

A Comprehensive Survey on Human Video Generation: Challenges, Methods, and Insights

Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, Li Liu

Abstract—Human video generation is a dynamic and rapidly evolving task that aims to synthesize 2D human body video sequences with generative models given control conditions such as text, audio, and pose. With the potential for wide-ranging applications in film, gaming, and virtual communication, the ability to generate natural and realistic human video is critical. Recent advancements in generative models have laid a solid foundation for the growing interest in this area. Despite the significant progress, the task of human video generation remains challenging due to the consistency of characters, the complexity of human motion, and difficulties in their relationship with the environment. This survey provides a comprehensive review of the current state of human video generation, marking, to the best of our knowledge, the first extensive literature review in this domain. We start with an introduction to the fundamentals of human video generation and the evolution of generative models that have facilitated the field’s growth. We then examine the main methods employed for three key sub-tasks within human video generation: text-driven, audio-driven, and pose-driven motion generation. These areas are explored concerning the conditions that guide the generation process. Furthermore, we offer a collection of the most commonly utilized datasets and the evaluation metrics that are crucial in assessing the quality and realism of generated videos. The survey concludes with a discussion of the current challenges in the field and suggests possible directions for future research. The goal of this survey is to offer the research community a clear and holistic view of the advancements in human video generation, highlighting the milestones achieved and the challenges that lie ahead.

Index Terms—Human video generation, Digital human, Virtual avatar, Diffusion model, Generative methods, Survey.

I. INTRODUCTION

HUMAN video generation task aims to synthesize natural and realistic 2D human video sequences with generative models given control conditions such as text [1], [2], audio [3]–[6] and pose [7], [8]. These generated video sequences feature full-body or half-body human figures, including detailed motion representations of body parts and faces. Recently, this field has gained significant attention due to a wide range of potential applications, including film production, video games, AR/VR, human-robot interaction, digital humans, and accessible human-machine interaction.

Recently, human video generation has achieved rapid progress benefiting from advancements in generation methods, *i.e.*, Variational Autoencoders (VAE) [9], Generative Adversarial Networks (GAN) [10], and Diffusion Models [11]. However, studying such a video synthesis problem is known to be

challenging for the following reasons. Firstly, the appearance consistency of humans along the time sequence is a significant obstacle in this task. Secondly, the deformation of the human body that people are sensitive to in a synthesized video is hard to avoid, *i.e.*, finger abnormalities, as shown in Fig. 1. Thirdly, the complexity of human motion video extends beyond just modeling the face; it also involves accurately modeling body motion and maintaining background consistency and harmony with body parts. Additionally, the demand for human motion generation often includes a context as the condition, such as text description, audio signals, pose sequences, ensuring temporal alignment with these conditional signals is crucial for producing a coherent and realistic human video.

In response to the rapid development and emerging challenges of human video generation, we present a comprehensive survey of this field to help the community keep track of its progress.

In summary, the main contributions of this survey are fourfold:

- We have carefully specified the boundaries of human video generation, offering a comprehensive analysis of recent advancements within this domain. We have categorized these advancements into three primary groups based on the modality driving the generation process: text-driven, audio-driven, and pose-driven. To our knowledge, this is the first survey that provides a systematic and focused examination of this particular field.
- We thoroughly examine the challenges and hurdles in human video generation through massive related methods and an extensive inventory of relevant datasets, challenges, evaluation metrics, and commercial projects. This paper guides readers in selecting suitable baselines or solutions for their unique applications. Additionally, our findings offer valuable insights into enhancing current methodologies.
- Drawing from our detailed literature review and in-depth analysis, we have identified several promising directions for future development in human motion generation.
- We also provide a continuously updated GitHub repository that includes the latest developments in the field, as well as links to awesome works and datasets. We aim to provide the research community the most cutting-edge information and provide easy access to important research works, datasets, and applications. For more details, please visit our [repository link](#)

The survey is organized as follows. In Section II, we discuss the comparison with the previous survey. Section III covers

All authors are with the Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511458, China.

All authors are equal contributions.

Corresponding author: Li Liu, avrillliu@hkust-gz.edu.cn.

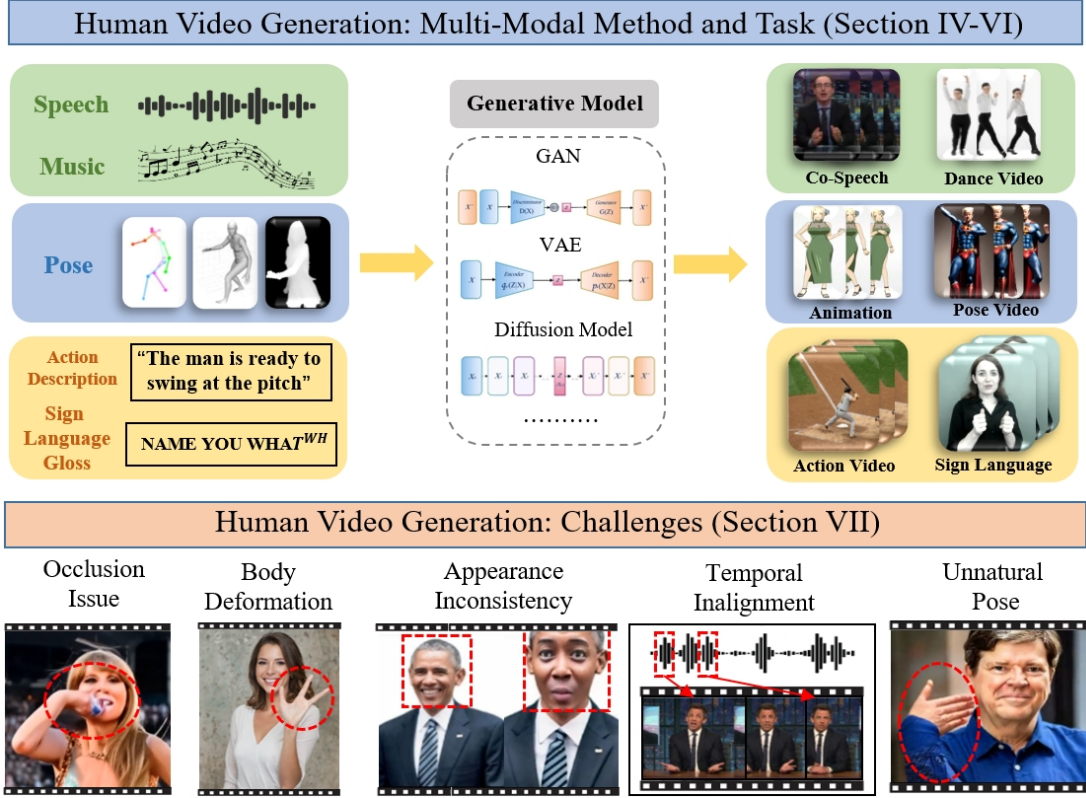


Fig. 1: An overview of typical Multi-Condition human video generation methods and challenges.

the fundamentals of the task, including the commonly used datasets and the evaluation metrics. In Sections IV-VI, we summarize existing approaches for human motion generation based on different conditional signals, respectively, including text, audio, and pose. Finally, we draw conclusions and provide insights for this field in Section VIII.

II. COMPARISONS WITH PREVIOUS SURVEYS

To the best of our knowledge, this survey is the first to focus directly on the human video generation task. Although several surveys have been conducted on video or motion generation, the differences between our survey and existing ones are mainly in the following three aspects.

1) Different Scope. This survey focuses on human video generation, which is a 2D video generation task that uses a generative model to input text, audio, posture, or other modal data and uses full-body or half-body characters, including hands and faces as generated subjects. Compared with the general video generation task that many previous surveys [12]–[15] have focused on, this paper details the unique challenges and developments of human generation. Additionally, surveys [16], [17] concentrated solely on the talking head task, which focuses only on the generation of the head. However, the scope of this survey pays additional attention to hands, thus extending to the generation of half-body and full-body. Furthermore, the work by Zhu *et al.* [18] explicitly addresses motion generation, emphasizing human poses rather than video generation.

2) Video Perspective. This paper especially discusses human generation challenges from a video perspective. In contrast, previous human generation surveys [19], [20] focused on the problems in image generation.

3) New Insight. To explore and solve the special challenges in human video generation and improve the generation quality, this paper provides a comprehensive analysis of the human video generation task through detailed methods and challenge discussion, as well as summarizes extra relevant datasets, evaluation metrics, and existing commercial projects. Our goal is to offer readers a clear and concise insight into the factors contributing to a successful human video generation and to answer the question, “What Makes a Good Human Video Generation?”

III. DATASET AND MATRIX

A. Metrics

Comparing different methods in this field requires appropriate and comprehensive evaluation metrics. However, evaluating generated human videos presents significant challenges due to factors such as the one-to-many mapping nature, the subjectivity inherent in human evaluations, and the complexity of high-level conditional signals. To address these challenges, this section provides an overview of the most commonly used evaluation metrics, highlighting their advantages and limitations. The details of the metrics are shown in Table. I

We summarize that the evaluation of generated human videos in this field covers several critical aspects: Image Quality, Video Quality, Consistency, Diversity, Aesthetics, and

Category	Metric	Description
Image Quality	L1 Error	Measures the absolute pixel-level difference between predicted and ground truth frames.
	Peak Signal-to-Noise Ratio (PSNR) [21]	Quantifies similarity between generated and real images in dB.
	Structural Similarity Index (SSIM) [22]	Evaluate structural similarity considering luminance, contrast, and structure.
	Learned Perceptual Image Patch Similarity (LPIPS) [23]	A deep learning-based metric evaluating visual similarity. Lower LPIPS indicates higher similarity.
	Fréchet Inception Distance (FID) [24]	Compares the feature distribution between generated and real samples. Lower FID indicates better quality.
Video Quality	Kernel Video Distance (KVD) [25]	Measures the distribution distance between generated and real video sequences.
	Fréchet Video Distance (FVD) [25]	Measures the distance between the distributions of generated and real videos.
	Average Content Distance (ACD) [26]	Assesses action sequence consistency in generated videos, especially for gesture generation tasks.
	Warping Error (WE) [27]	Obtain the optical flow of each two frames, then calculate the pixel-wise differences between the warped image and the predicted image.
	Fréchet Gesture Distance (FGD) [28]	Measure the distribution gap between real and generated gestures in the feature space.
	Fréchet Template Distance (FTD) [28]	similar to the FGD , measuring the distribution similarity between the generated ones and the real ones in the feature space
	Fréchet Inception Distance for Videos (FID-VID) [29]	Measures the distribution distance between generated and real video frames, incorporating both spatial and temporal features. Lower FID-VID indicates better quality.
Consistency	Beat Consistency (BC)	Assesses temporal consistency in videos content with audio.
	CLIP-I score [1]	Measures the face structural similarity between the reference image and the generated video.
	Beat Alignment Score (BAS) [30]	Evaluate the motion-music correlation in terms of the similarity between the kinematic beats and music beats.
	Frame Consistency (FC) [31]	Assesses temporal consistency in videos by calculating cosine similarity between feature vectors of consecutive frames.
Diversity	Inception Score (IS) [32]	Measures the diversity and clarity of generated images and sometimes used for video quality.
	Diversity (Div) [33]	Calculates feature distance between generated gestures on average.
Aesthetics	Dover Score [34]	Measures the overall quality of the generated video from both technical and aesthetic perspectives
Pose Accuracy	Percentage of Correct Keypoints (PCK) [35]	Measures the proportion of keypoints that are correctly localized within a specified threshold distance from the ground truth keypoints.
	Average Keypoint Distance (AKD) [36]	Evaluates the accuracy of human keypoints in generated videos by comparing distances to real keypoints.
	Missing Keypoint Rate (MKR) [37]	Measures the proportion of missed keypoints during detection or generation.

TABLE I: Commonly Used Evaluation Metrics for Human Video Generation.

Action Accuracy. Each of these categories is essential for a comprehensive assessment of the performance of different methods.

Image Quality focuses on the visual fidelity of individual frames, evaluating pixel-level differences, structural similarity, and perceptual similarity to ensure frames closely match real ones.

Video Quality extends this evaluation to the temporal domain, assessing the coherence and realism of frame sequences to capture the dynamic nature of real-world actions.

Temporal Consistency is to ensure that the generated content maintains a natural flow and synchronization over time, which is crucial for applications involving synchronized audio and video.

Diversity is to evaluate the variety and richness of the generated content, ensuring the model can produce a wide range of realistic videos.

Action Accuracy is to assess the precision of human actions and movements within the videos, which is vital for applications where the correctness of these actions is paramount. Together, these metrics provide a comprehensive framework for evaluating the performance and quality of methods in human video generation.

B. Datasets

Recently, various datasets have been utilized in the research on human video generation, encompassing a diverse array of scenes, actions, and backgrounds. The primary datasets include videos of dance, fashion, and daily activities, sourced from widely accessible platforms such as *TikTok* and *YouTube*. These datasets provide diverse data to support the training and evaluation of existing methods. The details of the datasets are shown in Table. II

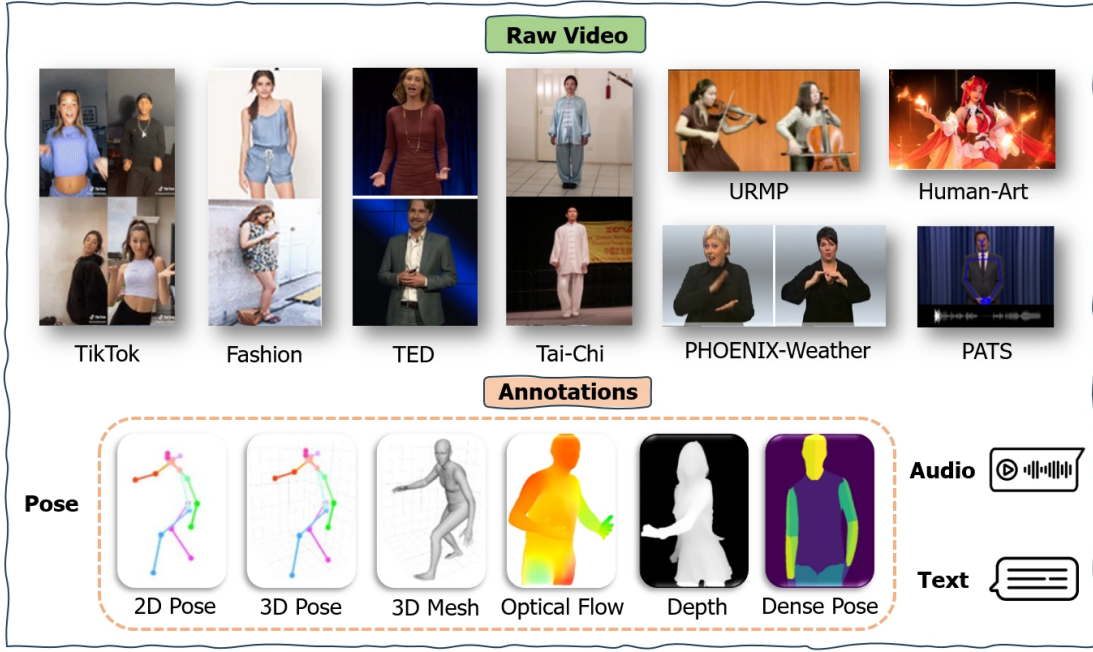


Fig. 2: Some examples of human video datasets and annotation formats.

Category	Dataset Name	Year	Data Size	Modality	Source
Human Action	ASTS [38]	2005	90 videos	Video/Mask	Link
	UCF-101 [39]	2012	~13k videos	Video	Link
	human3.6m [40]	2014	3.6M frames	Video/2D-Pose	Link
	NTU RGB+D [41]	2016	~114k videos	Video/3D-pose/Depth	Link
	TaiChi [36]	2019	3k videos	Video	Link
	HAR [42]	2023	~1k videos	Video	Link
	3D People Synthetic [43]	2023	~22k videos	Video/Masks/Pose/Depth/Mesh/OF	Link
Human Dance	MSP-Avatar [44]	2023	74 videos	Video/Audio/Motion	Link
	EverybodyDance [45]	2019	105 videos	Video/2D-Pose	Link
	AIST++ [30]	2021	10k videos	Video/Audio/3D-Pose	Link
	TikTok [46]	2021	340 videos	Video/Depth/Mesh	Link
	DanceIt [47]	2021	154 videos	Video/Audio/2D-Pose	Link
	TikTok-v4 [7]	2023	350 videos	Video/2D-Pose	Link
	Disco [48]	2024	700k frames	Video/2D-Pose/Mask	Link
Music Performance	Sub-URMP [49]	2017	~81 frames	Video/Audio	Link
	URMP [50]	2018	44 videos	Video/Musical Score/Audio	Link
Human Fashion	DeepFashion [51]	2016	~800k frames	Video/Mask/Text	Link
	Fashion [52]	2019	600 videos	Video	Link
	Fashion-Text2Video [53]	2023	600 videos	Video/Text	Link
Human Art	HumanArt [54]	2023	50k frames	Video/Text/2D-Pose	Link
Body Language	MS-ASL [55]	2018	25k videos	Video/Text	Link
	PHOENIX14T [56]	2018	~68K frames	Video/Text	Link
	How2sign [57]	2021	~35K frames	Video/Text/2D-Pose/Depth	Link
	Bold [58]	2023	~10k videos	Video//Text/Audio/3D-Pose	Link
	MCCS-2023 [59], [60]	2023	~ 4k videos	Video/2D-Pose/3D-Pose/Text/Audio	Link
	Speech2gesture [61]	2019	60k	Video/Audio/2D-Pose	Link
	Pats [62]	2020	84k videos	Videos/Text/Audio/2D-Pose	Link
	TED gesture [63]	2021	~2k videos	Video/Text/2D-Pose/Audio	Link
	Ted-talk [64]	2021	~3k videos	Video	Link

TABLE II: Dataset Information for human video generation.

For video generation tasks, effectively representing pose and motion information in videos is crucial. In this section, we will introduce common pose annotation formats, their characteristics, and commonly used methods.

2D Pose uses keypoints to form a skeletal graph for recognizing and analyzing human poses in 2D. Data is typically formatted as a set of (x, y) coordinates for each joint. Common methods: OpenPose [65], DwPose [66], PoseNet [67], HRNet

[68], StackPose [69].

3D Pose adds depth to 2D poses, providing 3D coordinates (x, y, z) for detailed human pose information. Common methods: ExPose [70], Alphapose [71], MotionBERT [72].

3D Mesh uses polygonal meshes to represent human surface shapes for realistic models. Data formats often include vertices and faces of the mesh. Common methods: SMPL [73], SMPL-X [74].

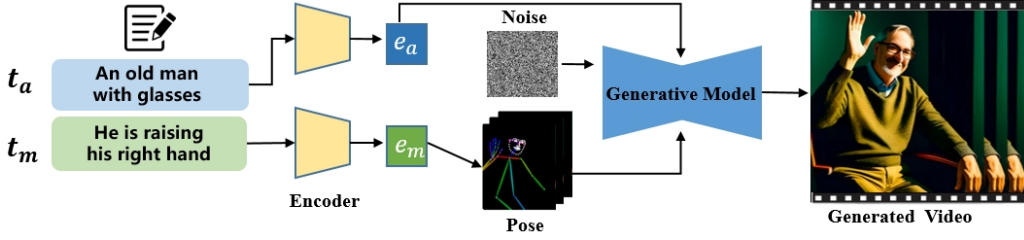


Fig. 3: An overview of text to human video generation approaches.

Optical Flow represents motion vectors of pixels to describe motion direction and speed in videos. Data is typically stored as a 2D field of vectors. Common methods: MMFlow [75], FlowNet [76], RAFT [77].

Depth creates a depth map showing the distance of each pixel from the camera, useful for 3D reconstruction and AR. Data is usually in the form of depth images where each pixel value corresponds to the distance from the camera. Common methods: vid2depth [78], monodepth2 [79], Depth Anything [80].

Dense Pose maps 3D body surface coordinates to each pixel for detailed pose information. Data includes UV coordinates for each pixel mapped to a 3D body model. Common methods: DensePose [81].

IV. TEXT TO HUMAN VIDEO GENERATION

In the following sections IV-VI, we will focus on the methods of human video generation based on different condition signals. Firstly, we will introduce the text-driven human video generation methods.

Text can describe specific appearances, scenes, and styles, providing a rich source of information for generative models to control the generated content. Recent generative methods such as stable diffusion [82] and Sora [14] have shown that using text as input to generate images and videos has achieved impressive results.

However, different from the general video generation tasks which focus on the coherence of the video, human video generation requires precise control over both the appearance and movement of the human body. Existing methods approach this challenge from two main angles: using text to maintain appearance and extracting semantic information from text to control poses. The overview of existing research in text-driven human video generation is shown in Fig. 3.

A. Text-driven Human Appearance Control

To control the appearance of the human body in the generated video, there are two approaches: one is to directly provide **reference images**, and the other is to use input **text descriptions** to control the generated human appearance. Here, we discuss the text-driven human appearance control methods. To ensure the **consistency of appearance** in generated videos with the textual descriptions while preserving identity details during frames, ID-Animator [1] leverages a pre-trained text-to-video (T2V) model with a lightweight face adapter to

encode identity-relevant embeddings. Text descriptions guide the generation of human videos and control the character's appearance in the video. Similarly, [2] uses text descriptions to provide semantic information about the content of the characters, ensuring the generated videos align with the textual descriptions.

B. Text-driven Human Motion Control

Existing methods for precisely controlling the motion of the human body in generated videos typically follow two approaches:

1) One approach follows a **two-stage pipeline**. It first generates corresponding poses based on the semantics of the input text according to the task and then uses these generated poses to guide the motion. More details about the pose-guided generation methods in the second stage can be referred in Section VI. For this type of task, it is necessary to establish a connection between text and poses to control motion in a video. HMTV [83] uses descriptive text to generate initial 3D human motion and control camera angles, ensuring dynamic and realistic video outputs. The text guides the actions and camera movements in the video, providing precise control over the character's movements and the viewer's perspective. For the Sign Language Production task, SignSynth [84] uses a Gloss2Pose network to generate sign language poses and a GAN to create high-quality sign language videos. Similarly, H-DNA [85] translates spoken sentences into sign language videos by first generating sign gesture poses and then using a GAN to produce the final video. In SignLLM [86], text descriptions are converted into gloss (an intermediate sign language representation) and then mapped to poses, which are rendered into Sign Language videos. Here, the semantics of the text are captured to align with the described human pose. In Cued Speech [87], [88] Generation task, [89] first leveraged a Large Language Model (LLM) to convert text into a descriptive gloss and then used the gloss to generate a fine-grained pose.

2) The other approach directly uses text as a prompt to guide the generation of video actions. For instance, Text2Performer [53] involves the motion text and a motion encoder. motion text describes the movement, such as “*She is swinging to the right.*” The model implicitly models these descriptions by separately representing appearance and motion, thereby generating high-quality videos with consistent appearance and actions.

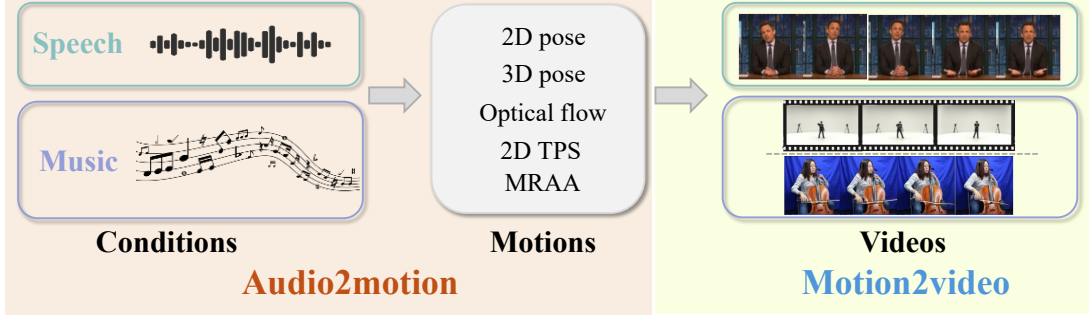


Fig. 4: An overview of audio to human video generation approaches. Example images adapted from [4], [90], [91].

Condition	Method	Venue	Model	Motion Feature	Dataset
Speech	speech2gesture [61]	CVPR 2019	GAN	2D pose	Speech2Gesture [61]
	Speech2video [92]	ACCV 2020	LSTM	3D pose	Self-collection
	Qian <i>al.</i> [3]	ICCV 2021	VAE	2D pose	Speech2Gesture [61]
	ANGIE [62]	NeurIPS 2022	VQ-VAE	MRAA	PATS [62]
	DR ² [93]	WACV 2024	VAE	3D pose	Speech2Gesture [61], Self-collection
	DiffTED [94]	CVPR 2024	Diffusion	2D TPS	TED-talks [64]
	He <i>al.</i> [4]	CVPR 2024	Diffusion	2D TPS, optical flow	PATS [62]
Music	Islam <i>al.</i> [6]	RAAICON 2019	GAN	2D pose	Self-collection
	DanceIt [47]	TIP 2021	GAN	2D pose	Self-collection
	Dabfusion [90]	ArXiv 2024	Diffusion	Optical flow	AIST++ [30]
	Zhu <i>al.</i> [5]	ICPR 2021	CNN	keypoint	Sub-URMP [49]
	Music2Play [91]	CAC 2023	LSTM	2D pose, optical flow	URMP [50]

TABLE III: Summary of works related to audio to human video generation.

V. AUDIO TO HUMAN VIDEO GENERATION

In addition to textual descriptions, human video generation from audio signals has also been explored in this survey. In this section, we mainly discuss two main subtasks: **speech-driven human video** and **music-driven human video**. Speech-driven human video generation aims to generate a sequence of human gestures based on input speech audio, which requires the generated human motion to be harmonious with the audio, not only in terms of high-level semantics but also emotion and rhythm. While music-driven human video generation focuses on synthesizing the video of a person dancing or playing a certain instrument guided by a given music clip, which especially lies in the low-level beat alignment. In this scenario, the direct conversion of audio into video poses a complex challenge. Previous research has often followed a two-stage pipeline, including audio-to-motion and motion-to-video, as illustrated in Fig. 4.

A. Speech-driven Human Video Generation

Many existing works have concentrated on generating talking videos, primarily focusing on the head region [95], [96]. In contrast, our review focuses on works that include body gestures [3], [4], [61], [92]–[94]. To the best of our knowledge, all of these works fall under the field of **co-speech gesture** video generation. Given the importance of motion representation for the final video, we review these works from the perspective of motion generation.

In speech-driven human video generation, some methods [61], [92], [93] synthesize talking videos from sequences of 2D skeletons [3], [61] or 3D models [92], [93], with the rendering process being separate from the generation of the

gestures. However, hand-crafted structural human priors like 2D/3D skeletons completely discard appearance information around key points, making precise motion control and video rendering highly challenging. Additionally, the pre-training of pose estimators relies on hand-crafted annotations, leading to error accumulation and often resulting in jitters. To alleviate these issues, ANGIE [62] utilizes an unsupervised feature, MRAA [64], to model body motion. A VQ-VAE [97] is then used to quantize common patterns, followed by a GPT-like network that predicts discrete motion patterns to generate gesture videos. However, MRAA, being a coarse modeling of motion, is linear and fails to represent complex-shaped regions, limiting the quality of gesture videos generated by ANGIE. Additionally, directly associating covariance with speech is inappropriate. To address these challenges, DiffTED and He *al.* propose decoupling motion from gesture videos while preserving critical appearance information of body regions. They use the learned 2D keypoints of the Thin-plate Spline (TPS) motion model [37] as targets for generation and leverage the TPS motion model to render the keypoints into images. Additionally, motivated by the success of recent diffusion models [11], DiffTED and He *al.* propose a diffusion-based approach to generate diverse gesture sequences.

B. Music-driven Human Video Generation

Music-driven human video generation uniquely intersects motion synthesis and music interpretation, aiming to create human motions synchronized with input music beat. This extends beyond general motion synthesis, as beat-aligned motions are complex to animate [90]. We have explored two sub-tasks, *i.e.*, **music-to-dance** and **music-to-performance**.

To achieve beat sensing motion generation, some music-to-dance video generation works [6], [90] explicitly detect beat from music audio, or design a matching phase learns the relationship between these two different modalities [47]. Islam *al.* [6] perform beat detection and repeated pattern extraction from input music first and then generate mathematical models of a person dancing and convert them into realistic images of the target person. Dabfusion [90] applies a beat extractor to explicitly disentangle beat features from music. These beat features are then used to guide the production of latent optical flows, followed by backward flow estimation to generate the output video. Differently, DanceIt [47] learns the relationship between these two different modalities at the first matching phase, then retrieves a sequence of pose fragments for each music audio and performs spatial-temporal alignment at the generation phase.

For music-to-performance video generation, it is challenging to generate high-dimensional temporal consistent videos from low-dimensional audio modality. Zhu *al.* [5] propose a multi-staged framework that first generates the coarse video from given audio and then makes refinements by integrating intra-frame structure information from predicted keypoints and temporal information for final performance video generation. Music2Play [91] gains a sequence of poses in an autoregressive way and, estimates the dense flow field information from the pair of poses, finally fuses multi-modal information (audio, flow, and image) to synthesize the output frame.

VI. POSE TO HUMAN VIDEO GENERATION

As illustrated in Fig. 5, existing research in pose-driven human video generation has often followed a common pipeline. In the task of pose-driven human video generation, various pose types, including **skeleton pose, dense pose, depth, mesh, and optical flow** (as shown in Tab. IV), serve as common guiding modalities along with the more traditional text and speech inputs. According to the number of conditional poses, we can divide the existing pose guided human video generation methods into two categories. The first category uses only a single type of pose, which is recorded as **single-condition pose-guided methods**. The second category uses different types of pose signals, which are referred to as **multi-condition poses-guided methods**.

A. Single-condition Pose-guided Methods

Among all types of conditional signals, the most common are skeleton pose and dense pose. Early pose-guided human video generation methods [26], [52], [75], [99]–[106] based on GANs primarily utilized conditional adversarial networks such as CGAN [107], pix2pix [108], and pix2pixHD [109]. These methods extracted skeleton poses using OpenPose [65] or StackPose [69] methods, or extracted dense pose using the DensePose method, and used the extracted skeleton pose or dense pose as conditional signal into CGAN or pix2pix generation models.

With the development of conditional generation models, current methods [8], [29], [110]–[112] mostly utilize stable diffusion (SD) [113] or Stable Video Diffusion (SVD) [114],

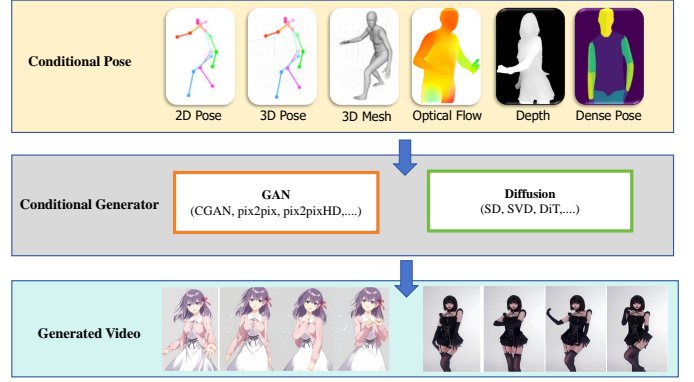


Fig. 5: An overview of pose-guided human video generation approaches. Examples come from [98] and [8].

[115] as the backbone for video generation models. For instance, the MagicPose [29] injects pose features into the diffusion model by ControlNet [116]. In contrast to directly utilizing ControlNet, methods such as MotionFollower [110], MimicMotion [111], AnimateAnyone [8], and UniAnimate [112] extract skeleton poses from target video frames using Dw-Pose [117] or OpenPose [65]. To align the extracted skeleton poses with the noise in the latent space and effectively leverage pose guidance during denoising processing, they design lightweight neural networks (composed of only a few convolutional layers) as pose guider.

Unlike the above skeleton pose-guided video generation diffusion models, methods like DreamPose [118] and MagicAnimate [119] utilize the DensePose [81] method to extract dense pose and directly concatenate dense pose and noise into the denoising UNet by ControlNet. Different from these types of 2D poses (skeleton pose and dense pose), Human4DiT [120] extracts corresponding 3D mesh maps using SMPL [121]. Inspired by the work of Sora and other variants [14], [122], Human4DiT [120] regards Diffusion Transformer as the backbone for video generation.

B. Multi-condition Poses-guided Methods

In addition to the single conditional pose-based human video generation, the recent success of SD [113] and SVD [114], [115] has laid the foundation for multi-conditional pose-guided human video generation. Most existing pose-guided methods use either skeleton pose or dense pose as the conditional input. However, these single-condition pose-guided methods often exhibit poor generalization to complex backgrounds and suffer from occlusion issues between different bodies and parts of the same individual.

To address the poor generalization considering the **complex backgrounds**, DISCO [48] presents an innovative model architecture featuring disentangled control over background and skeleton pose, thereby improving the compositionality of dance generation. This architecture enables the integration of both seen and novel subjects, backgrounds, and poses from diverse sources. Follow-Your-Pose v2 [138] integrates an optical flow guide with other condition guiders to enhance background

Model	Method	Venue	Condition	Extractor	Dataset
GAN	Cai <i>et al.</i> [101]	ECCV 2018	SK	SP [69]	Human3.6M [40]
	Yang <i>et al.</i> [99]	ECCV 2018	SK	OP [65]	[39], [123], [124]
	Yang <i>et al.</i> [125]	ICMEW 2019	SK	OP [65]	self-collection
	Chan <i>et al.</i> [103]	ICCV 2019	SK	OP [65]	EverybodyDance [45]
	DwNet [52]	BMVC 2019	DS	DP [81]	Fashion [52]
	Naoya <i>et al.</i> [104]	ECCVW 2020	SK	OP [65]	Human3.6M [40]
	Yoon <i>et al.</i> [100]	CVPR 2021	DS	DP [81]	3D-people [43]
	SGW-GAN [106]	ArXiv 2021	SK	OP [65]	MS-ASL [55]
Diffusion	DreamPose [118]	ICCV 2023	DS	DP [81]	Fashion [52]
	LEO [126]	ArXiv 2023	OF	LIA [127]	[128]–[130]
	DreaMoving [131]	ArXiv 2023	SK, DP	DwP [117], ZoeDepth [132]	self-collection
	DisCo [48]	CVPR 2024	BG, SK	G-SAM [133], OP [65]	TikTok [46]
	Animate Anyone [8]	CVPR 2024	SK	DP [81], OP [65]	self-collection
	MagicPose [29]	ICML 2024	SK	OP [65]	TikTok [46]
	MagicAnimate [119]	CVPR 2024	DS	DP [81]	TikTok [46], TED-talks [64]
	Champ [134]	ArXiv 2024	DP, NM, SM, SK	SMPL [121]	self-collection
	PoseAnimate [135]	ArXiv 2024	SK	OP [65]	Training-Free
	Liu <i>et al.</i> [125]	ArXiv 2024	SK, BG	DwP [117], HM [136]	self-collection
	MotionFollower	ArXiv 2024	SK	DwP [117]	self-collection
	Human4DiT [120]	ArXiv 2024	MS	SMPL [121]	self-collection
	VividPose [98]	ArXiv 2024	SK, MS	DwP [117], SMPL-X [137]	TikTok [46]
	UniAnimate [112]	ArXiv 2024	SK	DwP [117]	TikTok [46], Fashion [52]
	FYP v2 [138]	ArXiv 2024	SK, DP, OF	DwP [117], DA [80], MF [75]	self-collection
	MimicMotion [111]	ArXiv 2024	SK	DwP [117]	self-collection

TABLE IV: List of Methods Focusing on Pose Guided Human Video Generation. SP, DP, OP and DwP represent StackPose [69], DensePose [81], OpenPose [65] and DwPose [117]. G-SAM, DA, HM and MF represent Grounded-SAM [133], Depth Anything [80], Human Matting [136] and MMFlow [75]. SK, DS, MS, and OF represent skeleton pose, dense pose, mesh, and optical flow. BG, NM, and SM represent the background, normal map, and semantic map.

stability. Liu *et al.* [125] separates the motion representations of the foreground and background, animating human figures with pose-based motion while modeling background motion using sparse tracking points to capture natural interactions between the figure’s activity and environmental changes.

To tackle the **occlusion issues**, Follow-Your-Pose v2 [138] addresses occlusions in multi-character animation with a depth guider, and improves character appearance learning with a reference pose guider. VividPose [98] introduces depth and mesh information, particularly in conjunction with the SMPL-X [137] model, which helps the system to better handle occlusions and complex movements that are common in human pose sequences. DreaMoving [131] integrates depth information and skeleton pose, helping the model to understand the spatial relationships between different parts of the body and the environment. The depth information is useful for handling occlusions as it allows the model to determine which body parts are in front of or behind others.

VII. CHALLENGES

In this section, we summarize the key challenges in the human video generation task, discuss the special challenges existing in the models guided by the particular modality, and explain the common problems faced by this task and related video generation tasks. Representative challenges include:

1) Occlusion Issue. In the collected videos, overlapping body parts or multiple people occlusion is common, but most models cannot handle the problem of mutual influence well [98], [138].

2) Body Deformation. Ensuring that generated video features such as body shape, face, and hands adhere to typical

human characteristics is a significant obstacle in this task. One common example of this issue is the occurrence of malformed hands [139].

3) Appearance Inconsistency. The generation of human videos also requires that the various features of the human appearance, including face, body, clothing, accessories, etc., be consistent in the generated videos. However, most models cannot achieve utterly satisfactory consistency.

4) Background Influence. When generating videos with the human body in the foreground, the consistency of the background and the harmony with the foreground human body is also a major challenge. Poor background control will affect the quality of human generation and bring additional jitter and distortion.

5) Temporal Inalignment. In models guided by temporal signals, especially the audio-to-human video generation models, the synchronization of lips and voice is a significant challenge to improving the quality.

6) Unnatural Pose. Current generated human video often suffers from the unnatural pose problem. The specific manifestations of this problem include the inconsistency between the generated video and the inputted guided pose, as well as the naturalness of the movements in the generated videos.

In addition to the representative challenges mentioned above, in text- or audio-driven models, due to the one-to-many mapping nature in the dataset, meaning that a single input text or audio can correspond to several valid outputs. As a result, attempting to directly match the input with a single ‘correct’ gesture can lead to an unreliable and biased association. This approach hinders the model’s ability to capture and learn the variations present within the data [3].

It should be noted that since human video generation

is essentially a branch of video generation, the efficiency challenges brought by the common use of diffusion models, the challenges of multi-view generation, and the challenges of high-resolution generation still have a significant impact on the generation quality.

VIII. CONCLUSION AND DISCUSSION

A. Conclusion

In this survey, we provide a comprehensive overview of recent advancements in human video generation. Despite the rapid progress in this field, significant challenges remain that warrant further exploration. We summarize available dataset resources and commonly used evaluation metrics. Subsequently, we classify the existing researches based on conditional signals (*i.e.* text, audio and pose) and discuss each category in detail.

B. Discussion

In this section, we aim to discuss in detail the factors influencing the quality of human video generation, excluding dataset scale. To this end, we will focus on three aspects: generation paradigm, backbone, and condition pose.

- **Generation Paradigm.** Compared to pose-driven methods (which can be regarded as one-stage methods), text and audio-driven methods can be divided into one-stage and two-stage approaches. The former directly uses input text or audio as prompts to guide human video generation, while the latter generates poses from the input text or audio and then uses these generated poses as signals to guide human video generation. The introduction of various pose types, such as skeleton poses, in two-stage methods, provides additional geometric and semantic information, enhancing the accuracy and realism of video motions. This makes two-stage methods significantly more effective than one-stage methods, albeit at the cost of some efficiency.
- **Backbone.** Diffusion models, such as SD and SVD, are widely used in various generative tasks, including human video generation, due to their superior performance and diversity. However, unlike GANs, which generate samples in a single sampling step, diffusion models require multiple sampling steps, thereby increasing the time cost for training and inference.
- **Condition Pose.** Different types of conditional poses work because they provide complementary information. The most common skeleton pose accurately describes the spatial information of the human body in the frame and the relative positions of body parts. However, it captures discrete pose changes rather than continuous motion details, providing limited temporal coherence. In contrast, optical flow inherently includes temporal information, capturing changes between consecutive frames and providing continuous motion trajectories in the feature space. This allows the model to generate videos with smooth transitions between frames, avoiding jumps or discontinuities. Moreover, the skeleton pose does not include background and detail modeling, whereas depth

maps capture distance information between human body and the background, along with surface details and depth changes. 3D meshes offer detailed geometric structures of object surfaces that skeleton poses lack. In summary, different types of poses provide complementary spatiotemporal information, and there is no unified pose type that fulfills all requirements. Different scenarios and problems may require different poses.

C. Future Work

We outline several promising future directions from various perspectives, aiming to inspire new breakthroughs in human video generation research.

- **Large-Scale High-Quality Human Video Datasets.** Existing public datasets, including those in the fields of human action and human dance, are relatively small in scale. Collecting high-quality human video datasets is both challenging and expensive. However, a large-scale, high-quality human video dataset is crucial for developing a foundational model for human video generation.
- **Long Video Generation.** Current human video generation methods typically produce videos lasting only several seconds. Generating videos that extend to several minutes or even hours presents a significant challenge. Therefore, future research should focus on the generation of long-duration human videos.
- **Photorealistic Video Generation.** As previously mentioned, challenges such as occlusion, body deformation, pose unnaturalness, and appearance inconsistency can result in low-quality video generation. Resolving these visual and aesthetic issues to ensure that the generated human body movements follow real-world physical laws is a major challenge. Creating videos with highly realistic visual effects remains a difficult task.
- **Human Video Diffusion Efficiency.** Diffusion models have become the backbone for human video generation tasks. However, the heavy training costs and deployment requirements of video diffusion models pose significant challenges. Reducing training costs and scaling down model size are crucial issues. Therefore, exploring the efficiency of video diffusion models is a valuable direction for future research.
- **Fine-Grained Controllability.** Existing multimodal-driven human video generation methods, even when incorporating additional, conditional signals such as 3D mesh and depth map alongside skeleton pose, still lack fine-grained control over specific body parts, particularly hands, and face. Future research could focus on achieving fine-grained, controllable generation of these detailed human body regions.
- **Interactivity.** In addition to exploring fine-grained controllability, future work could further investigate interactive controllability. This would allow users to manipulate elements such as arm movements or facial expressions through simple actions like clicking, ultimately generating human videos that meet user satisfaction.

REFERENCES

- [1] X. He, Q. Liu, S. Qian, X. Wang, T. Hu, K. Cao, K. Yan, M. Zhou, and J. Zhang, "Id-animator: Zero-shot identity-preserving human video generation," *arXiv*, 2024.
- [2] Y. Ma, Y. He, X. Cun, X. Wang, S. Chen, X. Li, and Q. Chen, "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in *AAAI*, 2024.
- [3] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *ICCV*, 2021.
- [4] X. He, Q. Huang, Z. Zhang, Z. Lin, Z. Wu, S. Yang, M. Li, Z. Chen, S. Xu, and X. Wu, "Co-speech gesture video generation via motion-decoupled diffusion model," in *CVPR*, 2024.
- [5] H. Zhu, Y. Li, F. Zhu, A. Zheng, and R. He, "Let's play music: Audio-driven performance video generation," in *ICPR*, 2021.
- [6] M. S. Islam, M. S. Rahman, and M. A. Amin, "Beat based realistic dance video generation using deep learning," in *RAAICON*, 2019.
- [7] D. Chang, Y. Shi, Q. Gao, J. Fu, H. Xu, G. Song, Q. Yan, X. Yang, and M. Soleymani, "Magicdance: Realistic human dance video generation with motions & facial expressions transfer," *arXiv*, 2023.
- [8] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *CVPR*, 2024.
- [9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.
- [11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.
- [12] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: a review," *CSUR*, vol. 55, no. 2, pp. 1–25, 2022.
- [13] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," *arXiv*, 2023.
- [14] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, L.-H. Lee, T.-H. Kim, C. S. Hong, and C. Zhang, "Sora as an agi world model? a complete survey on text-to-video generation," *arXiv*, 2024.
- [15] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, "A survey on long video generation: Challenges, methods, and prospects," *arXiv*, 2024.
- [16] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?: A survey and benchmark," *arXiv*, 2020.
- [17] Z. Chen *et al.*, "A survey on talking head generation," *Journal of Computer-Aided Design & Computer Graphics*, 2023.
- [18] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [19] F. Liao, X. Zou, and W. Wong, "Appearance and pose-guided human generation: A survey," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–35, 2024.
- [20] T. Sha, W. Zhang, T. Shen, Z. Li, and T. Mei, "Deep person generation: A survey from the perspective of face, pose, and cloth synthesis," *ACM Computing Surveys*, vol. 55, no. 12, 2023.
- [21] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *ICPR*, 2010.
- [22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.
- [24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.
- [25] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv*, 2018.
- [26] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *CVPR*, 2018.
- [27] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," in *CVPR*, 2024.
- [28] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.
- [29] D. Chang, Y. Shi, Q. Gao, H. Xu, J. Fu, G. Song, Q. Yan, Y. Zhu, X. Yang, and M. Soleymani, "Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion," in *ICML*, 2023.
- [30] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *ICCV*, 2021.
- [31] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023.
- [32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *NeurIPS*, 2016.
- [33] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou, "Learning hierarchical cross-modal association for co-speech gesture generation," in *CVPR*, 2022.
- [34] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *ICCV*, 2023.
- [35] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.
- [36] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *NeurIPS*, 2019.
- [37] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *CVPR*, 2022.
- [38] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.
- [39] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv*, 2012.
- [40] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
- [41] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.
- [42] S. M. Rajput, M. Bilal, and A. Habib, "Human activity recognition (har - video dataset)," 2023.
- [43] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3dpeople: Modeling the geometry of dressed humans," in *ICCV*, 2019.
- [44] N. Sadoughi, Y. Liu, and C. Busso, "Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents," in *FG*, 2015.
- [45] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, 2019.
- [46] Y. Jafarian and H. S. Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *CVPR*, 2021.
- [47] X. Guo, Y. Zhao, and J. Li, "Danceit: music-inspired dancing video synthesis," *IEEE Transactions on Image Processing*, vol. 30, pp. 5559–5572, 2021.
- [48] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," in *CVPR*, 2024.
- [49] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *ACM MM Workshops*, 2017.
- [50] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.
- [51] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016.
- [52] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "Dwnet: Dense warp-based network for pose-guided human video generation," *arXiv*, 2019.
- [53] Y. Jiang, S. Yang, T. L. Koh, W. Wu, C. C. Loy, and Z. Liu, "Text2performer: Text-driven human video generation," in *ICCV*, 2023.
- [54] X. Ju, A. Zeng, W. Jianan, X. Qiang, and Z. Lei, "Human-art: A versatile human-centric dataset bridging natural and artificial scenes," in *CVPR*, 2023.
- [55] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv*, 2018.
- [56] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *CVPR*, 2018.
- [57] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: a large-scale

- multimodal dataset for continuous american sign language,” in *CVPR*, 2021.
- [58] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, “Arbee: Towards automated recognition of bodily expression of emotion in the wild,” *International journal of computer vision*, vol. 128, pp. 1–25, 2020.
- [59] L. Liu and L. Liu, “Cross-modal mutual learning for cued speech recognition,” in *ICASSP*, 2023.
- [60] L. Liu, L. Liu, and H. Li, “Computation and parameter efficient multi-modal fusion transformer for cued speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [61] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, “Learning individual styles of conversational gesture,” in *CVPR*, 2019.
- [62] X. Liu, Q. Wu, H. Zhou, Y. Du, W. Wu, D. Lin, and Z. Liu, “Audio-driven co-speech gesture video generation,” in *NeurIPS*, 2022.
- [63] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, “Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots,” in *ICRA*, 2019.
- [64] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, “Motion representations for articulated animation,” in *CVPR*, 2021.
- [65] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.
- [66] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” in *ICCV*, 2023.
- [67] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocation,” in *ICCV*, 2015.
- [68] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [69] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” *arXiv*, 2016.
- [70] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, “Monocular expressive body regression through body-driven attention,” in *ECCV*, 2020.
- [71] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, “Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2022.
- [72] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, “Motionbert: A unified perspective on learning human motion representations,” in *ICCV*, 2023.
- [73] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: A skinned multi-person linear model,” in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.
- [74] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body capture: 3d hands, face, and body from a single image,” in *CVPR*, 2019.
- [75] M. Contributors, “Mmflow: Openmmlab optical flow toolbox and benchmark,” 2021.
- [76] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *CVPR*, 2017.
- [77] Z. Teed and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” in *ECCV*, 2020.
- [78] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints,” in *CVPR*, 2018.
- [79] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in *ICCV*, 2019.
- [80] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth anything: Unleashing the power of large-scale unlabeled data,” in *CVPR*, 2024.
- [81] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *CVPR*, 2018.
- [82] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.
- [83] T. Kim, C. Kang, J. Park, D. Jeong, C. Yang, S.-J. Kang, and K. Kong, “Human motion aware text-to-video generation with explicit camera control,” in *WACV*, 2024.
- [84] S. Stoll, S. Hadfield, and R. Bowden, “Signsynth: Data-driven sign language video generation,” in *ECCV*, 2020.
- [85] B. Natarajan, E. Rajalakshmi, R. Elakkiya, K. Kotecha, A. Abraham, L. A. Gabralla, and V. Subramaniyaswamy, “Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation,” *IEEE Access*, vol. 10, pp. 104 358–104 374, 2022.
- [86] S. Fang, L. Wang, C. Zheng, Y. Tian, and C. Chen, “Signllm: Sign languages production large language models,” *arXiv*, 2024.
- [87] R. O. Cornett, “Cued speech,” *American annals of the deaf*, pp. 3–13, 1967.
- [88] L. Liu and G. Feng, “A pilot study on mandarin chinese cued speech,” *American Annals of the Deaf*, vol. 164, no. 4, pp. 496–518, 2019.
- [89] W. Lei, L. Liu, and J. Wang, “Bridge to non-barrier communication: Gloss-prompted fine-grained cued speech gesture generation with diffusion model,” *arXiv*, 2024.
- [90] X. Wang, H. Wang, D. Liu, and W. Cai, “Dance any beat: Blending beats with visuals in dance video generation,” *arXiv*, 2024.
- [91] R. Jia and S. Pang, “Music2play: Audio-driven instrumental animation,” in *CAC*, 2023.
- [92] M. Liao, S. Zhang, P. Wang, H. Zhu, X. Zuo, and R. Yang, “Speech2video synthesis with 3d skeleton regularization and expressive body poses,” in *ACCV*, 2020.
- [93] C. Zhang, C. Wang, Y. Zhao, S. Cheng, L. Luo, and X. Guo, “Dr2: Disentangled recurrent representation learning for data-efficient speech video synthesis,” in *WACV*, 2024.
- [94] S. Hogue, C. Zhang, H. Daruger, Y. Tian, and X. Guo, “Diffited: One-shot audio-driven ted talk video generation with diffusion-based co-speech gestures,” in *CVPR*, 2024.
- [95] Y. Gao, Y. Zhou, J. Wang, X. Li, X. Ming, and Y. Lu, “High-fidelity and freely controllable talking head video generation,” in *CVPR*, 2023.
- [96] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, “Efficient emotional adaptation for audio-driven talking-head generation,” in *ICCV*, 2023.
- [97] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” *NeurIPS*, 2017.
- [98] Q. Wang, Z. Jiang, C. Xu, J. Zhang, Y. Wang, X. Zhang, Y. Cao, W. Cao, C. Wang, and Y. Fu, “Vividpose: Advancing stable video diffusion for realistic human image animation,” *arXiv*, 2024.
- [99] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, “Pose guided human video generation,” in *ECCV*, 2018.
- [100] J. S. Yoon, L. Liu, V. Golyanik, K. Sarkar, H. S. Park, and C. Theobalt, “Pose-guided human animation from a single image in the wild,” in *CVPR*, 2021.
- [101] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, “Deep video generation, prediction and completion of human action sequences,” in *ECCV*, 2018.
- [102] L. Yang, Z. Zhao, S. Wang, S. Wang, S. Ma, and W. Gao, “Disentangled human action video generation via decoupled learning,” in *ICMEW*, 2019.
- [103] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” in *ICCV*, 2019.
- [104] N. Fushishita, A. Tejero-de Pablos, Y. Mukuta, and T. Harada, “Long-term human video generation of multiple futures using poses,” in *ECCV*, 2020.
- [105] X. Sun, H. Xu, and K. Saenko, “Twostreamvan: Improving motion modeling in video generation,” in *WACV*, 2020.
- [106] C. Kissel, C. Kümmel, D. Ritter, and K. Hildebrand, “Pose-guided sign language video gan with dynamic lambda,” *arXiv*, 2021.
- [107] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv*, 2014.
- [108] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017.
- [109] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *CVPR*, 2018.
- [110] S. Tu, Q. Dai, Z. Zhang, S. Xie, Z.-Q. Cheng, C. Luo, X. Han, Z. Wu, and Y.-G. Jiang, “Motionfollower: Editing video motion via lightweight score-guided diffusion,” *arXiv*, 2024.
- [111] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, and F. Zou, “Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance,” *arXiv*, 2024.
- [112] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, “Unianimate: Taming unified video diffusion models for consistent human image animation,” *arXiv*, 2024.
- [113] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.
- [114] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, “Stable video diffusion: Scaling latent video diffusion models to large datasets,” *arXiv*, 2023.

- [115] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “Align your latents: High-resolution video synthesis with latent diffusion models,” in *CVPR*, 2023.
- [116] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *ICCV*, 2023.
- [117] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” in *ICCV*, 2023.
- [118] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, “Dreampose: Fashion image-to-video synthesis via stable diffusion,” in *ICCV*, 2023.
- [119] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, “Magicanimate: Temporally consistent human image animation using diffusion model,” in *CVPR*, 2024.
- [120] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, “Human4dit: Free-view human video generation with 4d diffusion transformer,” *arXiv*, 2024.
- [121] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: A skinned multi-person linear model,” in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023.
- [122] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, “Latte: Latent diffusion transformer for video generation,” *arXiv*, 2024.
- [123] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.
- [124] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *CVPR workshops*, 2010.
- [125] J. Liu, K. Yu, M. Feng, X. Guo, and M. Cui, “Disentangling foreground and background motion for enhanced realism in human video generation,” *arXiv*, 2024.
- [126] Y. Wang, X. Ma, X. Chen, C. Chen, A. Dantcheva, B. Dai, and Y. Qiao, “Leo: Generative latent image animator for human video synthesis,” *arXiv*, 2023.
- [127] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, “Latent image animator: Learning to animate images via latent space navigation,” *arXiv*, 2022.
- [128] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” *NeurIPS*, 2019.
- [129] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” *arXiv*, 2018.
- [130] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, “Celebv-hq: A large-scale video facial attributes dataset,” in *ECCV*, 2022.
- [131] M. Feng, J. Liu, K. Yu, Y. Yao, Z. Hui, X. Guo, X. Lin, H. Xue, C. Shi, X. Li *et al.*, “Dreamoving: A human video generation framework based on diffusion models,” *arXiv*, 2023.
- [132] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, “Zoedepth: Zero-shot transfer by combining relative and metric depth,” *arXiv*, 2023.
- [133] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv*, 2023.
- [134] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, “Champ: Controllable and consistent human image animation with 3d parametric guidance,” *arXiv*, 2024.
- [135] B. Zhu, F. Wang, T. Lu, P. Liu, J. Su, J. Liu, Y. Zhang, Z. Wu, Y.-G. Jiang, and G.-J. Qi, “Poseanimate: Zero-shot high fidelity pose controllable character animation,” *arXiv*, 2024.
- [136] J. Liu, Y. Yao, W. Hou, M. Cui, X. Xie, C. Zhang, and X.-s. Hua, “Boosting semantic human matting with coarse annotations,” in *CVPR*, 2020.
- [137] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, “Smpler-x: Scaling up expressive human pose and shape estimation,” *NeurIPS*, 2024.
- [138] J. Xue, H. Wang, Q. Tian, Y. Ma, A. Wang, Z. Zhao, S. Min, W. Zhao, K. Zhang, H.-Y. Shum *et al.*, “Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control,” *arXiv*, 2024.
- [139] W. Lu, Y. Xu, J. Zhang, C. Wang, and D. Tao, “Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting,” *arXiv*, 2023.