)\times 4h\times 4w\times 8}$. Accordingly, we apply patchification along the temporal and spatial dimensions: the temporal axis is divided into $f+1$ segments, and the spatial dimensions into $h \times w$ patches, yielding pose tokens $\mathcal{P}' \in \mathbb{R}^{(f+1)\times h\times w\times (64\times 8)}$.

These pose tokens are projected into the latent space via a learned projection and fused with the high-resolution latent feature z^{hr}, producing a pose-aware latent representation $z'^{\mathrm{hr}} = z^{\mathrm{hr}} + \mathrm{Proj}(\mathcal{P}')$. The resulting z'^{hr} serves as the

input to the generator, enhancing both visual fidelity and the temporal consistency of motion in the generated video.

Refiner. To further enhance temporal consistency, we utilize the initial reference frame as a visual anchor. The Refiner module leverages the reference image I_{ref}, pose conditional features P, and the low-resolution degraded latent feature z_{deglr} to generate high-resolution video frames. Since the model is enhanced with temporal degradation during training and introduces pose information as a control signal that is more direct and structurally informative than audio, it can effectively maintain long-term identity consistency with the assistance of the reference image.

Unlike previous methods [15, 38] that rely on structure-aligned reference networks, we adopt a prefix-latent reference strategy to ensure identity consistency and enable high-quality long-sequence continuation. This strategy fully exploits the 3D global attention mechanism in the DiT architecture, allowing the model to directly extract identity features from the prefix latent. Specifically, we denote the high-resolution latent sequence as $\{z_i^{\text{hr}}\}_{i=0}^{f}$, where $z_0^{\text{hr}} = E(I_{\text{ref}})$ is the prefix latent extracted from the reference image using a pretrained 3D VAE encoder $E(\cdot)$, and z_1^{hr} to z_m^{hr} represent motion latents from preceding segments. As the first frame is not temporally compressed, z_0^{hr} preserves more detailed information crucial for identity preservation.

During forward diffusion, we inject noise $\epsilon_i \sim \mathcal{N}(0, I)$ only into the future latents:

$$z_{i,t}^{\text{hr}} = \begin{cases} z_i^{\text{hr}}, & 0 \le i \le m, \\ (1-t)\cdot \epsilon_i + t \cdot z_i^{\text{hr}}, & m < i \le f \end{cases} \tag{6}$$

so that frames 0 through m remain noise-free to provide stable identity and motion guidance, and their noise predictions are excluded from the loss to maintain reference stability and preserve identity consistency.

Formally, the training objective minimizes the velocity prediction error over the noised subset:

$$\mathcal{L}_{\text{ref}} = \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0,I),\, t\sim\mathcal{U}(0,1)}\, \mathbf{w} \cdot \Big\| f_\theta\big(\{\mathbf{z}_{i,t}^{\text{hr}}\}_{i=0}^{f},\, \mathbf{z}^{\text{deglr}},\, P',\, I_{\text{ref}},\, t\big) - \big\{\mathbf{z}_{i,1}^{\text{hr}} - \boldsymbol{\epsilon}_i\big\}_{i=0}^{f} \Big\|_2^2 \tag{7}$$

where $\mathbf{w} = \{w_i\}_{i=0}^{f}$ is a mask vector defined as

$$w_i = \begin{cases} 1, & i > m, \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

To achieve the continuity of motion in long video generation during inference, we take the first m latents of a new chunk from the last m latents of the previous chunk, ensuring smooth motion transitions between chunks.

3.3. Hand-Specific Reward Feedback Learning

Previous models primarily focus on body and facial movements but overlook detailed hand modeling, leading to unnatural hand distortions in generated videos. To address this, we introduce a hand-specific correction strategy that explicitly targets these artifacts.

The human visual system is highly sensitive to hand structures, with clear perceptual boundaries for distortions such as incorrect finger count, unnatural articulation, or broken textures. Motivated by this, we introduce a preference fine-tuning strategy that directly targets hand realism. By optimizing the diffusion model using reward scores from a pretrained image-level evaluator, we significantly improve the structural fidelity and visual quality of generated hands.

Specifically, we first manually constructed a dataset of 10,000 paired image data of hand structures. Leveraging this carefully domain-specific curated dataset, we performed fine-tuning on the open-source MPS [40] model to enhance its initial capability in capturing hand structural characteristics. Building on this, we leverage the pretrained image-level reward model to assess hand realism. To adapt it for video, we decode the low-resolution latent sequence $\{z_{m,1}^{\text{lr}}\}_{m=0}^{f}$ into RGB frames and randomly select one frame X_i^{lr} for evaluation. The training objective becomes:

$$\mathcal{L}_{\text{hand}}(\theta) = \mathbb{E}_{c\sim p(c)}\, \mathbb{E}_{X_i^{\text{lr}} \sim \mathcal{D}(z_{i,1}^{\text{lr}})} \left[T - r_{\text{hand}}(X_i^{\text{lr}}, c) \right] \tag{9}$$

where X_i^{lr} is a decoded frame randomly sampled from the low-resolution latent trajectory, $r_{\text{hand}}(\cdot)$ denotes the pretrained reward model's assessment of hand quality, and T denotes the threshold for hand quality. This approach introduces fine-grained, hand-specific supervision without additional annotations, effectively enhancing anatomical plausibility and reducing common distortions in generated human videos.

4. Experiment

4.1. Implementation Details

Datasets. Our data processing pipeline is as follows: First, we employ SceneDetect [2] for temporal cropping of the raw videos. Next, we use YOLO [18] to track the single person, obtain corresponding spatiotemporal bounding boxes, and perform spatiotemporal cropping. Additionally, videos are filtered based on criteria including video quality, aesthetics, motion amplitude, hand clarity, mouth clarity, and the proportion of the person within the frame. Ultimately, this process yields 7,700 hours of single-person video clips, which is used to train the pose-guider refiner. Building on this dataset, SyncNet [7] is further employed to assess the synchronization between audio and mouth move-

Table 1. **Quantitative Comparison of Audio-Driven Animation Methods on EMTD and HDTF**. $*$ denotes methods limited to talking-head animation. InfinityHuman achieves SOTA results across benchmarks.(§4.2)

	Method	Video Quality				Lip Sync		ID	Hand Stability	
		FID↓	FVD↓	IQA↑	ASE↑	SYNC↑	SYND↓	FSIM↑	HKC↑	HKV
HDTF	SadTalker* [41]	147.73	862.83	1.72	1.07	**8.87**	**6.71**	**0.93**	-	-
HDTF	AniPortrait* [33]	96.12	645.72	1.96	1.15	7.64	7.79	0.85	-	-
HDTF	V-Express* [30]	119.45	748.57	1.32	1.16	7.92	7.96	0.89	-	-
HDTF	EchoMimic* [6]	167.17	757.38	1.61	1.19	6.71	8.23	0.82	-	-
HDTF	HyAva [5]	100.10	662.61	1.52	1.06	7.22	8.98	0.85	-	-
HDTF	Hallo3 [8]	74.10	250.12	1.95	1.14	7.31	9.30	0.91	-	-
HDTF	MultiTalk [20]	85.01	404.45	1.78	1.13	8.76	7.69	0.84	-	-
HDTF	OmniAvatar [10]	131.69	705.14	1.67	1.10	8.81	7.76	0.78	-	-
HDTF	Ours	**69.28**	**239.05**	**2.11**	**1.22**	8.59	7.53	0.89	-	-
EMTD	Fantasy [32]	133.73	1307.20	2.11	1.12	1.11	12.88	0.59	0.57	8.0
EMTD	HyAva [5]	139.39	2160.92	1.76	1.18	4.89	9.37	0.67	0.75	29.2
EMTD	Hallo3 [8]	104.51	1256.10	2.31	1.48	4.26	10.22	0.73	0.77	6.3
EMTD	MultiTalk [20]	103.68	1040.43	2.07	1.30	6.34	8.47	0.71	0.79	14.6
EMTD	OmniAvatar [10]	82.54	1104.99	2.16	1.31	5.40	9.13	0.72	0.86	28.7
EMTD	Ours	**60.71**	**979.88**	**2.48**	**1.59**	**6.56**	**7.97**	**0.84**	**0.90**	16.0

ments, filtering to obtain 1,800 hours of clips for training low-resolution audio-driven video generation, where each clip is 4 seconds. For the hand-specific reward model, training data pairs are constructed to evaluate the hand distortion dimension. The resulting dataset contains 10,000 high-quality annotated samples, created by 10 professional annotators who labeled and filtered 40,000 candidate images.

Training. To train audio-driven low-resolution video generation, we begin with a pretrained Goku-I2V [3] model. For video generation training conditioned on multiple modalities, we include reference images, first frames, audio, and text as modal conditions. A multiple conditions dropout strategy is applied during training to enhance robustness. Specifically, text and audio are dropped with a 10% probability independently. Meanwhile, the reference image and first frame are each dropped with a 20% probability.

To train pose-guided refiner, we also use Goku-I2V as pretrained base model. We adopt the training strategy from Humandit [9], exposing the model to a range of resolutions to enable effective learning across diverse video qualities and sizes. Our conditioning modalities include pose extracted via Sapines [19], first-frame reference images, and low-resolution 3D VAE latents. During training, a dropout mechanism is applied: both pose and low-resolution latents are dropped with a 20% probability.

Both two models are trained using 128 NVIDIA GPUs with a learning rate of 5e-5. For LR-A2V inference, the we apply audio and text classifier-free guidance (CFG) [13] set to 6.5 and 30 denoising steps. For PG-Refiner, we apply pose CFG set to 1.5 and 20 denoising steps. Furthermore, we distill the PG-Refiner into a 1-step model while preserving output quality, enabling ultra-fast low-resolution generation and efficient high-resolution refinement with minimal steps. Detailed inference speed comparisons are provided in the appendix.

4.2. Comparison with State-of-the-Art Methods

Evaluation Metrics. To evaluate our model, we use a comprehensive video quality metric combining FID [12] for image quality, FVD [29] for video dynamics, and Q-align [34] for visual quality (IQA) and aesthetic appeal (AES). Lip-sync accuracy is assessed using Sync-C and Sync-D [7], while identity consistency is measured with FaceSIM [16, 37]. For hand evaluation, we use average Hand Keypoint Confidence (HKC) and Hand Keypoint Variance (HKV).

Test Datasets & Baselines. For evaluation, we use the EMTD [25] dataset, which contains 110 720P speech videos covering the upper body and hands. The longest video lasts 74 seconds, with 23.64% of the videos exceeding 15 seconds, making it well-suited for assessing audio-driven portrait video generation in high-resolution, long-duration scenarios. To further evaluate the generalization ability of our method, we additionally select 100 samples from the HDTF [42] dataset at a resolution of 512×512 as a talking-face test set. We also conduct a user test, detailed in the appendix.

We compare InfinityHuman with human animation methods, including FantasyTalking [32], Hallo3 [8], HunyuanAvatar [5], MultiTalk [20], and OmniAvatar [10], evaluated on the EMTD dataset. Since OmniHuman [23] is limited to 15-second videos and lacks long-form continuation support, it is excluded from the long-video evaluation. For

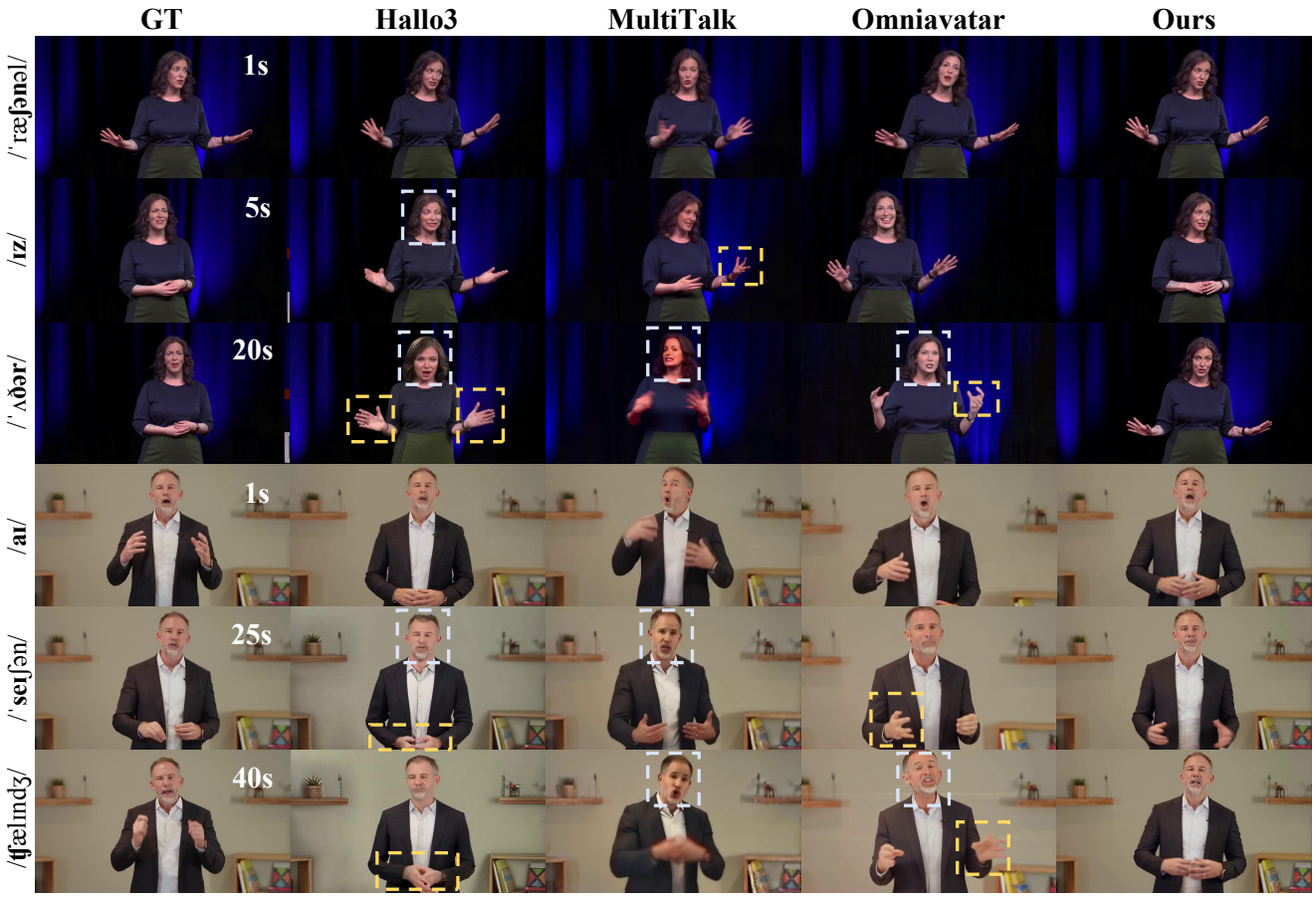

Figure 4. **Qualitative Results of Audio-Driven Animation Methods on EMTD.** Yellow and blue boxes highlight hand distortions and face ID mismatches, respectively. The results demonstrate the superiority of InfinityHuman in maintaining identity consistency, lip-sync accuracy, and visual fidelity during long-duration generation. Please zoom in for details. (§4.2)

Method	FID↓	FVD↓	FSIM↑	HKC↑
w/o refiner	109.54	876.49	0.79	0.85
w/o lr cond	91.92	1001.00	0.86	0.85
w/o pose cond	156.74	1163.75	0.83	0.83
w/o hand refl	86.32	844.57	**0.86**	0.85
ours	**91.74**	**758.98**	0.88	**0.87**

Table 2. **Quantitative Ablation Study.** Demonstrating the effectiveness of the pose-guided refiner and its corresponding conditions, including low-resolution video latent condition, pose guidance condition (§3.2), and hand-specific refl (§3.3).

short-video comparison, please refer to the appendix. In addition, we evaluate our method on the talking face generation benchmark HDTF [42], comparing it with methods such as SadTalker [41], Aniportrait [33], V-express [30], EchoMimic [6], and other representative full-body models.

Qualitative Results & Quantitative Results. For quantitative comparison, as shown in Table 1, our method achieves the best results in both FID and FVD across the audio-driven head and full-body animation tasks. Specifically, on the EMTD dataset, our model achieves an FID of 60.71 and an FVD of 979.88, outperforming the previous best results of 82.54 (OmniAvatar) and 1040.43 (MultiTalk), respectively. Notably, in full-body animation, our method achieves stronger identity consistency, with a FaceSIM of 0.84 (vs. 0.73 for Hallo3). It also delivers better hand motion quality, reaching the highest HKC of 0.90. These improvements demonstrate that our model generates videos that are both more visually realistic and exhibit better temporal coherence.

Additionally, we conduct a qualitative evaluation, as illustrated in Figure 4. Our method demonstrates the ability to generate highly consistent and visually coherent animations over long sequences, maintaining a strong alignment with the corresponding audio. For instance, in the 40-second case, our approach ensures consistent id preservation and color harmony throughout the video. In contrast, other methods exhibit noticeable discrepancies in skin tone, hair color, and facial shapes, especially during long video continuations.

Our method also excels in hand generation, particularly when handling complex hand movements. While other models often struggle with severe distortions or unnatural gestures, our method ensures stable and realistic hand movements, even in challenging poses like hand crossing. This further underscores the superiority of our approach in managing intricate visual dynamics.

Figure 5. **Visualization of Ablation Study.** Demonstrating the effects of key components on animation quality.

4.3. Ablation Study And Discussion

Ablation on the pose-guided refiner. By removing the pose-guided refiner, we directly decode videos from the low-resolution generator on a subset of the EMTD dataset to evaluate its effectiveness. As shown in Table 2, the overall generation quality significantly degrades, with FID increasing from 91.74 to 109.54 and FSIM dropping from 0.88 to 0.79. As illustrated in Figure 5, the degradation is particularly evident in blurred facial details and reduced temporal consistency. These results highlight the critical role of the refiner in recovering visual quality, enhancing temporal stability, and preserving identity over long sequences.

Furthermore, given that the refiner relies on multiple conditional inputs with non-trivial interdependencies, we conduct a deeper analysis of their individual contributions and guidance strength. As shown in Figure 5, omitting either the pose information or low-resolution latent features after training leads to color shifts and structural degradation in long-term video generation. This suggests that both inputs serve as essential references: the pose offers accurate structural constraints, while the low-resolution latent helps preserve overall semantic content and stylistic consistency.

Ablation on the hand-specific reward feedback learning. We also assess the impact of the hand-specific reward feedback (refl) mechanism on generation performance. As shown in Table 2, removing it from the full model results in a decline in hand keypoint accuracy, with HKC decreasing from 0.87 to 0.85. Qualitatively, more artifacts and discontinuities appear in the hand regions, especially in sequences involving complex or high-speed gestures. These findings demonstrate that the hand-specific reward plays a vital role in improving the realism, stability, and audio synchronization of hand motion, particularly under challenging gestural conditions.

4.4. Long-Form Video Stability

To explore the stability of our model on long-form video generation, we segment a subset of output into consecutive 10-second clips and compute cumulative metrics over time (i.e., 10s, 20s, 30s, etc.). This progressive evaluation enables us to analyze how performance evolves as video duration increases.

As shown in Table 3, our model maintains stable performance throughout extended video lengths. Specifically, key metrics such as FID, FVD, FSIM, and Sync show minimal degradation, indicating strong temporal consistency and robustness. In contrast, baseline models tend to suffer from more noticeable quality drops as duration increases.

Table 3. **Long-Form Video Stability Evaluation.** Cumulative metrics over increasing durations on the subset dataset.

Duration	FID↓	FVD↓	FSIM↑	HKC↑	Sync-C↑
10s	36.83	1015.36	0.8357	0.9224	7.23
20s	37.07	1156.05	0.8323	0.9062	7.36
30s	35.02	1315.40	0.8266	0.8991	7.62
40s	35.92	1260.71	0.8154	0.9007	7.81
50s	35.50	945.84	0.8057	0.9059	7.46

5. Conclusion and Future Work

We present InfinityHuman, a coarse-to-fine framework for high-fidelity, long-duration, audio-driven full-body human animation. By introducing a pose-guided refiner and a hand-specific reward mechanism, our approach effectively addresses key challenges in visual consistency, lip-sync accuracy, and hand motion realism. Extensive experiments on EMTD and HDTF demonstrate that InfinityHuman achieves state-of-the-art performance across multiple metrics.

A limitation of our current framework is that it is trained solely on continuous single-person footage, which restricts its ability to handle multi-person interactions and complex scene transitions such as shot changes or cuts. Extending InfinityHuman to support multi-person generation and scene transitions is an important direction for future work.

References

[1] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024. 3

[2] Castellano Brandon. Pyscenedetect. https://github.com/Breakthrough/PySceneDetect/, 2024. 5

[3] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23516–23527, 2025. 6

[4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 3

[5] 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. 2, 6

[6] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2410, 2025. 6, 7

[7] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 5, 6

[8] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025. 2, 6

[9] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025. 6

[10] 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. 2, 6

[11] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025. 3

[12] 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

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[14] 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 gestures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1922–1931, 2024. 2, 3

[15] 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. 5

[16] Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 6

[17] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 3

[18] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics yolo. https://github.com/ultralytics/ultralytics, 2023. 5

[19] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 6

[20] Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025. 2, 6

[21] Xiaodi Li, Zongxin Yang, Ruijie Quan, and Yi Yang. Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. *Advances in Neural Information Processing Systems*, 37:79868–79888, 2024. 3

[22] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 2024. 2, 3

[23] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13847–13858, 2025. 2, 3, 6

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[25] 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. 2, 3, 6

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[27] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 3

[28] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 3

[29] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[30] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 2, 3, 6, 7

[31] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 3

[32] 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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9891–9900, 2025. 2, 6

[33] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 6, 7

[34] 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. 6

[35] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 3

[36] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22963–22974, 2025. 3

[37] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025. 6

[38] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 5

[39] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025. 3

[40] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024. 5

[41] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 2, 3, 6, 7

[42] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3661–3670, 2021. 2, 6, 7