A Survey on Human Motion Video Generation

Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, *Senior Member, IEEE*, Zhensong Zhang, Minglei Li, Jian Yang, Fei Ma, Zhiyong Wu, *Member, IEEE*, Changpeng Yang, Zonghong Dai, and Fei Richard Yu, *Fellow, IEEE*

***Abstract*—Human motion video generation has garnered significant research interest due to its broad applications, enabling innovations such as photorealistic singing heads or dynamic avatars that seamlessly dance to music. However, existing surveys in this field focus on individual methods, lacking a comprehensive overview of the entire generative process. This paper addresses this gap by providing an in-depth survey of human motion video generation, encompassing over ten sub-tasks, and detailing the five key phases of the generation process: input, motion planning, motion video generation, refinement, and output. Notably, this is the first survey that discusses the potential of large language models in enhancing human motion video generation. Our survey reviews the latest developments and technological trends in human motion video generation across three primary modalities: vision, text, and audio. By covering over two hundred papers, we offer a thorough overview of the field and highlight milestone works that have driven significant technological breakthroughs. Our goal for this survey is to unveil the prospects of human motion video generation and serve as a valuable resource for advancing the comprehensive applications of digital humans. A complete list of the models examined in this survey is available in Our Repository.**

***Index Terms*—Human Motion Video Generation, Video Generation, Multi-Conditions Driven, AIGC**

I. Introduction

HUMAN motion video generation refers to utilizing generative networks to synthesize frame sequences that plan and depict human poses and motions, based on various inputs such as vision cues, text prompts, and audio signals. Early works in this field are limited to creating cartoon characters or human models lacking realistic textures [1], [2]. With the advent of general text-to-video generation models [3]–[5], the field has expanded to produce videos with realistic textures and human-like quality.

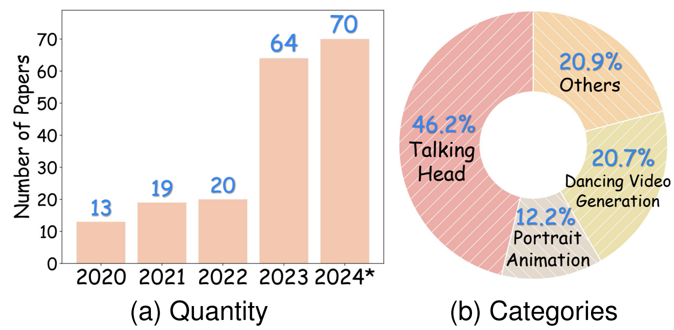

Fig. 1. Quantity of papers in the four categories reviewed in this survey, showing rapid growth in human motion video generation, emphasizing key areas such as talking head and dance videos. (2024* denotes the period from Jan. to Aug. in 2024.)

Recently, research on human-centered video synthesis [6], [7] has flourished, gaining significant interest in areas such as talking head, portrait animation, and dance video generation, as shown in Fig. 1. To minimize the uncanny valley effect and enhance human-computer interactions, the generation of photorealistic human motion videos has emerged as a prominent topic, involving the creation of videos with human-like appearances, realistic motions, and natural expressions.

Existing reviews [10]–[14] often concentrate on specific subtasks within human motion video generation and do not provide a complete pipeline for generating human-centric videos. Therefore, we clearly define *five key phases* for the human motion video generation task, as illustrated in Fig. 2. These phases collaboratively transform reference inputs into high-fidelity, lifelike human motion videos, enabling real-time and cost-efficient applications. The process begins with identifying driving sources, which include vision cues, text prompts, or audio signals. Notably, generating facial regions and holistic human bodies often requires distinct frameworks. The second phase involves motion planning based on these inputs. While previous approaches [15], [16] use input conditions to design implicit feature mappings for human motion, recent studies [17]–[19] explore the use of large language models (LLMs)

H. Xue, X. Luo, and Z. Wu are with Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. (E-mail: xhw22@mails.tsinghua.edu.cn, luo-xy24@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn)

J. Liu is with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (E-mail: jz.liu@siat.ac.cn)

Z. Zhang is with Huawei Noah's Ark Lab, Shenzhen, China. (E-mail: zhangzhensong@huawei.com)

Z. Hu, M. Li, and C. Yang are with 01.AI, Beijing, China. (E-mail: huzh666295@gmail.com, liminglei29@gmail.com, yangchangpeng@01.ai)

X. Zhang is with School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. (E-mail: zhangxin0526@stu.xjtu.edu.cn)

Z. Dai is with Artificial Intelligence Innovation and Incubation (AI') Institute of Fudan University. (E-mail: zhdai23@m.fudan.edu.cn)

X. Xiang is with University of Chinese Academy of Sciences, Beijing, China. (E-mail: xiangxunzhi21@mails.ucas.ac.cn)

Y. Dai and J. Yang are with PCA Lab, Nanjing University of Science and Technology, China. (E-mail: daiy@njust.edu.cn, csjyang@njust.edu.cn)

F. Ma is with Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ). (E-mail: mafei@gml.ac.cn)

Fei Richard Yu is with the College of Computer Science and Software Engineering, ShenZhen University, ShenZhen, Guangdong, China (e-mail: yufei@szu.edu.cn)

F. Ma and Z. Wu are the corresponding authors.

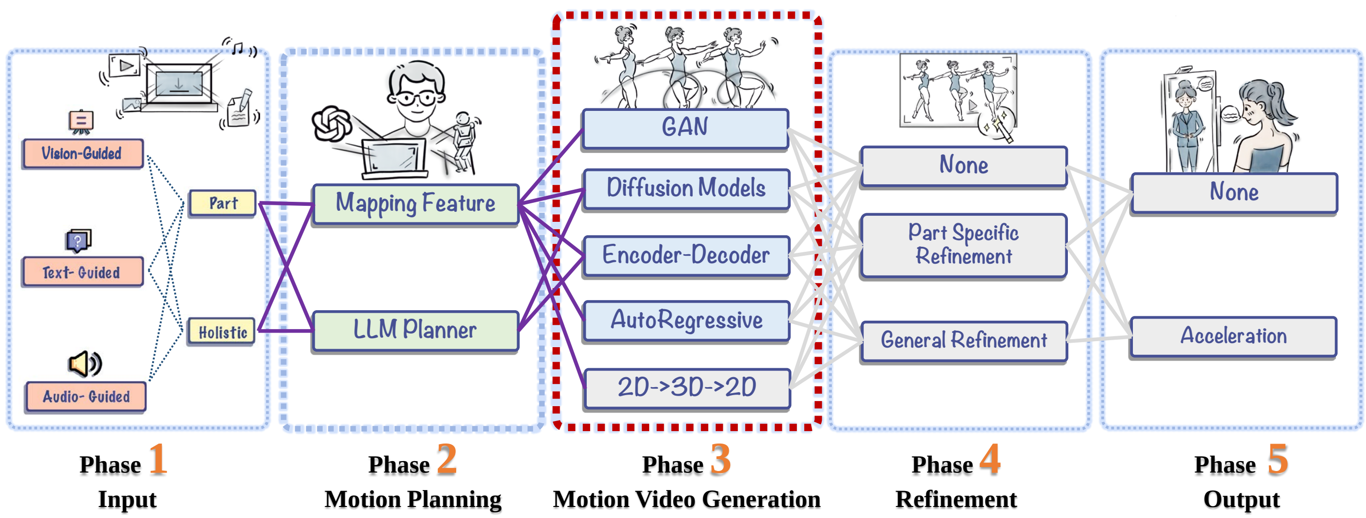

Fig. 2. Pipeline of generating human motion videos, which can be divided into five key phases. Initially, diverse input sources such as vision cues, text prompts, and audio signals are identified. Next, these inputs guide the planning of human motions, either through feature mapping or utilizing LLMs. The third phase focuses on modelling and translating these motion signals into video outputs according to the input conditions. Subsequently, the generated videos undergo refinement, with particular attention to optimizing details such as hand movements, synchronizing oral movements, adjusting gaze points, and enhancing the overall video quality. Finally, efforts focus on reducing production costs, enabling real-time streaming platform integration of digital humans, and incorporating practical functionalities. It is worth noting that this review does not cover the reconstruction of 3D assets from 2D images or video generation through rendering techniques such as NeRF [8] or 3DGS [9].

for motion planning. The third phase focuses on generating human motion videos, ensuring body consistency, precise movements, and high-quality output. The fourth phase involves enhancing the generated videos, optimizing hand movements, synchronizing mouth and teeth movements, adjusting eye gaze, and improving overall video quality. Finally, the fifth phase addresses the deployment of digital humans on real-time streaming platforms, integrating practical functionalities.

We identify three primary modalities that drive human motion video generation: vision, text, and audio. Given that many methods involve multiple modalities, with vision being ubiquitous and audio having a stronger influence than text when both are present, we categorize methods based on the following criteria: 1) If a method includes audio, it is classified as audio-driven, even if other modalities are also involved. 2) If a method is not in the audio-driven category but includes text, it is classified as text-driven. 3) A method is classified as vision-driven if it only utilizes the vision modality. This classification system effectively organizes over 200 papers into these three distinct categories, and the timeline of representative works for these three driving modalities is illustrated in Fig. 3.

For vision-based conditions, portrait animation primarily focuses on generating specific facial expressions. On a broader scale, pose-driven, video-driven dance video generation, and virtual Try-On techniques represent emerging research hotspots. It should be noted that we emphasize the distinction between portrait animation and talking head. Portrait animation relies on reference images and pose sequences, whereas talking head utilizes reference images alongside driving audio, often referred to as audio-driven portrait animation. Motion transfer, which spans both facial and holistic human animation, intersects with tasks like portrait animation, dance video generation, and pose2video. Text-driven methods include Text2Face, which generates facial animations from first-personal scripts or instructions, and Text2MotionVideo, which extends this concept to create holistic human motions from text prompts. While much of the current research is focused on Text2Motion3D, as discussed by Zhu et al. [20], our survey specifically addresses human motion video generation, placing 3D skeletal motions outside its scope. In audio-driven scenarios, we elaborate on the related studies in the order of the driving region from small to large, including audio-lip synchronization, head pose driving, and holistic human body motion generation.

Unlike previous surveys [10], [11] listed in Table I, our work offers a comprehensive definition of the five key phases in human motion video generation. Notably, we are the first to discuss the application and potential of LLMs in motion planning. Our survey covers a broader spectrum, providing detailed categorization of various methods. Within this framework, various sub-branches emerge, each focusing on specific aspects of human motion video generation, as summarized in Table II. Additionally, we collect 64 human-centered datasets to support related tasks, offering detailed information such as data duration and resolution. Moreover, we identify current challenges while offering insights into future research and development directions. The main contributions of this paper are summarized as follows:

- We decompose human motion video generation into five key phases, covering all subtasks across various driving sources and body regions. To the best of our knowledge, this is the first survey to offer such a comprehensive framework for human motion video generation.
- We provide an in-depth analysis of human motion video generation from both motion planning and motion generation perspectives, a dimension that has been underexplored in existing reviews.

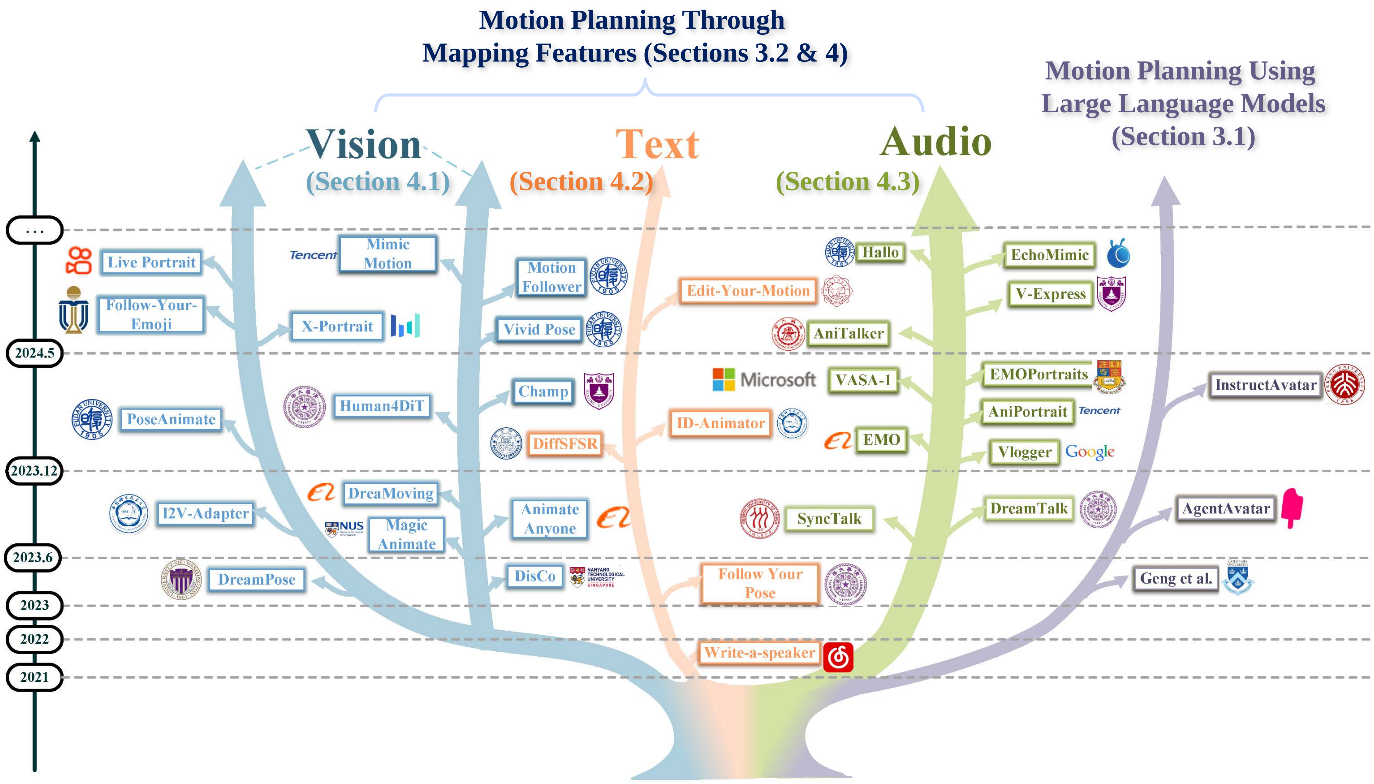

Fig. 3. Timeline of key advances in vision-, text-, and audio-driven human motion video generation methods.

- We clearly delineate established baselines and evaluation metrics, offering detailed insights into the key challenges shaping this field.
- We present a set of potential future research directions, aimed at inspiring and guiding researchers in the field of human motion video generation.

Given the significant advancements and extensive applications of human motion video generation, we present a comprehensive survey to help the community track its progress. The remaining sections of this survey are organized as follows: Section II establishes the foundational knowledge necessary for understanding human motion video generation. Next, Section III introduces innovative approaches utilizing LLMs for human motion planning. Following this, Section IV delves into the methodologies of motion modeling and video generation. In Section V, we briefly explore the strategies in the final two phases: refinement and output. In Section VI, we consolidate popular metrics and datasets. Finally, Section VII addresses the current challenges and provides potential future directions in human motion video generation. Note that all statistics referenced in this paper are current as of August 30, 2024.

TABLE I
SUMMARY OF RELATED SURVEYS.
Previous surveys have primarily concentrated on specific sub-tasks within human-centered video generation, often lacking a comprehensive perspective and omitting the role of LLMs in the video generation process. In contrast, our survey provides a more thorough coverage, firstly encompassing both human motion planning and video generation.

	Sha et al. [10]	Lei et al. [11]	Ours	
Generation Pipeline	⊘	⊘	✓	
Motion Planning	⊘	⊘	✓	
LLMs as Motion Planner	⊘	⊘	✓	
Motion Video Generation	✓	✓	✓	
Refinement	⊘	⊘	✓	
Structure Summary	✓	⊘	✓	
Benchmarks	✓	✓	✓	
Evaluation	⊘	⊘	✓	
Challenges	⊘	✓	✓	
	Sub-Tasks			Other Survey
Portrait Animation	✓	⊘	✓	⊘
Video-Driven Dance Video Generation	⊘	⊘	✓	⊘
Pose-Driven Dance Video Generation	✓	✓	✓	⊘
Try-On	✓	⊘	✓	Song et al. [12]
Text2Face	✓	⊘	✓	⊘
Text2MotionVideo	⊘	✓	✓	⊘
Talking Head	✓	⊘	✓	Meng et al. [13]
Audio Driven Holistic Human Motion Gerneration	⊘	✓	✓	⊘
Music-Driven Dance Video Generation	⊘	✓	✓	⊘
Multilingual Video Dubbing	⊘	⊘	✓	Bigioi et al. [14]

II. PRELIMINARIES

A. Generative Frameworks

Variational Autoencoder (VAE). VAE, introduced in 2013 by Kingma and Welling [64], has become a prominent generative model, recognized for its robust data representation capabilities. In the human motion video generation task, VAE and its variants are instrumental in encoding vision signals or reference images, facilitating the generation of corresponding videos [65], [66]. These factors lead to the adoption of variants, such as VQ-VAE [67]. However, VAE is susceptible to mode collapse and often produces samples that are less sharp compared to those generated by generative adversarial

TABLE II
FIVE PHASES IN THE HUMAN MOTION VIDEO GENERATION PIPELINE, EACH EMPHASIZING THE SIGNALS, REGIONS, AND MODELS EMPLOYED BY REPRESENTATIVE WORKS ACROSS VARIOUS TASKS.

Phase 1		Phase 2	Phase 3	Phase 4	Phase 5		
Input	Region	Motion Planning	Generation Model	Refinement	Acceleration		
Vision-driven Methods Text-driven Methods Audio-driven Methods	Part (Face) Holistic Human	Feature Mapping LLMs Planner	GAN Diffusion Model (DM) AutoRegressive (AR) Encoder-Decoder (ED)	None Hand Refinement Face Refinement	None Acceleration	Tasks	Related Works
Vision-driven	Part (Face)	Feature Mapping	GAN / ED	None / Face Refinement	None	Portrait Animation	OmniAvatar [21] GazeGANV2 [22] EDTN [23] OTAvatar [24]
Vision-driven	Part (Face)	Feature Mapping	DM	None / Face Refinement	None		Follow-Your-Emoji [25] LivePortrait [26] X-Portrait [27]
Vision-driven	Part (Face)	Feature Mapping	DM	None	Acceleration		MobilePortrait [28]
Vision-driven	Holistic Human	Feature Mapping	GAN / ED	None	None	Video-Driven Dance Video Generation	EDN [29] Human MotionFormer [30] FakeVideo [31]
Vision-driven	Holistic Human	Feature Mapping	DM	None	None		BTDM [32]
Vision-driven	Holistic Human	Feature Mapping	DM	None	None	Pose-Driven Dance Video Generation	DisCo [33] Animate Anyone [34] Follow-Your-Pose v2 [35]
Vision-driven	Holistic Human	Feature Mapping	DM (DiT)	None	None		Human4DiT [36]
Vision-driven	Holistic Human	Feature Mapping	DM	Hand Refinement	None		MimicMotion [37]
Vision-driven	Holistic Human	Feature Mapping	DM	None	Acceleration		I2V-Adapter [38]
Vision-driven	Holistic Human	Feature Mapping	DM	None	None	Try-On	Tunnel Try-On [39] ViViD [40] WildVidFit [41]
Vision-driven	Holistic Human	Feature Mapping	DM	None	None	Pose2Video	DreamPose [42] Make-Your-Anchor [43]
Text-driven	Part (Face)	Feature Mapping	GAN / ED	None	None	Text2Face	Write-a-speaker [44] NEUTART [45] Faces that Speak [6]
Text-driven	Part (Face)	Feature Mapping	DM	None	None		ID-Animator [46]
Text-driven	Part (Face)	LLMs Planner	DM	None	None		Geng et al. [17]
Text-driven	Holistic Human	Feature Mapping	DM	None	None	Text2MotionVideo	Edit-Your-Motion [47] Dancing Avatar [48] Follow-Your-Pose [25] DiffSFSR [49]
Text-driven	Holistic Human	Feature Mapping	ED	None	None		Text2Performer [50]
Audio-driven	Part (Face)	Feature Mapping	GAN / ED	None	None	Talking Head	StyleHEAT [51] PC-AVS [52] EDTalk [53]
Audio-driven	Part (Face)	LLMs Planner	ED	None	None		AgentAvatar [18]
Audio-driven	Part (Face)	Feature Mapping	DM	None	None		AniTalker [54] EchoMimic [55] LinguaLinker [56] EMO [57] Hallo [58]
Audio-driven	Part (Face)	LLMs Planner	DM	None	None		InstructAvatar [19]
Audio-driven	Part (Face)	Feature Mapping	DM	Face Refinement	None		Liang et al. [59]
Audio-driven	Part (Face)	Feature Mapping	AR	None	None		MakeItTalk [60]
Audio-driven	Part (Face)	Feature Mapping	AR	None	Acceleration		Live Speech Portraits [61]
Audio-driven	Holistic Human	Feature Mapping	DM	None	None	Audio-driven Holistic Human Motion Driving	Vlogger [62]
Audio-driven	Holistic Human	Feature Mapping	DM	None	None	Music-Driven Dance Video Generation	Dance-Any-Beat [63]

networks. To address these limitations, VAE is increasingly being combined with diffusion models to enhance the quality of generated outputs.

Generative Adversarial Networks (GANs). First proposed by Ian Goodfellow et al. in 2014 [68], GANs comprise two adversarial neural networks: a generator G and a discriminator D. The generator creates data that mimics real samples, while the discriminator distinguishes between real and generated data. Unlike the degradation in generation quality caused by the strong mathematical priors in VAE, GANs implicitly map feature relationships, resulting in higher-quality generated outputs. Variants such as StyleGAN [69], [70] achieve significant milestones, particularly in human motion video generation, including tasks like motion copy [29], [71] and talking head synthesis [6], [44]. However, GANs are limited by the lack of diversity in generated samples and are prone to mode collapse, primarily due to the challenges in balancing the training dynamics between the generator and discriminator.

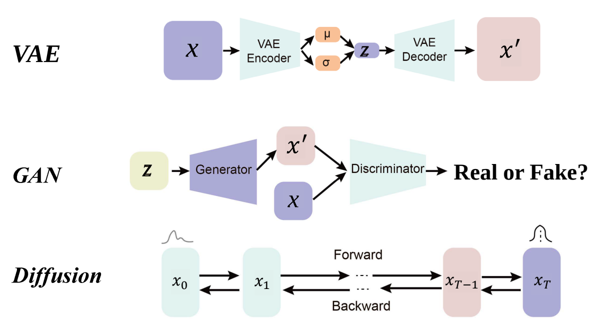

Fig. 4. Outlines of VAE, GAN, and diffusion models.

Diffusion Models (DMs). Traditional methods [72] often struggle with maintaining continuity and realism in generated videos, whereas diffusion models can progressively restore details in the target video, thereby better preserving motion continuity and visual quality [73]. Therefore, diffusion models [74] as shown in Fig. 4, attract widespread attention in the field of generative modelling [75]–[79]. Formally, DM processes can be interpreted as a sequence of denoising steps, represented by a denoising predictor $\epsilon_\theta(x_t, t)$, where $t = 1, 2, \cdots, T$. During training, $\epsilon_\theta(x_t, t)$ is optimized to predict a denoised version of x_t, where x_t is a noise version of the input x. The training objective can be simplified to

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{1}$$

where t is uniformly sampled from $1, 2, \cdots, T$. During inference, the input latent map $\mathbf{Z}_T$ is generated from a random Gaussian distribution. Given $\mathbf{Z}_T$, ϵ_θ predicts noise at each step t, conditioned on C. These models generate samples by progressively denoising an initially noisy input, and their training objective can be expressed as a reweighted variational lower bound [80], offering benefits such as distribution coverage, a stationary training objective, and easy scalability [78].

However, despite achieving high-quality sample generation, diffusion models incur significant computational overhead. Latent diffusion models (LDMs) apply the diffusion process in a latent space, significantly reducing computational costs. For example, Stable Video Diffusion (SVD) [5] and Animatediff [4] are classic text-to-video generation methods employing LDMs. Several studies [7], [81] on human motion video generation are based on this method, extending the text driven models to multi-condition driven ones.

B. Human Data Representations

Following Knap [82], we categorize human body posture representations into seven key types: mask, mesh, depth, normal, keypoint, semantics, and optical flow, as illustrated in Fig. 5.

Mask. Mask outlines the basic contours and occupied areas of characters, providing a coarse-grained layout prior. However, they lack detailed posture descriptions. Recent work has combined Grounding-DINO [83] with SAM [84] for person segmentation.

Mesh. Mesh offers a more detailed representation of body shape, including limb bending and occlusion. Currently, mesh-based representation is applied to overcome the loss of 3D spatial information in keypoint estimation, including examples of the Skinned Multi-Person Linear model (SMPL) [85], EDTN [23], OTAvatar [24], and DeCo [86].

Depth. Insufficient understanding of spatial relations limits existing methods' ability to accurately generate occluded body parts [35]. Incorporating depth cues can effectively leverage spatial information, aiding the model in learning the spatial relationships between characters. Depth Anything [87] leverages both labeled and unlabeled images to enhance monocular depth estimation, enabling existing methods to effectively generate depth maps with improved accuracy.

Normal. The normal condition critically emphasizes the orientation of the human body, ensuring precise alignment in the generation of realistic human poses for enhanced visual fidelity [88]. This alignment is paramount for maintaining the integrity of spatial relationships in animations, thereby significantly improving the believability and immersive quality of the rendered scenes. Spaiens [89] provides an efficient and versatile framework for calculating normal maps, facilitating the enhancement of surface detail in various applications.

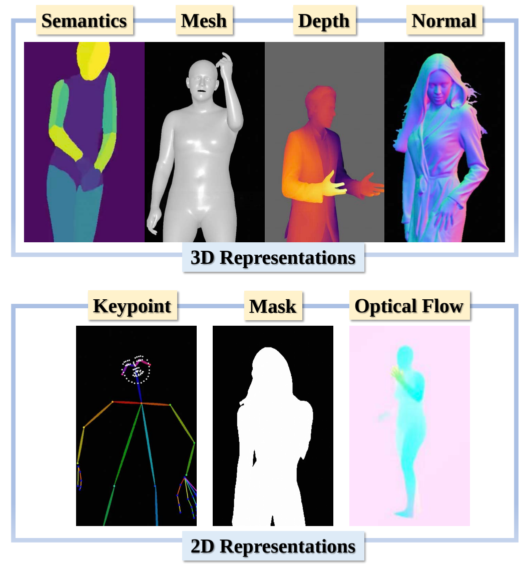

Fig. 5. Different human data representations.

TABLE III
COMPARISON OF 2D AND 3D METHODS.

	Body Shape Estimation Accuracy	Estimation Speed
2D Estimation	★★	★★★
3D Estimation	★★★	★★

Keypoint. Existing pose keypoint estimation methods typically employ models such as OpenPose [90] or DWPose [91]. Recent studies [34], [92] use DWPose [91] as the foundational keypoint-based human skeleton representation for modeling human motion. Additionally, optical flow [93], depth maps [94], and DensePose [95] are also common methods for representing 2D body data.

Semantics. The semantic condition prioritizes the semantic information of different body parts, facilitating the decoupling of distinct features. This approach enables targeted enhancements and independent manipulation of each body part, thereby refining the representation and interaction dynamics within human-centric models [96]. DensePose [95] is a framework that maps all human pixels of an image to the 3D surface of the human body, enabling precise localization and understanding of human poses.

Optical Flow. Directly optimizing the model on a noisy dataset often leads to background instability. To ensure that the generated video's background remains consistent with the reference images and avoids unintended variations, incorporating optical flow can provide the model with priors on illumination changes and motion dynamics [35].

Depending on whether it contains three-dimensional spatial information, we can divide the above representations into 2D and 3D categories. Keypoint and optical flow are categorized as 2D representations, which primarily capture spatial relationships and silhouette details. In contrast, mesh, depth map, normal map, and semantics are grouped as 3D representations. Table III compares the advantages and disadvantages of 2D and 3D estimation methods.

TABLE IV
SUMMARY OF RECENT WORKS ON MOTION PLANNING USING LARGE LANGUAGE MODELS.
We categorize them by driving regions (face and holistic body). Motion Codebook: a set of embeddings, each representing a distinct discrete motion token; DM: Diffusion Model; AR: Auto-Regressive; ED: Encoder-Decoder.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Task
2D						
Part (Face)	Jan. 26, 2023	Geng et al. [17]	+	Motion Descriptions	DM	Taking Head (Listener)
	Nov. 29, 2023	AgentAvatar [18]		Motion Descriptions	ED	Taking Head
	May. 24, 2024	InstructAvatar [19]	+ +	Latent	DM	Taking Head
3D						
Holistic Human	Aug. 21, 2023	Ng et al. [97]	+	Latent	AM	Text2Motion3D (Listener)
	Jun. 19, 2023	MotionGPT [98]		Latent	AM	Text2Motion3D
	Nov. 27, 2023	InterControl [99]		Latent	DM	Text2Motion3D
	Nov. 28, 2023	AvatarGPT [100]		Latent	AM	Text2Motion3D
	Dec. 7, 2023	MoMat-MoGen [101]		Latent	DM	Text2Motion3D
	Dec. 19, 2023	MotionScript [102]		Latent	DM	Text2Motion3D
	Dec. 22, 2023	PRO-Motion [103]		Latent	AM	Text2Motion3D
	Dec. 22, 2023	FineMoGen [104]		Latent	AM	Text2Motion3D

: Reference Real Images; : Text Prompts; : Audio; : Motion Codebook

III. Human Motion Planning

The human motion planning phase is critical in determining the specific motions that a virtual digital human performs. The motion sequence generated from input signals enables the virtual character to exhibit highly natural movements, align with human habits, and interact smoothly with surrounding objects, effectively mimicking human behavior.

Currently, human motion planning is primarily driven by two methodologies: one that leverages the power of LLMs for motion planning, and another that relies on the mapping of distinct features for motion generation. This section highlights the growing significance of LLMs in human motion video generation. By leveraging the inherent prior knowledge embedded within LLMs, these models can better comprehend semantic nuances and reason about emotions.

A. Motion Planning Using Large Language Models

In the realm of human motion video generation, studies by Geng et al. [17], AgentAvatar [18], and InstructAvatar [19] represent the current advancements in applying LLMs to motion planning. Notably, Geng et al. [17] pioneered the analysis of dialogue characteristics between two individuals in conversations using LLMs, thereby inferring the appropriate expressions for the listener. The other two studies focus on the generation of single-person human motion videos.

One of the pioneering efforts in this area, Geng et al. [17], demonstrates the innovative use of LLMs in human motion video generation. Their approach starts by inputting the speaker's scripts and conversational intention into LLMs, which then generate plausible reactions for the listener, such as a subtle smile. These generated motion descriptions are subsequently used to train a CLIP module [105], effectively integrating the LLM with motion generation. AgentAvatar [18] takes a broader approach by incorporating LLMs into the general context of human motion video generation. This method extends beyond dialogue-driven scenarios, enabling the planning and generation of human movements across a wide range of contexts. Their process begins with an environmental overview and avatar settings for the LLM-based planner. The planner generates detailed descriptions of facial movements, which are then passed to the driving engine to produce photorealistic video sequences. InstructAvatar [19] designs an automatic annotation pipeline to construct a rich dataset of instruction-video pairs. This dataset captures fine-grained facial details using Action Units (AUs) to describe facial muscle movements. AUs are extracted using an off-the-shelf model and refined through multimodal LLMs, which paraphrase AUs into natural textual descriptions. This process

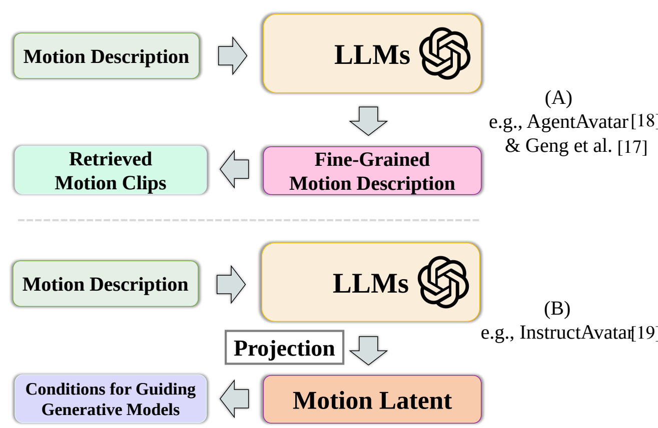

Fig. 6. Two common forms of human motion planning by LLMs. (A) LLMs generate fine-grained descriptions for retrieval, and (B) LLMs project descriptions into a latent space for guiding generative models.

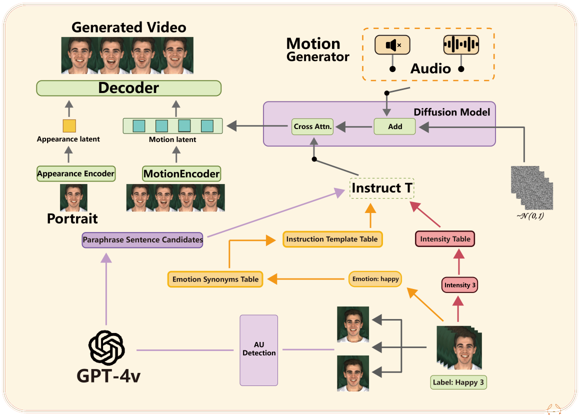

Fig. 7. Overview of InstructAvatar [19], which employs GPT-4 and diffusion models for video generation, producing expressive and dynamic videos that are synchronized with audio input.

not only enriches the dataset with detailed emotion and motion descriptions but also enhances the model's ability to generalize across various expressions and emotions. The architecture of InstructAvatar [19], as shown in Fig. 7, is underpinned by VAE [64] that disentangles motion from appearance, allowing for separate and focused manipulation of these elements. The motion generator, based on the Conformer architecture [106], employs a diffusion model to learn the mapping from audio and text instructions to the motion latent space. This model is further equipped with a two-branch cross attention mechanism that distinguishes between emotion and motion instructions, ensuring that the avatar maintains a consistent emotional state throughout the video while executing dynamic facial motions as directed by the text.

These studies [17]–[19] reveal two distinct motion planning strategies utilized by LLMs, as depicted in Fig. 6. In Fig. 6 (A), LLMs process motion descriptions to generate fine-grained motion descriptions, which are used to perform retrieval on a video database to obtain relevant clips. This retrieval-based approach is adopted by methods such as Geng et al. [17] and AgentAvatar [18]. In contrast, Fig. 6 (B) illustrates a generative strategy where the derived motion conditions serve as inputs to a generative model, as demonstrated by InstructAvatar [19].

LLMs are beginning to show promising results in 3D skeletal motion generation, acting as central orchestrators in the motion planning phase. For instance, FineMoGen [104], PRO-Motion [103], AvatarGPT [100], and MotionGPT [98] are capable of generating holistic human motions from text. MotionScript [102] facilitates the seamless transformation of motions into descriptive motion scripts and vice versa, establishing a robust bidirectional generation mechanism. Additionally, Ng et al. [97] introduce an innovative approach that leverages LLMs to synthesize 3D facial expressions tailored to audience characteristics. Furthermore, InterControl [99] and MoMat-MoGen [101] are beginning to explore the potential of LLMs in facilitating dyadic interaction tasks. All these works are summarized in Table IV. However, these efforts are limited to 3D skeletal motion and do not extend to the creation of authentic video content so far. The field of human motion video generation remains ripe for further investigation and advancement.

Currently, the generation of human motion videos predominantly relies on text as an intermediary to integrate LLMs with generative models. However, there remains a lack of exploration into more effective and novel intermediate representations. From the initial analysis, different challenges emerge:

- As shown in Table II, most existing works do not fully leverage LLMs as motion planners, leaving much of their potential untapped.
- Some works, such as Ng et al. [97], have explored intermediate representations beyond text, like codebooks. Future research could investigate additional modalities, including visual representations, skeletal signals, and others.
- Due to the substantial computational costs associated with LLMs during the reasoning process, which can degrade the interaction experience, optimizing the inference speed of the entire pipeline is a pressing issue for future work.
- Assessing the effectiveness of LLMs in motion planning is a crucial challenge that remains to be thoroughly addressed in future research.

B. Motion Planning through Mapping Features

In human motion video generation, the complexity of video synthesis has led most studies [109]–[111] to adopt implicit representations of input features. These approaches focus on learning the mapping between input conditions and motion, often generating diverse outcomes by incorporating a degree of stochastic noise. For instance, VASA-1 [112] encodes speech into a series of audio features and employs a motion latent diffusion model to map audio inputs to facial motion, thereby achieving audio-driven head pose and lip synchronization. A detailed discussion of these methods will be provided in Section IV.

IV. Motion Modeling and Video Generation

In the human motion video generation phase, our objective is to synthesize realistic videos that capture holistic human motion, grounded in the outcomes from the motion planning phase. To thoroughly examine the various generation approaches, we conduct comprehensive analyses on each specific sub-task. This section concentrates on video generation methods leveraging diffusion models, excluding other techniques like NeRF [8] and 3DGS [9]. We delve into three key areas: *vision guidance* (including portrait animation, dance video generation, TryOn, and pose-to-video), *text guidance* (covering Text2Face and Text2MotionVideo), and *audio-driven scenarios* (such as talking head generation and audio-driven holistic human motion generation).

A. Vision-Driven Human Motion Video Generation

Portrait Animation. Portrait animation focuses on breathing life into static images, typically portraits, using advanced animation techniques. The process begins with a static image, whether it is a photograph or a digital painting, and transforms it into an animated sequence that conveys the subject's

TABLE V
DETAILED REVIEW OF RECENT DEVELOPMENTS IN PORTRAIT ANIMATION.
This table showcases a spectrum of innovative approaches focusing on facial animations through advanced generative models. ✓indicates that the training and inference code are open-source, and × indicates that neither is publicly available. The same convention is applied in the subsequent tables. DM: Diffusion Model; ED: Encoder-Decoder; SIG.: SIGGRAPH; SIGA.: SIGGRAPH ASIA; 3D-P: 3D Parameterization.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Portrait Animation							
Part (Face)	Jun. 4, 2024	Follow-Your-Emoji [25]	+	KeyPoint	DM	×	SIGA.'24
	Jul. 5, 2024	LivePortrait [26]	+	KeyPoint	ED	✓	arXiv
	Jul. 9, 2024	MobilePortrait [28]	+	KeyPoint	DM	×	arXiv
	Oct. 16, 2023	EDTN [23]	+	3D-P	ED	×	ICASSP'24
	Mar. 26, 2023	OTAvatar [24]	+	3D-P	ED	✓	CVPR'23
	Mar. 27, 2023	OmniAvatar [21]	+	Latent	GAN	×	CVPR'23
	Dec. 4, 2023	GazeGANV2 [22]	+	Latent	GAN	✓	TIP'22
	Jun. 8, 2024	MegActor [107]	+	Latent	DM	×	arXiv
	May. 31, 2024	X-Portrait [27]	+	Latent	DM	×	SIG.'24
	Mar. 23, 2024	FaceOff [108]	+	Latent	ED	×	WACV'23

: Reference Real Images; : Driving Pose Video

emotional expressions. Recent advancements in this field are summarized in Table V.

OmniAvatar [21] exemplifies the use of geometric priors to guide the animation process, ensuring 3D consistency and detailed facial expressions. Follow-Your-Emoji [25] utilizes a diffusion-based framework for animating portraits with target landmark sequences. By using an expression-aware landmark to guide the animation, it ensures motion alignment and identity preservation while enhancing the portrayal of exaggerated expressions. Additionally, it employs a fine-grained facial loss function to improve the model's perception of subtle expressions and the reconstruction of the reference portrait's appearance. LivePortrait [26] introduces an efficient video-driven framework that balances computational efficiency with controllability, enabling rapid generation speeds.

Overall, portrait animation, which requires only a reference image to drive facial expressions based on input conditions, draws significant research focus. For instance, EDTN [23] and X-Portrait [27] address the challenge of cross-domain head reenactment, allowing human motions to be transferred to non-human domains, such as anime characters. MobilePortrait [28] focuses on real-time performance, offering one-shot solutions for talking face avatars with controllable rendering. FaceOff [108] presents a novel video-to-video face-swapping system that retains source expressions and identity while adapting to the pose and background of the target video. From numerous studies, we can identify different key challenges in portrait animation:

- Current facial driving techniques primarily focus on eye gaze, teeth, lip synchronization, and head posture. However, existing research often neglects the consistency of eye gaze and the finer details of teeth, failing to treat these aspects as an integrated whole.
- Maintaining identity consistency, especially in zero-shot identity preservation, remains a significant challenge.
- Most current studies are limited to single-person facial driving, and exploring methods for multi-person facial driving is a promising direction for future research.

Video-Driven Dance Video Generation. Video-driven dance video generation enables the transformation of a static individual into a dynamic dancer from a video source. This technology allows for the transfer of dance movements from a professional dancer to an amateur or non-dancer, enabling them to perform complex choreographies with lifelike fluidity.

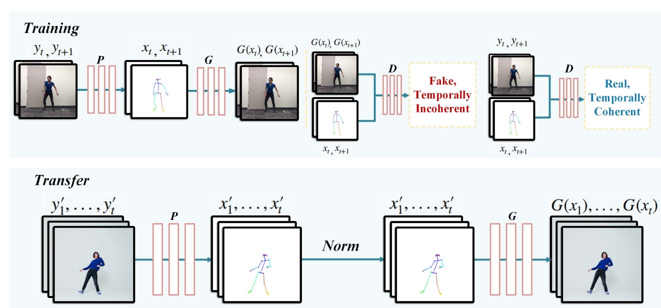

Fig. 8. Overview of Everybody Dance Now [29], depicting the two stages of the system: training and transfer.

Everybody Dance Now [29] introduces a method, illustrated in Fig. 8, for motion transfer using poses as an intermediate representation, enabling the transfer of dance performances onto subjects with different physical attributes.

BTDM [32] enforces temporal coherence by reducing motion ambiguity. Human MotionFormer [30], a hierarchical Vision Transformer [121] framework, is designed for transferring human motions by capturing both global and local perceptions to accurately match large and subtle motions. FakeVideo [31] employs perceptual loss and Gromov-Wasserstein loss [122] to bridge the gap between pose and appearance. It also introduces an episodic memory module to support continuous learning and uses facial geometrical cues to enhance facial details. The domain of human motion copy, particularly in video-driven dance video generation, presents key challenges that researchers and developers continue to address:

TABLE VI
COMPREHENSIVE OVERVIEW OF RESEARCH ON VIDEO-DRIVEN AND POSE-DRIVEN DANCE VIDEO GENERATION.
DM: Diffusion Model; ED: Encoder-Decoder; DiT: Diffusion Transformer; SIG.: SIGGRAPH; 3D-P: 3D Parameterization.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Video-Driven Dance Video Generation							
Holistic Human	Aug. 22, 2018	Chen et al. [29]	+	KeyPoint	GAN	✓	ICCV'19
	Jul. 02, 2023	BTDM [32]	+	Region	DM	×	arXiv
	Feb. 22, 2023	Human MotionFormer [30]	+	KeyPoint	ED	×	arXiv
	Jun. 24, 2024	FakeVideo [31]	+	KeyPoint	GAN	×	TDSC'24
Pose-Driven Dance Video Generation							
Holistic Human	Jun. 30, 2023	DisCo [33]	+	KeyPoint	DM	✓	CVPR'24
	Oct. 20, 2023	Dance-Your-Latents [81]	+	KeyPoint	DM	×	arxiv
	Nov. 18, 2023	MagicPose [113]	+	KeyPoint	DM	✓	ICML'24
	Nov. 27, 2023	MagicAnimate [96]	+	Region	DM	✓	CVPR'24
	Nov. 28, 2023	Animate Anyone [34]	+	KeyPoint	DM	×	CVPR'24
	Dec. 8, 2023	DreaMoving [114]	+ +	KeyPoint	DM	✓	arxiv
	Dec. 27, 2023	I2V-Adapter [38]	+	KeyPoint	DM	×	SIG.'24
	May. 26, 2024	Liu et. al [115]	+	KeyPoint	DM	×	arxiv
	May. 28, 2024	VividPose [92]	+	3D-P	DM	×	arxiv
	May. 30, 2024	MotionFollower [116]	+	KeyPoint	DM	×	arxiv
	Jun. 3, 2024	UniAnimate [117]	+	KeyPoint	DM	×	arxiv
	Jun. 5, 2024	Follow-Your-Pose v2 [35]	+	KeyPoint	DM	×	arxiv
	May. 27, 2024	Human4DiT [36]	+	3D-P	DiT	✓	arxiv
	Jan. 19, 2024	3DHM [118]	+	3D-P	DM	×	arxiv
	Mar. 21, 2024	Champ [88]	+	3D-P	DM	✓	ECCV'24
	Jul. 15, 2024	TCAN [119]	+	KeyPoint	DM	×	arxiv
	Jul. 1, 2024	MimicMotion [37]	+	KeyPoint	DM	×	arxiv
	Jul. 16, 2024	IDOL [120]	+	Region	DM	×	arxiv

: Reference Real Images; : Driving Pose Video; : Driving Real Video; : Text Prompts

- Generating realistic and detailed human body textures while maintaining temporal consistency remains a significant challenge. Existing methods often require large amounts of training data, which can be difficult to obtain, and they struggle to achieve both high-quality visual realism and temporal coherence.
- Addressing motion blur in dance movements, which typically involve high-speed actions, is an urgent issue. This blur can lead to models incorrectly learning human body structure, highlighting the need for effective solutions.
- Developing methods that require fewer training samples and can handle a wider variety of motions would significantly advance this field.

Pose-Driven Dance Video Generation. Pose-driven dance video generation focuses on synthesizing realistic and temporally coherent video frames that depict human images performing dance movements, where the recent works in this field are shown in Table VI. This process relies on target pose sequences, which define the desired movements and postures over time, and a reference image, which provides the visual appearance that the synthesized video should retain. To achieve this, most methods [33], [36], [37] rely on deep learning architectures, predominantly utilizing diffusion models, which are trained to understand the relationship between pose and appearance. And these generation frameworks based on diffusion models can be roughly divided into three categories: pure noise input, reference image plus noise input, and guided conditions plus noise input, as shown in Fig. 9.

(A) Pure Noise for Main Diffusion Branch. As illustrated in Fig. 9 (A), in this approach, the input of the main diffusion branch is pure noise, with the target image being embedded in one of two ways: either through ReferenceNet, which is another U-Net copy same as the main diffusion branch, or via a feature encoder. Simultaneously, the driving condition is typically encoded using ControlNet [123]. This generation framework is adopted by methods such as MagicPose [113], MagicAnimate [96], TCAN [119], and DreaMoving [114].

MagicAnimate [96] uses pure noise as the input for a diffusion model and applies an image-video joint training strategy to leverage diverse single-frame image data for augmentation. MagicPose [113] introduces a two-stage training strategy to disentangle human motions and appearance (e.g., facial expressions, skin tone, and dressing), consisting of the pre-training of an appearance-control block and learning appearance-disentangled pose control. TCAN [119] leverages a pre-trained ControlNet [123] and adapts the LoRA [124] technique to the U-Net [125], aligning the latent space between pose and appearance features. DreaMoving [114] argues that the cost of using the diffusion model for appearance coding is high. Thus, they replace the appearance encoder with a feature encoder with a multi-layer convolutional network, shown in Fig. 9 (A2). Pure noise input can produce a wider range of diverse outcomes, relying heavily on the generalization ability of pre-trained models. Even with pure noise as the input, the model can leverage prior knowledge to progressively denoise and generate images or videos that adhere to vision logic.

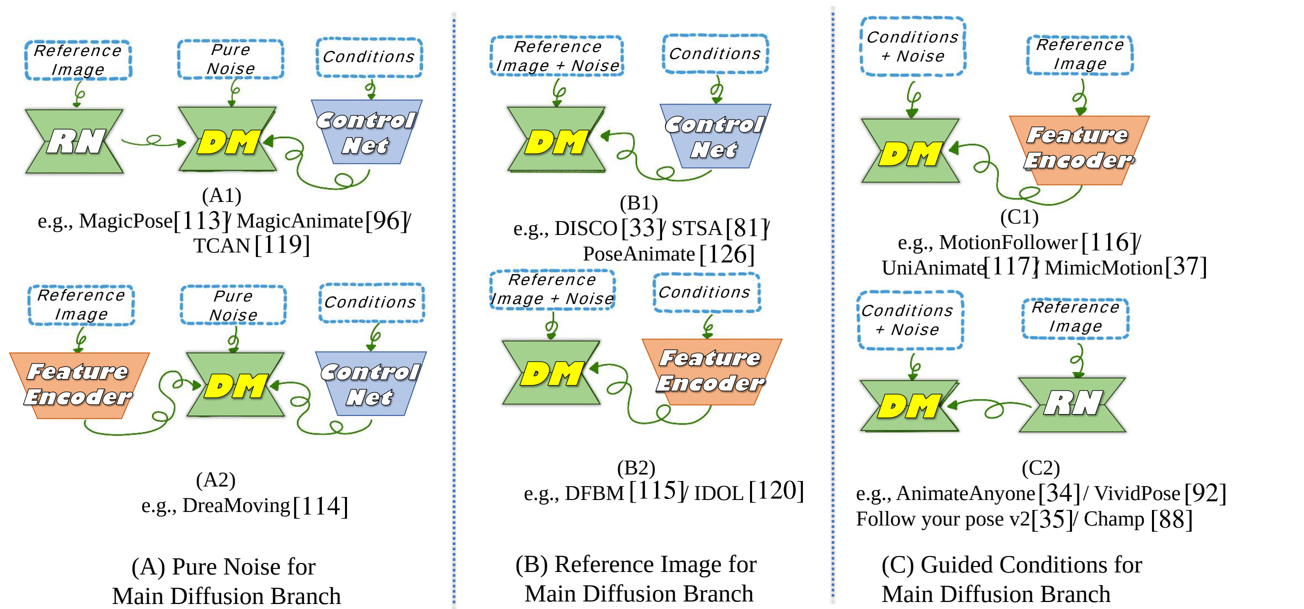

Fig. 9. Comparative overview of different generative frameworks based on diffusion models, where pure noise (A), a reference image (B), and guided conditions (C) are for the main diffusion branch. Each framework integrates unique components, such as ControlNet or feature encoders, to achieve diverse objectives in animation and pose generation objectives. The DM refers to the main diffusion model branch, while the ReferenceNet (RN) is a U-Net copy of the main diffusion model.

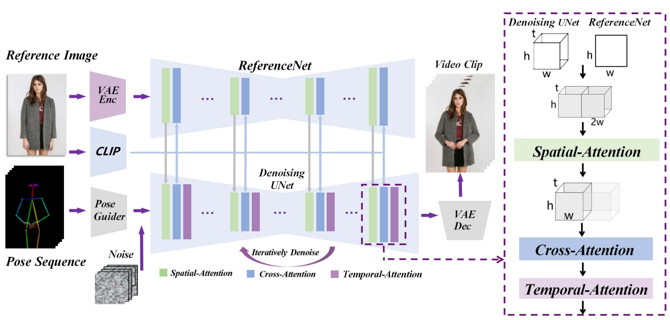

Fig. 10. Overview of Animate Anyone [34]. This framework integrates spatial, cross, and temporal attention mechanisms to generate video clips from reference images and pose sequences, with a focus on the iterative denoising process to produce high-fidelity animations.

(B) Reference Image for Main Diffusion Branch. Fig. 9 (B) illuminates the approach where noise is added to the reference image as input to the main diffusion branch, while the guided conditions are encoded using ControlNet [123] or a feature encoder. Disco [33], STSA [81], and PoseAnimate [126] utilize ControlNet to encode the driving signal. DFBM [115] and IDOL [120] employ feature encoders with multi-layer convolutional networks to encode the guided conditions.

Disco [33] is an early method for pose-driven dance video generation. It introduces a disentangled control architecture that separates the manipulation of human foreground, background, and pose. The human foreground is processed through a cross attention mechanism that utilizes local CLIP [105] image embeddings to capture fine-grained human semantics, while the pose is controlled via a dedicated ControlNet [123] branch. STSA [81] proposes spatial-temporal subspace-attention blocks that decompose the global space into a combination of regular subspaces, enabling efficient modelling of spatio-temporal consistency. PoseAnimate [126] addresses potential disruptions in character identity and background details by replacing self attention layers in the U-Net [125] architecture with a dual consistency attention mechanism. Similar to DreaMoving [114], DFBM [115] replaces ControlNet [123] with a feature encoder to learn both foreground and background dynamics using distinct motion representations. By adding noise to the reference image, the model can better preserve key identity features, such as the appearance of the characters, while introducing variations to enhance the diversity of the generated content.

(C) Guided Conditions for Main Diffusion Branch. Another design approach based on the diffusion models, as shown in Fig. 9 (C), involves adding noise to the guided conditions as the input to the main diffusion branch. In this approach, the reference image is used to encode appearance features either through a feature encoder (C1) or ReferenceNet (C2). For example, Animate Anyone [34], VividPose [92], Follow your pose v2 [35], and Champ [88] adopt the U-Net copy to encode the appearance features of the reference image, while MotionFollower [116], Unianimate [117], and MimicMotion [37] use the feature encoder for the reference image.

Animate Anyone [34], as shown in Fig. 10, has demonstrated remarkable success in pose-driven dance video generation. During the training stage, the model first conditions on individual video frames to prioritize spatial feature extraction and pose guidance, while the temporal layers are not considered. Subsequently, the temporal layer is seamlessly integrated, and the model is further refined using different video clips, while the rest of the network's weights remain fixed. This approach ensures a harmonious blend of spatial and temporal coherence. VividPose [92] utilizes an identity-

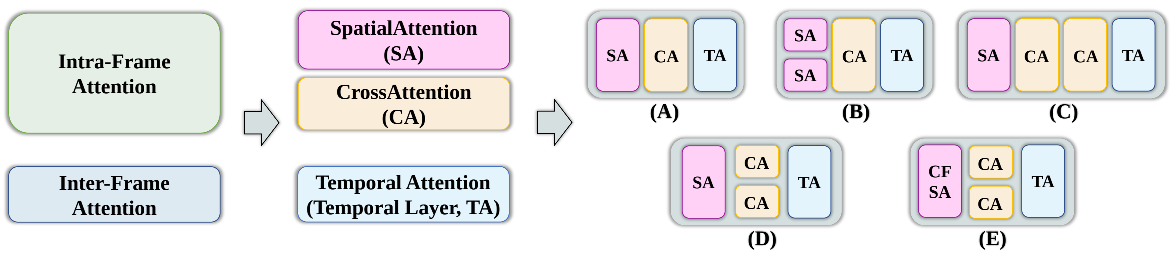

Fig. 11. Different attention fusion methods of diffusion-based vision-driven human motion video generation.

aware appearance controller that skillfully incorporates facial features, ensuring the preservation of the subject's identity across various poses without compromising other visual details. Follow Your Pose V2 [35] stabilizes the background by leveraging the guidance provided by optical flow. Champ [88] integrates a 3D human parametric model within a latent diffusion framework, using the SMPL model to establish a unified representation of body shape and pose.

To reduce network costs, MotionFollower [116] employs a feature encoder as a lightweight alternative to the time-consuming DDIM [77] inversion, which can lead to significant infidelity, as shown in Fig. 9 (C2). Unianimate [117] introduces the use of temporal Mamba [127], a state-space model, as an alternative to the traditional temporal Transformer, significantly enhancing efficiency and enabling the processing of longer video sequences with linear time complexity. MimicMotion [37] implements a hand region enhancement strategy that focuses on improving the visual quality of hand regions. In summary, we present an intuitive comparison Table VII to exhibit the differences between three distinct design schemes.

TABLE VII
COMPARISON OF DIFFERENT NOISE INPUT SCHEMES IS CONDUCTED FROM THREE PERSPECTIVES.

Method	Training Cost	Control Ability	ID Retention
Pure Noise	★★	★★	★★
Reference Image +Noise	★★★	★★	★★★
Guided Conditions +Noise	★★★	★★★	★★

Moreover, most methods for designing diffusion models primarily incorporate intra-frame and inter-frame attention mechanisms. Since the original diffusion model lacks a temporal dimension, many approaches introduce a temporal layer

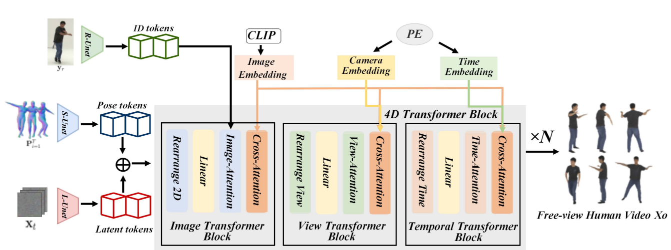

Fig. 12. Architecture of Human4DiT [36], combining ID and pose tokens with CLIP image embeddings to process. This network utilizes image, view, and temporal Transformer blocks to synthesize free-view human videos.

to implement inter-frame attention, thereby enhancing the consistency of video frames. Additionally, by analyzing the placement of spatial attention, self attention, and cross attention layers, we further classify different approaches into five variants, as shown in Fig. 11.

SA-CA-TA. This variant employs a single self attention layer followed by a cross attention layer to control single-frame image generation. Temporal attention layers are used to ensure multi-frame temporal continuity, with the features and semantics of the reference image injected into the noise predictor of the main diffusion model branch through a CLIP [105] encoder or ControlNet [67]. Disco [33], DreaMoving [114], MotionFollower [116], and MimicMotion [37] adopt this approach.

SA&SA-CA-TA. Since image encoders alone may not capture fine-grained features in reference images, this method employs an additional ReferenceNet and integrates the self attention layer features of ReferenceNet with the main diffusion model branch. We define this category of methods as hierarchical self attention. MagicPose [113], Animate Anyone [34], MagicAnimate [96], TCAN [119], Follow-Your-Pose-V2 [35], and Champ [88] are based on this calculation method.

SA-CA-CA-TA. To better handle multiple control signals, which can be challenging with a single cross attention layer, this approach enhances controllability by adding an extra following cross attention layer. Hallo [58] uses this method, combining additional audio signals to generate facial and mouth movements.

SA-CA&CA-TA. Since cross attention layers typically focus on the global features of the reference image, important semantic details that are critical to human perception may be diminished. VividPose [92], inspired by IP-Adapter [128], uses a hierarchical cross attention layer to enhance facial semantics, thereby improving character identity consistency.

CFSA-CA&CA-TA. A direct way to enhance inter-frame correlation is to perform attention calculations between every pixel of the current frame and every pixel of other parts or all frames. Follow-Your-Pose-V1 [35] and PoseAnimate [126] extend the basic self attention mechanism to a cross-frame self attention (CFSA) configuration, enhancing temporal coherence in video-based pose animation. However, compared to temporal attention, cross-frame attention results in higher computational costs and increased memory consumption.

In addition to the generative approach of diffusion models, Human4DiT [36] explores a video generation framework centered around Transformers. Human4DiT employs a 4D

TABLE VIII
COMPREHENSIVE OVERVIEW OF TRY-ON AND POSE2VIDEO.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Try-On Video Generation							
Holistic Human	Apr. 26, 2024	Tunnel Try-On [39]	+	KeyPoint	DM	×	arxiv
	May. 20, 2024	ViViD [40]	+	Region	DM	✓	arxiv
	Jul. 16, 2024	WildVidFit [41]	+	KeyPoint	DM	×	arxiv
Pose2Video							
Holistic Human	Apr. 12, 2023	DreamPose [42]	+	Region	DM	✓	ICCV'23
	Mar. 25, 2024	Make-Your-Anchor [43]	+	3D-P	DM	✓	CVPR'24
	Apr. 21, 2024	PoseAnimate [126]	+	KeyPoint	DM	×	arxiv

: Reference Real Images; : Driving Pose Video; : Driving Real Video; : Reference Cloth Images

diffusion Transformer, as shown in Fig. 12, to capture intricate correlations across views, time, and spatial dimensions, enabling the generation of coherent and realistic human videos from a single reference image. In conclusion, our investigation reveals key insights:

- Existing solutions face challenges such as inter-frame jitter, low integrity of the holistic human body, blurriness, and afterimages. These methods often require large amounts of data for long-term training, leading to high costs.
- Techniques like separating foreground and background, mixed training with both images and videos, and incorporating 3D signals can enhance the integrity of the holistic human body in videos.
- Most current approaches rely on diffusion models based on the U-Net structure, benefiting from open-source pre-trained weights. However, new video generation backbones are not fully explored, such as those based on DiT [129] or VAR [130].

Try-On Video Generation. Try-On video generation is another intriguing video generation task. A recent survey [12] discussed the development of current virtual Try-On technology. We briefly mention recent Try-On methods, such as ViViD [40], Tunnel Try-On [39], and WildVidFit [41], which are summarized in Table VIII. Tunnel Try-On [39] utilizes a Kalman filter to smooth the motion within the focused tunnel, ensuring temporal coherence. ViViD [40] collects a new, diverse dataset, the largest dataset for video virtual Try-On tasks. WildVidFit [41] employs a diffusion guidance module that leverages pre-trained models to enhance temporal coherence without the need for explicit temporal training.

Pose2Video. Pose2Video extends video generation beyond dance scenarios, encompassing general human motion generation. Notably, Make-Your-Anchor [43] advances this field by incorporating precise torso and hand movements within a diffusion-based pose sequence generation framework, requiring only a one-minute video clip for effective training. DreamPose [42] generates human and fabric motions simultaneously by adapting the Stable Diffusion [73] model into a pose and image-guided video synthesis system. Despite these advancements, recent studies [126] in human motion video generation highlight ongoing challenges in maintaining character consistency and temporal coherence.

B. Text-Driven Human Motion Video Generation

Text-driven human motion video generation can be broadly categorized into Text2Face and Text2MotionVideo. Recent works in these two domains are respectively listed in Table IX and Table X.

Text2Face. Text2Face typically focuses on generating talking faces, where text can be used as first-person scripts to control facial movements [49] or as instructions to influence the style or content of the video [50]. Text2Face generation can be divided into two approaches: first-person scripts and third-person instructions. The first-person statement approach aims to control the facial movements to synchronize with the spoken text [6], while third-person instruction methods focus on learning the mapping between objective text descriptions and the generated video [131].

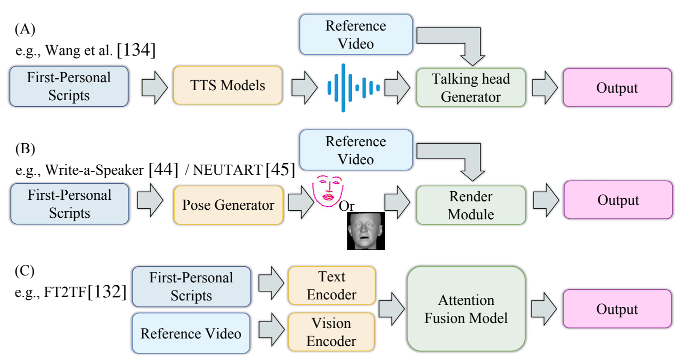

Fig. 13. Different pipelines of Text2Face when the input text is first-personal scripts.

As illustrated in Fig. 13, when text is utilized for first-person scripts, there are three distinct types of pipelines. Most methods generate intermediate outputs such as audios or landmarks, which subsequently control the production of talking head videos, as demonstrated in Fig. 13 (A) and (B). Notably, only the FT2FT [132] adopts an end-to-end pipeline that directly generates videos without relying on any intermediate outputs. Wang et al. [134] convert text to speech using a Text-To-Speech (TTS) model [135], effectively transforming the Text2Face task into a traditional talking head generation task. Write-a-Speaker. [44] design three parallel networks to generate 3D parameters from time-aligned text, which are then used to produce the final talking face videos. Taking it a step further,

TABLE IX

COMPREHENSIVE OVERVIEW OF TEXT2FACE. DM: DIFFUSION MODEL; ED: ENCODER-DECODER; 3D-P: 3D PARAMETERIZATION.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Text2Face							
Part (Face)	May. 7, 2021	Write-a-speaker [44]	+ T	KeyPoint	GAN	×	AAAI'21
	Dec. 11, 2023	NEUTART [45]	(video) + T	3D-P	GAN	✓	arXiv
	Jun. 3, 2023	VideoComposer [131]	+ T	Region	DM	✓	NeurIPS'24
	Apr. 23, 2024	ID-Animator [46]	+ T	Latent	DM	✓	arXiv
	Dec. 9, 2023	FT2TF [132]	+ T	Latent	ED	×	arXiv
	May. 16, 2024	Faces that Speak [6]	+ T	Latent	GAN	×	CVPR'24
	Mar. 1, 2020	LipGAN [133]	(video) + T	Latent	ED	✓	ACM MM'19
: Reference Real Images; T : Text Prompts							

NEUTART [45] simultaneously generates audio and employs a lipreading loss [136] to extract lip features. To address the challenges of two-stage generation, Diao et al. [132] propose a one-stage framework that directly generates face frames from text and reference images using multiscale cross attention.

When text is used as instructions, it typically influences the content and style of the generated video, with additional signals required to preserve the subject's identity. He et al. [46] address this by training an ID-adapter to control identity, while using text prompts to influence motion and background. Ma et al. [137] introduce a simple effective framework for generating subject-controllable videos, focusing on identity reinforcement and frame-wise correlation injection during initialization to ensure stable video outputs.

Text2MotionVideo. Text2MotionVideo tasks involve generating human motion videos where textual inputs primarily serve as instructions to influence the content or style of the video [25], [138], [139]. These tasks can be categorized into two types: ID-preservation, where the generated human retains the identity of the input, and ID-transfer, where the identity may change.

For ID-preserved video generation, the identity is primarily derived from the input images or videos. When other explicit conditions (e.g., keypoint) are used to control motion, the text's role is often limited, typically serving as a caption for the generated video. Specifically, Zuo et al. [47] employ a dual pipeline that combines the identity from the source video with motion information from the keypoints of a reference video, with the input text description fixed according to the source and target prompts. Wang et al. [131] propose a spatio-temporal condition encoder capable of encoding various conditional information, allowing for more flexible control. However, the input text must be closely related to the input image for better results. Zhang et al. [140] adopt a two-stage method that first converts multi-modal inputs into control signals (such as human pose, depth, and DensePose [95]) and then generate videos by these control signals.

In the absence of other explicit conditions, the text provides a general motion directive, which can guide the animation of a static human portrait or directly generate a human motion video from text, typically accomplished directly through text-to-video models [4]. Recent methods in this area demonstrate a trend toward more stable training, easier control, and finer granularity in identity preservation. VideoCrafter 1 [141] and VideoCrafter 2 [142] explore stable generation video generation model training schemes extended from diffusion model, using text, images, and low-quality videos to achieve high-quality video models [141], [142]. To enhance control, Guo et al. [4] propose a plug-and-play motion module that can be trained once and integrated into any personalized text-to-image (T2I) model, adding motion dynamics to reference images. Additionally, AnimateZero [143] offers spatial appearance control and temporal consistency by replacing the global time node of the original text-to-video model with a position correction window and utilizing intermediate latent embeddings from T2I generation. In terms of finer granularity, Renshuai et al. [49] refine fine-grained emotional expression, expanding from 8 sentiment indicators to 135 detailed emotional descriptions. They introduce a framework that simultaneously controls identity, expression, and background from multi-modal inputs, addressing fine-grained emotional challenges with maintaining identity.

In the context of identity transfer, current research primarily focuses on maintaining global features such as spatial and temporal consistency during video generation and editing, as well as utilizing lightweight networks to reduce training costs and enhance efficiency. For spatial consistency, Liu et al. [147] propose a dual-stream diffusion network that improves spatial variation consistency in video generation. Ren et al. [148] and Geyer et al. [25], [145] enhance temporal consistency by incorporating additional modules and enforcing semantic correspondence across frames. Additionally, Qin et al. [48] design an alignment strategy to ensure the consistency of facial features, clothing, and background. Deco [86] decouples appearance, motion, and background, making it easier to achieve consistent generation. Recent efforts have focused on integrating spatial and temporal capabilities. Khachatryan et al. [138] enhance latent embeddings with motion dynamics to preserve global spatial features, introducing cross-frame attention mechanisms to maintain consistent background timing. Eldesokey and Wonka [172] develop spatial latent alignment and pixel-wise guidance modules to improve temporal consistency. For cost reduction, Shi et al. introduced BIVDiff [139],

TABLE X

COMPREHENSIVE OVERVIEW OF TEXT-DRIVEN HUMAN MOTION VIDEO GENERATION. DM: DIFFUSION MODEL; ED: ENCODER-DECODER; 3D-P: 3D PARAMETERIZATION.

Driving Region	Publication Time	Paper Title	Input Signals	Motion Representation	Backbone	Open Source	Venue
			Text2MotionVideo				
Holistic Human	May. 8, 2024	Edit-Your-Motion [47]	(video) + + T	KeyPoint	DM	×	arXiv
	Aug. 15, 2023	Dancing Avatar [48]	+ T	KeyPoint	DM	×	arXiv
	Apr. 3, 2023	Follow-Your-Pose [25]	+ T	KeyPoint	DM	✓	AAAI'24
	Aug. 28, 2024	MagicAvatar [140]	+ T	KeyPoint	DM	×	arXiv
	Feb. 14, 2024	Magic-Me [137]	(video) + T	Latent	DM	✓	arXiv
	Apr. 7, 2024	DiffSFSR [49]	+ T	Latent	DM	×	CVPR'24
	Apr. 17, 2023	Text2Performer [50]	+ T	Latent	ED	✓	ICCV'23
	Apr. 14, 2024	LoopAnimate [144]	+ T	Latent	DM	×	arXiv
	Jul. 10, 2023	AnimateDiff [4]	+ T	Latent	DM	✓	arXiv
	Dec. 6, 2023	AnimateZero [143]	+ T	Latent	DM	×	arXiv
	Jul. 19, 2023	TokenFlow [145]	(video) + T	Latent	DM	✓	arXiv
	Mar. 23, 2023	Text2Video-Zero [138]	(video) + T	Latent	DM	✓	ICCV'23
	Feb. 2, 2023	Dreamix [146]	(video) + T	Latent	DM	×	arXiv
	Dec. 5, 2023	BIVDiff [139]	(video) + T	Latent	DM	✓	CVPR'24
	Dec. 30, 2023	DSDN [147]	(video) + T	Latent	DM	×	arXiv
	Feb. 22, 2024	Customize-A-Video [148]	T	Latent	DM	×	arXiv
	Dec. 12, 2023	LatentMan [149]	T	3D-P	DM	✓	arXiv

: Reference Real Images; : Driving Pose Video; : Driving Real Video; T : Text Prompts

a training-free framework utilizing image diffusion models for video synthesis, while Yang et al. proposed ZeroSmooth [173], a training-free video interpolation method for generative video diffusion models. These advancements highlight the growing emphasis on achieving effective identity transfer in video generation. Based on our investigation, we discover the following challenges:

- Text2Face merits further exploration because texts are more accessible than audio. However, research in this area remains limited, and an end2end framework is still lacking.
- In Text2MotionVideo, the role of text is often constrained when other explicit conditions are present, requiring it to precisely describe the motions specified by those conditions [47], [131], [140]. If the motions are complex and difficult to articulate through text, the quality of the generated output may be compromised.
- Additionally, training-free methods are starting to gain attention.

C. Audio-Driven Human Motion Video Generation

Our primary focus is on generating human motion based on audio-driven conditions. By inputting audio and reference signals (such as images or videos), we infer human motion videos that correspond to the audio signals. Modeling human motion from audio involves tackling several significant research challenges, such as accurately capturing lip movements, head poses, audio-driven holistic human motion, and producing fine-grained animations. Furthermore, we investigate and summarize the paradigms of audio used in human motion video generation, and provide clear instructions for the use of speech signals, as shown in Fig. 14.

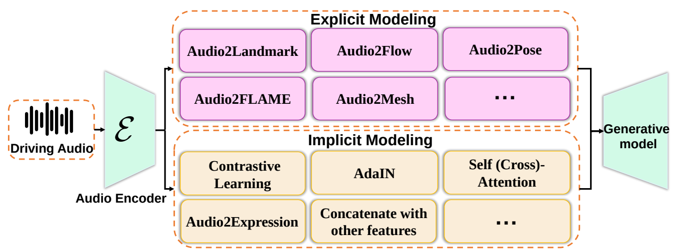

Fig. 14. Paradigm summary of audio-driven human motion video generation. The explicit modeling methods (upper) can provide more efficient and controllable generated output, while the implicit modeling (below) can achieve better temporal synchronization.

Lip Synchronization. Early research primarily focused on lip motion. Wen et al. [150] map the audio to facial expression and re-render the synthetic video through a regional mask. Wav2lip [154] synchronizes the given audio with the lip movements of characters in a video, achieving realistic lip-synchronization effects. Chen et al. [151] employ a cascaded GANs approach to model the synchronization between speech and lip motion, effectively decoupling content-related representations from non-content-related signals in speech, thereby enhancing robustness to diverse facial shapes and viewpoints. Additionally, Cudeiro et al. [152] propose explicit modeling of the relationship between lip motion and audio speech signals using a 3D model structure.

Head Pose Driving. Focusing solely on lip motions fails to produce a realistic visual effect of the speaker. Greenwood et al. [155] are pioneers in using a Bi-LSTM model [174]

TABLE XI
COMPREHENSIVE OVERVIEW OF LIP SYNCHRONIZATION AND HEAD POSE DRIVING.
DM: Diffusion Model; ED: Encoder-Decoder; AM: Autoregressive Model; 3D-P: 3D Parameterization.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Lip Synchronization							
Part (Face)	Sep. 17, 2020	AudioDVP [150]	🎙 + 👤 (video)	Region	ED	✓	TVCG'20
	May. 9, 2019	ATVGnet [151]	🎙 + 👤	KeyPoint	AM	✓	CVPR'19
	May. 8, 2019	VOCA [152]	🎙 + 3D Template	KeyPoint	ED	✓	CVPR'19
	Jan. 10, 2023	Bigioi et al. [153]	🎙 + 👤	Latent	DM	✓	IVC'24
	Nov. 11, 2020	Wav2lip [154]	🎙 + 👤	Latent	ED	✓	ACM MM'20
Head Pose Driving							
Part (Face)	Aug. 20, 2017	Greenwood et al. [155]	🎙	Latent	AM	×	ISCA'17
	Apr. 27, 2020	MakeItTalk [60]	🎙 + 👤	KeyPoint	AM	✓	TOG'20
	Sep. 22, 2021	Live Speech Portraits [61]	🎙 + 👤	KeyPoint	AM	✓	TOG'21
	Jan. 3, 2022	DFA-NeRF [156]	🎙 + 👤 (video)	KeyPoint	ED	✓	arxiv
	Jan. 10, 2023	DiffTalk [15]	🎙 + 👤	Latent	DM	✓	CVPR'23
	May. 15, 2023	IP-LAP [16]	🎙 + 👤 + landmark	Muliti-Conditions	Transformer	✓	CVPR'23
	May. 1, 2023	GeneFace++ [66]	🎙 + 👤	KeyPoint + 3D-P	ED	✓	arxiv
	Mar. 16, 2022	StyleHEAT [51]	🎙 + 👤 + 🕺	Region	GAN	✓	ECCV'22
	Feb. 20, 2023	SD-NeRF [157]	🎙 + 👤 (video)	3D-P	ED	×	TMM'23
	Mar. 26, 2024	AniPortrait [158]	🎙 + 👤	KeyPoint + 3D-P	DM	✓	arxiv
	Jun. 17, 2024	MyTalk [159]	🎙 + 👤	KeyPoint	DM	×	arxiv
	Jun. 12, 2024	Liang et al. [59]	🎙 + 👤 + 🆃	KeyPoint + 3D-P	DM	×	arxiv
	Jun. 27, 2024	RealTalk [160]	🎙 + 👤 (video)	3D-P	Transformer	×	arxiv
	Mar. 20, 2021	AD-NeRF [161]	🎙 + 👤	3D-P + Latent	ED	✓	ICCV'21
	Jan. 19, 2022	SSP-NeRF [162]	🎙 + 👤 (video)	Latent	ED	✓	ECCV'22
	Apr. 22, 2021	PC-AVS [52]	🎙 + 👤 + 🕺	Latent	GAN	✓	CVPR'21
	Jan. 6, 2023	Diffused Heads [163]	🎙 + 👤	Latent	DM	✓	CVPR'24
	Mar. 30, 2023	DAE-Talker [110]	🎙 + 👤	Latent	DM	×	ACM MM'23
	Nov. 26, 2023	GAIA [164]	🎙 + 👤	KeyPoint + Latent	DM	×	ICLR'24
	Dec. 9, 2023	R2-Talker [165]	🎙 + 👤	KeyPoint	ED	✓	arxiv
	May. 6, 2024	AniTalker [54]	🎙 + 👤	Latent	ED	✓	arxiv
	Jul. 12, 2024	EchoMimic [55]	🎙 + 👤	KeyPoint	DM	✓	arxiv
	Jul. 29, 2024	LinguaLinker [56]	🎙 + 👤	Latent	DM	×	arxiv
	Nov. 22, 2022	RAD-NeRF [166]	🎙 + 👤	Latent	ED	✓	arxiv
	May. 4, 2023	Xu et al. [111]	🎙 + 👤	Latent	Transformer	×	CVPR'23
	Apr. 2, 2024	EDTalk [53]	🎙 + 👤	Latent	GAN	✓	ECCV'24
	Nov. 29, 2023	SyncTalk [167]	🎙 + 👤 (video)	3D-P	ED	✓	CVPR'24
	Apr. 23, 2024	TalkingGaussian [168]	🎙 + 👤 (video)	3D-P	ED	✓	ECCV'24
	Apr. 28, 2024	GaussianTalker [169]	🎙 + 👤	3D-P	ED	✓	ACM MM'24
	Dec. 10, 2021	FaceFormer [2]	🎙	KeyPoint	Transformer	✓	CVPR'22
	Sep. 15, 2023	Delbosc et al. [170]	🎙	Latent	GAN	×	ICMI '23
	Oct. 17, 2023	CorrTalk [171]	🎙	Latent	ED	×	TCSVT'24

👤 : Reference Real Images; 🕺 : Driving Real Video; 🆃 : Text Prompts; 🎙 : Audio

to predict head pose motions from audio. Zhou et al. [60] explicitly model the speaker's head movements through facial landmarks, training a Transformer [175] to capture long-term dependencies through adversarial learning, thereby generating natural head poses. Lu et al. [61] propose an autoregressive probabilistic model conditioned on historical head poses and speech representations to predict the motion distribution at the current time. Key head poses are sampled from this predicted probabilistic model. Their approach focuses on real-time audio-driven head pose video generation, capable of producing head pose videos at 30 frames per second.

Since audio alone does not provide cues for head pose and global head movements, directly inferring head movements from audio can lead to significant mismatches between the head pose and the motion of the edited head in the target video. Therefore, Ji et al. [176] propose a target-adaptive facial synthesis technique that bridges the pose gap between the inferred landmarks and the target video portrait in 3D space. By employing a precisely crafted 3D perception keypoint alignment algorithm, 2D landmarks can be accurately projected onto the target video.

For more controllability, recent studies [59], [158], [159] have proposed a two-stage framework: audio-to-landmark and landmark-to-video. In the first stage, audio features are extracted using wav2vec [191] and converted into 2D facial landmark sequences. In the second stage, a diffusion model and motion modules transform these sequences into temporally consistent and realistic portrait animations. Similar two-stage approaches [160], [165] are employed to refine facial expression prediction and achieve high-fidelity facial rendering. Furthermore, Meng et al. [13] elaborately summarize the latest advancements regarding the talking head task. To provide a

TABLE XII
COMPREHENSIVE OVERVIEW OF AUDIO-DRIVEN HOLISTIC HUMAN DRIVING AND FINE-GRAINED ANIMATION METHODS.
DM: Diffusion Model; ED: Encoder-Decoder; AM: Autoregressive Model; DiT: Diffusion Transformer; SIG.: SIGGRAPH; 3D-P: 3D Parameterization.

Driving Region	Publication Time	Paper	Input Signals	Motion Representation	Backbone	Open Source	Venue
Fine-Grained Style and Emotion-Driven Animation							
Part (Face)	May. 19, 2021	EVP [176]	🎙️ + 👤 (video)	KeyPoint	ED	✓	CVPR'21
	Jun. 10, 2023	StyleTalk [177]	🎙️ + 👤 (video)	3D-P	ED	Only Inference	AAAI'23
	Jan. 16, 2024	Real3D-Portrait [178]	🎙️ + 👤	3D-P	ED	✓	ICLR'24
	Dec. 15, 2023	DreamTalk [179]	🎙️ + 👤 (video)	3D-P	DM	Only Inference	arXiv
	Jun. 4, 2024	V-Express [180]	🎙️ + 👤	KeyPoint	DM	✓	arXiv
	Jul, 21, 2021	Eskimez et al. [181]	🎙️ + 👤	Latent	GAN	✓	TMM'21
	Nov. 22, 2022	SadTalker [182]	🎙️ + 👤	Latent + 3D-P	GAN	Only Inference	CVPR'23
	Nov. 28, 2022	Bai et al. [183]	🎙️ + 👤	3D-P	GAN	✓	CVPR'23
	May. 9, 2023	StyleSync [184]	🎙️ + 👤 (video)	Latent	GAN	×	CVPR'23
	Feb. 27, 2024	EMO [57]	🎙️ + 👤	Latent	DM	×	arXiv
	Mar. 4, 2024	FaceChain-ImagineID [185]	🎙️ + 👤	Latent	DM	✓	CVPR'24
	Apr. 29, 2024	EMOPortraits [186]	🎙️ + 👤	Latent	GAN	×	CVPR'24
	May. 15, 2024	SPEAK [187]	🎙️ + 👤	Latent	GAN	×	arXiv
	Jun. 16, 2024	Hallo [58]	🎙️ + 👤	Latent	DM	✓	arXiv
	Apr. 16, 2024	VASA-1 [112]	🎙️ + 👤 (video)	Latent	DiT	×	arXiv
	Jan. 5, 2023	chen et al. [188]	🎙️	Latent	ED	✓	ICMEW'23
	Jan. 28, 2024	Media2Face [189]	🎙️	Latent	DM	×	SIG.'24
Audio-driven Holistic Human Driving							
Holistic Human	Mar. 16, 2024	VLOGGER [62]	🎙️ + 👤	3D-P	DM	×	arXiv
	Dec. 15, 2022	ANGIE [190]	🎙️ + 👤	Latent	ED	×	NeurIPS'22
	May. 15, 2024	Dance-Any-Beat [63]	🎙️ + 👤	Region	DM	×	arXiv

👤 : Reference Real Images; 🎙️ : Audio

broader perspective, Table XI offers a comprehensive overview of the diverse works in the field of head pose driving.

Audio-Driven Holistic Human Driving. Previous audio-driven human motion video generation methods have typically focused on facial expressions and lip synchronization, neglecting the generation of head, upper body, and hand movements synchronized with audio. This limitation reduces the effectiveness of these videos in conveying richer human communication, as the absence of body language and gestures makes the synthesized videos appear less natural and realistic. Vlogger [62] addresses this gap by generating realistic and temporally coherent videos based on a single input image and audio sample. It employs a Transformer-based network that inputs Mel spectrograms to predict a series of 3D facial expressions and body pose parameters, capturing the complex mapping relationship between audio signals and human movements. This method not only includes head movements, gaze, and lip motions but also generates upper body and hand gestures, thus advancing audio-driven synthesis technology to a new level.

Similarly, some work has focused on holistic motion video generation using only speech to provide motion information. ANGIE [190] employs a unified framework to generate speaker image sequences driven by speech audio. The key insight is that co-speech gestures can be decomposed into common motion patterns and subtle rhythmic dynamics. Specifically, Dance Any Beat [63] pioneers the task of music-driven dance video generation. Their method is based on the LFDM [192] framework, which generates corresponding dance movements from music using just a single human photo.

Fine-Grained Style and Emotion-Driven Animation. Controlling facial actions based on emotions derived from audio presents the primary challenge of extracting emotions, as emotional information is intricately entangled with other factors like speech content. Ji et al. [176] address this by proposing a cross-reconstruction emotion disentanglement technique for audio, which extracts two independent latent spaces: a duration-independent space that encodes emotion without regard to content, and a duration-dependent space that encodes the speech content of the audio. Additionally, DreamTalk [179] not only manages the generation of stylized facial animations but also enables fine-grained control over facial expressions through style-aware lip experts and style predictors.

Recently, emotion and speech are frequently used simultaneously as conditions [111], [181], [184], [185], [193], providing implicit motion representations for generating human motion videos. To address the complexity of integrating audio with diffusion models due to the ambiguity in mapping audio to facial expressions, EMO [57] introduces stable control mechanisms, including a speed controller and a face region controller, to enhance generation stability. Furthermore, VASA-1 [112] integrates several additional signals to make the generative modeling more manageable and improve the controllability of the generation process. A comprehensive overview of these innovative approaches is provided in Table XII. Indeed, addressing the complexity of audio-driven human motion video generation faces these challenges:

- Existing methods struggle with generalization, resulting in poor performance when dealing with speakers from different multilingual datasets.

- Another significant challenge is the accurate interpretation of input signals while maintaining precise vision control to ensure output consistency.
- The complexity of audio and motion signals often leads to issues such as jitter, incoherence, or desynchronization in the generated results. For instance, maintaining synchronization between the audio signal and visual output (e.g., hand and lip movements) is crucial.

Multilingual Video Dubbing. Multilingual video dubbing is another intriguing task in human motion video generation, where videos are translated from one language to another. The speech content in the source language is transcribed into text, translated, and then automatically synthesized into speech in the target language, retaining the original speaker's voice. The vision content is aligned with the translated audio by synthesizing the speaker's lip movements, creating a seamless audio-visual experience in the target language.

Yang et al. [194] are among the first to tackle this task. For a given video to be translated, the content is first transcribed into text, which is then translated into the target language. A TTS model [135], specifically trained for the speaker, generates the corresponding audio. This translated audio is then combined with the original video frames. However, since the translated audio often differs in duration from the original, the speeds of the video and audio must be adjusted to match. The merged video is subsequently input into a lip-synchronization model to predict the mouth movements. Finally, additional data processing steps integrate the predicted mouth frames with the original video frames to produce the complete video. Furthermore, DINet [195] achieves face visually dubbing on high-resolution videos with a deformation part and an inpainting part. Bigioi et al. [14] further elaborate on the challenges in this field, highlighting the complexities of achieving realism, cross-lingual adaptability, and overcoming limitations in data diversity and generalization, which continue to drive research and development efforts.

V. Refinement and Output

The domain of human motion video generation has recently attracted considerable attention due to its extensive potential applications. However, the current generation frameworks are still in their nascent stages and exhibit limited control capabilities. To address these challenges, advanced refinement strategies have been introduced. This section details both the refinement phase (Step 4) and the output phase (Step 5).

A. Refinement

Refinement processes are crucial for enhancing the output of generation models can be broadly classified into two categories: part specific refinement and general refinement. Part-specific refinement targets particular, whereas general refinement aims to enhance overall video quality.

Part Specific Refinement involves targeted enhancements of specific body parts, such as the mouth, eyes, teeth, and hands, which are commonly affected by generation errors. This refinement is essential in addressing the limitations of generation models in these delicate regions. Targeted restoration methods, falling under this broader strategy, are utilized to correct inaccuracies in these specific regions. These methods [22], [26] include the development of specialized loss functions which are tailored to targeted regions during the training phase, and the application of pre-trained networks for post-processing improvements. For instance, the MimicMotion [37] exemplifies the use of advanced pose guidance mechanisms and design a hand region enhance method to enhance the precision of subtle movements, minimize distortions, and improve overall motion fidelity. In addition, post-processing pipelines often incorporate tools like Codeformer [196] or the approach by Feng et al. [197] to eliminate facial artifacts in the generated outputs.

General Refinement techniques include super-resolution, frame rate enhancement, and denoising networks. These collectively work to enhance the resolution, frame rate, and overall clarity of videos, thereby significantly improving the quality of viewing.

B. Output

Real-time generation of human motion videos remains a challenging and relatively unexplored domain. While GAN-based methods such as those developed by Guo et al. [26] and Jiang et al. [28] demonstrate some capacity for real-time performance, primarily in applications like talking head and portrait animations, they face issues with training instability and lower video quality. The focus on diffusion models has increased due to their exceptional capability for generating high-quality videos. Nevertheless, the high computational costs for training and inference present significant challenges for real-time applications. Although these models hold great promise, research on cost optimization and inference acceleration is still in its nascent stages, posing a considerable barrier to their practical, real-time use. Emerging works such as Kodaira et al. [198] and Liang et al. [199] explore innovative real-time video editing techniques using stream-based diffusion, presenting a promising direction for integrating these methods with human motion video generation. Further advancements in model distillation, as seen in the work of Sauer et al. [200] and Zhai et al. [201], are anticipated to enhance model sampling speeds significantly, paving the way for real-time video generation capabilities in the near future.

VI. Evaluation

The evaluation metrics for generation tasks provide an indication of the quality of the generated content. In addition to subjective evaluations, we summarize the commonly used objective metrics for video generation.

A. Evaluation Metrics

Appropriate evaluation metrics are essential for comparing different approaches and advancing the field. However, a uniform evaluation system for generated human motion videos is still lacking. In this section, we summarize the different aspects of commonly used evaluation metrics (Fig. 15).

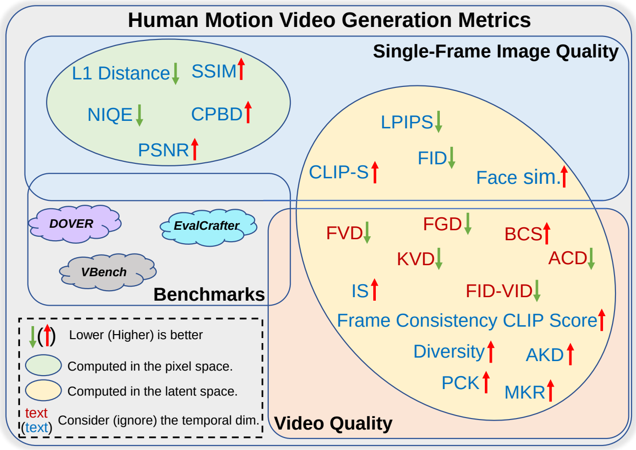

Fig. 15. Overview of commonly used metrics in human motion video generation. The diagram offers a systematic overview of the role various metrics play in the collective evaluation of video quality, emphasizing the complexity and multidimensional aspects of assessing the performance of generated videos.

1) Single-Frame Image Quality: In the video generation task, the goal of measuring the quality of each generated frame is to ensure that the video not only appears coherent and smooth as a whole but also that each frame maintains high quality. Common metrics include L1 Distance, Fréchet Inception Distance (FID) [202], [203], Structural Similarity Index Measure (SSIM) [204], Peak Signal-to-Noise Ratio (PSNR) [205], and Learned Perceptual Image Patch Similarity (LPIPS) [206]. Additional metrics include:

CLIP Similarity Scores (CLIP-S). CLIP-S [105] measures the compatibility of image-text pairs, with higher scores indicating higher compatibility.

Face Similarity. This metric gauges the model's ability to preserve the identity information of the reference image [113].

Cumulative Probability of Blur Detection (CPBD). CPBD [207], [208] is a perceptual metric for assessing image blur. It quantifies the degree of blur by accumulating the blur detection probability based on the human vision system's perception.

Natural Image Quality Evaluator (NIQE). NIQE [209] is a reference-free image quality assessment method that models the statistical features of natural images, enabling effective assessment of image quality.

2) Video Quality Assessment: A comprehensive evaluation of video generation quality involves examining temporal consistency, motion coherence, and overall visual appeal. We start by outlining the key metrics used for assessing video quality, with the evaluation results summarized in Table XIII.

Kernel Video Distance (KVD). KVD [210] is a metric used to evaluate the similarity between two video sequences. It measures the distance between the distributions of real and generated videos by projecting video frames onto a high-dimensional feature space and computing a statistical distance within this space.

Fréchet Video Distance (FVD). FVD [210] is an adaptation of FID for videos. This metric offers a more comprehensive evaluation of video generation models, considering both individual frame appearance and the entire video sequence's temporal coherence and dynamics. Fréchet Gesture Distance (FGD) [211] adapts the principle of the Fréchet Distance to the domain of gesture video generation. FID-VID [212] extends the well-known FID to measure the similarity between two video distributions by incorporating temporal information.

Average Content Distance (ACD). ACD [213] measures the content consistency of a generated video. A smaller ACD indicates that the generated frames in a video are perceptually more similar, reflecting better content consistency.

3) Video Characteristics Assessment: **Frame Consistency CLIP Score.** This metric [214] is commonly used in video editing to measure the coherence of edited videos. It is calculated by computing CLIP image embeddings for all frames of the edited videos and reporting the average cosine similarity between all pairs of video frames.

Inception Score (IS). IS [215] is a popular evaluation metric for diversity, widely used in image generation and sometimes applied to evaluate video diversity.

Diversity. Diversity [216] evaluates variations among generated video gestures. It uses the same feature extractor as FGD to map synthesized gestures into latent feature vectors and calculates the average feature distance for evaluation.

Beat Consistency Score (BCS). BCS [217], [218] is a metric proposed for assessing the correlation between motion and audio beats.

Pose Accuracy. Pose accuracy is a crucial metric in evaluating the effectiveness of human pose estimation techniques. Percentage of Correct keypoint (PCK), which measures the proportion of keypoints that are correctly localized within a specified threshold distance from the ground truth keypoints [219]. Average Keypoint Distance (AKD), which evaluates the accuracy of human keypoints in videos by comparing the distances to real keypoints [220]. Missing Keypoint Rate (MKR), which measures the proportion of missed keypoints during detection or generation [221].

4) Video Evaluators and Benchmarks: Finally, we introduce video evaluators and benchmarks, which are essential tools for objectively assessing the quality of generated videos. These evaluators help quantify various aspects of video generation, ensuring that the generated content meets both technical and aesthetic standards.

Disentangled Objective Video Quality Evaluator. DOVER [222] is a subjectively inspired video quality evaluator with two branches focusing on aesthetic and technical perspectives.

VBench. VBench [223] evaluates video generation quality across 16 different dimensions, including subject identity inconsistency, motion smoothness, temporal flickering, and spatial relationships. Each dimension is assessed using tailored prompts and methods, offering a comprehensive analysis of model performance.

EvalCrafter. EvalCrafter [224] is a novel framework and pipeline for exhaustively evaluating the performance of generated videos in terms of visual qualities, content qualities, motion qualities, and text-video alignment.

TABLE XIII
QUANTITATIVE PERFORMANCES OF FIVE OPEN-SOURCE METHODS ON THE TIKTOK TEST DATASET FOR POSE-GUIDED DANCE VIDEO GENERATION.

	SSIM (↑)	PSNR (↑)	LPIPS (↑)	L1 (↓)	FID (↓)	FID-VID (↓)
DisCo [33]	0.6646	14.1169	0.3913	0.1282	87.6365	79.9397
Champ [88]	0.7022	16.1076	0.3267	0.0991	62.5922	**20.1171**
MagicPose [113]	0.7422	17.2470	0.2780	0.0845	**49.4882**	49.8016
Animate Anyone [34]	0.7392	17.8921	0.2687	**0.0779**	62.4994	21.0533
MagicAnimate [96]	**0.7553**	**18.1637**	**0.2510**	0.0829	57.9520	24.1416

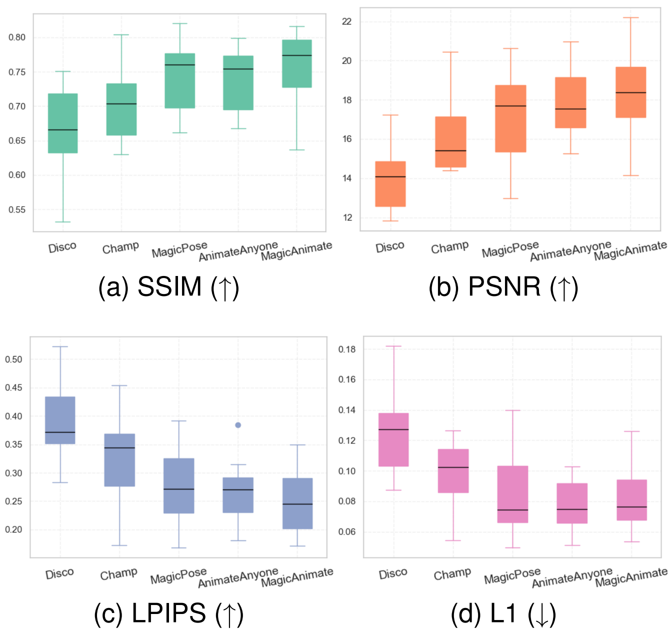

Fig. 16. Quantitative performances of five open-source methods across different matrices for pose-guided dance video generation.

5) Comparative Analysis: This section centers on the task of pose-guided dance video generation. Through quantitative comparative experiments on the TikTok test dataset with 10 videos [225], we benchmark five related open-source methods: Disco [33], Champ [88], MagicPose [113], Animate Anyone [34], and MagicAnimate [96]. Since Animate Anyone [34] has not been officially open-sourced, the version used in our experiments is an unofficial implementation[1]. We assess the performances of the five methods using six key evaluation metrics: L1, PSNR [205], SSIM [204], LPIPS [206], FID [203], and FID-VID [202], [203], as illustrated in Fig. 16 and Table XIII. All the test videos are standardized to a resolution of 512×512.

The experimental results indicate that no method performs outstandingly in all metrics. For instance, Champ performs best in FID-VID (20.1171) but underperforms in SSIM (0.7022). MagicAnimate excels in SSIM (0.7553), PSNR (18.1637), and LPIPS (0.2510), underscoring its superiority in maintaining structural integrity and perceptual quality in generated images. In terms of the L1 metric, Animate Anyone achieves a superior score (0.0779), highlighting its effectiveness in producing outputs closely aligned with the ground truth. For the FID metric, MagicPose attains an impressive FID score (49.4882), showcasing its superior fidelity in generating individual frame images. Overall, MagicAnimate demonstrates the best overall performance, achieving state-of-the-art results across multiple metrics.

[1]https://github.com/MooreThreads/Moore-AnimateAnyone

B. Datasets

Video generation models require a large quantity of data to learn the prior distribution of human knowledge, making data a crucial component. Since crawling data to train human motion models is sensitive considering current privacy issues, to facilitate the training of human action generation models in future research, we carefully collect 64 publicly available human-related video datasets to support future research in this area. Table XIV provides an overview of these datasets along with their respective download links. Unlike other dataset overviews [10], [11], we include a concise one-sentence description for each dataset, enabling researchers to quickly identify the most relevant data for their work.

VII. CHALLENGES AND FUTURE WORK

Human motion video generation presents several challenges, including data availability, signal understanding, motion planning, and the quality of the generated video.

Lack of Data. The field of human motion video generation is hindered by limited data availability due to privacy concerns, poor quality, and high collection costs. This scarcity weakens model robustness and compromises real-world reliability. Expanding datasets is essential for training models to better recognize, understand, and replicate human behavior, ultimately improving the quality and diversity of generated videos.

Motion Planner. Current motion planning relies heavily on pre-existing data distributions, which limits its ability to grasp the deeper semantic layers of human actions. This approach constrains adaptability and sophistication. To advance motion planning, we must shift from purely statistical methods to those that incorporate meaning, context, and intent. Leveraging LLMs can enhance this process by analyzing complex human movement patterns, enabling the generation of more realistic and contextually appropriate human motion in video, which is treated as a promising direction [17]–[19].

Lack of Photorealism. In human motion video generation, numerous challenges, as illustrated in Fig. 17, require sophisticated solutions. The fidelity of generated human forms, particularly in the face and hands, is crucial for realism and expressiveness.

TABLE XIV

OVERVIEW OF KEY ATTRIBUTES ACROSS VARIOUS HUMAN VIDEO DATASETS.

This table categorizes 64 datasets based on their duration, video resolution (High: >1024 pixels, Mid: 512–1024 pixels, and Low: <512 pixels), number of unique individuals ("ID"), camera movement (Static or Moving), presence of single or multiple persons ("Person"), and whether the backgrounds are clean. The descriptions offer insights into the specific focus or application of each dataset, including tasks such as emotion recognition, action synthesis, and multimodal human interaction analysis. This comprehensive overview highlights the diversity and distinct characteristics of each dataset, facilitating the selection process for research in human motion analysis and related fields.

Item	Dataset	Year	Hours	Resolution	ID	Camera	Person	Clean Background	Description
1	NExTVideo [226]	2024	65	Mid	>1000	Moving	Multiple	⊘	A dataset for video grounding extended from NExt-QA
2	LOVEU-TGVE-2023 [227]	2023	0.6	Low	76	Static	Single	✓	Text-driven video editing competition
3	3DYoga90 [228]	2023	80	High	>200	Moving	Single	✓	A Yoga dataset
4	CAP [229]	2023	1000	Low	780	Static	Single	⊘	A large dataset for fine-grained activity classfication
5	MEVID [230]	2023	144	Mid	176	Moving	Single	⊘	A dataset for large-scale, video person re-identification
6	VideoChatGPT [231]	2023	30	Low	>1000	Moving	Multiple	⊘	A dataset for video-based conversation
7	VFHQ [232]	2022	1800	Mid	16000	Static	Multiple	⊘	A high-quality video face dataset
8	HEADSET [233]	2022	20	Mid	27	Static	Multiple	⊘	A multimodal XR dataset with facial expressions
9	OpenASL [234]	2022	288	Low	>200	Static	Single	✓	An open-domain sign language translation dataset
10	BRACE BRACE [235]	2022	3	Mid	81	Moving	Multiple	✓	A breakdancing competition dataset for dance motion synthesis
11	CelebV-HQ [236]	2022	68	Mid	>1000	Static	Single	⊘	A face video dataset
12	LSA-T [237]	2022	22	High	103	Static	Multiple	⊘	The first continuous Argentinian sign language dataset
13	CelebV-Text [238]	2022	279	Mid	>1000	Static	Single	⊘	A large dataset of facial text-video pairs
14	MUSIC-AVQA [239]	2022	150	Mid	>1000	Moving	Multiple	⊘	A dataset for audio-visual question answering task
15	Vico [240]	2022	1.5	Low	92	Static	Single	⊘	A dataset of face-to-face conversation
16	TED-Dataset [93]	2021	3	Low	411	Static	Single	⊘	Videos of presentations that include movement and audio
17	TikTok Dataset [225]	2021	1	Mid	>300	Static	Single	✓	Social media dance videos
18	VideoMatte240K [241]	2021	3	High	484	Static	Multiple	✓	High-resolution alpha matte and foreground video clips
19	FR-FS Dataset [242]	2021	1	Low	>10	Moving	Multiple	⊘	Fall recognition in figure skating
20	How2sign [243]	2021	80	High	11	Static	Single	⊘	A multiview continuous American sign language dataset
21	WhenAct [244]	2021	20	Low	171	Moving	Single	⊘	A dataset for action detection and localization
22	Writing in the Air [245]	2021	16.8	Mid	122	Static	Single	⊘	A dataset for writing in the air task
23	VPCD [246]	2021	23	Mid	>1000	Moving	Multiple	⊘	A multi-model dataset from movies and TV-shows
24	SEAM Match-RCNN [247]	2021	50	Mid	>1000	Moving	Multiple	⊘	A dataset for video-to-shop task
25	MFA-Wild [248]	2021	0.5	Low	201	Moving	Single	⊘	A dataset for negative emotion recognition
26	MEAD [249]	2020	40	High	60	Static	Single	✓	Talking-face videos
27	GSL [250]	2020	2	Mid	7	Static	Single	✓	A sign language dataset
28	Rhythmic gymnastics dataset [251]	2020	37	Mid	200	Moving	Multiple	⊘	A rhythmic gymnastics dataset
29	EDUVSUM [252]	2020	1	Low	98	Static	Single	⊘	Three popular e-learning platforms
30	KIT Bianual Actions Dataset [253]	2020	2	Mid	6	Static	Single	⊘	Daily action sequences
31	LSFP [254]	2020	90	Mid	100	Static	Single	✓	A dataset for French Belgian sign languages
32	MOD20 [255]	2020	5	Mid	>1000	Moving	Multiple	⊘	A dataset for human action recognition
33	EDN [256]	2019	3	High	100	Static	Multiple	⊘	A two-part motion transfer dataset
34	Taichi [220]	2019	1	Low	3000	Static	Single	✓	A dataset for 3D human pose estimation from multiple views
35	Fashion dataset [257]	2019	1	Mid	>300	Static	Single	⊘	A fashion dataset
36	IfAct [258]	2019	21	Mid	>150	Static	Multiple	⊘	For video transcripts and actions
37	MELD [259]	2018	10	Low	400	Static	Single	✓	A multimodal emotionlines dataset
38	CMU-MOSEI [260]	2018	150	Mid	1000	Static	Single	✓	Multimodal sentiment analysis and emotion recognition
39	OMGEmotionChallenge [261]	2018	15	Mid	500	Static	Single	✓	An emotion dataset
40	iQIYI [262]	2018	800	Mid	5000	Moving	Multiple	⊘	An iQIYI TV dataset
41	UCF101 [263]	2012	27	Low	>1000	Moving	Multiple	⊘	A lightweight dataset for action recognition
42	HMDB [264]	2011	20	Mid	51	Static	Single	⊘	A large human motion database
43	IEMOCAP [265]	2008	12	Mid	10	Moving	Multiple	✓	An acted, multimodal and multispeaker database
44	EPEMP [266]	2022	0.2	High	2	Static	Multiple	✓	A new lab-based person interaction dataset of professional dancers
45	PP-HumanSeg14K [267]	2022	0.1	High	>295	Static	Multiple	⊘	A large-scale teleconferencing video dataset
46	ViCo-X [268]	2022	0.8	High	10	Static	Multiple	✓	Face-to-face conversational videos performed by two qualified actors
47	SHORT VIDEO DATASET [269]	2019	378	High	80000	Moving	Multiple	⊘	A large-scale short video dataset
48	DeeperForensics-1.0 [270]	2020	16	High	100	Static	Multiple	⊘	Real-world face forgery
49	VideoForensicsHQ [271]	2021	1111	High	1737	Static	Single	✓	Face video forgery
50	STVD-FC [272]	2022	6540	Mid	6730	Static	Single	⊘	Political content analysis and fact-checking tasks
51	OULU-NPU [273]	2017	13.75	Low	4950	Static	Single	✓	Real access and attack videos
52	AUTSL [274]	2021	10	Low	31	Static	Single	⊘	A large-scale, multimode dataset that contains isolated Turkish sign videos
53	SENDv1 [275]	2021	5.6	Low	48	Static	Single	⊘	Self-paced, unscripted emotional narratives, annotated for emotional valence over time
54	The Jester dataset [276]	2019	36	Low	>1000	Static	Single	⊘	A dataset for human hand gesture recoginiton
55	Voxceleb [277]	2019	>2000	Low	>7000	Static	Single	⊘	A large scale audio-visual dataset of human speech
56	CMU-MOSEI [278]	2018	65	Mid	3228	Static	Single	⊘	A dataset of sentiment analysis and emotion recognition to date
57	CREMA-D [279]	2015	1	Mid	91	Static	Single	✓	A dataset for the study of multi-modal emotion expression and perception
58	LRW [280]	2016	0.3	Low	1000	Moving	Single	⊘	BBC lip reading in the wild
59	L-SVD [281]	2024	5.5	Mid	>20000	Static	Single	✓	A benchmark for emotion recognition
60	MMPD [282]	2023	11	Low	33	Static	Multiple	✓	Recordings from mobile phones of 33 subjects with eight descriptive labels
61	Casual Conversations v2 [283]	2023	674	High	5567	Static	Multiple	⊘	Diverse, large benchmark in audio, vision, and speech models
62	InternVid-10M-FLT [284]	2023	760	Low	>7M	Moving	Multiple	⊘	Publicly available web videos
63	AVCAffe [285]	2023	108	Low	106	Static	Single	✓	A large scale audio-visual dataset of cognitive load and affect for remote work
64	2K2K [286]	2023	2.2	Mid	2050	Static	Single	✓	2050 high-quality human scans

Maintaining visual consistency throughout a video is another challenge, requiring not only stable visual characteristics but also seamless integration of the subject within the environment. Ensuring that the image remains consistent over time is vital for creating an immersive and believable scene. Stability in both the foreground and background is essential, as any instability can disrupt the viewer's experience and diminish the video's credibility.

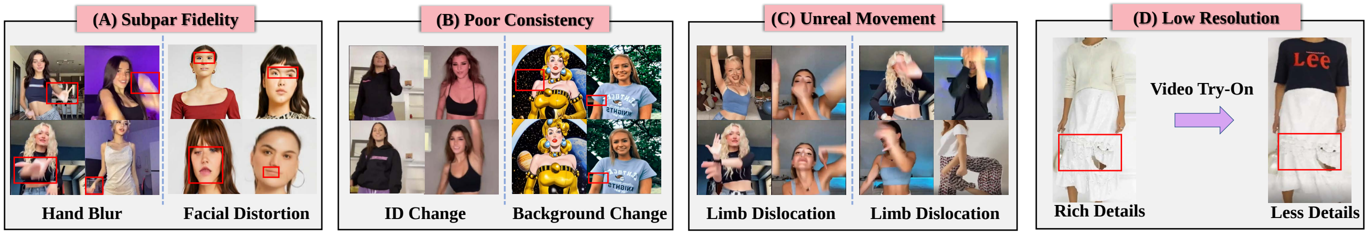

Fig. 17. Main challenges in human motion video generation: subpar fidelity with examples like hand blur and facial distortion, poor consistency with identity and background changes, unrealistic movements, and low resolution.

The plausibility of movements is equally critical. Regardless of how realistic the forms and backgrounds appear, if the depicted actions are not physically or logically sound, the overall effect is compromised. Movements must be natural, smooth, and contextually appropriate to the scenario. Additionally, resolution and image quality are key factors. Poor resolution can obscure important details, detracting from the scene's impact. High-definition, sharp visuals are necessary for delivering a clear and engaging experience.

Addressing these challenges is essential for producing videos that adhere to the physical laws of the real world. Achieving lifelike visual effects demands advancements in rendering technologies and an in-depth understanding of human anatomy and motion.

Expanding Duration and Refining Control. Most current methods for human motion video generation produce only short clips, usually lasting a few seconds. Extending this to longer videos, spanning minutes or even hours, remains a major challenge. Future research should aim to develop techniques that maintain coherence and quality over extended durations. Additionally, current multimodal approaches, even with signals like meshes and depth maps, often struggle with detailed control of specific body parts. Enhancing the realism and expressiveness of videos requires improved control over intricate areas such as hands and faces.

Real-Time Platform and Cost. Real-time streaming of virtual humans demands low latency to ensure smooth and natural interactions, while high latency disrupts communication and reduces lifelike responsiveness. High-quality graphics also need substantial bandwidth, with limited bandwidth leading to lower video quality and buffering issues. Additionally, the user interface must be highly responsive, providing immediate feedback to keep users engaged. While diffusion models [73] are pivotal in human motion video generation, their high computational demands necessitate the development of more efficient models to lower costs and improve accessibility.

Ethics. Creating digital humanoids introduces significant ethical concerns, particularly regarding the use of personal data and privacy. A robust ethical framework is needed to guide their development and integration, ensuring informed consent for biometric data and establishing accountability for their actions. Safeguarding privacy and addressing negative impacts are essential.

References

[1] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *TOG*, vol. 37, no. 4, pp. 1–14, 2018.

[2] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *CVPR*, 2022.

[3] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[4] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[5] 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 preprint arXiv:2311.15127*, 2023.

[6] Y. Jang, J.-H. Kim, J. Ahn, D. Kwak, H.-S. Yang, Y.-C. Ju, I.-H. Kim, B.-Y. Kim, and J. S. Chung, "Faces that speak: Jointly synthesising talking face and speech from text," in *CVPR*, 2024.

[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 preprint arXiv:2311.12052*, 2023.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *TOG*, vol. 42, no. 4, pp. 1–14, 2023.

[10] 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 Comput. Surv.*, vol. 55, no. 12, pp. 1–37, 2023.

[11] W. Lei, J. Wang, F. Ma, G. Huang, and L. Liu, "A comprehensive survey on human video generation: Challenges, methods, and insights," *arXiv preprint arXiv:2407.08428*, 2024.

[12] D. Song, X. Zhang, J. Zhou, W. Nie, R. Tong, and A.-A. Liu, "Image-based virtual try-on: A survey," *arXiv preprint arXiv:2311.04811*, 2023.

[13] M. Meng, Y. Zhao, B. Zhang, Y. Zhu, W. Shi, M. Wen, and Z. Fan, "A comprehensive taxonomy and analysis of talking head synthesis: Techniques for portrait generation, driving mechanisms, and editing," *arXiv preprint arXiv:2406.10553*, 2024.

[14] D. Bigioi and P. Corcoran, "Multilingual video dubbing—a technology review and current challenges," *Frontiers in Signal Processing*, vol. 3, p. 1230755, 2023.

[15] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in *CVPR*, 2023.

[16] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li, "Identity-preserving talking face generation with landmark and appearance priors," in *CVPR*, 2023.

[17] S. Geng, R. Teotia, P. Tendulkar, S. Menon, and C. Vondrick, "Affective faces for goal-driven dyadic communication," *arXiv preprint arXiv:2301.10939*, 2023.

[18] D. Wang, B. Dai, Y. Deng, and B. Wang, "Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents," *arXiv preprint arXiv:2311.17465*, 2023.

[19] Y. Wang, J. Guo, J. Bai, R. Yu, T. He, X. Tan, X. Sun, and J. Bian, "Instructavatar: Text-guided emotion and motion control for avatar generation," *arXiv preprint arXiv:2405.15758*, 2024.

[20] 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 TPAMI*, 2023.

[21] H. Xu, G. Song, Z. Jiang, J. Zhang, Y. Shi, J. Liu, W. Ma, J. Feng, and L. Luo, "Omniavatar: Geometry-guided controllable 3d head synthesis," in *CVPR*, 2023.

[22] J. Zhang, J. Chen, H. Tang, E. Sangineto, P. Wu, Y. Yan, N. Sebe, and W. Wang, "Unsupervised high-resolution portrait gaze correction and animation," *IEEE TIP*, vol. 31, pp. 5272–5286, 2022.

[23] T. Kang, J. Oh, J. Lee, S. Park, and J. Choo, "Expression domain translation network for cross-domain head reenactment," in *ICASSP*, 2024.

[24] Z. Ma, X. Zhu, G.-J. Qi, Z. Lei, and L. Zhang, "Otavatar: One-shot talking face avatar with controllable tri-plane rendering," in *CVPR*, 2023.

[25] Y. Ma, H. Liu, H. Wang, H. Pan, Y. He, J. Yuan, A. Zeng, C. Cai, H.-Y. Shum, W. Liu, *et al.*, "Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation," *arXiv preprint arXiv:2406.01900*, 2024.

[26] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "Liveportrait: Efficient portrait animation with stitching and retargeting control," *arXiv preprint arXiv:2407.03168*, 2024.

[27] Y. Xie, H. Xu, G. Song, C. Wang, Y. Shi, and L. Luo, "X-portrait: Expressive portrait animation with hierarchical motion attention," in *SIGGRAPH*, 2024.

[28] J. Jiang, G. Lin, Z. Rong, C. Liang, Y. Zhu, J. Yang, and T. Zhong, "Mobileportrait: Real-time one-shot neural head avatars on mobile devices," *arXiv preprint arXiv:2407.05712*, 2024.

[29] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, 2019.

[30] H. Liu, X. Han, C. Jin, L. Qian, H. Wei, Z. Lin, F. Wang, H. Dong, Y. Song, J. Xu, and Q. Chen, "Human motionformer: Transferring human motions with vision transformers," in *ICLR*, 2023.

[31] S. Wu, Z. Liu, B. Zhang, R. Zimmermann, Z. Ba, X. Zhang, and K. Ren, "Do as i do: Pose guided human motion copy," *IEEE Trans. Dependable Secur. Comput.*, no. 01, pp. 1–16, 2024.

[32] T. Adiya, J. S. Yoon, J. Lee, S. Kim, and H. Lim, "Bidirectional temporal diffusion model for temporally consistent human animation," in *ICLR*, 2024.

[33] 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.

[34] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *CVPR*, 2024.

[35] 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 preprint arXiv:2406.03035*, 2024.

[36] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, "Human4dit: Free-view human video generation with 4d diffusion transformer," *arXiv preprint arXiv:2405.17405*, 2024.

[37] 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 preprint arXiv:2406.19680*, 2024.

[38] X. Guo, M. Zheng, L. Hou, Y. Gao, Y. Deng, P. Wan, D. Zhang, Y. Liu, W. Hu, and Z. Zha, "I2v-adapter: A general image-to-video adapter for diffusion models," in *SIGGRAPH*, 2024.

[39] Z. Xu, M. Chen, Z. Wang, L. Xing, Z. Zhai, N. Sang, J. Lan, S. Xiao, and C. Gao, "Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos," *arXiv preprint arXiv:2404.17571*, 2024.

[40] Z. Fang, W. Zhai, A. Su, H. Song, K. Zhu, M. Wang, Y. Chen, Z. Liu, Y. Cao, and Z.-J. Zha, "Vivid: Video virtual try-on using diffusion models," *arXiv preprint arXiv:2405.11794*, 2024.

[41] Z. He, P. Chen, G. Wang, G. Li, P. H. Torr, and L. Lin, "Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models," *arXiv preprint arXiv:2407.10625*, 2024.

[42] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion," in *ICCV*, 2023.

[43] Z. Huang, F. Tang, Y. Zhang, X. Cun, J. Cao, J. Li, and T.-Y. Lee, "Make-your-anchor: A diffusion-based 2d avatar generation framework," in *CVPR*, 2024.

[44] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *AAAI*, 2021.

[45] G. Milis, P. P. Filntisis, A. Roussos, and P. Maragos, "Neural text to articulate talk: Deep text to audiovisual speech synthesis achieving both auditory and photo-realism," *arXiv preprint arXiv:2312.06613*, 2023.

[46] 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 preprint arXiv:2404.15275*, 2024.

[47] Y. Zuo, L. Li, L. Jiao, F. Liu, X. Liu, W. Ma, S. Yang, and Y. Guo, "Edit-your-motion: Space-time diffusion decoupling learning for video motion editing," *arXiv preprint arXiv:2405.04496*, 2024.

[48] B. Qin, W. Ye, Q. Yu, S. Tang, and Y. Zhuang, "Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model," *arXiv preprint arXiv:2308.07749*, 2023.

[49] R. Liu, B. Ma, W. Zhang, Z. Hu, C. Fan, T. Lv, Y. Ding, and X. Cheng, "Towards a simultaneous and granular identity-expression control in personalized face generation," in *CVPR*, 2024.

[50] Y. Jiang, S. Yang, T. L. Koh, W. Wu, C. C. Loy, and Z. Liu, "Text2performer: Text-driven human video generation," in *ICCV*, 2023.

[51] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, and Y. Yang, "Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan," in *ECCV*, 2022.

[52] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *CVPR*, 2021.

[53] S. Tan, B. Ji, M. Bi, and Y. Pan, "Edtalk: Efficient disentanglement for emotional talking head synthesis," in *ECCV*, 2024.

[54] T. Liu, F. Chen, S. Fan, C. Du, Q. Chen, X. Chen, and K. Yu, "Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding," *arXiv preprint arXiv:2405.03121*, 2024.

[55] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions," *arXiv preprint arXiv:2407.08136*, 2024.

[56] R. Zhang, Y. Fang, Z. Lu, P. Cheng, Z. Huang, and B. Fu, "Lingualinker: Audio-driven portraits animation with implicit facial control enhancement," *arXiv preprint arXiv:2407.18595*, 2024.

[57] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions," *arXiv preprint arXiv:2402.17485*, 2024.

[58] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, L. Van Gool, Y. Yao, and S. Zhu, "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," *arXiv preprint arXiv:2406.08801*, 2024.

[59] J. Liang and F. Lu, "Emotional conversation: Empowering talking faces with cohesive expression, gaze and pose generation," *arXiv preprint arXiv:2406.07895*, 2024.

[60] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makeittalk: Speaker-aware talking-head animation," *TOG*, vol. 39, no. 6, pp. 1–15, 2020.

[61] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: Real-time photorealistic talking-head animation," *TOG*, vol. 40, no. 6, pp. 1–17, 2021.

[62] E. Corona, A. Zanfir, E. G. Bazavan, N. Kolotouros, T. Alldieck, and C. Sminchisescu, "Vlogger: Multimodal diffusion for embodied avatar synthesis," *arXiv preprint arXiv:2403.08764*, 2024.

[63] X. Wang, H. Wang, D. Liu, and W. Cai, "Dance any beat: Blending beats with visuals in dance video generation," *arXiv preprint arXiv:2405.09266*, 2024.

[64] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[65] Z. Ye, Z. Jiang, Y. Ren, J. Liu, J. He, and Z. Zhao, "Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis," in *ICLR*, 2023.

[66] Z. Ye, J. He, Z. Jiang, R. Huang, J. Huang, J. Liu, Y. Ren, X. Yin, Z. Ma, and Z. Zhao, "Geneface++: Generalized and stable real-time audio-driven 3d talking face generation," *arXiv preprint arXiv:2305.00787*, 2023.

[67] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," in *NeurIPS*, 2017.

[68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[69] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[70] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020.

[71] J. Ma, X. Zhang, and S. Yu, "An identity-preserved framework for human motion transfer," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 3495–3509, 2024.

[72] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: A review," *ACM Comput. Surv.*, vol. 55, no. 2, 2022.

[73] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023.

[74] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[75] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *NeurIPS*, 2020.

[76] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021.

[77] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[78] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[79] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, p. 2814–2830, 2024.

[80] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[81] H. Fang, Z. Sun, Z. Huang, F. Tang, J. Cao, and S. Tang, "Dance your latents: Consistent dance generation through spatial-temporal subspace attention guided by motion flow," *arXiv preprint arXiv:2310.14780*, 2023.

[82] P. Knap, "Human modelling and pose estimation overview," *arXiv preprint arXiv:2406.19290*, 2024.

[83] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *ECCV*, 2024.

[84] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment anything," in *ICCV*, 2023.

[85] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866, 2023.

[86] X. Zhong, X. Huang, X. Yang, G. Lin, and Q. Wu, "Deco: Decoupled human-centered diffusion video editing with motion consistency," *arXiv preprint arXiv:2408.07481*, 2024.

[87] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[88] 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 preprint arXiv:2403.14781*, 2024.

[89] R. Khirodkar, T. Bagautdinov, J. Martinez, Z. Su, A. James, P. Selednik, S. Anderson, and S. Saito, "Sapiens: Foundation for human vision models," *arXiv preprint arXiv:2408.12569*, 2024.

[90] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE TPAMI*, vol. 43, no. 1, pp. 172–186, 2021.

[91] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *ICCVW*, 2023.

[92] 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 preprint arXiv:2405.18156*, 2024.

[93] A. Siarohin, O. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *CVPR*, 2021.

[94] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*, 2018.

[95] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018.

[96] 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.

[97] E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar, "Can language models learn to listen?," in *ICCV*, 2023.

[98] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang, "Motiongpt: Finetuned llms are general-purpose motion generators," in *AAAI*, 2024.

[99] Z. Wang, J. Wang, D. Lin, and B. Dai, "Intercontrol: Generate human motion interactions by controlling every joint," *arXiv preprint arXiv:2311.15864*, 2023.

[100] Z. Zhou, Y. Wan, and B. Wang, "Avatargpt: All-in-one framework for motion understanding planning generation and beyond," in *CVPR*, 2024.

[101] Z. Cai, J. Jiang, Z. Qing, X. Guo, M. Zhang, Z. Lin, H. Mei, C. Wei, R. Wang, W. Yin, *et al.*, "Digital life project: Autonomous 3d characters with social intelligence," in *CVPR*, 2024.

[102] P. J. Yazdian, E. Liu, L. Cheng, and A. Lim, "Motionscript: Natural language descriptions for expressive 3d human motions," *arXiv preprint arXiv:2312.12634*, 2023.

[103] J. Liu, W. Dai, C. Wang, Y. Cheng, Y. Tang, and X. Tong, "Plan, posture and go: Towards open-world text-to-motion generation," *arXiv preprint arXiv:2312.14828*, 2023.

[104] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, and Z. Liu, "Finemogen: Fine-grained spatio-temporal motion generation and editing," in *NeurIPS*, 2024.

[105] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[106] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[107] S. Yang, H. Li, J. Wu, M. Jing, L. Li, R. Ji, J. Liang, and H. Fan, "Megactor: Harness the power of raw video for vivid portrait animation," *arXiv preprint arXiv:2405.20851*, 2024.

[108] A. Agarwal, B. Sen, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Faceoff: A video-to-video face swapping system," in *WACV*, 2023.

[109] W. Zhong, J. Li, Y. Cai, L. Lin, and G. Li, "Style-preserving lip sync via audio-aware style reference," *arXiv preprint arXiv:2408.05412*, 2024.

[110] C. Du, Q. Chen, T. He, X. Tan, X. Chen, K. Yu, S. Zhao, and J. Bian, "Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder," in *ACM MM*, 2023.

[111] C. Xu, J. Zhu, J. Zhang, Y. Han, W. Chu, Y. Tai, C. Wang, Z. Xie, and Y. Liu, "High-fidelity generalized emotional talking face generation with multi-modal emotion space learning," in *CVPR*, 2023.

[112] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," *arXiv preprint arXiv:2404.10667*, 2024.

[113] D. Chang, Y. Shi, Q. Gao, J. Fu, H. Xu, G. Song, Q. Yan, X. Yang, and M. Soleymani, "Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion," in *ICML*, 2024.

[114] 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 dance video generation framework based on diffusion models," *arXiv preprint arXiv:2312.05107*, 2023.

[115] J. Liu, K. Yu, M. Feng, X. Guo, and M. Cui, "Disentangling foreground and background motion for enhanced realism in human video generation," *arXiv preprint arXiv:2405.16393*, 2024.

[116] 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 preprint arXiv:2405.20325*, 2024.

[117] 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 preprint arXiv:2406.01188*, 2024.

[118] B. Li, J. Rajasegaran, Y. Gandelsman, A. A. Efros, and J. Malik, "Synthesizing moving people with 3d control," *arXiv preprint arXiv:2401.10889*, 2024.

[119] J. Kim, M.-J. Kim, J. Lee, and J. Choo, "Tcan: Animating human images with temporally consistent pose guidance using diffusion models," *arXiv preprint arXiv:2407.09012*, 2024.

[120] Y. Zhai, K. Lin, L. Li, C.-C. Lin, J. Wang, Z. Yang, D. Doermann, J. Yuan, Z. Liu, and L. Wang, "Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation," *arXiv preprint arXiv:2407.10937*, 2024.

[121] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[122] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-wasserstein averaging of kernel and distance matrices," in *ICML*, 2016.

[123] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[124] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *ICLR*, 2022.

[125] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[126] 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 preprint arXiv:2404.13680*, 2024.

[127] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024.

[128] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[129] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023.

[130] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *arXiv preprint arXiv:2404.02905*, 2024.

[131] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," in *NeurIPS*, 2024.

[132] X. Diao, M. Cheng, W. Barrios, and S. Jin, "Ft2tf: First-person statement text-to-talking face generation," *arXiv preprint arXiv:2312.05430*, 2023.

[133] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar, "Towards automatic face-to-face translation," in *ACM MM*, 2019.

[134] Z. Wang, M. Dai, and K. Lundgaard, "Text-to-video: A two-stage framework for zero-shot identity-agnostic talking-head generation," *arXiv preprint arXiv:2308.06457*, 2023.

[135] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML*, 2022.

[136] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nat. Mach. Intell.*, vol. 4, no. 11, pp. 930–939, 2022.

[137] Z. Ma, D. Zhou, C.-H. Yeh, X.-S. Wang, X. Li, H. Yang, Z. Dong, K. Keutzer, and J. Feng, "Magic-me: Identity-specific video customized diffusion," *arXiv preprint arXiv:2402.09368*, 2024.

[138] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *ICCV*, 2023.

[139] F. Shi, J. Gu, H. Xu, S. Xu, W. Zhang, and L. Wang, "Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models," in *CVPR*, 2024.

[140] J. Zhang, H. Yan, Z. Xu, J. Feng, and J. H. Liew, "Magicavatar: Multimodal avatar generation and animation," *arXiv preprint arXiv:2308.14748*, 2023.

[141] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[142] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *CVPR*, 2024.

[143] J. Yu, X. Cun, C. Qi, Y. Zhang, X. Wang, Y. Shan, and J. Zhang, "Animatezero: Video diffusion models are zero-shot image animators," *arXiv preprint arXiv:2312.03793*, 2023.

[144] F. Wang, P. Liu, H. Hu, D. Meng, J. Su, J. Xu, Y. Zhang, X. Ren, and Z. Zhang, "Loopanimate: Loopable salient object animation," *arXiv preprint arXiv:2404.09172*, 2024.

[145] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "Tokenflow: Consistent diffusion features for consistent video editing," *arXiv preprint arXiv:2307.10373*, 2023.

[146] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[147] B. Liu, X. Liu, A. Dai, Z. Zeng, Z. Cui, and J. Yang, "Dual-stream diffusion net for text-to-video generation," *arXiv preprint arXiv:2308.08316*, 2023.

[148] Y. Ren, Y. Zhou, J. Yang, J. Shi, D. Liu, F. Liu, M. Kwon, and A. Shrivastava, "Customize-a-video: One-shot motion customization of text-to-video diffusion models," *arXiv preprint arXiv:2402.14780*, 2024.

[149] A. Eldesokey and P. Wonka, "Text2ac-zero: Consistent synthesis of animated characters using 2d diffusion," *arXiv preprint arXiv:2312.07133*, 2023.

[150] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE TVCG*, vol. 26, no. 12, pp. 3457–3466, 2020.

[151] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *CVPR*, 2019.

[152] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *CVPR*, 2019.

[153] D. Bigioi, S. Basak, M. Stypułkowski, M. Zieba, H. Jordan, R. McDonnell, and P. Corcoran, "Speech driven video editing via an audio-conditioned diffusion model," *Image Vis. Comput.*, vol. 142, 2024.

[154] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM MM*, 2020.

[155] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder," in *ISCA*, 2017.

[156] S. Yao, R. Zhong, Y. Yan, G. Zhai, and X. Yang, "Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering," *arXiv preprint arXiv:2201.00791*, 2022.

[157] S. Shen, W. Li, X. Huang, Z. Zhu, J. Zhou, and J. Lu, "Sd-nerf: Towards lifelike talking head animation via spatially-adaptive dual-driven nerfs," *IEEE TMM*, 2023.

[158] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," *arXiv preprint arXiv:2403.17694*, 2024.

[159] R. Yu, T. He, A. Zeng, Y. Wang, J. Guo, X. Tan, C. Liu, J. Chen, and J. Bian, "Make your actor talk: Generalizable and high-fidelity lip sync with motion and appearance disentanglement," *arXiv preprint arXiv:2406.08096*, 2024.

[160] X. Ji, C. Lin, Z. Ding, Y. Tai, J. Yang, J. Zhu, X. Hu, J. Zhang, D. Luo, and C. Wang, "Realtalk: Real-time and realistic audio-driven face generation with 3d facial prior-guided identity alignment network," *arXiv preprint arXiv:2406.18284*, 2024.

[161] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *ICCV*, 2021.

[162] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, and B. Zhou, "Semantic-aware implicit neural audio-driven video portrait generation," in *ECCV*, 2022.

[163] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *WACV*, 2024.

[164] T. He, J. Guo, R. Yu, Y. Wang, J. Zhu, K. An, L. Li, X. Tan, C. Wang, H. Hu, H. Tao, S. Zhao, and J. Bian, "Gaia: Zero-shot talking avatar generation," in *ICLR*, 2024.

[165] Z. Ye, L. Zhang, D. Zeng, Q. Lu, and N. Jiang, "R2-talker: Realistic real-time talking head synthesis with hash grid landmarks encoding and progressive multilayer conditioning," *arXiv preprint arXiv:2312.05572*, 2023.

[166] J. Tang, K. Wang, H. Zhou, X. Chen, D. He, T. Hu, J. Liu, G. Zeng, and J. Wang, "Real-time neural radiance talking portrait synthesis via audio-spatial decomposition," *arXiv preprint arXiv:2211.12368*, 2022.

[167] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan, "Synctalk: The devil is in the synchronization for talking head synthesis," in *CVPR*, 2024.

[168] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting," in *ECCV*, 2024.

[169] H. Yu, Z. Qu, Q. Yu, J. Chen, Z. Jiang, Z. Chen, S. Zhang, J. Xu, F. Wu, C. Lv, *et al.*, "Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting," *arXiv preprint arXiv:2404.14037*, 2024.

[170] A. Delbosc, M. Ochs, N. Sabouret, B. Ravenet, and S. Ayache, "Towards the generation of synchronized and believable non-verbal facial behaviors of a talking virtual agent," in *ICMI*, 2023.

[171] Z. Chu, K. Guo, X. Xing, Y. Lan, B. Cai, and X. Xu, "Corrtalk: Correlation between hierarchical speech and facial activity variances for 3d animation," *IEEE TCSVT*, 2024.

[172] A. Eldesokey and P. Wonka, "Latentman: Generating consistent animated characters using image diffusion models," in *CVPR*, 2024.

[173] S. Yang, Y. Zhang, X. Cun, Y. Shan, and R. He, "Zerosmooth: Training-free diffuser adaptation for high frame rate video generation," *arXiv preprint arXiv:2406.00908*, 2024.

[174] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *ICANN*, 2005.

[175] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[176] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *CVPR*, 2021.

[177] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *AAAI*, 2023.
[178] Z. Ye, T. Zhong, Y. Ren, J. Yang, W. Li, J. Huang, Z. Jiang, J. He, R. Huang, J. Liu, *et al.*, "Real3d-portrait: One-shot realistic 3d talking portrait synthesis," *arXiv preprint arXiv:2401.08503*, 2024.
[179] Y. Ma, S. Zhang, J. Wang, X. Wang, Y. Zhang, and Z. Deng, "Dreamtalk: When expressive talking head generation meets diffusion probabilistic models," *arXiv preprint arXiv:2312.09767*, 2023.
[180] C. Wang, K. Tian, J. Zhang, Y. Guan, F. Luo, F. Shen, Z. Jiang, Q. Gu, X. Han, and W. Yang, "V-express: Conditional dropout for progressive training of portrait video generation," *arXiv preprint arXiv:2406.02511*, 2024.
[181] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Trans. Multimedia*, vol. 24, 2021.
[182] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023.
[183] Y. Bai, Y. Fan, X. Wang, Y. Zhang, J. Sun, C. Yuan, and Y. Shan, "High-fidelity facial avatar reconstruction from monocular video with generative priors," in *CVPR*, 2023.
[184] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu, *et al.*, "Stylesync: High-fidelity generalized and personalized lip sync in style-based generator," in *CVPR*, 2023.
[185] C. Xu, Y. Liu, J. Xing, W. Wang, M. Sun, J. Dan, T. Huang, S. Li, Z.-Q. Cheng, Y. Tai, *et al.*, "Facechain-imagineid: Freely crafting high-fidelity diverse talking faces from disentangled audio," in *CVPR*, 2024.
[186] N. Drobyshev, A. B. Casademunt, K. Vougioukas, Z. Landgraf, S. Petridis, and M. Pantic, "Emoportraits: Emotion-enhanced multimodal one-shot head avatars," in *CVPR*, 2024.
[187] C. Cai, G. Guo, J. Li, J. Su, C. He, J. Xiao, Y. Chen, L. Dai, and F. Zhu, "Listen, disentangle, and control: Controllable speech-driven talking head generation," *arXiv preprint arXiv:2405.07257*, 2024.
[188] Y. Chen, J. Zhao, and W.-Q. Zhang, "Expressive speech-driven facial animation with controllable emotions," in *ICMEW*, 2023.
[189] Q. Zhao, P. Long, Q. Zhang, D. Qin, H. Liang, L. Zhang, Y. Zhang, J. Yu, and L. Xu, "Media2face: Co-speech facial animation generation with multi-modality guidance," in *SIGGRAPH*, 2024.
[190] X. Liu, Q. Wu, H. Zhou, Y. Du, W. Wu, D. Lin, and Z. Liu, "Audio-driven co-speech gesture video generation," in *NeurIPS*, vol. 35, 2022.
[191] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
[192] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *CVPR*, 2023.
[193] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "End-to-end generation of talking faces from noisy speech," in *ICASSP*, 2020.
[194] Y. Yang, B. Shillingford, Y. Assael, M. Wang, W. Liu, Y. Chen, Y. Zhang, E. Sezener, L. C. Cobo, M. Denil, Y. Aytar, and N. de Freitas, "Large-scale multilingual audio visual dubbing," *arXiv preprint arXiv:2011.03530*, 2020.
[195] Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, and Y. Ding, "Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video," in *AAAI*, 2023.
[196] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," in *NeurIPS*, 2022.
[197] R. Feng, C. Li, and C. C. Loy, "Kalman-inspired feature propagation for video face super-resolution," in *ECCV*, 2024.
[198] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuhori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer, "Streamdiffusion: A pipeline-level solution for real-time interactive generation," *arXiv preprint arXiv:2312.12491*, 2023.
[199] F. Liang, A. Kodaira, C. Xu, M. Tomizuka, K. Keutzer, and D. Marculescu, "Looking backward: Streaming video-to-video translation with feature banks," *arXiv preprint arXiv:2405.15757*, 2024.
[200] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.
[201] Y. Zhai, K. Lin, Z. Yang, L. Li, J. Wang, C.-C. Lin, D. Doermann, J. Yuan, and L. Wang, "Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation," *arXiv preprint arXiv:2406.06890*, 2024.
[202] 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," in *NeurIPS*, 2017.
[203] G. Parmar, R. Zhang, and J.-Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation," in *CVPR*, 2022.
[204] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
[205] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *ICPR*, 2010.
[206] 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.
[207] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *QoMEX*, 2009.
[208] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," *IEEE TIP*, vol. 20, no. 9, pp. 2678–2683, 2011.
[209] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.
[210] 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 preprint arXiv:1812.01717*, 2018.
[211] 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," *TOG*, vol. 39, no. 6, 2020.
[212] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[213] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *CVPR*, 2018.
[214] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023.
[215] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image gans," in *WACV*, 2021.
[216] 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.
[217] B. Li, Y. Zhao, and L. Sheng, "Dancenet3d: Music based dance generation with parametric motion transformer," *arXiv preprint arXiv:2103.10206*, 2021.
[218] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao, "Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," in *ICCV*, 2021.
[219] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE TPAMI*, vol. 35, no. 12, pp. 2878–2890, 2013.
[220] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *NeurIPS*, 2019.
[221] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *CVPR*, 2022.
[222] 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.
[223] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *CVPR*, 2024.
[224] 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.
[225] Y. Jafarian and H. S. Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *CVPR*, 2021.
[226] J. Xiao, A. Yao, Y. Li, and T.-S. Chua, "Can i trust your answer? visually grounded video question answering," in *CVPR*, 2024.
[227] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *ICCV*, 2023.
[228] S. Kim, "3dyoga90: A hierarchical video dataset for yoga pose understanding," *arXiv preprint arXiv:2310.10131*, 2023.
[229] J. Byrne, G. Castanon, Z. Li, and G. Ettinger, "Fine-grained activities of people worldwide," in *WACV*, 2023.
[230] D. Davila, D. Du, B. Lewis, C. Funk, J. Van Pelt, R. Collins, K. Corona, M. Brown, S. McCloskey, A. Hoogs, and B. Clipp, "Mevid: Multi-view extended videos with identities for video person re-identification," in *WACV*, 2023.
[231] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[232] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan, "Vfhq: A high-quality dataset and benchmark for video face super-resolution," in *CVPRW*, 2022.

[233] F. G. Lohesara, D. R. Freitas, C. Guillemot, K. Eguiazarian, and S. Knorr, "Headset: Human emotion awareness under partial occlusions multimodal dataset," *TVCG*, vol. 29, no. 11, pp. 4686–4696, 2023.

[234] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu, "Open-domain sign language translation learned from online video," in *EMNLP*, 2022.

[235] D. Moltisanti, J. Wu, B. Dai, and C. C. Loy, "Brace: The breakdancing competition dataset for dance motion synthesis," in *ECCV*, 2022.

[236] 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.

[237] P. Dal Bianco, G. Ríos, F. Ronchetti, F. Quiroga, O. Stanchi, W. Hasperué, and A. Rosete, "Lsa-t: The first continuous argentinian sign language dataset for sign language translation," in *IBERAMIA*, 2022.

[238] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu, "Celebv-text: A large-scale facial text-video dataset," in *CVPR*, 2023.

[239] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022.

[240] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei, "Responsive listening head generation: A benchmark dataset and baseline," in *ECCV*, 2022.

[241] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *CVPR*, 2021.

[242] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *ACM MM*, 2021.

[243] 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.

[244] O. Ignat, S. Castro, Y. Zhou, J. Bao, D. Shan, and R. Mihalcea, "When did it happen? duration-informed temporal localization of narrated actions in vlogs," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 3s, 2022.

[245] U.-H. Kim, Y. Hwang, S.-K. Lee, and J.-H. Kim, "Writing in the air: Unconstrained text recognition from finger movement using spatio-temporal convolution," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1386–1398, 2022.

[246] A. Brown, V. Kalogeiton, and A. Zisserman, "Face, body, voice: Video person-clustering with multiple modalities," in *ICCV*, 2021.

[247] M. Godi, C. Joppi, G. Skenderi, and M. Cristani, "Movingfashion: a benchmark for the video-to-shop challenge," in *WACV*, 2022.

[248] R. Javadi and A. Lim, "The many faces of anger: A multicultural video dataset of negative emotions," in *FG*, 2021.

[249] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020.

[250] K. Papadimitriou and G. Potamianos, "Sign language recognition via deformable 3d convolutions and modulated graph convolutional networks," in *ICASSP*, 2023.

[251] L.-A. Zeng, F.-T. Hong, W.-S. Zheng, Q.-Z. Yu, W. Zeng, Y.-W. Wang, and J.-H. Lai, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *ACM MM*, pp. 2526–2534, 2020.

[252] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Classification of important segments in educational videos using multimodal features," *arXiv preprint arXiv:2010.13626*, 2020.

[253] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robot. Autom. Lett.*, vol. 5, no. 1, pp. 187–194, 2020.

[254] J. Fink, B. Frénay, L. Meurant, and A. Cleve, "Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition," in *IJCNN*, 2021.

[255] A. G. Perera, Y. W. Law, T. T. Ogunwa, and J. Chahl, "A multiviewpoint outdoor dataset for human action recognition," *IEEE Trans. Human-Machine Syst.*, vol. 50, no. 5, pp. 405–413, 2020.

[256] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, 2019.

[257] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "Dwnet: Dense warp-based network for pose-guided human video generation," *arXiv preprint arXiv:1910.09139*, 2019.

[258] O. Ignat, L. Burdick, J. Deng, and R. Mihalcea, "Identifying visible actions in lifestyle vlogs," in *ACL*, 2019.

[259] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[260] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *AAAI*, 2018.

[261] P. Barros, N. Churamani, E. Lakomkin, H. Sequeira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *IJCNN*, 2018.

[262] Y. Liu, B. Peng, P. Shi, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin, J. Jiang, Y. Fan, *et al.*, "iqiyi-vid: A large dataset for multi-modal person identification," *arXiv preprint arXiv:1811.07548*, 2018.

[263] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[264] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011.

[265] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[266] W. Guo, X. Bie, X. Alameda-Pineda, and F. Moreno-Noguer, "Multi-person extreme motion prediction," in *CVPR*, 2022.

[267] L. Chu, Y. Liu, Z. Wu, S. Tang, G. Chen, Y. Hao, J. Peng, Z. Yu, Z. Chen, B. Lai, and H. Xiong, "Pp-humanseg: Connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset," in *WACV*, 2022.

[268] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei, "Vico-x: Multimodal conversation dataset." https://project.mhzhou.com/vico, 2022.

[269] Q.-Y. Jiang, Y. He, G. Li, J. Lin, L. Li, and W.-J. Li, "Svd: A large-scale short video dataset for near-duplicate video retrieval," in *ICCV*, 2019.

[270] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," *arXiv preprint arXiv:2001.03024*, 2020.

[271] G. Fox, W. Liu, H. Kim, H.-P. Seidel, M. Elgharib, and C. Theobalt, "Video-forensics-hq: Detecting high-quality manipulated face videos," in *ICME*, 2021.

[272] F. Rayar, M. Delalandre, and V.-H. Le, "A large-scale tv video and metadata database for french political content analysis and fact-checking," in *CBMI*, 2022.

[273] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *FG*, 2017.

[274] O. M. Sincan, J. Junior, C. Jacques, S. Escalera, and H. Y. Keles, "Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, results and future research," in *CVPR*, 2021.

[275] D. C. Ong, Z. Wu, Z.-X. Tan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling emotion in complex stories: The stanford emotional narratives dataset," *IEEE Trans. Affective Comput.*, vol. 12, no. 3, pp. 579–594, 2021.

[276] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *ICCVW*, 2019.

[277] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, 2019.

[278] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.

[279] M. K. Keutmann, S. L. Moore, A. Savitt, and R. C. Gur, "Generating an item pool for translational social cognition research: Methodology and initial validation," *Behav. Res. Methods*, vol. 47, pp. 228–234, 2015.

[280] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[281] X. Y. U. o. W. M. Peiran L, Linbo T, "L-svd: A comprehensive video dataset for emotion recognition." https://github.com/PeiranLi0930/emotionnet, 2024.

[282] J. Tang, K. Chen, Y. Wang, Y. Shi, S. Patel, D. McDuff, and X. Liu, "Mmpd: Multi-domain mobile video physiology dataset," *EMBC*, 2023.

[283] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The casual conversations v2 dataset," in *CVPR*, 2023.

[284] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[285] P. Sarkar, A. Posen, and A. Etemad, "Avcaffe: a large scale audio-visual dataset of cognitive load and affect for remote work," in *AAAI*, 2023.
[286] S.-H. Han, M.-G. Park, J. H. Yoon, J.-M. Kang, Y.-J. Park, and H.-G. Jeon, "High-fidelity 3d human digitization from single 2k resolution images," in *CVPR*, 2023.