Human Motion Generation: A Survey

Wentao Zhu, *Graduate Student Member, IEEE*, Xiaoxuan Ma, *Graduate Student Member, IEEE*, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, *Fellow, IEEE*, and Yizhou Wang, *Member, IEEE*

(Survey Paper)

***Abstract*—Human motion generation aims to generate natural human pose sequences and shows immense potential for real-world applications. Substantial progress has been made recently in motion data collection technologies and generation methods, laying the foundation for increasing interest in human motion generation. Most research within this field focuses on generating human motions based on conditional signals, such as text, audio, and scene contexts. While significant advancements have been made in recent years, the task continues to pose challenges due to the intricate nature of human motion and its implicit relationship with conditional signals. In this survey, we present a comprehensive literature review of human motion generation, which, to the best of our knowledge, is the first of its kind in this field. We begin by introducing the background of human motion and generative models, followed by an examination of representative methods for three mainstream sub-tasks: text-conditioned, audio-conditioned, and scene-conditioned human motion generation. Additionally, we provide an overview of common datasets and evaluation metrics. Lastly, we discuss open problems and outline potential future research directions. We hope that this survey could provide the community with a comprehensive glimpse of this rapidly evolving field and inspire novel ideas that address the outstanding challenges.**

***Index Terms*—Human motion, generative model, deep learning, literature survey.**

I. Introduction

HUMANS plan and execute body motions based on their intention and the environmental stimulus [1], [2]. As an essential goal of artificial intelligence, generating human-like motion patterns has gained increasing interest from various research communities, including computer vision [3], [4], computer graphics [5], [6], multimedia [7], [8], robotics [9], [10], and human-computer interaction [11], [12]. The goal of human motion generation is to generate natural, realistic and diverse human motions that can be used for a wide range of applications, including film production, video games, AR/VR, human-robot interaction, and digital humans.

With the rise of deep learning [17], recent years have witnessed a rapid development of various generation methods, e.g., Autoregressive models [18], Variational Autoencoders (VAE) [19], Normalizing Flows [20], Generative Adversarial Networks (GAN) [21], and Denoising Diffusion Probabilistic Models (DDPM) [22]. These methods have demonstrated great success across different domains, including text [23], [24], image [25], [26], [27], video [28], [29], [30], and 3D objects [31], [32]. On the other hand, the remarkable progress in human modeling [33], [34], [35] makes it easier to extract human motion from videos [36], [37], [38] and construct large-scale human motion datasets [39], [40], [41], [42]. Consequently, the community has gained increasing interest in data-driven human motion generation over the past few years.

Nonetheless, human motion generation presents a complex challenge that extends beyond the mere application of deep generative models to human motion datasets. First, human motion is highly non-linear and articulated, subject to physical and bio-mechanical constraints. Additionally, human brains possess specialized neural mechanisms for perceiving biological motion [2], [43] and are sensitive to even slightly unnatural kinematics [44], [45]. As a result, high visual quality is required for generated motions in terms of naturalness, smoothness, and plausibility. Second, the demand for human motion generation often includes a context as the conditional signal, such as text description, background audio, or surrounding environments, as shown in Fig. 1. Generated motion should not only be plausible in itself but also harmonious with the conditional signal. Third, human motion serves as an essential nonverbal communication medium, reflecting various underlying factors, such as goals, personal styles, social norms, and cultural expressions [46]. Ideally, motion generation models should learn to capture subtle variations and the semantic connection with the conditional signals.

In light of the rapid development and emerging challenges, we present a comprehensive survey of this field to help the

Manuscript received 20 July 2023; revised 20 October 2023; accepted 5 November 2023. Date of publication 8 November 2023; date of current version 6 March 2024. This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0114900, and in part by the National Natural Science Foundation of China under Grant 62176006. Recommended for acceptance by J. Gall. *(Wentao Zhu, Xiaoxuan Ma, and Dongwoo Ro contributed equally to this work.) (Corresponding authors: Feng Gao; Qi Tian; and Yizhou Wang.)*

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, and Jinlu Zhang are with the Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100871, China (e-mail: wtzhu@pku.edu.cn; maxiaoxuan@pku.edu.cn; dwro0121@gmail.com; cihai@pku.edu.cn; jinluzhang@stu.pku.edu.cn).

Jiaxin Shi and Qi Tian are with the Huawei Cloud Computing Technologies Company, Ltd., Shenzhen, Guangdong 518129, China (e-mail: shijx12@gmail.com; tian.qi1@huawei.com).

Feng Gao is with the School of Arts, Peking University, Beijing 100871, China (e-mail: gaof@pku.edu.cn).

Yizhou Wang is with the Center on Frontiers of Computing Studies, School of Computer Science, Institute for Artificial Intelligence, Peking University, Beijing 100871, China (e-mail: Yizhou.Wang@pku.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2023.3330935

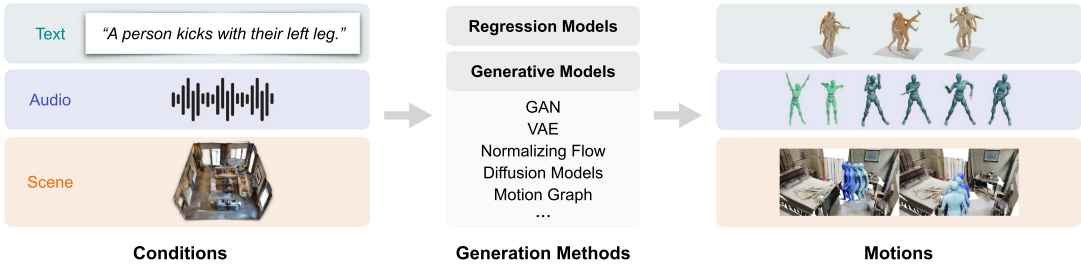


Fig. 1. An overview of typical human motion generation approaches. Example images adapted from [13], [14], [15], [16].

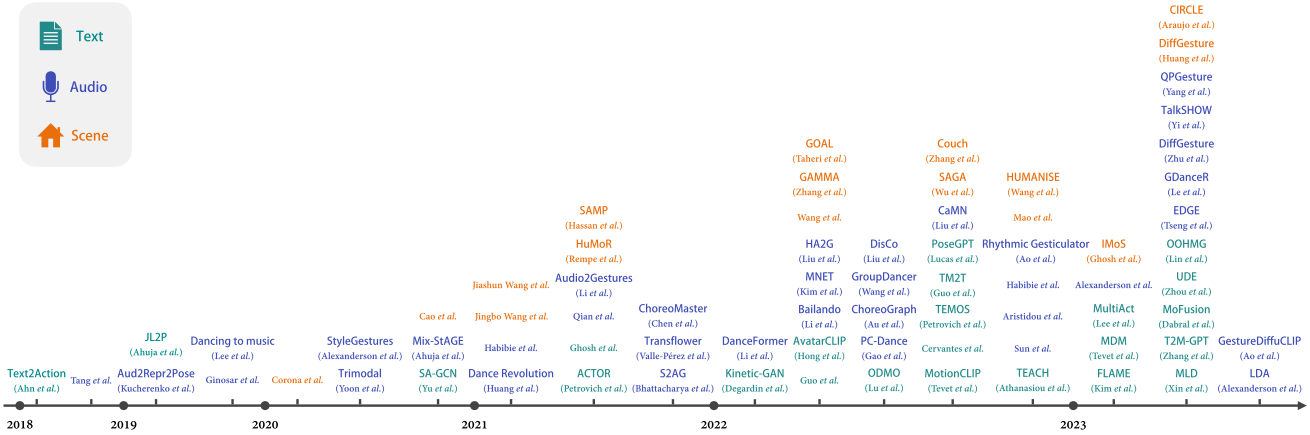


Fig. 2. Recent advances of human motion generation methods with different conditions.

community keep track of its progress. In Fig. 2, we summarize the development of human motion generation methods in recent years. The rest of the survey is organized as follows. In Section II, we discuss the scope of this survey. Section III covers the fundamentals of the task, including representations of human motion, motion data collection techniques, and various generative methods. In Sections IV, V, and VI, we summarize existing approaches for human motion generation based on different conditional signals respectively, including text, audio, and scene. Section VII introduces the commonly used datasets and their properties. Section VIII summarizes the evaluation metrics from various perspectives. Finally, we draw conclusions and provide some future directions for this field in Section IX.

II. Scope

This survey focuses on the generation of human motion based on given conditional signals. We primarily discuss text, audio, and scene conditions. Some works also propose to generate human motion based on other conditions (e.g., others' motion [47]). With respect to the generation target, we incorporate different types of human motion representations, such as sequences of 2D/3D body keypoint, joint rotations, and parametric human body models [33], [34]. We do not cover methods on human motion completion (e.g., motion prediction, motion interpolation) or human motion editing (e.g. motion retargeting, motion style transfer). For reviews on these approaches, we direct readers to [48], [49], [50], [51]. Additionally, we do not discuss works on generating human motion using physical simulation environments (e.g., character control, locomotion); please refer to [52] for a summary of such methods. This survey serves as a complement to existing survey papers that focus on human pose estimation [53], [54], motion capture [55], [56], and deep generative models [57], [58], [59].

III. Preliminaries

A. Motion Data

We first introduce the human motion data representations, then discuss various human motion data collection techniques and their characteristics.

1) Motion Data Representation: Human motion data can be effectively represented by the sequence of human body poses over the temporal dimension. More specifically, we categorize the data representations into *keypoint-based* and *rotation-based*. It is worth noting that a conversion is possible between these two types of representations. We can transition from joint rotations to keypoints using forward kinematics (FK), and inversely, from keypoints to joint rotations using inverse kinematics (IK).

Keypoint-Based: The human body is represented by a set of keypoints, which are specific points on the body that correspond to anatomical landmarks, such as joints or other significant

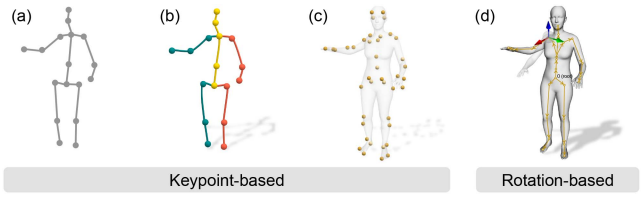


Fig. 3. Typical human pose and shape representations with the same pose in (a) 2D keypoints, (b) 3D keypoints, (c) 3D marker keypoints, and (d) rotation-based model.

Fig. 4. Human motion data collection methods. (a) Examples of marker-based motion capture setup where (left) optical markers [65] or (right) IMUs [66] are attached to the subject's body surface. (b) Example of the markerless multiview motion capture system [41]. (c) Pseudo-labeling pipeline involves using pose or mesh estimators to generate pseudo labels [67]. (d) Example interface for manual collection using MikuMikuDance (MMD) resources.

locations. Each keypoint is represented by its 2D/3D coordinates in the pixel or world coordinate system, as shown in Fig. 3(a) and (b). The motion data is then represented as a sequence of keypoint configurations over time. Keypoint-based representations can be directly derived from motion capture systems and exhibit great interpretability. However, in order to use keypoint-based data for animation or robotics, it is usually necessary to solve the inverse kinematics (IK) problem and convert the keypoints to rotations. Recently, some work [60], [61], [62] propose to represent the human pose with more landmarks on the surface of the human body, i.e., body markers, as shown in Fig. 3(c). Compared to traditional skeleton keypoints, body markers provide more comprehensive information in terms of body shapes and limb twists.

Rotation-Based: Human pose could also be represented by joint angles, i.e., the rotation of the body parts or segments, relative to their parent in a hierarchical structure. Most studies consider 3D joint rotations in $\mathrm{SO}(3)$ and the rotations can be parameterized using various formats, such as Euler angles, axis angles, and quaternions. Based on the joint angle, some works [33], [34] model the human with statistical mesh models that further capture the shape of the body and the deformations that occur during movement. A widely-used statistical body model is the Skinned Multi-Person Linear (SMPL) model [33].

The SMPL model is parametrized by a set of pose and shape parameters, which can be used to generate a 3D mesh of a human body in a specific pose and shape, as shown in Fig. 3(d). Pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{K\times 3}$ of each joint are defined by the relative rotation with respect to its parent in a standard skeletal kinematic tree with $K = 24$ joints. For simplicity, we include the root orientation as part of the pose parameters for the root joint in our formulation. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ indicate the body shape configurations, such as height. Given the pose and shape parameters, the model deforms accordingly and generates a triangulated mesh comprising $N = 6890$ vertices as $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{N\times 3}$. The deformation process $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ is differentiable with respect to the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters. Once the final mesh is obtained, sparse 3D keypoints can be mapped from vertices through a pretrained linear regressor. Other models such as SMPL-X [34] extends SMPL [33] model and constructs a comprehensive model, wherein the body, face, and hands are modeled jointly. In addition to SMPL-based linear models, alternative modeling approaches have been explored, such as GHUM [63] and STAR [64]. To ensure conciseness, we employ the shorthand term "Rot." in the tables below to encompass both joint-based 3D rotations and their applications in statistical human models (e.g. SMPL), without delving into intricate differentiation between the two.

2) Motion Data Collection: There are four main approaches to collecting human motion data: (i) *marker-based* motion capture, (ii) *markerless* motion capture, (iii) *pseudo-labeling*, and (iv) *manual annotation*.

Marker-based motion capture involves placing small reflective markers or Inertial Measurement Units (IMUs) at specific locations in the subject's body and then tracking the movement of these markers in a 3D space. See Fig. 4(a) for illustration. This data can then be used to obtain 3D keypoints by applying forward kinematics [39] or a parametric body mesh such as SMPL [33] with the help of MoSh [68]. Optical markers provide more accurate data than IMUs, but are less portable and are typically used in indoor environments, while IMUs can be used in outdoor settings.

Markerless motion capture solutions track the movement of the subject's body without the need for markers from one or multiple cameras and use computer vision algorithms (e.g., [69], [70], [71]) to get the 3D motion by exploiting multi-view geometry, as shown in Fig. 4(b). Multiple RGB or RGB-D cameras will be set up and synchronized during the capture process. This solution is less accurate than marker-based motion capture, but is more convenient and can be used in a wider range of settings.

Pseudo-labeling of human motion is primarily intended for in-the-wild captured monocular RGB images or videos. This involves predicting 2D or 3D human keypoints with existing human pose estimators such as OpenPose [72] and VideoPose3D [37], or fits body model to image evidence to generate pseudo 3D mesh labels, e.g., by using SMPLify-X [67]. See Fig. 4(c). However, pseudo-labels tend to have more errors compared to motion capture systems.

Manual annotation involves designing human motion with an animation engine manually, typically using a team of skilled artists. Fig. 4(d) shows an example engine interface of MikuMikuDance (MMD). While this approach can produce high-quality animations, it is expensive, time-consuming, and not scalable.

B. Motion Generation Methods

We roughly classify human motion generation methods into two classes. The first class of methods is based on *regression models* to predict human motion using features encoded

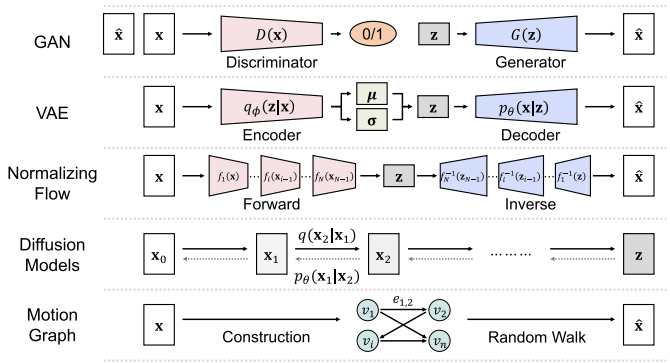


Fig. 5. Overview of different generative models.

from input conditions. They fall into the supervised learning paradigm and aim to establish a direct mapping from input conditions to target motions. The other class of methods base on *generative models*. They focus on modeling the underlying distribution of motion (or joint distribution with conditions) in an unsupervised manner. Typical deep generative models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Normalizing Flows, and Denoising Diffusion Probabilistic Models (DDPMs). In addition to the general generative models, a task-specific model, motion graph, has also been widely used especially in the field of computer graphics and animation. Fig. 5 shows an overview of different generative models. In the following, we will briefly go over commonly used generative models in motion generation.

Generative Adversarial Networks: GANs [21] are a class of generative models composed of two neural networks: the generator G and the discriminator D. The generator produces synthetic data from a noise vector $\mathbf{z}$ to deceive the discriminator. Conversely, the discriminator tries to differentiate between real data and synthetic data produced by the generator. This dynamic between the generator and the discriminator can be viewed as a zero-sum or min-max game. The loss function representing their interaction can be formulated as

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(D(\mathbf{x}))] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))], \quad (1)$$

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(D(G(\mathbf{z})))]. \quad (2)$$

With the rise of deep learning, various deep learning-based GANs have been proposed. Models such as DCGAN [73], PGGAN[74], and StyleGAN [75], [76] have demonstrated remarkable achievements and potential. These advancements in GANs have contributed significantly to the field of generative models, especially in the generation of synthetic data. However, GANs face several challenges, including training instability, convergence problems, and mode collapse.

Variational Autoencoders: VAEs [19] are notable generative models that provide robust solution for data representation. They address the challanges of interactable likelihood by using feed-forward model, denoted as $q_\phi(\mathbf{z}|\mathbf{x})$, to approximate the intractable posterior. The primary optimization goal is to minimize the KL divergence between this approximation and the original posterior. VAEs adopt the Evidence Lower Bound (ELBO) as loss function

$$\text{ELBO} = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\log(p_\theta(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})). \quad (3)$$

VAEs efficiently generate and infer new samples due to the feed-forward mode of $q_\phi(\mathbf{z}|\mathbf{x})$. Additionally, the reparameterization trick enables differentiable sample generation and the utilization of a reconstruction-based loss function, ultimately enhancing training efficiency and stability. These advantages have led to the widespread adoption of VAEs variants, such as CVAE [77], LVAE [78], and VQ-VAE [79], in various fields and drive advances in generative models. However, VAEs are subject to the risk of posterior collapse and may produce less sharp samples compared to GANs.

Normalizing Flows: GANs and VAEs implicitly learn the probability density of data. They can hardly calculate the exact likelihood. In contrast, Normalizing Flows [20] is a class of generative models that explicitly learn the data distribution $p(\mathbf{x})$ and allows for tractable probability density estimation. These models employ a series of invertible transformations $\{f_i\}_{1:N}$ to map a simple prior distribution $p(\mathbf{z}_0)$ (e.g., a standard Gaussian) to a complex data distribution $p(\mathbf{x})$

$$\mathbf{z}_i = f_i\left(\mathbf{z}_{i-1}\right) \quad (4)$$

$$\mathbf{x} = \mathbf{z}_N = f_N \circ f_{N-1} \circ \cdots \circ f_1(\mathbf{z}_0). \quad (5)$$

The density of the target distribution can be obtained by applying the change of variables theorem

$$\log p(\mathbf{z}_i) = \log p(\mathbf{z}_{i-1}) - \log \left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right| \quad (6)$$

$$\log p(\mathbf{x}) = \log p(\mathbf{z}_0) - \sum_{i=1}^{K} \log \left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|, \quad (7)$$

where $\det$ is short for the determinant of a square matrix. Normalizing Flows can be typically trained by maximizing the log-likelihood of the observed data. Owing to the invertible transformation, Normalizing Flows offer flexibility, exact likelihood computation, and easy data sampling. However, they require a large number of transformations to model complex distributions and can be computationally expensive and difficult to train.

Diffusion Models: Diffusion models [22], [80], [81] define a forward diffusion process that gradually adds a small amount of Gaussian noise to the input data $\mathbf{x}_0$ in T steps, producing a series of noisy samples $\{\mathbf{x}_t\}_{1:T}$. Noise is scheduled by $\{\beta_t\}_{1:T}$.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (8)$$

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \quad (9)$$

As $T \to \infty$, $\mathbf{x}_T$ is actually a Gaussian distribution. If we know the reverse transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, then we can sample from a Gaussian prior $\mathbf{x}_t \sim \mathcal{N}(0, \mathbf{I})$ and run the diffusion process in reverse to get a sample from the real data distribution $p(\mathbf{x}_0)$. However, since $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ depends on the entire dataset and is hard to estimate, we train a neural network p_θ to match

the posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, a tractable Gaussian, instead of $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \tag{10}$$

p_θ is learned by optimizing the ELBO like VAE. In practice, diffusion models are able to produce high-quality samples and can benefit from stable training. However, it relies on a long Markov chain of reverse diffusion steps to generate samples, so it can be computationally expensive and slower than GANs and VAEs.

Motion Graph: Motion graph [82], [83], [84] can be represented mathematically as a directed graph $G = \langle V, E \rangle$ where V denotes the set of nodes or vertices, and E denotes the set of directed edges or transitions. Each node $v \in V$ represents a pose or keyframe, and each directed edge $e \in E$ connects two vertices (v_1, v_2) and represents a feasible transition between the corresponding poses. Motion graphs are first constructed based on a collection of motion clips. To ensure smooth transitions, the algorithm identifies compatible poses within the motion clips and connects them with edges, forming a graph that can be traversed to generate new motion sequences. After constructing the motion graph, a random walk $W = (v_1, v_2, \ldots, v_n)$ can be performed on the graph, starting from an initial node and following the directed edges. The output motion sequence is a concatenation of the poses corresponding to the traversed nodes, ensuring smooth transitions between consecutive poses. Meanwhile, further constraints can be incorporated as optimization objectives [85], [86]. This process effectively creates new motion sequences that were not explicitly present in the original dataset, but are consistent with the overall characteristics of the data.

IV. Text-Conditioned Motion Generation

Text possesses a remarkable ability to convey various types of actions, velocities, directions, and destinations, either explicitly or implicitly. This characteristic makes the text an appealing condition for generating human motion. This section aims to elucidate the topic of text-conditioned human motion generation tasks (see Table I top block), which can be primarily divided into two categories: *action-to-motion* and *text-to-motion*.

A. Action to Motion

The action-to-motion task is focused on generating human motion sequences based on specific action categories, such as 'Walk', 'Kick', or 'Throw'. These actions are often represented using techniques like one-hot encoding, which simplifies the motion generation process. Compared to text-to-motion tasks which deal with the complexities of natural language processing, this representation provides a more straightforward task due to limited and well-defined action classes.

Yu et al. [88] introduce SA-GAN, which leverages a self-attention-based graph convolutional network (GCN) with GAN architecture. They also propose to enhance the generative capabilities through the use of two discriminators - one frame-based and the other sequence-based. In a similar vein, Kinetic-GAN [92] combines the strengths of GAN and GCN, and further utilizes latent space disentanglement and stochastic variations to generate high-quality and diverse human motions. Guo et al. [7] introduce Action2Motion, a per-frame VAE architecture based on Gated Recurrent Units (GRU) to generate motion sequences. Similarly, ACTOR [90] employs a sequence-level CVAE model that uses transformers as a backbone for generating motion sequences non-autoregressively. This non-autoregressive approach allows for the one-shot generation of motion sequences. ODMO [94] adopts a novel strategy of applying contrastive learning within a low-dimensional latent space, thus generating hierarchical embeddings of motion sequences. The model initially creates motion trajectories before generating the motion sequences, thus benefiting trajectory control. Furthermore, PoseGPT [95] utilizes an auto-regressive transformer to encode human motion into quantized latent representations, subsequently employing a GPT-like model for next motion index predictions within this discrete space. Cervantes et al. [98] introduce a method that uses implicit neural representation (INR) and a fitted conditional Gaussian mixed model (GMM). This method controls the length and action classes of the sequence by extracting representations from the variational distributions of each training sequence. In addition, MDM [14] utilizes a diffusion model to predict samples at each diffusion step, rather than just noise. MLD [101] draws inspiration from the Latent Diffusion Model (LDM) [169] to employ latent-level diffusion and VAE for motion generation.

While these methods have greatly advanced the field of action-to-motion, they primarily excel at generating single-action motions. The transition to generating complex sequences that involve multiple actions remains a challenge and often requires additional post-processing to connect disparate actions. To this end, a recent work, MultiAct [99], leverages past motion to recurrently generate long-term multi-action 3D human motion and proposes a novel face-front canonicalization methodology to ensure the local coordinate system shares the ground geometry in each recurrent step.

B. Text to Motion

The text-to-motion task aims to generate human motion sequences from natural language descriptions, leveraging the vast expressive power of language. In contrast to action-to-motion, which utilizes limited predefined labels, text-to-motion has the potential to produce a wider variety of motions based on diverse textual descriptions. Nonetheless, the challenge lies in accurately converting the intricacies of text into corresponding movements, requiring a profound understanding of both linguistic nuances and physical motion dynamics.

Text2Action [102] first leverages GAN to generate a variety of motions from a given natural language description. Some other methods have explored the potential of learning a joint embedding of text and motion. For instance, JL2P [104] uses a GRU-based text encoder and motion encoder-decoder to map the text into a corresponding human motion. Ghosh et al. [105] further develop a two-stream encoder-decoder model for co-embedding text and body movements, while also employing a GAN structure to generate more natural motions. Guo et al. [3]

TABLE I
REPRESENTATIVE WORKS OF HUMAN MOTION GENERATION

Method	Venue	Representation	Model	Condition	Dataset
Action2Motion [7]	MM 2020	Rot.	VAE	Action class	[7], [87]
SA-GCN [88]	ECCV 2020	Kpts. (2D, 3D)	GAN	Action class	[87], [89]
ACTOR [90]	ICCV 2021	Rot.	VAE	Action class	[7], [87], [91]
Kinetic-GAN [92]	WACV 2022	Kpts. (3D)	GAN	Action class	[87], [89], [93]
ODMO [94]	MM 2022	Kpts. (3D)	VAE	Action class, Trajectory†	[7], [91]
PoseGPT [95]	ECCV 2022	Rot.	VAE	Action class, Duration, Past motion†	[7], [96], [97]
Cervantes *et al.* [98]	ECCV 2022	Rot.	Regression	Action class	[7], [87], [91]
MultiAct [99]	AAAI 2023	Kpts. (3D) / Rot.	VAE	Action class, Past motion†	[96]
MDM [14]	ICLR 2023	Kpts. (3D) / Rot.	Diffusion	Action class / Text	[3], [7], [91], [100]
MLD [101]	CVPR 2023	Kpts. (3D) / Rot.	Diffusion, VAE	Action class / Text	[3], [7], [91], [100]
Text2Action [102]	ICRA 2018	Kpts. (3D)	GAN	Text	[103]
JL2P [104]	3DV 2019	Kpts. (3D)	Regression	Text	[100]
Ghosh *et al.* [105]	ICCV 2021	Kpts. (3D)	GAN	Text	[100]
Guo *et al.* [3]	CVPR 2022	Kpts. (3D) / Rot.	VAE	Text, POS	[3], [100]
AvatarCLIP [106]	TOG 2022	Rot.	VAE	Text	[40]
MotionCLIP [107]	ECCV 2022	Rot.	Regression	Text	[96]
TEMOS [108]	ECCV 2022	Kpts. (3D) / Rot.	VAE	Text	[100]
TM2T [109]	ECCV 2022	Kpts. (3D) / Rot.	VAE	Text	[3], [100]
TEACH [110]	3DV 2022	Rot.	VAE	Text, Past motion†	[96]
FLAME [111]	AAAI 2023	Kpts. (3D) / Rot.	Diffusion	Text	[3], [96], [100]
T2M-GPT [112]	CVPR 2023	Kpts. (3D) / Rot.	VAE	Text	[3], [100]
OOHMG [113]	CVPR 2023	Rot.	VAE	Text	[40], [96]
UDE [114]	CVPR 2023	Rot.	Diffusion, VAE	Text / Music	[3], [115]
MoFusion [116]	CVPR 2023	Kpts. (3D)	Diffusion	Text / Music	[3], [96], [115]
Dance with Melody [117]	MM 2018	Kpts. (3D)	Regression	Music	[117]
Dancing to Music [118]	NeurIPS 2019	Kpts. (2D)	GAN	Music	[118]
Dance Revolution [119]	ICLR 2021	Kpts. (2D)	Regression	Music	[119]
AI Choreographer [115]	ICCV 2021	Rot.	Regression	Music, Past motion	[115]
Transflower [120]	TOG 2021	Kpts. (3D)	Normalizing Flow	Music, Past motion	[115], [120], [121], [122]
ChoreoMaster [86]	TOG 2021	Rot.	Motion Graph	Music	[86]
DanceFormer [123]	AAAI 2022	Rot.	GAN	Music	[123]
Bailando [124]	CVPR 2022	Kpts. (3D)	VAE	Music, Past motion	[115]
MNET [4]	CVPR 2022	Rot.	GAN	Music, Past motion, Style code	[115]
PC-Dance [8]	MM 2022	Rot.	Motion Graph	Music, Anchor pose†	[8]
ChoreoGraph [125]	MM 2022	Kpts. (3D)	Motion Graph	Music	[115]
GroupDancer [126]	MM 2022	Rot.	Regression	Music	[126]
Sun *et al.* [127]	NeurIPS 2022	Rot.	VAE	Music, Past motion	[115]
Aristidou *et al.* [128]	TVCG 2022	Rot.	Regression	Music	[115], [128]
EDGE [15]	CVPR 2023	Rot.	Diffusion	Music	[115]
GDanceR [129]	CVPR 2023	Rot.	Regression	Music	[129]
Ginosar *et al.* [130]	CVPR 2019	Kpts. (2D)	GAN	Speech	[130]
Aud2Repr2Pose [11]	IVA 2019	Kpts. (3D)	Regression	Speech	[131]
Mix-StAGE [132]	ECCV 2020	Kpts. (2D)	GAN	Speech, Style code	[130], [132]
StyleGestures [133]	CGF 2020	Rot.	Normalizing Flow	Speech, Past motion	[134]
Trimodal Context [135]	TOG 2020	Kpts. (3D)	GAN	Speech, Text, Speaker, Past motion	[135], [136]
Habibie *et al.* [137]	IVA 2021	Kpts. (3D)	GAN	Speech	[130], [137]
S2AG [138]	MM 2021	Kpts. (3D)	GAN	Speech, Text, Speaker, Past motion	[136], [139]
Qian *et al.* [140]	ICCV 2021	Kpts. (2D)	VAE	Speech, Template vector	[130]
Audio2Gestures [141]	ICCV 2021	Kpts. (2D) / Rot.	VAE	Speech	[130], [134]
HA2G [142]	CVPR 2022	Kpts. (3D)	GAN	Speech, Text, Speaker, Past motion	[135], [136]
DisCo [143]	MM 2022	Rot.	GAN	Speech, Past motion	[130], [134]
CaMN [144]	ECCV 2022	Rot.	GAN	Speech, Text, Speaker, Expressions, Emotions, Semantics	[130], [144]
Habibie *et al.* [145]	TOG 2022	Kpts. (3D)	GAN	Speech, Start pose, Control†	[130], [137]
Rhythmic Gesticulator [146]	TOG 2022	Rot.	VAE	Speech, Text, Speaker, Past motion	[134], [136], [146]
DiffGesture [147]	CVPR 2023	Kpts. (3D)	Diffusion	Speech	[135], [136]
TalkSHOW [148]	CVPR 2023	Rot.	VAE	Speech, Speaker	[148]
QPGesture [149]	CVPR 2023	Rot.	VAE	Speech, Text, Anchor pose, Control†	[144]
LDA [5]	TOG 2023	Rot.	Diffusion	Speech / Music / Path, Style code†	[120], [121], [139], [150], [151]
GestureDiffuCLIP [6]	TOG 2023	Rot.	Diffusion, VAE	Speech, Text, Style prompt	[144], [150]
Corona *et al.* [152]	CVPR 2020	Kpts. (3D)	Regression	Scene (object), Past motion	[65], [153]
Cao *et al.* [154]	ECCV 2020	Kpts. (3D)	VAE	Scene (image), Past motion	[154], [155]
Wang *et al.* [156]	CVPR 2021	Rot.	VAE	Scene (mesh), Start pose, End pose, Sub-goal	[13], [155]
Wang *et al.* [157]	CVPR 2021	Kpts. (3D)	GAN	Scene (image), Start pose	[154], [155]
HuMoR [158]	ICCV 2021	Rot.	VAE	Scene (ground), Past motion	[40], [155], [159]
SAMP [160]	ICCV 2021	Rot.	VAE	Scene (interactive object), Action class	[160]
Wang *et al.* [161]	CVPR 2022	Rot.	VAE	Scene (mesh), Action class	[13], [155]
GAMMA [162]	CVPR 2022	Kpts. (3D)	RL	Scene (goal)	[40]
GOAL [163]	CVPR 2022	Rot.	VAE	Scene (object), Start pose	[97]
SAGA [164]	ECCV 2022	Kpts. (3D)	VAE	Scene (object), Start pose	[97]
Couch [66]	ECCV 2022	Rot.	Regression	Scene (chair, contact), Start pose	[66]
Mao *et al.* [165]	NeurIPS 2022	Kpts. (3D)	Regression	Scene (point cloud), Past motion	[154], [155]
HUMANISE [16]	NeurIPS 2022	Rot.	VAE	Scene (point cloud), Text	[16]
IMoS [166]	EUROGRAPHICS 2023	Rot.	VAE	Scene (object), Text	[97]
SceneDiffuser [167]	CVPR 2023	Rot.	Diffusion	Scene (point cloud, goal)	[155]
CIRCLE [168]	CVPR 2023	Rot.	Regression	Scene (point cloud, goal), Start pose	[168]

"Kpts." and "Rot." denotes keypoints and 3D rotations, respectively. † denotes optional condition.

propose a VAE-based approach that utilizes a length estimation module and a word-level attention module at each frame to produce diverse multi-length motions. Additionally, TEMOS [108] learns the joint distribution of the motion and the text through a VAE with Transformer layers, enabling the generation of varied motion sequences. TEACH [110] further employs past motions as supplementary inputs to the encoder module, which enables the generation of more natural and coherent motion sequences, especially when dealing with several sequences of text inputs.

While the above methods pay attention to generating motion based on a given dataset, they may encounter inherent challenges when it comes to zero-shot generation. To address this challenge,

MotionCLIP [107] utilizes a Transformer-based autoencoder and aligns the motion latent space with the text and image space of a pre-trained vision-language model CLIP [170] to enhance the zero-shot generation ability. AvatarCLIP [106] also employs CLIP [170] and a reference-based motion synthesis method to generate diverse animations from natural language descriptions. Furthermore, OOHMG [113] uses a text-pose generator to obtain text-consistent poses, which are then fed as masked prompts into a pre-trained generator. This allows for efficient full-motion reconstruction, eliminating the need for paired data or online optimization. It is worth noting that while these methods utilize text as input, they only employ short text that primarily consists of the action class name.

In recent years, there has been a growing interest in VQ-VAE and Diffusion models inspired by their remarkable success in the field of text-to-image generation. For instance, TM2T [109] exploits VQ-VAE to train the text-to-motion and motion-to-text modules in tandem. Similarly, T2M-GPT [112] applies a GPT-like transformer architecture for motion sequence generation, combining VQ-VAE with an Exponential Moving Average (EMA) and code reset strategy. FLAME [111] proposes to concatenate the motion-length token, language pooler token, time-step token, and motion embeddings, which are then utilized by the diffusion model to generate variable-length and diverse motions. MDM [14] and MLD [101], already introduced in the action-to-motion section, also apply the aforementioned methods for text-to-motion generation. Several works further explore motion generation from various conditions. For instance, MoFusion [116] utilizes a diffusion model with 1D U-Net style Transformer module to generate human motion sequences from either natural language or audio input. Furthermore, Zhou et al. [114] introduce UDE, a framework that discretizes motion sequences into latent codes, maps conditions into a unified space, predicts the quantized codes using a GPT-style transformer, and generates motions via a diffusion model.

V. Audio-Conditioned Motion Generation

In addition to textual descriptions, human motion generation from audio signals has also been explored. Unlike text, audio signals typically do not provide explicit depictions of the corresponding human motions, resulting in a higher degree of freedom for the generative task. Meanwhile, the generated human motion should be harmonious with the audio in terms of both high-level semantics and low-level rhythms. In this section, we mainly discuss two subtasks of increasing attention: *music-to-dance* and *speech-to-gesture*. The audio conditions can be represented by raw audio waveform, spectrogram, and mel-frequency cepstrum coefficients (MFCC). To enhance controllability, some works incorporate additional conditions such as style code or textual transcripts. Please refer to Table I middle block for a summary of the methods.

A. Music to Dance

The music-to-dance generation task aims to generate corresponding dance moves given an input music sequence. One straightforward idea is to approach the problem using fully-supervised *regression models*, similar to sequence-to-sequence translation. For instance, Tang et al. [117] employ an LSTM autoencoder to extract acoustic features and translate them to motion features. AI Choreographer [115] utilizes a full-attention cross-modal Transformer (FACT) and predicts N future motion frames in an auto-regressive manner. GroupDancer [126] proposes an additional dancer collaboration stage to select active dancers to generate multi-person dance. GDancer [129] introduces global-local motion representations to ensure both local coherency and global consistency. The above methods adopt a fully supervised learning perspective to minimize the distance between the predicted and ground truth motions. Nevertheless, for a given music sequence, there exists a wide variety of plausible dancing motions. Simple reconstruction supervision does not adequately address this one-to-many mapping relationship.

From a generative perspective, GAN-based methods [118], [123] apply adversarial learning to regularize the distance between generated and real motion data manifolds. MNET [4] additionally incorporates a music style code for the generator and designs a multi-task discriminator to perform per-style classification. Transflower [120] utilizes normalizing flow to express the complex probability distributions over valid motions. Bailando [124] first quantizes 3D motions using a VQ-VAE codebook, then leverages an actor-critic Generative Pretrained Transformer (GPT) to compose coherent sequences from the learned latent codes. EDGE [15] builds upon the diffusion model and formulates the task as a motion denoising problem conditioned on music. Another class of approaches is based on the classical motion graph framework, which casts motion generation as solving for an optimal path in a pre-constructed graph. ChoreoMaster [86] proposes to learn a shared embedding space of music and dance, then integrate the learned embeddings and expert knowledge into the graph-based motion synthesis framework. PC-Dance [8] further achieves controllable motion generation by incorporating anchor poses as additional inputs. ChoreoGraph [125] utilizes motion segment warping to address rhythm alignment issues, reducing motion nodes in the graph and computation cost.

While most methods utilize short music-dance clips for training, an important user demand is to generate perpetual dance for an entire song. However, long-sequence generation tends to incur error accumulation issues that result in freezing motions. To overcome this challenge, Huang et al. [119] propose a curriculum learning approach that progressively transitions from a teacher-forcing scheme to an autoregressive scheme as training advances. Sun et al. [127] employ a VQ-VAE to learn a low-dimensional manifold, which effectively denoises the motion sequences. They also develop a past-future motion dynamics bank to provide explicit priors about future motions. Aristidou et al. [128] address the problem from three levels, including pose, motif, and choreography, to generate long dances that maintain a genre-specific global structure.

B. Speech to Gesture

The speech-to-gesture generation (or co-speech gesture synthesis) task aims to generate a sequence of human gestures based on input speech audio and, in some cases, text transcripts. Co-speech gestures play a crucial role in non-verbal communication,

conveying the speaker's information and emotions, fostering intimacy, and enhancing trustworthiness [171]. Existing research works for this task generally focus on upper-body motion, as lower-body movement tends to be static.

Some studies generate speech gestures from text transcripts [136], [172], [173]. A greater number of research works focus on speech audio conditions. For instance, Ginosar et al. [130] collect a speech video dataset of person-specific gestures and train a generative model with adversarial loss. Aud2Repr2Pose [11] first constructs a motion autoencoder and then trains a speech encoder to map the speech audio to motion representations. StyleGestures [133] adapts MoGlow [174] and further exerts directorial control over the styles of the generated motions. Recognizing that speech cannot fully determine gestures, Qian et al. [140] propose to learn a set of gesture template vectors to model the general appearance of generated gestures. Audio2Gestures [141] disentangles motion representation into audio-motion shared and motion-specific information to reflect the one-to-many mapping between audio and motion. Habibie et al. [137] apply an audio encoder and three separate decoders for face, body, and hand respectively. DisCo [143] first clusters the motion sequences into content and rhythm segments, then trains on the content-balanced data distribution. Habibie et al. [145] propose to first search for the most plausible motion from the database using the k-Nearest Neighbors (k-NN) algorithm, then refine the motion. DiffGesture [147] utilizes diffusion models with a cross-modal Transformer network and explores classifier-free guidance to balance diversity and gesture quality.

Nevertheless, co-speech gestures could have a significant inter-person variability due to individual personalities. The aforementioned methods do not explicitly consider speaker identities, necessitating separate models for each speaker and hindering transfer to general scenarios. Furthermore, these methods are limited to modeling either text or audio of speech and fail to combine both modalities. Motivated by these deficiencies, Yoon et al. [135] propose a generation framework that considers the trimodal context of text, audio, and speaker identity. Bhattacharya et al. [138] further enhance generation quality in terms of affective expressions with an affective encoder and an MFCC encoder. Mix-StAGE [132] learns unique style embeddings for each speaker and generates motions for multiple speakers simultaneously. HA2G [142] employs a hierarchical audio learner to extract audio representations and a hierarchical pose inferer to blend features between audio and body parts. Liu et al. [144] develop a Cascaded Motion Network (CaMN) that further considers facial expressions, emotions, and semantic meaning based on a large-scale dataset. Rhythmic Gesticulator [146] draws inspiration from linguistic theory and explicitly models both the rhythmic and semantic relations between speech and gestures. TalkSHOW [148] employs an autoencoder for face motions, and a compositional VQ-VAE for body and hand motions based on speech audio and speaker identity. QPGesture [149] introduces a quantization-based and phase-guided motion matching framework using VQ-VAE and Levenshtein distance. LDA [5] demonstrate style control using classifier-free guidance for diffusion models in both music-to-dance, speech-to-gesture, and path-driven locomotion. GestureDiffuCLIP [6] adapts a latent diffusion model for speech gesture generation and enables controlling with style prompt (text, motion, or video).

VI. Scene-Conditioned Motion Generation

Human motion is goal-oriented and influenced by the surrounding scene layout, with individuals moving their bodies to interact with the environment while being constrained by its physical properties. The scene-to-motion generation task aims to generate reasonable human motions consistent with the scene context and has been a long-standing problem in computer graphics and computer vision. This survey primarily focuses on data-driven scene-conditioned motion generation methods as discussed in Section II, and does not cover methods based on physical simulation [175], [176], [177], [178], [179]. Prior to human motion generation, some works have also proposed to synthesize static human poses given a scene condition [180], [181], [182], [183], [184], which will not be further discussed as they also fall outside of the scope of this survey. In the following, we discuss existing approaches from two perspectives: *scene representation* and *generation pipeline*. Please refer to Table I bottom block.

A. Scene Representation

Current methods exploit various options of scene representations, including 2D images [154], [157], point clouds [16], [165], [167], [168], mesh [156], [161], 3D objects [66], [152], [160], [163], [164], [166] and a specific goal position [156], [162], [167], [168]. Cao et al. [154] and Wang et al. [157] use RGB images as the scene constraints which are fused implicitly by extracting features from the image. Many works [16], [156], [161], [165], [167], [168] use point clouds or mesh to represent the scene, e.g. a room with furniture, and often extract scene features using PointNet [185] to serve as condition signal. For 3D objects, the configurations include 3D positions of the object [152], [163], object type [152], [166] and voxel representation of the object [66], [160]. For example, Corona et al. [152] represent the object using its 3D bounding box with its object type (e.g., cup) as a one-hot vector, and introduce a directed semantic graph to jointly parameterize the human motion and the object. They use Recurrent Neural Networks (RNN) to generate the human motion to interact with the object. COUCH [66] aims to generate controllable, contact-driven human-chair interactions and represents the chair using an occupancy voxel grid, which accurately captures the spatial relation between the person and the chair. Another typical example using 3D objects as scene conditions involves works that generate whole-body grasping motion [163], [164], [166], where 3D object positions [163], [166] or point cloud [164] are provided. Some works give a goal position [156], [162], [167], [168] to guide motion generation. For example, GAMMA [162] uses Reinforcement Learning to learn a policy network to synthesize plausible motions given the goal position on the ground. SceneDiffuser [167] proposes a generic framework for diverse 3D scene understanding tasks and uses diffusion models [22] to generate plausible human motions given the point cloud scene and the goal.

Note that most of the methods take more than one type of scene representation as input, and many of them take the past motion or the start pose [66], [152], [154], [156], [157], [158], [165], [168] as input together. There also emerge some methods that generate motion with extra language instructions [16] or action labels [160], [161]. For example, HUMANISE [16] incorporates language descriptions (e.g., walk to the table) to generate human motions in the scene. IMoS [166] integrates intended action instructions (e.g., drink) to generate a controllable whole-body grasping motion given the object positions and type.

B. Generation Pipeline

Most existing methods propose a multi-stage pipeline. One common pipeline is to first predict the goal position [154], [157] or goal interaction anchor [66], [160], [161], then plan a path or trajectory and finally infill the motion along the trajectory [66], [154], [157], [160], [161], [165], [167]. For example, Cao et al. [154] propose a three-stage motion generation method given the scene as 2D images, which first predicts a 2D goal, then plans a 2D and 3D path, and finally generates the 3D motion along the path via the VAE model. Similar to Cao et al. [154], Wang et al. [157] use an RGB image as the scene context, and synthesize human future motion by first generating the trajectory and then guiding the motion generation. They further add a discriminator branch to emphasize the consideration of the scene context. SAMP [160] also adopts a multi-stage pipeline which first estimates a goal position and the interaction direction of the object then plans a 3D path given the start body pose, and finally generates reasonable human motions with an autoregressive VAE. Compared to SAMP [160], which only models the coarse human-object interaction in the final frame, Mao et al. [165] propose to use per-joint contact maps to provide more detailed contact information for every human body joint at each future frame to promote the generation quality. Wang et al. [161] first predict diverse human-scene interaction anchors, then incorporate the standard A* algorithm with scene-aware random exploration for diverse path planning. Finally, a VAE-based framework is used to synthesize anchor poses and complete the motion. GOAL [163] and SAGA [164] aim to generate whole-body grasping motion and propose two-step approaches whereby the ending grasping pose is generated first, followed by the generation of whole-body motion. Different from GOAL [163], SAGA [164] further captures both the diversity of grasping ending poses and the in-between motions by using VAE models.

Some other works utilize given goal positions instead of predicting them. For example, Wang et al. [156] propose a hierarchical pipeline that uses the VAE model [19] to generate a static human body on each given sub-goal and generate the in-between human motions for each pair of sub-goal in the scene using bi-directional LSTM [186]. By stitching these motion clips, long-term human motion is synthesized. A recent paradigm proposed by CIRCLE [168] is to first initialize the motion using linear interpolation with a given start pose and a goal position, and then propose a scene-aware motion refinement module to generate the final motion. The scene feature is extracted from the 3D point cloud and fused into the refinement module.

VII. Datasets

In this section, we discuss the datasets for human motion generation. Commonly used datasets can be categorized on the basis of their accompanying conditional signals. We introduce the datasets with paired human motion and conditional signals of text, audio, and scene, respectively. In Table II, we summarize the key properties of these datasets and also include the large-scale human motion datasets that do not have extra conditional signals for reference.

A. Text-Motion Datasets

KIT-Motion Language [100] is a paired dataset consisting of motion and language data. Motion data are collected via optical marker-based systems, while language data consists of annotations describing each motion datum.

UESTC [91] includes motion data captured in three modalities - RGB videos, depth and skeleton sequences - using a Microsoft Kinect V2 sensor. The dataset comprises 15 action categories for both standing and sitting position, and 25 categories for standing only, totaling 40 distinct categories.

NTU-RGB+D 120 [87] is an extension of the NTU-RGB+D [93] dataset, with 60 additional classes and 57600 additional RGB+D video samples. The dataset contains 120 different action categories, representing a mix of daily and health-related activities.

HumanAct12 [7], derived from the PHSPD [188], presents a specialized collection of 3D motion clips, segmented into a spectrum of actions typical of human behavior. The dataset includes daily motions such as walk, run, sit, and warm-up, and is categorized into 12 motion classes and 34 sub-classes.

BABEL: Bodies, Action and Behavior with English Labels [96], provides text labels for motion sequences from the comprehensive motion dataset AMASS [40]. The dataset provides labels on two unique levels: sequence-level for entire sequences and frame-level for individual frames. It covers over 28k sequences and 63k frames across 250 motion categories.

HumanML3D [3] is a dataset derived from the combination of the HumanAct12 [7] and AMASS [40] datasets, and it includes three distinct text descriptions corresponding to each motion sequence. The dataset covers a wide range of activities in daily life, sports, acrobatics, and arts.

B. Audio-Motion Datasets

The audio-motion datasets can be categorized into *controlled* and *in-the-wild* based on the data collection techniques discussed in Section III-A2. The *controlled* audio-motion pairs are obtained by motion capture systems (marker-based, markerless) or manual annotation. On the contrary, *in the wild* audio-motion pairs are typically obtained by searching and downloading online videos with specific keywords and utilizing an off-the-shelf pose estimator to extract human motion. Although *in-the-wild* data offer a higher motion diversity and are more scalable, the extracted motions tend to be less accurate.

TABLE II
DATASETS FOR HUMAN MOTION GENERATION

Name	Venue	Collection	Representation	Subjects	Sequences	Frames	Length	Condition	Remarks
Human3.6M [39]	TPAMI 2014	Marker-based	Kpts. (3D)	11	-	3.6M	5.0h	-	15 actions
CMU Mocap [65]	Online 2015	Marker-based	Rot.	109	2605	-	9h	-	6 categories, 23 subcategories
AMASS [40]	ICCV 2019	Marker-based	Rot.	344	11265	-	40.0h	-	Unifies 15 marker-based MoCap datasets
HuMMan [42]	ECCV 2022	Markerless	Rot.	1000	400K	60M	-	-	500 actions
KIT Motion Language [100]	Big data 2016	Marker-based	Kpts. (3D)	111	3911	-	10.3h	Text	6.3k Text descriptions
UESTC [91]	MM 2018	Markerless	Kpts. (3D)	118	25.6K	-	83h	Text	40 Action classes
NTU-RGB+D [87]	TPAMI 2019	Markerless	Kpts. (3D)	106	114.4K	-	74h	Text	120 Action classes
HumanAct12 [7]	MM 2020	Markerless	Kpts. (3D)	12	1191	90K	6h	Text	12 Action classes
BABEL [96]	CVPR 2021	Marker-based	Rot.	344	-	-	43.5h	Text	260 Action classes
HumanML3D [3]	CVPR 2022	Marker-based & Markerless	Kpts. (3D)	344	14.6K	-	28.5h	Text	44.9K Text descriptions
Tang *et al.* [117]	MM 2018	Marker-based	Kpts. (3D)	-	61	907K	1.6h	Music	4 genres
Lee *et al.* [118]	NeurIPS 2019	Pseudo-labeling	Kpts. (2D)	-	361K	-	71h	Music	3 genres
Huang *et al.* [119]	ICLR 2021	Pseudo-labeling	Kpts. (2D)	-	790	-	12h	Music	3 genres
AIST++ [115]	ICCV 2021	Markerless	Rot.	30	1,408	10.1M	5.2h	Music	10 genres
PMSD [120]	TOG 2021	Marker-based	Kpts. (3D)	8	-	-	3.8h	Music	4 genres
ShaderMotion [120]	TOG 2021	Marker-based	Kpts. (3D)	8	-	-	10.2h	Music	2 genres
Chen *et al.* [86]	TOG 2021	Manual annotation	Rot.	-	-	160K	9.9h	Music	9 genres
PhantomDance [123]	AAAI 2022	Manual annotation	Rot.	-	260	795K	3.7h	Music	13 genres
MMD-ARC [8]	MM 2022	Manual annotation	Rot.	-	213	-	11.3h	Music	-
MDC [126]	MM 2022	Manual annotation	Rot.	-	798	-	3.5h	Music	2 genres
Aristidou *et al.* [128]	TVCG 2022	Marker-based	Rot.	32	-	-	2.4h	Music	3 genres
AIOZ-GDANCE [129]	CVPR 2023	Pseudo-labeling	Rot.	>4000	-	-	16.7h	Music	7 dance styles, 16 music genres
Trinity [134]	IVA 2018	Pseudo-labeling	Kpts. (2D)	1	23	-	4.1h	Speech	Casual talks
TED-Gesture [136]	ICRA 2019	Pseudo-labeling	Kpts. (3D)	-	1,295	-	52.7h	Text	TED talks
Speech2Gesture [130]	CVPR 2019	Pseudo-labeling	Kpts. (2D)	10	-	-	144h	Speech	TV shows, Lectures
TED-Gesture++ [135]	TOG 2020	Pseudo-labeling	Kpts. (3D)	-	1,766	-	97.0h	Speech, Text	Extension of [136]
PATS [132]	ECCV 2020	Pseudo-labeling	Kpts. (2D)	25	-	-	251h	Speech, Text	Extension of [130]
Speech2Gesture-3D [137]	IVA 2021	Pseudo-labeling	Kpts. (3D)	6	-	-	33h	Speech	Videos from [130]
BEAT [144]	ECCV 2022	Marker-based	Rot.	30	2508	30M	76h	Speech, Text, Emotion	8 emotions, 4 languages
Chinese Gesture [146]	TOG 2022	Marker-based	Rot.	5	-	-	4h	Speech, Text	Chinese
ZEGGS [150]	CGF 2023	Marker-based	Rot.	1	67	-	2.3h	Speech, Style	19 Styles
SHOW [148]	CVPR 2023	Pseudo-labeling	Rot.	-	-	-	27h	Speech	Videos from [130]
WBHM [153]	ICAR 2015	Marker-based	Rot.	43	3704	691K	7.68h	Object	41 objects
PiGraph [182]	TOG 2016	Markerless	Kpts. (3D)	5	63	0.1M	2h	Scene, Object	30 scenes, 19 objects
PROX [155]	ICCV 2019	Markerless	Rot.	20	60	0.1M	1h	Scene, Object	12 indoor scenes
i3DB [159]	SIGGRAPH 2019	Pseudo-labeling	Kpts. (3D)	1	-	-	-	Scene, Object	15 scenes
GTA-IM [154]	ECCV 2020	Marker-based	Kpts. (3D)	50	119	1M	-	Scene	Synthetic, 10 indoor scenes
GRAB [97]	ECCV 2020	Marker-based	Rot.	10	1334	1.6M	-	Object	51 objects
HPS [187]	CVPR 2021	Marker-based	Rot.	7	-	300K	-	Scene	8 large scenes, some > 1000 m^2
SAMP [160]	ICCV 2021	Marker-based	Rot.	1	-	185K	0.83h	Scene, Object	7 objects
COUCH [66]	ECCV 2022	Marker-based	Rot.	6	>500	-	3h	Scene, Chairs	3 chairs, hand interaction on chairs
HUMANISE [16]	NeurIPS 2022	Marker-based	Rot.	-	19.6K	1.2M	-	Scene, Object, Text	643 scenes
CIRCLE [168]	CVPR 2023	Marker-based	Rot.	5	>7K	4.3M	10h	Scene	9 scenes

"Kpts." and "Rot." denotes keypoints and 3D rotations, respectively.

1) Controlled Datasets: Tang et al. [117] pioneers to capture 3D dance and corresponding music of 4 types (waltz, tango, cha-cha, and rumba).

AIST++ [115] is constructed from the AIST Dance Video DB [189]. They utilize multi-view videos to estimate the camera parameters, 3D keypoints, and SMPL parameters.

PATS: Pose-Audio-Transcript-Style [120] dataset consists of synchronized audio and recordings of various dancers and dance styles.

ShaderMotion [120] extract dance from a social VR platform where the avatars' motion are retargeted from the participants with a 6-point tracking system.

Aristidou et al. [128] invites a group of professional dancers for motion capture and features long sequences of music-dance pairs.

Trinity [134] is a multi-modal dataset of conversational speech, containing 4 hours of audio, motion, and video data from one actor. Precise 3D motion is obtained with marker-based motion capture (MoCap) systems.

BEAT: Body-Expression-Audio-Text dataset [144] is a large-scale semantic and emotional dataset for conversational gestures synthesis, which features rich frame-level emotion and semantic relevance annotations. It also includes facial expressions and multi-lingual speeches.

Chinese Gesture[146] is a Chinese speech gesture dataset that allows for the exploration of cross-language gesture generation.

In addition to MoCap-based solutions, several works also propose to extract the audio-motion pairs from character animation resources produced by animators. For example, Chen et al. [86] and *MMD-ARC* [8] utilize MikuMikuDance (MMD) resources from the anime community. *PhantomDance* [123] recruits a team of experienced animators instructed by professional dancers to create the dance motions. *MDC. Multi-Dancer Choreography* [126] dataset focuses on group dance and they invite the dancers to arrange motion phrases and annotate the temporal dancer activation sequences.

2) In-the-Wild Datasets: Lee et al. [118] collects dance videos from the Internet with keywords (ballet, Zumba, and hip-hop) and extract 2D body keypoints with OpenPose [72].

Huang et al. [119] addresses the lack of a long-term dance generation dataset. It features one-minute music-dance pairs from the Internet.

AIOZ-GDANCE [129] collects in-the-wild group dancing videos along with music and fits SMPL sequences to the tracked 2D keypoints using a temporal extension of SMPLify-X [67]. They manually fix the incorrect cases for 2D keypoints and 3D motion, and use human annotations for multi-person relative depth.

TED-Gesture [136] is a co-speech gesture of TED talks that contains videos and English transcripts (along with timestamps for phrases). The authors use OpenPose [72] to extract 2D poses, then design a neural network to convert 2D poses into 3D poses.

Speech2Gesture [130] is a speaker-specific gesture dataset. It is based on the unlabeled in-the-wild videos of television shows and university lectures. The pseudo ground truth is obtained with an off-the-shelf 2D pose estimation algorithm [72]. The dataset contains 10 speakers with diverse motion styles, including television show hosts, university lecturers, and televangelists, and therefore enables studying person-specific motion generation.

TED-Gesture++ [135] extends TED-Gesture [136] with more videos, featuring synchronized video, speech audio, and transcribed English speech text. The 3D body keypoints are obtained with a temporal 3D pose estimation method [37].

PATS: Pose-Audio-Transcript-Style [132] extends [130] to more speakers including 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists. Similarly, they extract the skeletal keypoints with OpenPose [72]. In addition, PATS provides the transcripts corresponding to motion and audio signals.

Speech2Gesture-3D [137] annotates the Speech2Gesture dataset [130] with state-of-the-art 3D face [190], [191], body [192] and hand [193] pose estimation algorithms. Some videos and subjects from [130] are excluded due to low resolution and poor 3D reconstruction results.

SHOW: Synchronous Holistic Optimization in the Wild [148] fits SMPL-X [67] parametric model with hand gestures and facial expressions on the Speech2Gesture dataset [130]. It improves SMPLify-X [67] with advanced regression-based approaches [194], [195], [196].

C. Scene-Motion Datasets

WBHM: Whole-Body Human Motion [153] contains 3D whole-body motion data of multiple individuals and objects collected by the Vicon motion capture system. The motion data considers not only the motions of the human subject but the positions and motions of objects with which the subject is interacting as well. 41 different objects with 3D models are included, such as stairs, cups, food, *etc*.

PiGraph: Prototypical interaction Graphs [182] scans real-world 3D scenes [197] and uses Kinect.v2 to capture people's skeletons when they interact with the environments. All objects in the 3D scenes are manually annotated with semantic labels. Multiple interactions are annotated as well.

PROX: Proximal Relationships with Object eXclusion [155] contains RGB-D videos of people interacting with real indoor environments, captured by the Kinect-One sensor. The poses of the objects are also captured using attached markers and each object has the CAD model.

i3DB [159] contains several human-scene interactions with annotated object locations and 3D human motion for each captured RGB video. Each object has a class label, such as a chair or table. The 3D human motion is obtained from the estimated 2D motion [198] with manual corrections.

GTA-IM: GTA Indoor Motion [154] is a large-scale synthetic dataset that captures human-scene interactions based on the Grand Theft Auto (GTA) gaming engine. The dataset is equipped with RGB-D videos, 3D human poses, scene instance labels, *etc*. Note that the motions in the GTA game engine are from a marker-based motion capture system.

GRAB: GRasping Actions with Bodies [97] is a large-scale dataset capturing real-world whole-body grasps of 3D objects with Vicon motion capture system. Full-body human motion, object motion, in-hand manipulation, and contact areas are included in the annotation.

HPS: Human POSEitioning System [187] captures 3D humans interacting with large 3D scenes (300-1000 m^2, up to 2500 m^2), with images captured from a head-mounted camera coupled with the 3D pose and location of the subject in a pre-scanned 3D scene.

SAMP: Scene-Aware Motion Prediction [160] is a rich and diverse human-scene interaction dataset, collected by the high-resolution optical marker MoCap system with 54 Vicon cameras. Several types of objects such as sofas and armchairs are used during motion capture.

COUCH [66] captures humans interacting with chairs in different styles of sitting and free movement. The dataset is collected with IMUs and Kinects and features multiple subjects, real chair geometry, accurately annotated hand contacts, and RGB-D images.

HUMANISE [16] is a large-scale and semantic-rich synthetic dataset by aligning the captured human motion sequences in AMASS dataset [199] with the scanned indoor scenes in ScanNet dataset [200]. Each motion segment has rich semantics about the action type and the corresponding interacting objects, specified by the language description.

CIRCLE [168] collects 10 hours of both right and left-hand reaching motion across 9 scenes, captured both in the real world (with the Vicon system) and the VR app. Diverse motions are included such as crawling, bending, *etc*.

The above datasets contain not only the scene but also the human motion. Meanwhile, there are also some datasets that only contain scenes and are often used as test sets, such as *Matterport3D* [13], *Replica* [201], and *ScanNet* [200].

VIII. Evaluation Metrics

Proper evaluation metrics are vital to compare different methods and drive the progress of the field. However, evaluation of synthesized human motion is a non-trivial problem due to the one-to-many mapping nature, human evaluation subjectivity, and high-level cues of the conditional signals. In this section, we summarize the commonly-used evaluation metrics from different aspects, and discuss their strengths and limitations. See Table III for a summary.

A. Fidelity

The fidelity metrics aim to evaluate the general quality of the generated motions in terms of naturalness, smoothness, plausibility, *etc*.

1) Comparison With Ground-Truth: In assessing the quality of generated motion, comparing it to the ground truth serves as the most straightforward approach.

Distance: Most works [11], [16], [104], [105], [108], [110], [111], [114], [117], [130], [135], [137], [138], [140], [141], [146], [152], [154], [156], [158], [160], [163], [164], [165], [166], [167], [168] employ the distance metrics to measure the

TABLE III
EVALUATION METRICS FOR HUMAN MOTION GENERATION

Fidelity	Comparison with Ground-truth	**Distance**: [104], [105], [108], [110], [111], [114], [11], [117], [123], [130], [135], [137], [138], [140], [141], [146], [148], [16], [152], [154], [156], [158], [160], [163], [164], [165], [166], [167], [168] **Accuracy**: [130], [141]
	Naturalness	**Motion space**: [105], [108], [110], [111], [114], [149], [158], [160], [161], [164] **Feature space**: [3], [7], [14], [88], [90], [92], [94], [98], [99], [101], [109], [111], [112], [114], [116], [4], [6], [8], [15], [86], [115], [118], [119], [120], [123], [124], [125], [127], [128], [132], [135], [138], [140], [142], [143], [144], [146], [147], [148], [149], [157], [166]
	Physical Plausibility	**Foot sliding**: [163], [164], [168] **Foot-ground contact**: [15], [158], [162], [164]
Diversity	Intra-motion	**Variation**: [141], [143], [145] **Freezing rate**: [127]
	Inter-motion	**Coverage**: [3], [7], [14], [90], [94], [98], [99], [101], [109], [112], [114], [4], [8], [15], [86], [115], [118], [119], [123], [124], [127], [142], [147], [66] **Multi-modality**: [3], [7], [14], [90], [94], [98], [99], [101], [109], [112], [114], [118], [119], [141] [16], [160], [161], [164], [166], [167]
Condition Consistency	Text-Motion	**Accuracy**: [3], [7], [14], [90], [94], [98], [99], [101], [107], [109], [112], [113], [114], [116] **Distance**: [3], [14], [101], [109], [111], [112], [113]
	Audio-Motion	**Beat**: [4], [15], [115], [118], [119], [123], [124], [125], [127], [129], [142], [143], [144], [146], [147], [149] **Semantics**: [6], [144]
	Scene-Motion	**Non-collison score**: [156], [157], [160], [161], [163], [164], [167], [168] **Human-scene contact**: [66], [156], [161], [163], [164], [167]
User Study	Subjective Evaluation	**Preference**: [3], [7], [14], [104], [108], [109], [4], [5], [6], [15], [115], [118], [119], [123], [124], [127], [128], [132], [135], [140], [143], [144], [148], [166] **Rating**: [106], [8], [86], [117], [120], [125], [126], [128], [130], [131], [137], [138], [141], [142], [145], [146], [147], [149], [16], [156], [157], [161], [162], [163], [164]

Text-, audio-, and scene-conditioned motion generation works, respectively.

difference between synthesized motion and ground truth motion. Li et al. [123] utilize Normalized Power Spectrum Similarity (NPSS) [202] for evaluating long-term motion synthesis capabilities. NPSS operates in the frequency domain and is less sensitive to frame misalignment compared to MSE. Meanwhile, Normalized Directional Motion Similarity (NDMS) [203] is proposed to measure the similarity of the motion direction and the ratio of motion magnitudes in the motion prediction field.

Accuracy: As direct distance computation alone might not provide a thorough evaluation, some works [130], [141] further compute the Percentage of Correct 3D Keypoints (PCK) [204] which has been a popular evaluation metric of pose estimation. To compute PCK, the proportion of accurately generated joints is determined, with a joint deemed accurate if its distance to the target remains within a predefined threshold.

However, the ground truth represents only one feasible outcome for the given conditional input, with countless alternative solutions being potentially adequate. Consequently, relying solely on ground truth comparisons for motion generation evaluation may lack comprehensive coverage.

2) Naturalness: Motion quality evaluates the naturalness of the generated motion, which is usually measured by comparing the generated motion manifold with the real motion manifold. Existing metrics can be categorized into *motion space* and *feature space*, based on the space used for evaluation.

Motion Space: Some approaches measure the distribution distance based on geometric statistics in the motion space. For example, some works [105], [108], [110], [111], [114] report Average Variance Error (AVE) which compute the difference between the variance of the real motion and the synthesized motion. QPGesture [149] measures the Hellinger distance [205] between the speed-distribution histograms. Some works [149], [158] also compares the higher-order derivatives of joint positions (acceleration, jerk). SAMP [160] and Wang et al. [161] calculate the Fréchet distance (FD) of the two distributions based on the pose rotations. In motion prediction literature, Power Spectrum Entropy (PSEnt) and KL divergence (PSKL) [206] are used for computing the distribution distance. SAGA [164] utilizes PSKL-J [206], [207] to measure the acceleration distribution of generated and real motion to evaluate motion smoothness.

Feature Space: The second category is to compute the distribution distance in the feature space using a standalone neural network as motion feature extractor. To this end, some works compute Fréchet Inception Distance (FID) using an auxiliary action classifier [3], [14], [90], [94], [99], [101], [109], [111], [112], [116], [118], [119], [166] or an autoencoder [5], [6], [8], [86], [135], [138], [140], [142], [143], [144], [146], [147], [149], [157]. The metric can be extended by disentangling the motion feature into two aspects of geometric (pose) and kinetic (movement) [5], [15], [115], [120], [123], [124], [125], [127]. These works take advantage of the well-designed motion feature extractors [3], [7], [208], [209], [210] to calculate the feature distance. Kim et al. [4] further train a dance genre classifier to extract the style features and calculate corresponding FID. Except for FID, several other metrics are employed to compute the distribution distance between generation and real, including Inception Score (IS) [132], [211], chi-square distance [128],

Maximum Mean Discrepancy (MMD) [88], [92], Mean Maximum Similarity(MMS) [98], Canonical correlation analysis (CCA) [149], [212], and realistic score [148].

Although these metrics are intuitive, there exist several critical challenges. Their evaluation of naturalness highly depends on dataset distribution and the effectiveness of the pretrained motion feature extractor, which may not be comprehensive to reflect overall motion quality. For instance, EDGE [15] illustrates that the prevailing FID score is inconsistent with human evaluations, questioning the effectiveness of the common practice.

3) Physical Plausibility: Physical plausibility refers to the degree to which a generated motion is in accordance with the physical rules, particularly relevant to foot-ground interactions: (1) foot sliding, and (2) foot-ground contact.

Foot Sliding: Some work [163], [164], [168] measure the foot skating artifacts of the generated motion. For example, SAGA [164] defines skating as when the heel is within a threshold of the ground and the heel speed of both feet exceeds a threshold. CIRCLE [168] reports the percentage of frames in a sequence with foot sliding.

Foot-Ground Contact: Previous work has proposed several different metrics. For example, EDGE [15] reports the physical foot contact score (PFC). SAGA [164] reports a non-collision score which is defined as the number of body mesh vertices above the ground divided by the total number of vertices. HuMoR [158] reports the binary classification accuracy of person-ground contacts and the frequency of foot-floor penetrations of the generated motion. GAMMA [162] computes the contact score by setting a threshold height from the ground plane and a speed threshold for skating. However, at present, there is a lack of a standardized metric for quantifying physical plausibility. Various methods may employ disparate parameter choices and even design distinct evaluation approaches. Consequently, there is a potential requirement for the development of a more robust and universally applicable metric that effectively measures the degree of physical plausibility.

B. Diversity

Another important goal is to generate various human motions and avoid repetitive contents. To this end, researchers measure the generation results from different levels: diversity within single motion sequence (intra-motion diversity) and diversity among different motion sequences (inter-motion) diversity.

1) Intra-Motion Diversity: Long-sequence motion generation tends to incur the "freezing" problem [119], [127]. To evaluate "non-freezability" and discriminate the static motions, some works measure the intra-motion diversity metrics.

Variation: For example, some studies [141], [143] split the generated motions into equal-lengthed non-overlapping motion clips and calculate their average pairwise distance. Habibie et al. [145] measure the temporal position and velocity variations.

Freezing Rate: Sun et al. [127] propose to calculate the temporal differences of the pose and translation parameters and report a freezing rate.

2) Inter-Motion Diversity: To evaluate the inter-motion diversity of the generated motion manifold, the existing metrics can be categorized into *coverage* and *multi-modality*.

Coverage: The coverage of generated motion manifold is usually evaluated by first sampling N different conditional signals on the validation set, then computing the diversity of the generated motions. For example, [3], [7], [8], [14], [86], [90], [94], [98], [99], [101], [109], [112], [114], [118], [119], [129], [142], [147], [149] reports the average feature distance of the model results. Similar to FID, the feature distance can be divided into geometric, kinetic [15], [115], [124], [127], and style [4]. Some works [66], [123], [149] also calculate diversity in the motion space.

Multi-Modality: Given the same conditional signal, the probabilistic generative methods could generate a distribution over the plausible corresponding motions. The multi-modality metrics aim to evaluate the variations of the distribution. The common practice is to first sample N different conditional signals on the validation set, then generate M motions for each condition, and calculate the average pairwise distance for each condition. Existing methods report the average feature distance [3], [7], [14], [90], [94], [98], [99], [101], [109], [112], [114], [118], [119], [148], [160], [164], [166] or average pose distance [16], [141], [160], [161], [167]. ODMO [94] also uses normalized APD (n-APD) [213], which is determined by the ratio of APD values between generated motions and ground truth. Yuan et al. [213] also utilize average displacement error (ADE), final displacement error (FDE), multi-modal ADE (MMADE), and multi-modal FDE (MMFDE) based on the multi-modal nature of the problem in motion prediction. Some work further evaluates the generation diversity at levels of interaction anchors or planned paths [161].

C. Condition Consistency

The above metrics all focus on the properties of the generated motion itself, while it is also crucial to evaluate the consistency between the generated motion and the corresponding conditional signals. As these evaluation metrics highly correlate with the condition types, we will discuss them according to different tasks.

1) Text-Motion Consistency. Accuracy: In assessing the consistency between the generated motions and the corresponding texts in action-to-motion tasks, various existing approaches leverage recognition accuracy [7], [90], [94], [98], [99], [107] to evaluate the generation results. This metric is based on a pretrained action recognition model and determines whether the generated motions can be correctly identified as their corresponding action classes. The use of recognition accuracy provides a high-level view of how well the generated samples fit within the expected action class, given the textual description. In addition, some methods [3], [14], [101], [109], [112], [113], [114], [116] use R-Precision to assess the correspondence between the generated motions and their associated descriptions. This metric calculates and ranks the Euclidean distances among the features and averages the accuracy of the top-k results, offering a granular measure of text-motion consistency.

Distance: On the other hand, some methods delve deeper into the feature-level distance to measure the text-motion consistency. For instance, Multimodal Distance [3], [14], [101], [109], [112] quantifies the disparity between the feature from

a given description and the motion feature from the generated result, providing a direct measure of the feature-level alignment between the text and the motion. Similarly, Motion CLIP Score (mCLIP) [111], [113] utilizes cosine similarity to capture the closeness between text features and motion features in CLIP space, providing a quantifiable measure of how well the modalities align. Flame [111] further leverage Mutual Information Divergence (MID) [214] to measure the alignment between different modalities.

Nonetheless, these metrics are significantly influenced by the performance of the pretrained models, as well as the quality and distribution of data used for their training. Consequently, these metrics may have limitations in their ability to offer an objective evaluation.

2) Audio-Motion Consistency: Beat: Existing methods typically assess the degree to which kinematic beats of generated motions align with the input audio beats. To achieve this, beat coverage and hit rate [118], [119], [146] represents the ratio of aligned beats to all beats. Li et al. [115] proposes a beat alignment score calculated with beat distances and is followed by [4], [127], [129], [143], [144]. Some later works [5], [15], [123], [124], [125], [149] further refine the score definition by emphasizing music beat matching. In addition, studies [142], [147] suggest using mean angle velocity instead of position velocity.

Semantics: To further evaluate the semantic consistency, Liu et al. [144] propose Semantic-Relevant Gesture Recall (SRGR), which weighs PCK based on semantic scores of the ground truth data. They suppose that it is more in line with subjective human perception than the L1 variance. GestureDiffuCLIP [6] proposes semantic score (SC) to measure the semantic similarity between generated motion and transcripts in their joint embedding space.

At present, most evaluation metrics primarily focus on the basic connection between audio and motion, often neglecting subtler and cultural connections like style and emotion. For instance, hip-hop music and ballet dance may not be considered harmonious by human standards even with well-aligned beats. The same applies to a speech in a sad tone accompanied by cheerful gestures. Unfortunately, these nuances have not been fully addressed by existing audio-motion consistency metrics.

3) Scene-Motion Consistency: We distinguish physical plausibility (Section VIII-A3) and scene-motion consistency by dividing the scene into the ground and other objects. The scene-motion consistency refers to the agreement of the generated motion with the given scene condition (except for ground). There are mainly two perspectives to evaluate the consistency: (1) non-collision score, and (2) human-scene contact.

Non-collision score is a metric used to evaluate the safety and physical plausibility of a generated motion colliding with other objects or obstacles in the environment [156], [157], [160], [161], [163], [164], [167], [168]. For example, Wang et al. [157] compute human-scene collision as the intersection points between a human motion represented as a cylinder model and the point cloud of the given scenes. The non-collision ratio is defined as the ratio between the number of human motions without human-scene collision and all sampled motions. Some work [163], [164] uses body-scene penetration for this metric. For example, SAGA [164] measures the interpenetration volumes between the body and object mesh, and GOAL [163] reports the volume of the penetrations (cm^3).

Human-scene contact focuses on the contact areas to evaluate the scene-motion consistency [66], [156], [161], [163], [164], [167], and has different definitions considering different scene conditions. SAGA [164] measures the ratio of body meshes being in minimal contact with object meshes to evaluate the grasp stability. COUCH [66] focuses on how well the synthesized motion meets the given contacts by using Average Contact Error (ACE) as the mean squared error between the predicted hand contact and the corresponding given contact, as well as Average Contact Precision (AP@k), where a contact is considered as correctly predicted if it is closer than k cm.

There are some other metrics that aim to evaluate how well the generated motion reaches the final goal state, such as the execution time [160], the success rate of the character reaching the goal within several attempts [162], the body-to-goal distance [16], [162], [168]. The execution time [160] is the time required to transition to the target action label from an idle state. HUMANISE [16] and CIRCLE [168] assess the body-to-goal distance to evaluate how well the generated motion interacts with or reaches the correct object. In summary, various novel metrics have been proposed to measure scene-motion consistency, reflecting the diverse and intricate nature of scene representations. However, these metrics are often specifically crafted to address unique aspects within their respective research contexts, potentially restricting their broad generality and universal applicability.

D. User Study

User study, or subjective evaluation, serves as an essential component in evaluating generated motions, as it can uncover aspects of motion quality that may not be captured by objective metrics alone. First, humans are highly sensitive to minor artifacts in biological motion, such as jittering and foot skating [44], [45]. Second, current objective metrics are unable to encompass nuanced cultural aspects of generated motions, e.g., aesthetics and emotional impact. Existing methods design user studies that focus on one or several of the aforementioned aspects (quality, diversity, consistency) using *preference* or *rating*.

Preference: Many studies employ user study with pairwise preference comparisons between their generation results and baselines or GT. Specifically, participants observe a pair of human motions and respond to questions such as, "Which motion corresponds better to the textual description?", "Which dance is more realistic, regardless of music?", "Which dance matches the music better in terms of style?", or "Which motion best satisfies scene constraints?", *etc*. Subsequently, researchers calculate a win rate for their method against the baselines. The preference-based user studies offer a direct evaluation between the compared methods; however, they may be insufficient for comparing multiple methods. To address this, EDGE [15] performs pairwise comparisons among all generation methods and uses the Elo Rating [215] to represent their generation quality simultaneously.

Rating: Another prevalent user study approach involves instructing volunteers to provide explicit scores for the generation results. Participants are typically shown multiple motion generations and asked to assign a score (e.g., from 1 to 5) for each motion. Some studies further require a separate score for each aspect (quality, diversity, consistency).

IX. Conclusion and Future Work

In this survey, we provide a comprehensive overview of recent advancements in human motion generation. We begin by examining the fundamental aspects of this problem, specifically focusing on human motion and generation methods. Subsequently, we classify research studies based on their conditional signals and discuss each category in detail. Furthermore, we provide a summary of available dataset resources and commonly used evaluation metrics. Despite the rapid progress in this field, significant challenges remain that warrant future exploration. In light of this, we outline several promising future directions from various perspectives, hoping to inspire new breakthroughs in human motion generation research.

Data: Different from images or video, collecting high-quality human motion data is much more difficult and expensive, which leads to a trade-off in data quantity and data quality. Furthermore, the variability in motion representations and conditional signals hinders the broad applicability of existing datasets. To address these issues, future research could investigate the use of heterogeneous data sources, integrating their benefits through weakly-supervised learning approaches [216], [217] or multi-modal foundation models [107], [218].

Semantics: It is worth noting that human motion is more than just the movement of body parts; it also serves as a crucial non-verbal communication tool, conveying semantic information within cultural and societal contexts. Capturing the semantic relationship between human motion and conditional signals (e.g. high-level text descriptions, music/speech styles, and environment affordances) is essential for visually appealing and aesthetically pleasing results that align with human perception. A specific challenge in this field is how to equip generative models with prior knowledge of human motion semantics. Some studies [6], [107] adopt the pretrained foundation models with language priors. We believe that future research could delve deeper into exploring semantic connections from various perspectives, encompassing data, methodology, and evaluation.

Evaluation: As discussed in Section VIII, appropriate evaluation metrics for human motion are crucial yet challenging. Although various objective evaluation metrics have been explored, they all possess inherent limitations and cannot supplant subjective user studies [15]. Future work could focus on devising more principled objective evaluation metrics that not only align closely with human perception but also maintain interpretability.

Controllability: The ability to control the generation content is important in real-world applications, which has been a popular topic in image generative models [219], [220], [221]. Some recent works explore controllable human motion generation with joint mask [15] or style prompt [6]. We believe that future works could further explore controllability to create more user-friendly experiences, e.g., interactive and fine-grained editing [222].

Interactivity: The interactive nature of human motion is important but has not been fully explored yet. Most current studies focus primarily on generating single-human motion within static environments. Future works could delve into human motion generation in the context of human-human and human-environment interactions. Examples of potential areas of exploration include motion generation for closely interacting social groups (e.g., conversation, group dance, *etc*.) and motion generation in dynamic, actionable scenes [223], [224], [225].

References

[1] B. Hommel, "Toward an action-concept model of stimulus-response compatibility," *Adv. Psychol.*, vol. 118, pp. 281–320, 1997.

[2] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Rev. Neurosci.*, vol. 2, no. 8, pp. 561–567, 2001.

[3] C. Guo et al., "Generating diverse and natural 3D human motions from text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5152–5161.

[4] J. Kim, H. Oh, S. Kim, H. Tong, and S. Lee, "A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3490–3500.

[5] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–20, 2023.

[6] T. Ao, Z. Zhang, and L. Liu, "GestureDiffuCLIP: Gesture diffusion model with CLIP latents," *ACM Trans. Graph.*, vol. 42, 2023, Art. no. 42.

[7] C. Guo et al., "Action2Motion: Conditioned generation of 3D human motions," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2021–2029.

[8] J. Gao, J. Pu, H. Zhang, Y. Shan, and W.-S. Zheng, "PC-dance: Posture-controllable music-driven dance synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1261–1269.

[9] Y. Nishimura, Y. Nakamura, and H. Ishiguro, "Long-term motion generation for interactive humanoid robots using GAN with convolutional network," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2020, pp. 375–377.

[10] G. Gulletta, W. Erlhagen, and E. Bicho, "Human-like arm motion generation: A review," *Robotics*, vol. 9, no. 4, p. 102, 2020.

[11] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proc. Int. Conf. Intell. Virtual Agents*, 2019, pp. 97–104.

[12] T. Yin, L. Hoyet, M. Christie, M.-P. Cani, and J. Pettré, "The one-man-crowd: Single user generation of crowd motions using virtual reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 5, pp. 2245–2255, May 2022.

[13] A. Chang et al., "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 667–676.

[14] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *Proc. Int. Conf. Learn. Representations*, 2023.

[15] J. Tseng, R. Castellon, and K. Liu, "EDGE: Editable dance generation from music," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 448–458.

[16] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, "HUMANISE: Language-conditioned human motion generation in 3D scenes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 14959–14971.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[18] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 932–938.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[20] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1530–1538.

[21] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, Art. no. 574.
[23] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[24] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 27730–27744.
[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4396–4405.
[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8107–8116.
[27] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 852–863.
[28] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," 2022, *arXiv:2204.03458*.
[29] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, "StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3626–3636.
[30] S. Yu et al., "Generating videos with dynamics-aware implicit generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2022.
[31] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3D using 2D diffusion," in *Proc. Int. Conf. Learn. Representations*, 2023.
[32] J. Gao et al., "GET3D: A generative model of high quality 3D textured shapes learned from images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 31841–31854.
[33] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, 2015.
[34] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
[35] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, 2017, Art. no. 245.
[36] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2640–2649.
[37] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7753–7762.
[38] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5253–5263.
[39] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[40] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5441–5450.
[41] Z. Yu et al., "HUMBI: A large multiview dataset of human body expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2987–2997.
[42] Z. Cai et al., "HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 557–577.
[43] E. Grossman et al., "Brain areas involved in perception of biological motion," *J. Cogn. Neurosci.*, vol. 12, no. 5, pp. 711–720, 2000.
[44] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *J. Vis.*, vol. 2, no. 5, pp. 371–387, 2002.
[45] S. Shimada and K. Oki, "Modulation of motor area activity during observation of unnatural body movements," *Brain Cogn.*, vol. 80, no. 1, pp. 1–6, 2012.
[46] N. S. Sutil, *Motion and Representation: The Language of Human Movement*. Cambridge, MA, USA: MIT Press, 2015.
[47] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10873–10883.
[48] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3D human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022.
[49] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *Int. J. Robot. Res.*, vol. 39, no. 8, pp. 895–935, 2020.
[50] A. Haarbach, T. Birdal, and S. Ilic, "Survey of higher order rigid body motion interpolation methods for keyframe animation and continuous-time trajectory estimation," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 381–389.
[51] S. M. A. Akber, S. N. Kazmi, S. M. Mohsin, and A. Szczęsna, "Deep learning-based motion style transfer tools, techniques and future challenges," *Sensors*, vol. 23, no. 5, p. 2597, 2023.
[52] L. Mourot, L. Hoyet, F. Le Clerc, F. Schnitzler, and P. Hellier, "A survey on deep learning for skeleton-based human animation," *Comput. Graph. Forum*, vol. 41, no. 1, pp. 122–157, 2022.
[53] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, 2020, Art. no. 102897.
[54] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–41, 2022.
[55] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 231–268, 2001.
[56] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2/3, pp. 90–126, 2006.
[57] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7327–7347, Nov. 2022.
[58] M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models," *Adv. Robot.*, vol. 36, no. 5/6, pp. 261–278, 2022.
[59] Z. Shi, S. Peng, Y. Xu, Y. Liao, and Y. Shen, "Deep generative models on 3D representations: A survey," 2022, *arXiv:2210.15663*.
[60] Y. Zhang, M. J. Black, and S. Tang, "We are more than our joints: Predicting how 3D bodies move," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3372–3382.
[61] M. Zanfir et al., "THUNDR: Transformer-based 3D human reconstruction with markers," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 12971–12980.
[62] X. Ma, J. Su, C. Wang, W. Zhu, and Y. Wang, "3D human mesh estimation from virtual markers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 534–543.
[63] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "GHUM & GHUML: Generative 3D human shape and articulated pose models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6184–6193.
[64] A. A. A. Osman, T. Bolkart, and M. J. Black, "STAR: A sparse trained articulated human body regressor," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 598–613.
[65] J. Hodgins, "CMU graphics lab motion capture database," 2015. [Online]. Available: http://mocap.cs.cmu.edu/
[66] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll, "COUCH: Towards controllable human-chair interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 518–535.
[67] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
[68] M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *ACM Trans. Graph.*, vol. 33, no. 6, 2014, Art. no. 220.
[69] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7718–7727.
[70] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards multi-camera 3D human pose estimation in wild environment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 197–212.
[71] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, "Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 142–159.
[72] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.
[73] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2016.

[74] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018.

[75] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[76] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[77] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3483–3491.

[78] C. K. Sønderby, T. Raiko, L. Maaløe, S. R. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3738–3746.

[79] A. Van Den Oord et al., "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6309–6318.

[80] L. Weng, "What are diffusion models?," 2021. [Online]. Available: lilianweng.github.io

[81] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[82] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 483–490, 2002.

[83] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," in *Proc. 29th Annu. Conf. Comput. Graph. Interactive Techn.*, 2002, pp. 491–500.

[84] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 473–482, 2002.

[85] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Dancing-to-music character animation," *Comput. Graph. Forum*, vol. 25, no. 3, pp. 449–458, 2006.

[86] K. Chen et al., "ChoreoMaster: Choreography-oriented music-driven dance synthesis," *ACM Trans. Graph.*, vol. 40, no. 4, Jul. 2021, Art. no. 145.

[87] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[88] P. Yu, Y. Zhao, C. Li, J. Yuan, and C. Chen, "Structure-aware human-action generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–34.

[89] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[90] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3D human motion synthesis with transformer VAE," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10965–10975.

[91] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng, "A large-scale RGB-D database for arbitrary-view human action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1510–1518.

[92] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi, and H. Proença, "Generative adversarial graph convolutional networks for human action synthesis," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1150–1159.

[93] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.

[94] Q. Lu, Y. Zhang, M. Lu, and V. Roychowdhury, "Action-conditioned on-demand motion generation," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 2249–2257.

[95] T. Lucas *, F. Baradel *, P. Weinzaepfel, and G. Rogez, "PoseGPT: Quantization-based 3D human motion generation and forecasting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 417–435.

[96] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "BABEL: Bodies, action and behavior with English labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 722–731.

[97] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 581–600.

[98] P. Cervantes, Y. Sekikawa, I. Sato, and K. Shinoda, "Implicit neural representations for variable length human motion generation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 356–372.

[99] T. Lee, G. Moon, and K. M. Lee, "MultiAct: Long-term 3D human motion generation from multiple action labels," in *Proc. Assoc. Adv. Artif. Intell.*, 2023, pp. 1231–1239.

[100] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big Data*, vol. 4, no. 4, pp. 236–252, 2016.

[101] C. Xin et al., "Executing your commands via motion diffusion in latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18000–18010.

[102] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2Action: Generative adversarial synthesis from language to action," in *Proc. Int. Conf. Robot. Autom.*, 2018, pp. 5915–5920.

[103] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288–5296.

[104] C. Ahuja and L. Morency, "Language2Pose: Natural language grounded pose forecasting," in *Proc. Int. Conf. 3D Vis.*, 2019, pp. 719–728.

[105] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek, "Synthesis of compositional animations from textual descriptions," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 1396–1406.

[106] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–19, 2022.

[107] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "MotionCLIP: Exposing human motion generation to clip space," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 358–374.

[108] M. Petrovich, M. J. Black, and G. Varol, "TEMOS: Generating diverse human motions from textual descriptions," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 480–497.

[109] C. Guo, X. Zuo, S. Wang, and L. Cheng, "TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 580–597.

[110] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "TEACH: Temporal action compositions for 3D humans," in *Proc. Int. Conf. 3D Vis.*, 2022, pp. 414–423.

[111] J. Kim, J. Kim, and S. Choi, "FLAME: Free-form language-based motion synthesis & editing," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 8255–8263.

[112] J. Zhang et al., "T2M-GPT: Generating human motion from textual descriptions with discrete representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14730–14740.

[113] J. Lin et al., "Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23222–23231.

[114] Z. Zhou and B. Wang, "UDE: A unified driving engine for human motion generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5632–5641.

[115] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3D dance generation with AIST," 2021, *arXiv:2101.08779*.

[116] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, "MoFusion: A framework for denoising-diffusion-based motion synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9760–9770.

[117] T. Tang, J. Jia, and H. Mao, "Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1598–1606.

[118] H.-Y. Lee et al., "Dancing to music," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 322.

[119] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, "Dance revolution: Long-term dance generation with music via curriculum learning," in *Proc. Int. Conf. Learn. Representations*, 2021.

[120] G. Valle-Pérez, G. E. Henter, J. Beskow, A. Holzapfel, P.-Y. Oudeyer, and S. Alexanderson, "Transflower: Probabilistic autoregressive dance generation with multimodal attention," *ACM Trans. Graph.*, vol. 40, Dec. 2021, Art. no. 195.

[121] O. Alemi, J. Françoise, and P. Pasquier, "GrooveNet: Real-time music-driven dance movement generation using artificial neural networks," *Networks*, vol. 8, no. 17, pp. 26–31, 2017.

[122] A. Holzapfel, M. Hagleitner, and S. Pashalidou, "Diversity of traditional dance expression in crete: Data collection, research questions, and method development," in *Proc. 1st Symp. ICTM Study Group Sound Movement Sci.*, 2020, pp. 16–18.

[123] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "DanceFormer: Music conditioned 3D dance generation with parametric motion transformer," in *Proc. Assoc. Adv. Artif. Intell.*, 2022, pp. 1272–1279.

[124] L. Siyao et al., "Bailando: 3D dance generation via actor-critic GPT with choreographic memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11050–11059.
[125] H. Y. Au, J. Chen, J. Jiang, and Y. Guo, "ChoreoGraph: Music-conditioned automatic dance choreography over a style and tempo consistent dynamic graph," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3917–3925.
[126] Z. Wang et al., "GroupDancer: Music to multi-people dance synthesis with style collaboration," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1138–1146.
[127] J. Sun, C. Wang, H. Hu, H. Lai, Z. Jin, and J.-F. Hu, "You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 9995–10007.
[128] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 8, pp. 3519–3534, Aug. 2023.
[129] N. Le, T. Pham, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen, "Music-driven group choreography," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8673–8682.
[130] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3492–3501.
[131] K. Takeuchi, S. Kubota, K. Suzuki, D. Hasegawa, and H. Sakuta, "Creating a gesture-speech dataset for speech-based automatic gesture generation," in *Proc. Int. Conf. Hum.- Comput. Interact.*, 2017, pp. 198–202.
[132] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency, "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 248–265.
[133] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," *Comput. Graph. Forum*, vol. 39, no. 2, pp. 487–496, 2020.
[134] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, 2018, pp. 93–98.
[135] Y. Yoon et al., "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Trans. Graph.*, vol. 39, no. 6, 2020, Art. no. 222.
[136] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 4303–4309.
[137] I. Habibie et al., "Learning speech-driven 3D conversational gestures from video," in *Proc. 21st ACM Int. Conf. Intell. Virtual Agents*, 2021, pp. 101–108.
[138] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, "Speech2AffectiveGestures: Synthesizing co-speech gestures with generative adversarial affective expression learning," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 2027–2036.
[139] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA challenge 2020," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 11–21.
[140] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 11077–11086.
[141] J. Li et al., "Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 11293–11302.
[142] X. Liu et al., "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10462–10472.
[143] H. Liu et al., "DisCo: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3764–3773.
[144] H. Liu et al., "BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 612–630.
[145] I. Habibie et al., "A motion matching-based framework for controllable gesture synthesis from speech," in *Proc. ACM SIGGRAPH Conf.*, 2022, Art. no. 46.
[146] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings," *ACM Trans. Graph.*, vol. 41, no. 6, Nov. 2022, Art. no. 209.
[147] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10544–10553.
[148] H. Yi et al., "Generating holistic 3D human motion from speech," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 469–480.
[149] S. Yang et al., "QPGesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2321–2330.
[150] S. Ghorbani, Y. Ferstl, D. Holden, N. F. Troje, and M.-A. Carbonneau, "ZeroEGGS: Zero-shot example-based gesture generation from speech," *Comput. Graph. Forum*, vol. 42, no. 1, pp. 206–216, 2023.
[151] I. Mason, S. Starke, and T. Komura, "Real-time style modelling of human locomotion via feature-wise transformations and local motion phases," in *Proc. ACM Comput. Graph. Interactive Techn.*, vol. 5, no. 1, May 2022, Art. no. 6.
[152] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-aware human motion prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6992–7001.
[153] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The kit whole-body human motion database," in *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 329–336.
[154] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 387–404.
[155] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2282–2292.
[156] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang, "Synthesizing long-term 3D human motion and interaction in 3D scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9401–9411.
[157] J. Wang, S. Yan, B. Dai, and D. Lin, "Scene-aware generative network for human motion synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12206–12215.
[158] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "HuMoR: 3D human motion model for robust pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 11488–11499.
[159] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra, "iMapper: Interaction-guided scene mapping from monocular videos," in *Proc. ACM SIGGRAPH Conf.*, 2019, pp. 1–15.
[160] M. Hassan et al., "Stochastic scene-aware motion prediction," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 11374–11384.
[161] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai, "Towards diverse and natural scene-aware 3D human motion synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20460–20469.
[162] Y. Zhang and S. Tang, "The wanderings of odysseus in 3D scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20481–20491.
[163] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "GOAL: Generating 4D whole-body motion for hand-object grasping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13263–13273.
[164] Y. Wu et al., "SAGA: Stochastic whole-body grasping with contact," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 257–274.
[165] W. Mao, M. Liu, R. I. Hartley, and M. Salzmann, "Contact-aware human motion forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 7356–7367.
[166] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "IMoS: Intent-driven full-body motion synthesis for human-object interactions," in *Proc. Eurograph. Conf.*, 2023, pp. 1–12.
[167] S. Huang et al., "Diffusion-based generation, optimization, and planning in 3D scenes," 2023, *arXiv:2301.06015*.
[168] J. P. Araújo et al., "CIRCLE: Capture in rich contextual environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21211–21221.
[169] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[170] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[171] S. Van Mulken, E. Andre, and J. Müller, "The persona effect: How substantial is it?," in *Proc. People Comput. XIII*, 1998, pp. 53–66.
[172] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro, "A speech-driven hand gesture generation method and evaluation in Android robots," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3757–3764, Oct. 2018.

[173] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, "Text2Gestures: A transformer-based network for generating emotive body gestures for virtual agents," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2021, pp. 1–10.

[174] G. E. Henter, S. Alexanderson, and J. Beskow, "MoGlow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–14, 2020.

[175] S. I. Park, H. J. Shin, and S. Y. Shin, "On-line locomotion generation based on motion blending," in *Proc. ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, 2002, pp. 105–111.

[176] H. P. Shum, T. Komura, M. Shiraishi, and S. Yamazaki, "Interaction patches for multi-character animation," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–8, 2008.

[177] S. Agrawal and M. van de Panne, "Task-based locomotion," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[178] S. Starke, H. Zhang, T. Komura, and J. Saito, "Neural state machine for character-scene interactions," *ACM Trans. Graph.*, vol. 38, no. 6, 2019, Art. no. 209.

[179] D. Holden, T. Komura, and J. Saito, "Phase-functioned neural networks for character control," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[180] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert, "From 3D scene geometry to human workspace," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1961–1968.

[181] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3D scenes by learning human-scene interaction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14708–14718.

[182] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "PiGraphs: Learning interaction snapshots from observations," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, 2016.

[183] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang, "PLACE: Proximity learning of articulation and contact in 3D environments," in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 642–651.

[184] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. Tang, "Generating 3D people in scenes without people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6194–6204.

[185] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[186] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[187] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human poseitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4318–4329.

[188] S. Zou et al., "3D human shape reconstruction from a polarization image," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 351–368.

[189] S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto, "AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 501–510.

[190] P. Garrido et al., "Reconstruction of personalized 3D face rigs from monocular video," *ACM Trans. Graph.*, vol. 35, no. 3, pp. 1–15, 2016.

[191] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, pp. 200–215, 2011.

[192] D. Mehta et al., "XNect: Real-time multi-person 3D motion capture with a single RGB camera," *ACM Trans. Graph.*, vol. 39, no. 4, 2020, Art. no. 82.

[193] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu, "Monocular real-time hand shape and motion capture using multi-modal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5346–5355.

[194] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 792–804.

[195] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, 2021.

[196] H. Zhang et al., "PyMAF-X: Towards well-aligned full-body model regression from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12287–12303, Oct. 2023.

[197] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–11, 2013.

[198] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2500–2509.

[199] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5442–5451.

[200] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.

[201] J. Straub et al., "The replica dataset: A digital replica of indoor spaces," 2019, *arXiv:1906.05797*.

[202] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12116–12125.

[203] J. Tanke, C. Zaveri, and J. Gall, "Intention-based long-term human motion anticipation," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 596–605.

[204] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[205] T. Kucherenko et al., "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proc. Int. Conf. Multimodal Interaction*, 2020, pp. 242–250.

[206] A. Hernandez, J. Gall, and F. Moreno-Noguer, "Human motion prediction via spatio-temporal inpainting," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7134–7143.

[207] S. Zhang, Y. Zhang, F. Bogo, M. Pollefeys, and S. Tang, "Learning motion priors for 4D human body capture in 3D scenes," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 11343–11353.

[208] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 677–685, 2005.

[209] K. Onuma, C. Faloutsos, and J. K. Hodgins, "Fmdistance: A fast and effective distance function for motion capture data," in *Proc. Eurographics Conf.*, 2008, pp. 83–86.

[210] D. Gopinath and J. Won, "Fairmotion - Tools to load, process, and visualize motion capture data," 2020. [Online]. Available: https://github.com/facebookresearch/fairmotion

[211] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[212] N. Sadoughi and C. Busso, "Speech-driven animation with meaningful behaviors," *Speech Commun.*, vol. 110, pp. 90–100, 2019.

[213] Y. Yuan and K. Kitani, "DLow: Diversifying latent flows for diverse human motion prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–364.

[214] J.-H. Kim, Y. Kim, J. Lee, K. M. Yoo, and S.-W. Lee, "Mutual information divergence: A unified metric for multimodal generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 35072–35086.

[215] A. E. Elo, *The Rating of Chessplayers, Past and Present*. Tomar, Portugal: Arco Pub., 1978.

[216] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Learning human motion representations: A unified perspective," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 15085–15099.

[217] Y.-L. Li et al., "From isolated islands to pangea: Unifying semantic space for human action understanding," 2023, *arXiv:2304.00553*.

[218] L. Xue et al., "ULIP-2: Towards scalable multimodal pre-training for 3D understanding," 2023, *arXiv:2305.08275*.

[219] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 185.

[220] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "GAN-control: Explicitly controllable GANs," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 14083–14093.

[221] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6038–6047.

[222] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM SIGGRAPH Conf.*, 2023, Art. no. 78.

[223] M. Savva et al., "Habitat: A platform for embodied AI research," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347.

[224] A. Szot et al., "Habitat 2.0: Training home assistants to rearrange their habitat," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 251–266.

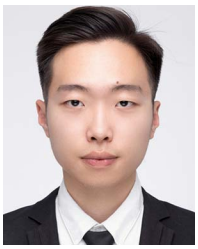
[225] C. Li et al., "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Proc. Conf. Robot Learn.*, PMLR, 2023, pp. 80–93.

Wentao Zhu (Graduate Student Member, IEEE) received the bachelor's degrees in computer science and economics from Peking University, in 2020. He is currently working toward the PhD degree in computer science with Peking University. His current research interests mainly include computer vision and machine learning. Specifically, he is interested in building human-centric AI systems that can perceive, understand, and interact with human beings.

Xiaoxuan Ma (Graduate Student Member, IEEE) received the bachelor's degree in computer science from Peking University, in 2018, and the master's degree in computer science from Peking University, in 2021. She is currently working toward the PhD degree in computer science with Peking University. Her current research interests include computer vision and machine learning.

Dongwoo Ro received the bachelor's degree in computer science and technology from the Harbin Institute of Technology, in 2020, and the master's degree in computer applied technology from Peking University, in 2023. His current research interests include computer vision and machine learning.

Hai Ci received the bachelor's degree in computer science from Nankai University, in 2017, and the doctoral degree in computer application technology from Peking University, in 2022. His current research interests include computer vision and machine learning.



Jinlu Zhang received the bachelor's degree in computer science from Shandong University, in 2020, and the master's degree in computer application technology from Wuhan University, in 2023. His current research interests include computer vision and machine learning.

Jiaxin Shi received the bachelor's and PhD degrees from the Department of Computer Science and Technology, Tsinghua University, in 2016 and 2021, respectively. He is a senior researcher with Huawei Cloud Computing Technologies Company, Ltd. He visited MReaL Lab of Nanyang Technological University from 2018 to 2019. His research interests include natural language processing, vision-language reasoning, and large-scale pre-training.

Feng Gao received the BS degree in computer science from University College London, in 2007, and the PhD degree in computer science from Peking University, in 2018. He was a post-doctoral research fellow with the Future Laboratory, Tsinghua University, from 2018 to 2020. He joins Peking University as Assistant Professor since 2020. His research interest is on the intersection of computer science and art, including but not limit on artificial intelligence and painting art, deep learning and painting robot, *etc*.

Qi Tian (Fellow, IEEE) received the BE degree in electronic engineering from Tsinghua University, in 1992, the MS degree in ECE from Drexel University, in 1996, and the PhD degree in ECE from the University of Illinois at Urbana-Champaign (UIUC), in 2002. He is currently the chief scientist of artificial intelligence with Huawei Cloud. He was a tenured associate professor from 2008–2012 and a tenure-track assistant professor from 2002–2008. During 2008–2009, he took one-year Faculty leave with Microsoft Research Asia (MSRA) as lead researcher with the Media Computing Group. His research interests include multimedia information retrieval, computer vision, pattern recognition and published more than 360 refereed journal and conference papers.

Yizhou Wang (Member, IEEE) received the bachelor's degree in electrical engineering from Tsinghua University, in 1996, and the PhD degree in computer science from the University of California at Los Angeles (UCLA), in 2005. He is currently an Endowed Boya professor and vice director of CFCS with Peking University. He joined Xerox Palo Alto Research Center (Xerox PARC) as a research staff from 2005 to 2007. He has published more than 200 papers and obtained a number of best paper awards. His research interests include computational vision, cognitive computing, medical image analysis, and computational arts.