Information Fusion 135 (2026) 104435

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal generative AI for human motion understanding and generation: A survey and way forward

Muhammad Islam [a,b], Tao Huang [a,b,*], Euijoon Ahn [a,b], Usman Naseem [c]

[a] College of Science and Engineering, James Cook University, Cairns, 4878, QLD, Australia
[b] Centre for AI and Data Science Innovation, James Cook University, Cairns, 4878, QLD, Australia
[c] School of Computing, Macquarie University, Sydney, 2113, NSW, Australia

ARTICLE INFO

Keywords:
Multimodal GenAI
Multimodal LLM
Multimodal diffusion model
Multimodal transformer
Text to human motion
Autoregressive model
Motion understanding
Motion generation

ABSTRACT

This paper presents an in-depth survey of the use of multimodal Generative Artificial Intelligence (GenAI) with autoregressive Large Language Models (LLMs) for human motion understanding and generation, offering insights into emerging methods and architectures and their potential to advance realistic and versatile motion synthesis. Focusing exclusively on text and motion modalities, this research investigates how textual descriptions can guide the generation of complex, human-like motion sequences. The paper explores various generative approaches, including multimodal autoregressive LLMs, multimodal diffusion, and multimodal transformers and their variants, and analyzes their strengths and limitations with respect to motion quality, computational efficiency, and adaptability. It highlights recent advances in text-conditioned motion generation, where textual inputs are used to control and refine motion outputs with greater precision. The use of LLMs further enhances these models by enabling semantic alignment between instructions and motion, improving coherence and contextual relevance. This systematic survey underscores the transformative potential of text-to-motion GenAI and LLM architectures in applications such as healthcare, humanoids, gaming, animation, and assistive technologies, while addressing ongoing challenges and research directions to guide future developments in human-centric GenAI.

1. Introduction

Human motion generation using multimodal generative AI particularly through the fusion of autoregressive Large Language Models (LLMs) has emerged as a promising research frontier, driving advancements toward the broader goal of Artificial General Intelligence (AGI) [1]. Recent examples, such as SORA [2], GPT-4V [3], and Dall-E [4], showcase advancements in generative modeling conditioned on signals to produce motion, video, and images. Human motion generation [5], in particular, has emerged as a crucial field, driven by these Generative Artificial Intelligence (GenAI) advancements and the increasing demand for realistic, human-like movements in applications like film production, video games, robotics, healthcare, autonomous vehicle perception, and virtual training environments. This survey focuses on the intersection of text-conditioned motion generation using LLMs, diffusion models, and transformer models, which represent significant breakthroughs in enabling machines to produce human motions from natural language descriptions. The rise of generative models in vision [6] such as Generative Adversarial Networks (GANs) [7], Variational Autoencoders (VAEs) [8], Normalizing Flows [9], Autoregressive (AR) [10] and Diffusion Models [11,12] has been central to improving the quality and diversity of generated motions. These models, while excelling in image generation and other domains, have begun to make strides in the challenging task of generating human motions that are plausible, natural, controllable, and contextually appropriate from text prompts [13,14].

This survey paper provides a comprehensive and insightful review of human motion understanding and generation, drawing on more than 200 research articles from prestigious and top-rated journals and conferences, including CVPR, ICCV, ECCV, SIGIR, NIPS, TIPS, TPAMI, and 3DV over recent years. We also critically reviewed the methodologies and backbone architectures, highlighting the technical challenges posed by human motion's inherent complexity and how recent innovations in LLMs like GPT-4 [15], Llama-3 [16], BERT [17], and T5 [18] are being leveraged to condition motion understanding based on text. By integrating diffusion models [5], GANs [19], and VAEs [20] variants into text-to-motion pipelines, we explore how AI can move closer to mimicking human-like movements by creating digital human actions and how to achieve generalized and robust motions with low computational cost. The work also tracks and highlights progress in motion understanding and generation based on condition signals, trends, and challenges, and

* Corresponding author.
E-mail addresses: muhammad.islam1@jcu.edu.au (M. Islam), tao.huang1@jcu.edu.au (T. Huang), euijoon.ahn@jcu.edu.au (E. Ahn), Usman.Naseem@jcu.edu.au (U. Naseem).

https://doi.org/10.1016/j.inffus.2026.104435
Received 30 June 2025; Received in revised form 24 April 2026; Accepted 27 April 2026
Available online 30 April 2026

examines how a unified model helps achieve the required goal. This work will serve as a valuable resource for researchers and practitioners interested in advancing the state-of-the-art in motion understanding and generation through multimodal AI approaches. The need for this survey is crucial to understand the human motion generation from low-level to current multimodal practices in GenAI using Multimodal LLM. Some recent survey papers are broader, covering video [21] and images [22], or are based on downstream subtopics of human motion, such as prediction [23]. Some surveys are intended to provide a general understanding of generative methodologies, such as GAN, VAE, and diffusion [8]. Similarly, one survey covers general perspectives and condition-based motion generation, excluding multimodal LLMs and transformers [24]. Recently, a survey on multimodal LLMs and multimodal generative AI includes text, audio, video, and images, but does not explicitly address human motion generation [25]. Our work is the first and only attempt to cover all areas of Human Motion Understanding and Generation (**HMUG**). Table 1 showcases a comparative analysis between our work and other surveys.

Survey methodology and scope

To ensure a comprehensive and transparent literature survey, we adopted a systematic methodology similar to that used in the related surveys summarized in Table 1, covering the search strategy, keyword formulation, paper selection, and taxonomy construction. This section details the process for identifying, filtering, and organizing the relevant literature on multimodal generative AI for human motion generation.

Search Strategy

We conducted a systematic literature search across multiple academic databases and research platforms, including *Google Scholar, IEEE Xplore, ACM Digital Library, arXiv,* and *SpringerLink*. These sources were selected to ensure broad coverage of peer-reviewed journal articles, conference proceedings, and high-quality preprints, which are particularly important in the rapidly evolving field of generative AI.

The search primarily focused on studies published between **2020 and 2025**. This time window corresponds to the emergence and maturation of transformer-based architectures, diffusion models, and large language models, which have fundamentally reshaped research in text-to-motion generation and multimodal human motion modeling. A small number of earlier foundational studies published before 2020 were also included where relevant.

Keyword Design and Query Formulation

Search queries were designed to reflect the thematic structure of this survey and were grouped into multiple categories corresponding to the major components of the human motion generation pipeline. The keywords covered five primary dimensions: *(i) human motion representation, (ii) text-to-motion generation, (iii) generative modeling paradigms, (iv) multimodal large language models, and (v) motion understanding, evaluation, and applications*.

Representative keywords include *"text-to-motion," "text-conditioned motion generation," "human motion representation," "SMPL," "SMPL-X," "motion diffusion models," "VAE-based motion synthesis," "GAN-based motion generation," "vector-quantized models,"* and *"unified human motion modeling."* To capture recent advances in multimodal reasoning, LLM-specific terms such as *"GPT," "LLaMA," "BERT," "T5,"* and *"vision-language models"* were also incorporated. In addition, task-level keywords including *"motion retrieval," "motion understanding," "motion captioning,"* and *"multimodal motion learning"* were used to identify works beyond pure generation. A detailed, section-wise list of related keywords used in the search process is summarized in Fig. 1.

Inclusion and Exclusion Criteria

Papers were included in this survey if they addressed human motion generation, understanding, or retrieval under textual conditioning and employed learning-based generative paradigms such as variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion models, or large language model-based architectures. Eligible studies were required to provide experimental validation on established motion datasets or benchmark protocols and to be published in top-tier journals or conferences, as described in the Introduction section and reflected throughout the tables in terms of venue, within the defined time window of 2020-2025.

Studies were excluded if they focused exclusively on low-level pose estimation, classical animation techniques without generative learning components, or non-human motion domains, or if they lacked sufficient technical detail to support meaningful comparison. To maintain a focused scope, works incorporating conditioning modalities beyond text such as audio, scene context, images, or videos, were also excluded. Similarly, motion generation tasks involving human-human, human-object, human-scene, or mixed interaction settings were considered out of scope, as these topics are comprehensively addressed in existing dedicated surveys [22,24,26].

Paper Screening and Selection Process

The initial search yielded a large collection of 471 papers. A multi-stage screening process was employed to ensure relevance and quality. First, duplicate records across databases were removed. Next, titles and abstracts were screened to eliminate clearly irrelevant works. The remaining 226 papers underwent full-text review to verify that their methods aligned with the survey scope and inclusion criteria. This structured screening and deduplication process ensures the auditability of the aggregated tables and guarantees that the conclusions drawn in this survey are traceable to clearly defined and verifiable sources.

Taxonomy Construction and Organization

Based on the selected literature, we constructed a unified taxonomy to systematically organize existing methods and research directions. The taxonomy categorizes prior work along multiple axes, including motion data representation, multimodal text encoding, generative modeling paradigms (autoregressive, VAE/GAN-based, diffusion-based), unified architectures, datasets, evaluation metrics, and application domains.

The overall organization of the survey, illustrated in Fig. 1, reflects this taxonomy and provides a structured view of the research landscape in multimodal human motion generation. This organization not only facilitates comparison across methods but also highlights emerging trends, open challenges, and opportunities for future research.

2. Preliminary of human motion data

We begin our discussion by introducing how human pose and motion data are represented and processed. This includes physical and mathematical formulations, as well as text-based conditioning and motion data collection techniques, which are essential for training generative models.

2.1. Pose data representation

In the context of 3D human motion, several popular representations are commonly employed, each offering unique advantages depending on the specific task [27]. These include point clouds, which model the human body as a set of individual points distributed in 3D space; voxels, which capture volumetric information by partitioning space into a grid of cubes; meshes, which represent surface geometry using interconnected vertices and edges; and depth maps, which encode distance information from a specific viewpoint. Neural fields and hybrid representations combine these traditional approaches with neural networks to model continuous 3D geometry and appearance, providing more flexible and detailed descriptions of human motion. These representations form the basis for developing accurate motion generation models, allowing for the simulation of complex human movements with high fidelity. Understanding these representations is crucial for analyzing human motion in various fields, from computer vision to biomechanics.

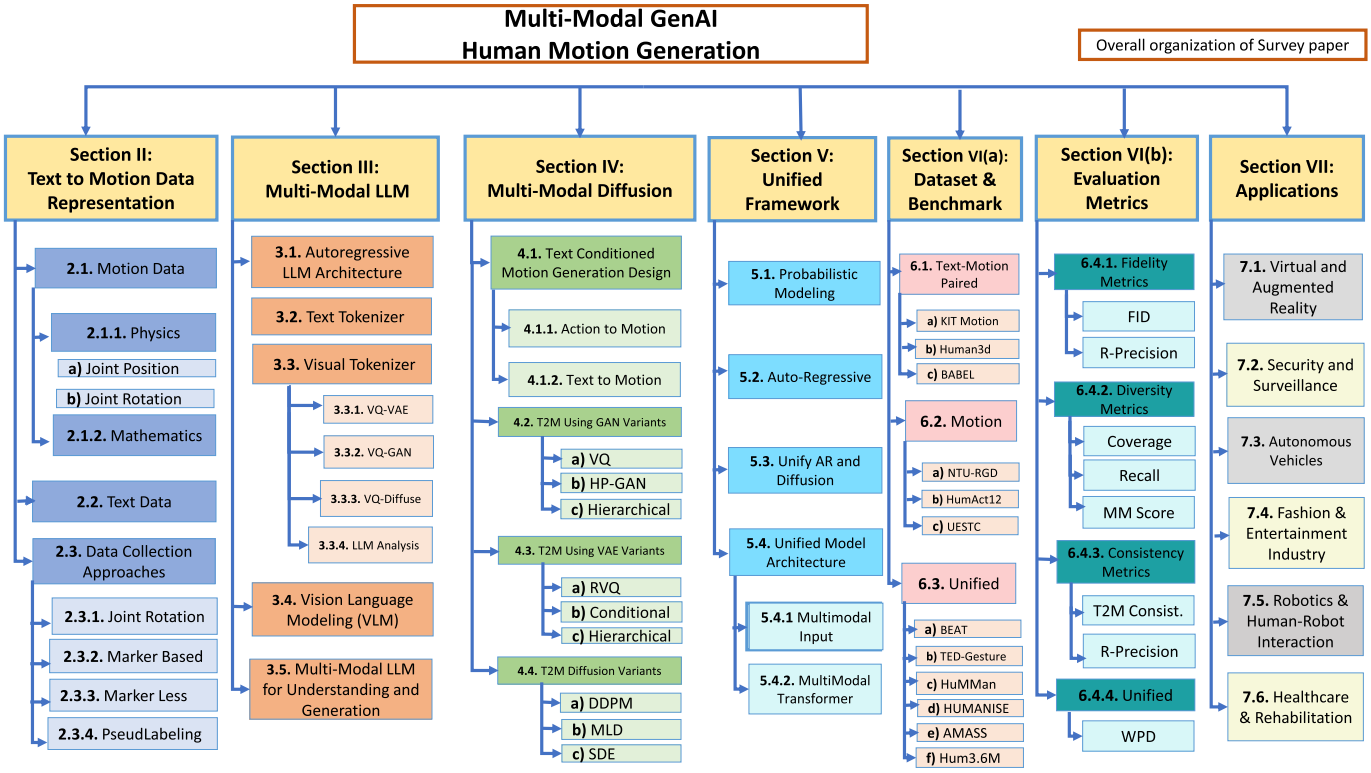


Fig. 1. Overall flow diagram illustrating variants of multimodal generative models, multimodal LLMs, and unified frameworks, along with their transformer backbone architectures.

2.1.1. Physical representation

This section presents the physical and intuitive perspective of human motion representation, focusing on observable aspects of joint positions and rotations and their interpretability in real-world tasks.

In motion analysis, physical representation is essential for accurately modeling, predicting, and generating human movements. Two key strategies to address are the overall representation of the body structure and the use of the kinematic chain to understand human motion [28]. The second one is the motion laws [29], which depend on the motion of the body-specific part and its relevance to other body parts in a spatial-temporal environment [23].

Two further key components of physical representation are joint positions [30,31] and joint rotations, which together provide a comprehensive understanding of human body motion dynamics. Below, we explore these two concepts and their importance in human motion analysis, prediction, and generation.

Joint Positions Refer to the 2D or 3D coordinates of specific key points on the human body, such as knees, elbows, and shoulders. These positions form the skeletal structure and are often used in motion analysis to track and represent body movements over time. By mapping the relative positions of these joints, motion systems can capture how different body parts move through space, making Joint Positions (JPs) a robust foundation for motion representation. This key-point-based approach is widely used in motion capture systems and has been adopted in various models, such as SkelNet [28], CHA [32], MGCN [33], and AM-GAN [34]. These models rely on joint positions to effectively track and predict human motion in real time. Joint positions are intuitive and provide direct visual data, making them highly interpretable and useful for applications such as animation, robotics, and biomechanics. However, JPs alone may not capture the full range of motion, especially for the angular behavior of body parts. This is where joint rotations play a vital role, often necessitating the conversion of joint positions into joint rotations via techniques such as inverse kinematics (IK). Recent advancements have expanded this keypoint-based representation by introducing body markers [35], which provide more detailed insights into body shapes and limb twists than traditional skeleton-based keypoint representations [36,37]. This extended approach enables a richer and more comprehensive analysis of motion, providing a better understanding and interpretation of complex body movements.

Joint Rotations While joint positions emphasize the spatial layout of the biomechanical structure, joint rotations offer insights into the angular motion of joints, describing how one joint rotates relative to its parent in a hierarchical structure. Joint Rotations (JR) are critical for capturing the true complexity of human motion, including limb twisting and joint bending. These rotations can be parameterized using formats such as Euler angles, axis angles, or quaternions, each of which provides a different representation of joint movements. In rotation-based models, such as the Skinned Multi-Person Linear (SMPL) model [38], the body is represented through a combination of pose and shape parameters, where the pose parameters (joint rotations) describe the angular relationship between joints. The SMPL model uses these parameters to generate a detailed 3D mesh of the human body [39], including 24 joints that define the skeleton's rotations. The shape parameters, on the other hand, capture individual body dimensions, such as height. This fusion of joint rotations and shape parameters offers a detailed and dynamic representation of human motion, widely utilized in applications such as virtual reality, animation, and biomechanics. Furthermore, SMPL-X [38,39], an extension of the SMPL model, incorporates facial and hand movements, providing a more comprehensive representation of body motion. Beyond SMPL, other statistical models, such as GHUM and STAR [40,41], have been developed to provide alternative approaches to human body modeling. Incorporating joint rotations enables models to capture dynamic properties, such as velocity and acceleration, thereby enhancing the precision and realism of motion predictions. By integrating both JPs and JRs, motion analysis systems achieve a more complete and nuanced representation of human motion.

2.1.2. Mathematical representation

This section presents the mathematical perspective on human motion representation, in which joint positions and rotations are encoded as formal algebraic, differential, graph, or trajectory-based structures for computational analysis and motion generation.

Mathematical representations play a crucial role in human motion analysis by providing abstract descriptions of human poses, allowing researchers to apply prior knowledge to motion generation and prediction. These representations, derived from motion capture data, map human poses to various mathematical spaces and distributions, simplifying feature extraction for machine learning models. The existing representations are encoded in mathematical structures such as algebra [42–44], differential encoding [33,45], graphs [45–48], and motion trajectories [49,50]. Key approaches include the use of joint positions and joint rotations, each of which provides a unique perspective on motion dynamics.

Joint Positions are a fundamental aspect of representing human motion mathematically, where the human body is modelled as a set of key points, typically corresponding to anatomical landmarks such as joints (e.g., knees, elbows, and shoulders). These positions, captured in 2D or 3D, provide the spatial coordinates necessary for tracking human movements. Motion capture systems often use these JPs to describe the motion of different body parts in space, making them a vital tool for real-time applications such as animation, robotics, and biomechanics. In mathematical models, JPs are utilized to create a skeletal structure, enabling algorithms to track motion trajectories over time. These positions are frequently parameterized in terms of algebraic structures, making it easier to capture dependencies between joints. For example, models such as SRNN, MGCN, and MT-GCN [33,51,52] rely on JP data to generate motion by analyzing how key points change over time. However, representing motion solely through JPs has limitations, as it cannot capture fine-grained angular motion and limb orientation. To address these gaps, JRs are often used in conjunction with JPs, allowing for a more comprehensive and accurate analysis of motion.

Joint Rotations provide a more intricate view of motion by describing the angular movement of body parts relative to each other. This rotation-based representation is crucial for capturing the full complexity of human motion, including twisting, bending, and rotating movements. JRs are typically represented using formats such as Euler angles, axis angles, or quaternions [42], each offering a mathematical framework for modeling rotational behavior without encountering discontinuities. One of the key models that utilizes JRs is the SMPL model, which represents the human body through a combination of JRs (pose parameters) and body dimensions (shape parameters). The SMPL model and its extensions, such as SMPL-X [38], are widely used for generating detailed 3D meshes that capture both the body's shape and its motion. This form of representation is critical in domains where realistic modeling is key, including virtual reality and biomechanics. Mathematically, JRs are often encoded into algebraic [42–44] or differential structures [33,45] to facilitate analysis. For example, the quaternion representation, commonly used in models such as QuaterNet [42], avoids singularities, making it a stable and efficient way to represent rotations. Additionally, some models use Lie groups [53] to capture the geometry of joint relationships, enabling more sophisticated motion generation. By combining JPs and JRs, mathematical models can produce a rich, dynamic representation of human motion that is well-suited for generation and predictive tasks, animation, and real-time motion tracking. In summary, mathematical representations of JPs and rotations underpin modern human motion generation models. Together, they provide a robust framework for understanding and simulating human motion, bridging the gap between abstract pose data and practical applications.

2.2. Text data representations

Textual representations are pivotal in text-to-motion understanding and generation tasks, especially when employing text-conditioned models. Text is typically represented as a sequence of words or tokens, which are converted into numerical vectors using word embeddings or more advanced models, such as transformers. Traditional embedding techniques, like Word2Vec [60] and GloVe [61], produce fixed, non-contextualized word representations, where semantically similar words are placed closer together in a high-dimensional space. In contrast, recent transformer-based models, such as BERT [62] and GPT [16], generate contextualized embeddings, meaning the representation of each word is influenced by its surrounding context within a sentence, enabling richer, more nuanced semantic understanding.

For instance, in models such as BERT [63], the embedding dimensionality typically ranges from 768 to over 1000, enabling the model to capture intricate relationships among words across different contexts. In this way, a sentence like "A person is walking slowly" is tokenized into individual words ("A," "person," "is," "walking," "slowly"), with each token represented as a vector in the embedding space. These vectors, such as "walking" $\rightarrow [v_1, v_2, \ldots, v_D]$, where D represents the embedding dimension, are learned by the model and capture the word's semantic meaning relative to its context. These vectors are then used as input to models that generate corresponding motion sequences, ensuring that the generated movement aligns with the described action. The recent emergence of LLMs has significantly contributed to the evolution of text-to-motion generation models, making them more varied and refined [64,65]. These LLMs excel at encoding rich semantic and contextual information, enabling them to guide motion generation models with greater precision, ensuring that the generated motion more closely matches the text description.

2.3. Ways of motion data collection

Human motion data can be collected through four main methods: marker-based motion capture, which uses reflective markers for precise tracking; **markerless** motion capture, which employs computer vision techniques for more flexible, natural settings; **pseudo-labeling**, where experts label data by hand, often used for fine-tuning, but is labor-intensive. Each method has its advantages and is selected based on the task's specific needs. **Marker-based** motion capture involves placing reflective markers or Inertial Measurement Units (IMUs) on specific points of a person's body to track movement in 3D space. This data can be processed using techniques such as forward kinematics or parametric body meshes (e.g., **SMPL**) to represent motion in detail [38,66,67]. Optical markers offer high precision but are limited to indoor settings, while IMUs are portable and usable outdoors. **Markerless** motion capture, on the other hand, relies on computer vision algorithms and multiple synchronized cameras to track body movement without the need for markers, making it more flexible but less accurate [68–70]. Pseudo-labelling is another technique for monocular RGB images or videos, in which 2D or 3D key points are estimated using models such as OpenPose or VideoPose3D; however, it tends to be less precise than motion capture systems [63,71,72]. Lastly, manual annotation involves creating human motion by hand through animation tools, a labor-intensive process that delivers high-quality results but lacks scalability.

3. Foundational multimodal LLM

Multimodal Large Language Models (MLLMs) have recently gained prominence and are now leading advances in visual understanding, particularly in motion generation tasks. In this section, we review and discuss recent research at the intersection of GenAI and MLLM. Though the discussion of MLLM is paramount, we first consider the LLM and its nature of work.

3.1. Autoregressive probabilistic LLM

LLM serves as the core module in MLLMs, processing inputs across multiple modalities, including textual user queries, visual information,

Table 1
The summary of related surveys. Existing studies have predominantly focused on specific tasks in human-centered motion, often lacking a unified multimodal perspective. They also tend to overlook architectural frameworks, model variants, backbone models, and the emerging role of LLMs in motion generation. In contrast, our survey provides a comprehensive overview of HMUG and categorizes models by architecture, backbone type, and application domain.

Survey	Link	Year	Venue	Data Rep.	MM LLM	Diff. Model	Unif. FW	Bench- mark	Eval. Metr.	Sub- tasks	Appli- cation	chall- enges
Content Based Management of Motion Data	[54]	2021	IEEE Acc.	✓	✗	✗	✗	✓	✓	✗	✗	✓
3D Human Motion Prediction: A Survey	[23]	2022	Neuro comp.	✓	✗	✗	✗	✓	✓	✗	✗	✗
Deep GenAI Models on 3D Representations: A Survey	[27]	2023	arXiv	✓	✗	✗	✗	✗	✗	✗	✓	✓
Human Motion Generation: A Survey	[24]	2023	TPAMI	✓	✗	✓	✗	✓	✓	✗	✗	✓
MultiModal Generative AI: LLM, Diffusion	[25]	2024	arXiv	✗	✓	✓	✓	✗	✗	✗	✗	✗
Human Motion Video Generation: A Survey	[55]	2024	arXiv	✓	✓	✓	✗	✗	✓	✓	✗	✓
Human Image Generation: A Comp Survey	[22]	2024	ACM CS	✗	✗	✓	✗	✓	✓	✗	✓	✓
A Survey of Cross-Modal Visual Content Generation	[21]	2024	CSVT	✗	✓	✓	✗	✓	✓	✗	✗	✓
Appearance and Pose-Guided Human Generation	[57]	2024	ACM CS	✓	✗	✓	✗	✓	✓	✓	✗	✗
Synthetic Data in Human Analysis: A Survey	[56]	2024	TPAMI	✓	✗	✗	✗	✓	✓	✓	✗	✗
Establishing a Unified Evaluation Framework for HMUG	[58]	2024	arXiv	✗	✗	✗	✗	✗	✓	✗	✗	✗
Autoregressive Models in Vision: A Survey	[59]	2024	arXiv	✗	✓	✗	✗	✗	✓	✗	✗	✓
Ours	–	**2026**	–	✓	✓	✓	✓	✓	✓	✓	✓	✓

and descriptive representations, and generating responses by interpreting these inputs. The fundamental characteristic of an LLM is its **autoregressive** nature, whereby the next token is predicted conditioned on the sequence of previously generated tokens. This process can be formulated as

$$P(z) = \prod_{t=1}^{T} P(z_t \mid z_1, z_2, \ldots, z_{t-1}). \tag{1}$$

In this formula, z represents a sequence of tokens, where each token z_t is predicted based on the previous tokens $z_1, z_2, \ldots, z_{t-1}$. This autoregressive approach, used in transformer-based LLMs, enables step-by-step sequence generation, where each token z_t depends on its preceding tokens.

A primary consideration is how LLMs can handle visual information, such as human motion, since they typically process text tokens as their primary input. To integrate visual data, recent works [65,73–76] align textual descriptions with both text and visual encoders from Vision Language Modeling (VLM) during pre-training. A prominent example of VLM is CLIP [73,77].

Some recent approaches represent motion sequences as a set of discrete visual tokens, treating visual motion as a "foreign language" [78] by using codebooks and vector quantization. This setup enables the use of both text and visual tokens, which LLMs can process in an autoregressive manner.

In the following subsections, we will discuss text tokenizers, VLMs, and visual tokenizers, with a focus on recent work in motion understanding and generation.

3.2. Text Tokenizer

Text tokenization converts raw text into smaller units, such as subwords, words, or characters, and maps them to numerical representations for model processing. Traditional tokenization methods, such as **Byte Pair Encoding (BPE)** and WordPiece, efficiently handle text sequences but struggle with capturing fine-grained motion semantics [17,62]. To bridge this gap, attention mechanisms and embedding strategies are crucial in text-based encoders, capturing contextual dependencies and ensuring semantic richness in multimodal tasks [79]. Self-attention enhances long-range dependencies, while embedding layers transform tokenized inputs into dense representations, facilitating better alignment between textual descriptions and motion features [80–82]. Recent advancements explore hierarchical and continuous tokenization approaches to further refine these representations, ensuring improved granularity and contextual awareness for motion reasoning and synthesis.

3.3. Visual Tokenizer

The mechanism of Visual Tokenizer (VT) is to convert video frames or images into a discrete set of tokens that the model can process for generating motion similar to a text tokenizer in **NLP**. For motion generation, the **VT** breaks down motion sequences into basic units, including body joint positions, limb movements, and key frames, enabling the model to understand and generate human motion patterns [17]. In early research on images or video frames, they are broken down into a series of patches and paired with a continuous linear embedding. The inspiration comes from LLM with language modeling. The discrete tokenizer gained fame

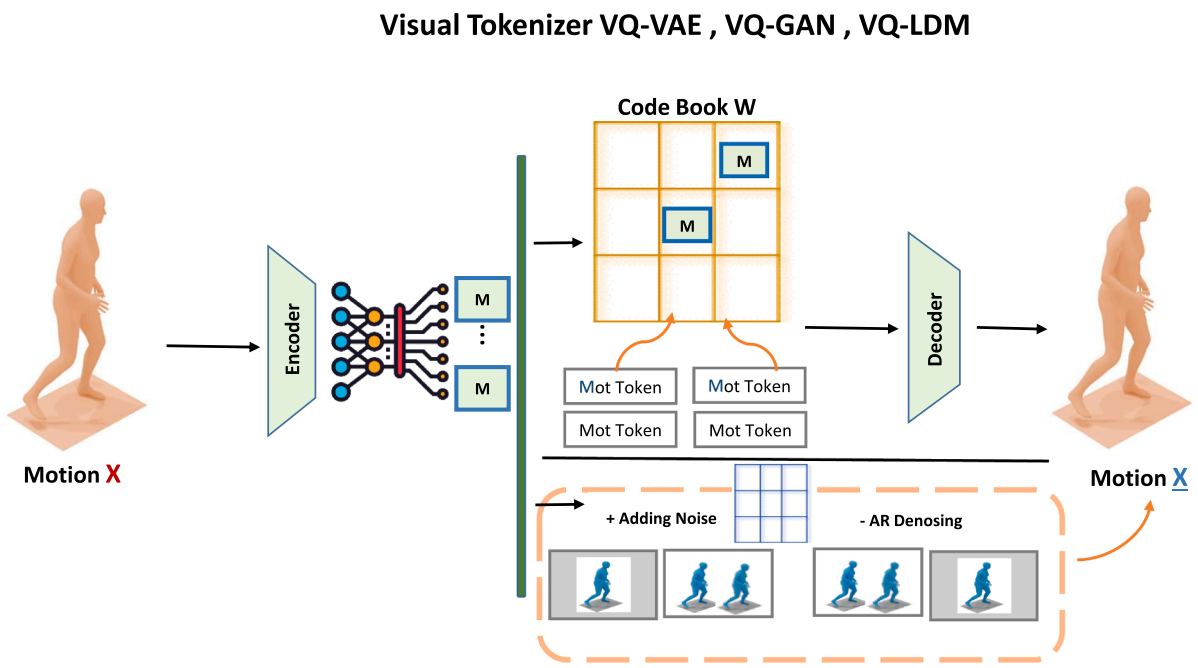


Fig. 2. The architecture of a visual tokenizer typically used in generative AI. The model converts visual inputs (e.g., images or video frames) into discrete tokens that can be processed by language or multimodal models, enabling effective cross-modal generation.

for its ability to transform motion frames into token sequences [63]. Fundamentally, visual tokenizers based on vector quantization with the codebook concept are VQ-VAEs [18,20,80], VQ-GAN [81,83], and VQ-Diffuse [84–86]. We simplify the concept of motion generation using each model and illustrate it in Fig. 2.

3.3.1. VQ-VAE

In the context of motion understanding and generation, **VQ-VAE** [8,20] operates similarly to an auto-encoder, consisting of an encoder $F(\cdot)$ and a decoder $G(\cdot)$. For a given motion sequence y, VQ-VAE first transforms it into a lower-dimensional continuous representation $F(y)$ using the encoder. This encoded representation is then discretized via a codebook $W = \{w_m\}_{m=1}^{M}$, which serves as a set of discrete motion prototypes, akin to word embeddings in language models. Each prototype $w_m \in \mathbb{R}^{d_m}$ represents a specific motion feature. Recent research has focused on architectural enhancements to VAEs, yielding improved variants such as VQ-VAE [81,83,87–89] and advances in backbone architectures with Residual Vector Quantization (RVQ-VAE) [90–93]. These improvements help generate more realistic motion while reducing computational complexity.

To convert the continuous vector into a discrete token, the closest match between the encoded vector $F(y)$ and the codebook entries is found using:

$$\text{Discrete}(F(y)) = w_r, \quad r = \arg\min_r ||F(y) - w_r||. \tag{2}$$

Once this discrete token w_r is determined, it is used to reconstruct the motion sequence via the decoder $G(w_r)$, generating the output $\hat{y} = G(w_r)$.

The training loss function for VQ-VAE in motion generation is as follows:

$$L = ||y - G(w_r)||^2 + ||\text{stopgrad}[F(y)] - w_r||^2 + \alpha||\text{stopgrad}[w_r] - F(y)||^2. \tag{3}$$

- The first term ensures that the reconstructed motion $G(w_r)$ closely matches the original motion sequence y.
- The second term pushes the motion token w_r to align with the encoded vector $F(y)$.
- The third term pulls the encoded vector $F(y)$ towards the selected discrete token w_r.

This objective enables VQ-VAE to efficiently transform motion sequences into discrete tokens, thereby facilitating their effective use in generating and reconstructing motion sequences.

3.3.2. VQ-GAN

In the context of motion generation, **VQ-GAN** [94] employs a GAN-based perceptual loss instead of the traditional L_2 reconstruction loss, enabling the model to learn a more expressive and detailed codebook. Recent advancements in GAN variants have demonstrated comparable performance in motion tokenization, styling, and reconstruction [95,96]. Despite their drawbacks, such as mode collapse, GANs offer significantly faster training times than diffusion models and present substantial opportunities for architectural improvements. The tokenization process for motion generation can be illustrated through an example.

Given a motion sequence z with dimensions $T \times M \times D$, where T is the time steps, M represents the number of joints or key points, and D is the feature dimension, the encoder $P(\cdot)$ compresses it into a lower-dimensional vector representation $P(z)$ of size $t \times m \times d_c$, where $t < T$, $m < M$, and d_c is the codebook dimension. This results in $t \times m$ vectors, each with dimensionality d_c.

For each encoded vector, the nearest neighbor in the codebook $Q = \{q_n\}_{n=1}^{N}$, where $q_n \in \mathbb{R}^{d_c}$, is identified for discretization:

$$\text{Discrete}(P(z)) = q_k, \quad k = \arg\min_k ||P(z) - q_k||. \tag{4}$$

This process produces a discrete sequence of length $t \times m$ that represents the motion in a compressed form. The decoder $G(q_k)$ then reconstructs the motion sequence using the selected discrete tokens:

$$\hat{z} = G(q_k). \tag{5}$$

The GAN perceptual loss encourages high-quality reconstructions, ensuring that the generated motion sequence $\hat{z}$ not only resembles the original motion z but also captures detailed, realistic movement dynamics, including complex motion patterns. This makes VQ-GAN well-suited for tasks such as motion generation, where detailed temporal and spatial coherence are required.

3.3.3. VQ-diffuse

VQ-Diffuse functions as a generative model that combines diffusion processes with visual tokenization techniques. The process begins with the encoder $E(\cdot)$, which takes an input image x of size $H \times W \times 3$ and maps it into a lower-dimensional latent representation $E(x)$ of size $h \times w \times n_c$, where $h < H$ and $w < W$. This latent representation consists of $h \times w$ tokens, each with dimension n_c.

To discretize these latent vectors, **VQ-Diffuse** [85] employs a codebook $Z = \{z_k\}_{k=1}^{K}$, analogous to word embeddings in NLP. For each latent vector $E(x)$, the model identifies the nearest token in the codebook, resulting in a discrete representation z_q. This is achieved through the following process:

$$z_q = \text{Discrete}(E(x)) = z_k, \quad k = \arg\min_k \|E(x) - z_k\|. \tag{6}$$

The discrete tokens are then subjected to a diffusion process [97,98], which iteratively refines the representation to generate high-quality outputs. The training objective of VQ-Diffuse can be framed as:

$$L = \sum_{t=1}^{T} \left(\|x - D(z_q^{(t)})\|^2 + \lambda \cdot \text{Loss}_{\text{diffusion}} \right), \tag{7}$$

where $D(\cdot)$ represents the decoder that reconstructs the image from the discrete tokens, λ is a balancing hyperparameter, and $\text{Loss}_{\text{diffusion}}$ represents the diffusion loss that promotes smooth transitions between the generated frames.

By utilizing this approach, VQ-Diffuse effectively captures the complex dynamics of motion through a refined understanding of visual tokens. It enables the generation of motion sequences that are not only coherent and realistic but also contextually relevant to the visual content they represent, thereby advancing motion generation capabilities and reducing computational cost.

3.3.4. Analysis of LLM with VQ-VAE, VQ-GAN, and VQ-diffuse

In motion generation, VQGAN and VQ-VAE serve as effective visual tokenizers, converting motion data, such as videos or sequences of key points representing human motion, into discrete tokens. Compared to continuous latent representations used in traditional autoencoders, these discrete tokenization approaches enable better alignment with the token-based structure of LLMs, thereby facilitating more seamless multimodal integration. This allows motion data to be represented in a format compatible with autoregressive modeling, enabling LLMs to generate motion sequences in a manner analogous to language generation. For example, recent LLM-driven motion generation frameworks leverage VQ-VAE-based tokenization to map motion sequences into discrete vocabularies, enabling prompt-based generation similar to text generation pipelines [63,73,99]. Such designs improve semantic coherence between text and motion, particularly in complex action or compositional motion scenarios. Furthermore, VQGAN and VQ-VAE enable compression of motion into latent representations suitable for diffusion-based generation. Compared to pixel-space diffusion, latent diffusion significantly reduces computational cost while preserving essential motion structure. Within this setting, VQ-Diffuse-style frameworks integrate discrete tokenization with diffusion to improve temporal consistency and diversity, addressing limitations such as mode collapse in GAN-based approaches.

Importantly, these approaches exhibit distinct trade-offs when integrated with LLMs. VQ-VAE-based methods offer stable training and efficient tokenization but may produce over-smoothed motion due to reconstruction objectives. In contrast, VQ-GAN improves perceptual quality and motion realism through adversarial learning, although it can suffer from instability and mode collapse. VQ-Diffuse further extends this paradigm by leveraging diffusion processes to generate more diverse and temporally coherent motion sequences, albeit at higher computational cost. Compared to standalone GAN or VAE-based pipelines, this hybrid integration with LLMs enables more controllable and semantically aligned motion synthesis from both textual and visual inputs. For example, in virtual character animation, LLM-guided token generation can produce context-aware motion sequences that adapt to narrative inputs, improving realism and interactivity. This synergy not only streamlines the generation process but also opens avenues for applications such as virtual reality, animation, and human-computer interaction, where generating fluid, context-aware motion is essential. Moreover, it suggests that discrete token-based representations are particularly advantageous for multimodal reasoning tasks, whereas continuous representations are likely to remain preferable for fine-grained motion reconstruction scenarios.

3.4. Vision language modeling (VLM)

A **VLM** integrates visual and textual modalities to support cross-modal understanding, retrieval, and generation tasks. Compared to single-modality encoders such as BERT [62], VLMs jointly learn visual perception and language semantics, enabling richer multimodal representations. Key components include vision encoders, language encoders, and fusion modules, as illustrated in Fig. 3. In contrast to early fusion strategies, modern VLMs such as CLIP, ViLT, and Flamingo improve cross-modal alignment through large-scale pretraining, enabling stronger generalization in tasks such as motion captioning, VQA, and cross-modal retrieval [17,73,77,100–102]. These models have also been adapted for HMUG, where motion is treated as an additional structured visual modality for text-conditioned generation and retrieval.

3.5. Multi-modal LLM in *HMUG*

After discussing the architectures of the MLLM, we will discuss how the current work in motion generation is based on these architectures. The researchers have conducted comprehensive work on image generation using LLM. Afterward, they started working on human motion and video generation. Specifically, in human motion generation, it requires a sequence of image frames; some motion LLMs also utilize audio, speech, and scene modalities. Though our focus is on LLM text-to-motion generation. Motion LLMs have higher computational complexity than other text modalities, such as image LLMs. Due to the considerable challenge of collecting high-quality motion datasets for training, early fusion techniques are often considered computationally expensive for human motion and video generation. Most existing work relies on alignment methods, such as cross-attention and contrastive learning, to synchronize heterogeneous data representations, thereby enhancing coherence in multimodal generation tasks. The architecture of both techniques is illustrated in detail in Fig. 3a.

Recently, researchers shared their work on Motion-Agent [90], A dialogue-based framework for generating human motion from textual descriptions. They used a simple adapter to fine-tune only 1–3% of the Motion LLM [64] modal parameters with pre-trained GPT-4, achieving comparable results. The base model, MotionLLM [99], is designed to understand human behavior from both human motion and video data. They introduced the MoVid benchmark dataset, bridging the gap between motion and video modalities. Both MotionLLM and MotionAgent showed fantastic results in generation, understanding, captioning, and reasoning. The limitations faced by these researchers are on video encoders. Another proposed work is the AvatarGPT [74] framework, inspired by the base work InstructGPT [103]. This work constructed a dataset and achieved SOTA results on low-level tasks, demonstrating promising performance on high-level tasks. This work serves as the interface for task planning, decomposition, generation, summarization, and understanding.

Another work proposed by the authors, MotionGPT [78], claims improved diversity and realism through GPT-3 prompting. They also transform the high-level text to low-level text using zero-shot prompting techniques. Another work on fine-tuned LLM with Motion general-purpose motion generators, MotionGPT [104]. They used a multi-control signal that includes text and single-frame poses, rather than a single-modality

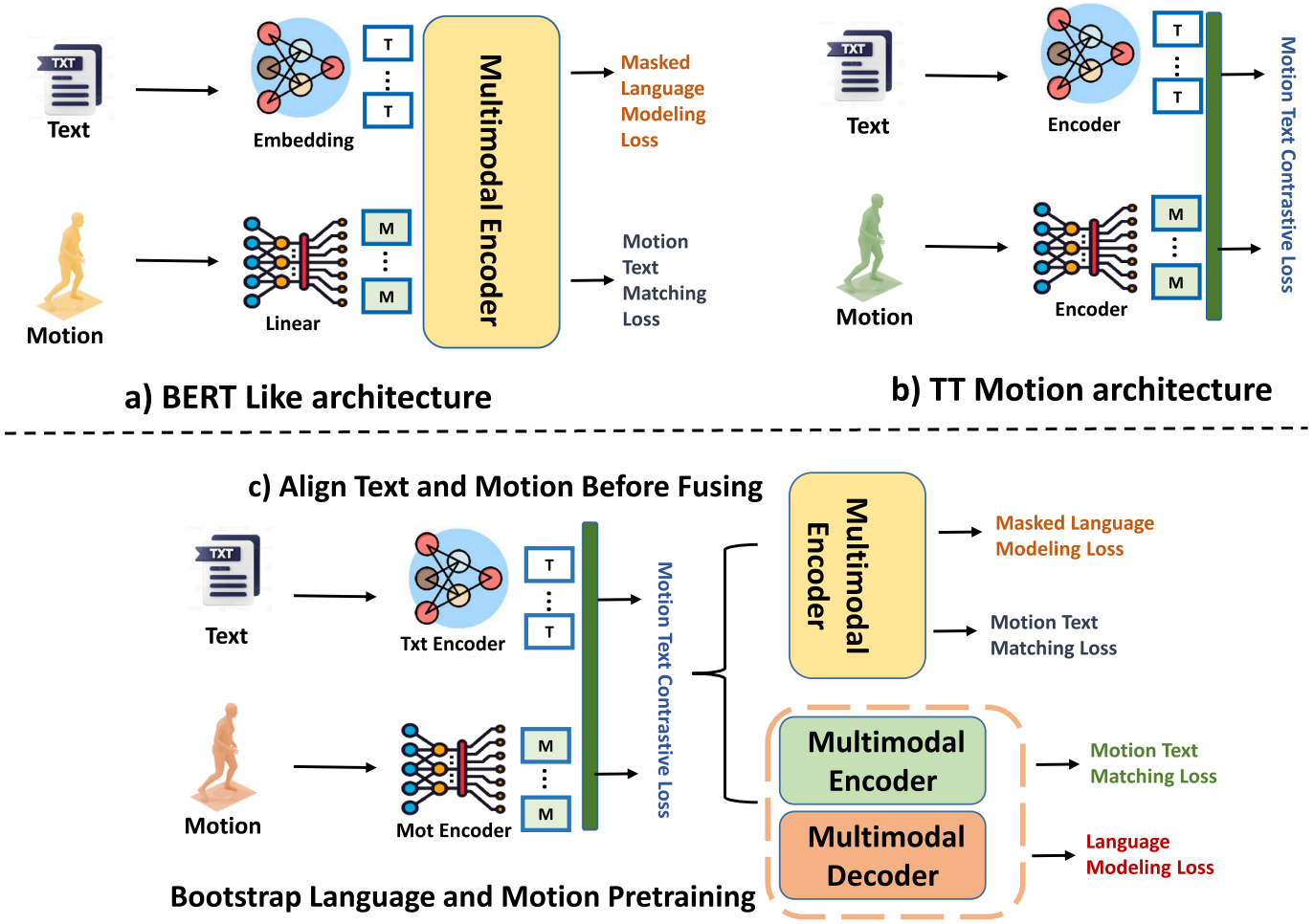


Fig. 3. An illustration of various architectures used in motion understanding and generation, including **a)** BERT-like architectures, **b)** contrastive learning frameworks, and **c)** CLIP-based pretraining models.

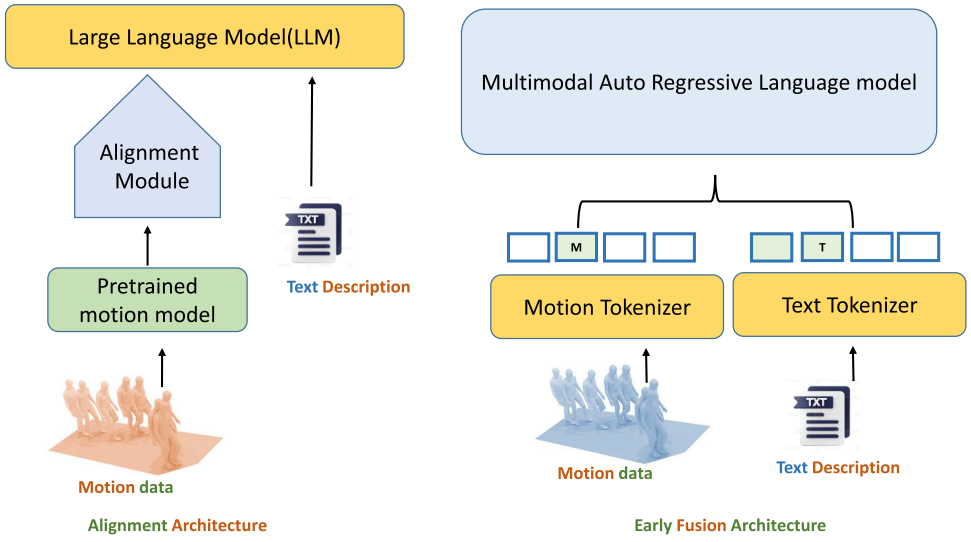


Fig. 4. A representation of early fusion and alignment architectures employing contrastive learning. The early fusion approach combines multimodal features at the input or intermediate levels, whereas the alignment architecture projects modalities into a shared embedding space, thereby facilitating cross-modal contrastive learning.

control signal. This is the uniqueness of their work. They also employed Low-Rank Adaptation **(LORA)**, similar to motionAgent, by freezing most of the LLM parameters, and obtained comparable results. These techniques are efficient for generating motion based on conditions. Recently, authors proposed MotionScript [105], an LLM-based approach for generating expressive 3D human motion descriptions. They also introduced a novel approach to motion text conversion and to NLP representations of human motion. The primary contribution of this work is the automatic generation of detailed motion captions using an LLM, achieving a higher level of granularity than T2M-GPT [80].

On the other hand, groundbreaking studies such as MotionChain [106] and TAAT [107] leverage LLMs, vision-language models, and human motion data to enhance motion generation tasks. The authors introduced a multimodal conversational motion controller that uses multimodal prompts to control virtual humans. The emergence of text-to-motion generation has opened up numerous compelling applications and has attracted significant research interest. However, Alert-Motion [75] raises security concerns about integrating LLMs into motion generation, highlighting the risk of malicious exploitation via adversarial inputs injected into black-box text-to-motion models. This risk is further amplified when such malicious motion generation is applied to humanoid controllers. To address these challenges, researchers propose LLM-based agents that iteratively refine and search for adversarial prompts, employing contrastive modules to ensure semantic and contextual relevance.

Currently, researchers are exploring real-world applications of human motion generation, including robotics [108,109], autonomous vehicles [110,111], and healthcare [112]. A recent study presented at the Intelligent Vehicle Symposium, Walk the Talk [113], introduced an LLM-based model for generating pedestrian motion. The authors constructed a dataset that captures pedestrian motion behavior during road crossings, thereby contributing to advancements in autonomous driving simulations. Another notable study on humanoid motion generation [114], published by Intel Labs, explores its applicability in virtual agents. Additionally, healthcare applications are emerging [115], such as an LLM-driven fitness coach chatbot [78], demonstrating the potential of motion-aware AI assistants. Table 2 presents a comparative study of LLMs on downstream tasks related to understanding and generation.

4. Preliminaries in multimodal diffusion

In this section, we first outline key concepts in human motion generation before delving into diffusion models. Human motion generation techniques can generally be grouped into two categories. The first category comprises regression-based methods that use supervised learning to directly predict motion from input features by mapping conditions to target motions. The second category is based on generative models, which aim to model the underlying motion distribution or its joint distribution with input conditions, typically using unsupervised learning. Prominent generative models include Generative Adversarial Networks (**GANs**), Variational Autoencoders (**VAEs**), Normalizing Flows (**NFs**), and Denoising Diffusion Probabilistic Models (**DDPMs**). Additionally, specialized models, such as motion graphs, are widely used in areas such as computer graphics and animation.

With this background, we will introduce the generative models above (Fig. 4) and diffusion probabilistic modeling, detailing its framework and available variants (Fig. 5). We will also discuss the latent diffusion model and its significance, as well as advanced models for generating text-to-motion outputs.

Generative Adversarial Networks Generative Adversarial Networks (GANs) [120] have gained popularity for human motion generation due to their capability to model complex data distributions. In this framework, GANs consist of two neural networks-a generator (G) and a discriminator (D)-that engage in an adversarial process. The generator's objective is to produce realistic human motion sequences from a random noise vector, z, while the discriminator's role is to differentiate

Table 2
A comprehensive comparison of technical papers in multimodal LLM in motion understanding and generation, showing the modalities being used, representations, backbone, and datasets. It also highlights the tasks and subtasks catered to in the research, with open-source availability. T: Text; M: Motion; AR: Autoregressive.

Method	Link	Submission Date	Input Modality	Multimodal LLM	Motion Rep.	Backbone	Datasets	Downstream Tasks	Open Source	Venue
MotionLLM	[64]	27-May-2024	T & M	GPT-4, VQ-VAE, Vicuna-7B	3D	Encoder Decoder	BABEL	Generation, Captioning, Reasoning	✓	ICLR
AvatarGPT	[74]	28-Nov-2023	T	Lama-13B, GPT2-large, T5	3D	Auto Regressive	HumanML3D	Generation, Understanding, Summarization	✓	CVPR
MotionGPT-3	[104]	19-Jun-2023	T	GPT-3, DistilBERT	2D, 3D	Encoder Denoiser	HumanML3D, BABEL	Retrieval, Generation	✓	CVPR
MotionGPT-2	[116]	29-Oct-2024	T & M	LLMs, VQ-VAE	3D	Encoder Decoder	HumanML3D, Motion X, KIT-ML	Generation, Captioning, Prediction	✗	ArXiv
T2M-GPT	[78]	24-Sep-2023	T & M	VQVAE, GPT	3D	AR Encoder Decoder	HumanML3D, KIT-ML	Reconstruction, Generation	✓	CVPR
MotionChain	[117]	2-Apr-2024	T & M	ChatGPT, VQ-VAE	3D	Encoder Decoder	TMR, HumanML3D	Generation, Reasoning, Translation	✗	ECCV
MotionScript	[105]	19-Dec-2023	M	GPT-4.0	3D	Encoder Decoder	HumanML3D, BABEL	Description Generation	✓	CVPR
WalkLLM	[118]	5-Jun-2024	T & M	Flan-T5, VQVAE	2D	Encoder Decoder	HumanML3D, KIT-ML	Pedestrian Generation	✗	IVS
MotionLLMVid	[99]	30-May-2024	T & M	GPT-4, LLaVA	3D	Vision Encoder Decoder	HumanML3D, Motion-X, H3DQA	Understanding, Reasoning, Generation	✓	ArXiv
PRO-Motion	[119]	22-Dec-2023	T	GPT-3.5, DDPM	2D, 3D	Auto Regressive	HumanML3D, Motion-X, PoseScript	Planning, P-Diffuser, Generation	✗	ECVA
FineMoGen	[113]	22-Dec-2023	T	LLM, Diffusion	3D	Encoder Decoder	MoGen, HumanML3D, KIT-ML	Generation, M-Editing	✓	NeurIPS
AlertMotion	[75]	1-Aug-2024	T & M	GPT-3.5, T2M	3D	Encoder Decoder	HumanML3D, AMASS	Adversarial Similarity, Naturality	✗	ArXiv

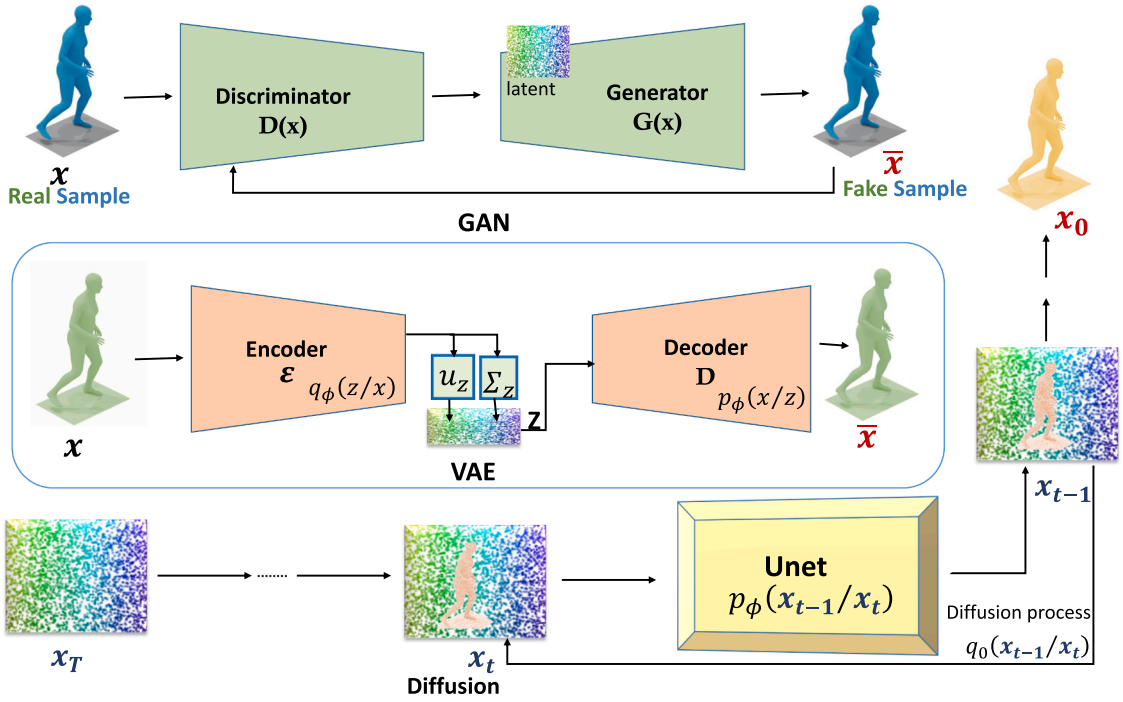


Fig. 5. An illustration of generative model architectures, including GAN, VAE, and diffusion models.

between real human motion data and the synthetic motion generated by the generator (G). The adversarial setup can be framed as a min-max game, with its objective captured by the following loss functions:

$$\min_G \max_D \left(\mathbb{E}_{x\sim p_{\text{data}}} \log D(x) + \mathbb{E}_{z\sim p_z} \log(1 - D(G(z))) \right). \tag{8}$$

Discriminator's Loss

$$L_D = -\mathbb{E}_{x\sim p_{\text{data}}} \log D(x) - \mathbb{E}_{z\sim p_z} \log(1 - D(G(z))), \tag{9}$$

Generator's Loss

$$L_G = -\mathbb{E}_{z\sim p_z(z)} \left[\log D(G(z)) \right]. \tag{10}$$

In these equations, $p_{\text{data}}(x)$ represents the real data distribution (i.e., the distribution of human motion), and $p_z(z)$ represents the noise distribution (typically a Gaussian distribution). The generator aims to minimize L_G, i.e., to produce realistic motion sequences that maximize the discriminator's belief that they are real. Meanwhile, the discriminator attempts to maximize L_D, thereby learning to more effectively distinguish real from generated motion sequences. The architecture of the GAN model is represented in Fig. 5.

GANs [121] have been successfully applied to generate various types of human motion [94,113], such as walking, running, and complex dance sequences. Some advanced GAN architectures incorporate temporal features using techniques such as Recurrent Neural Networks (RNNs) or 3D convolutional networks, thereby enabling them to model the time-dependent nature of motion data. For example, models like Temporal GAN (TGAN) [108,114] extend traditional GANs by incorporating 3D spatial-temporal convolutions to generate continuous motion over time, thereby improving realism. Additionally, variants like Deep Convolutional GAN (DCGAN) [109] enhance the modeling of spatial features in motion, Progressive Growing GAN (PGGAN) [110] allows for refined output through gradual complexity, and StyleGAN [95,112] provides control over different aspects of generated motion, enabling the creation of diverse and stylized sequences.

Remark. Despite their success, GANs in motion generation face challenges such as mode collapse, where the generator produces limited varieties of motion, and training instability due to the delicate balance between G and D. Recent work has focused on addressing these challenges. Still, they remain key areas of ongoing research.

Variational Autoencoders. VAEs have emerged as a prominent approach for generating human motion sequences, offering a robust framework for representing and reconstructing complex motion data. Unlike GANs, which rely on adversarial training, **VAEs** [100,122,123] utilize an encoder-decoder architecture to learn a probabilistic latent space, ensuring that the latent variables z are drawn from a smooth and continuous distribution, typically a Gaussian, shown in Fig. 5.

In motion generation, the encoder transforms human motion data x into a latent representation $z = E(x)$, and the decoder reconstructs the motion sequence $\hat{x} = D(z)$, aiming to closely approximate the original motion data.

VAEs employ a unique training objective that incorporates the Evidence Lower Bound (ELBO) to balance reconstruction accuracy and regularization of the latent space. The loss function can be expressed as:

$$L(\alpha, \beta; x) = -D_{KL}(q_\beta(z|x) \| p_\alpha(z)) + \mathbb{E}_{q_\beta(z|x)} \log p_\alpha(x|z). \tag{11}$$

This optimization ensures that the generated motions preserve high fidelity while enforcing that the latent space adheres to a known distribution, thereby reducing the risk of overfitting.

Remark. A common challenge faced by VAEs is the phenomenon of posterior collapse, where the latent variables become less informative, relying too heavily on the decoder. Despite this limitation, VAEs and their variants, such as Conditional VAE (CVAE) [11] and Vector Quantized VAE (VQ-VAE) [89,124], have proven effective in capturing both static and dynamic aspects of human motion, making them valuable tools for generating diverse and realistic motion sequences.

Diffusion Probabilistic Modeling. Diffusion models have shown great promise in generating human motion [5,99], leveraging their probabilistic framework to produce realistic motion sequences through a gradual denoising process, as illustrated in Fig. 4. The core idea is a forward diffusion process, in which Gaussian noise is added incrementally to human motion data over a series of iterations. This noise is governed by a predefined schedule and results in a set of noisy motion sequences. The goal is to learn to reverse this process, starting from random noise and gradually denoising it until realistic human motion is generated. The forward diffusion process can be formalized as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; (1 - \beta_t)x_{t-1}, \beta_t I), \tag{12}$$

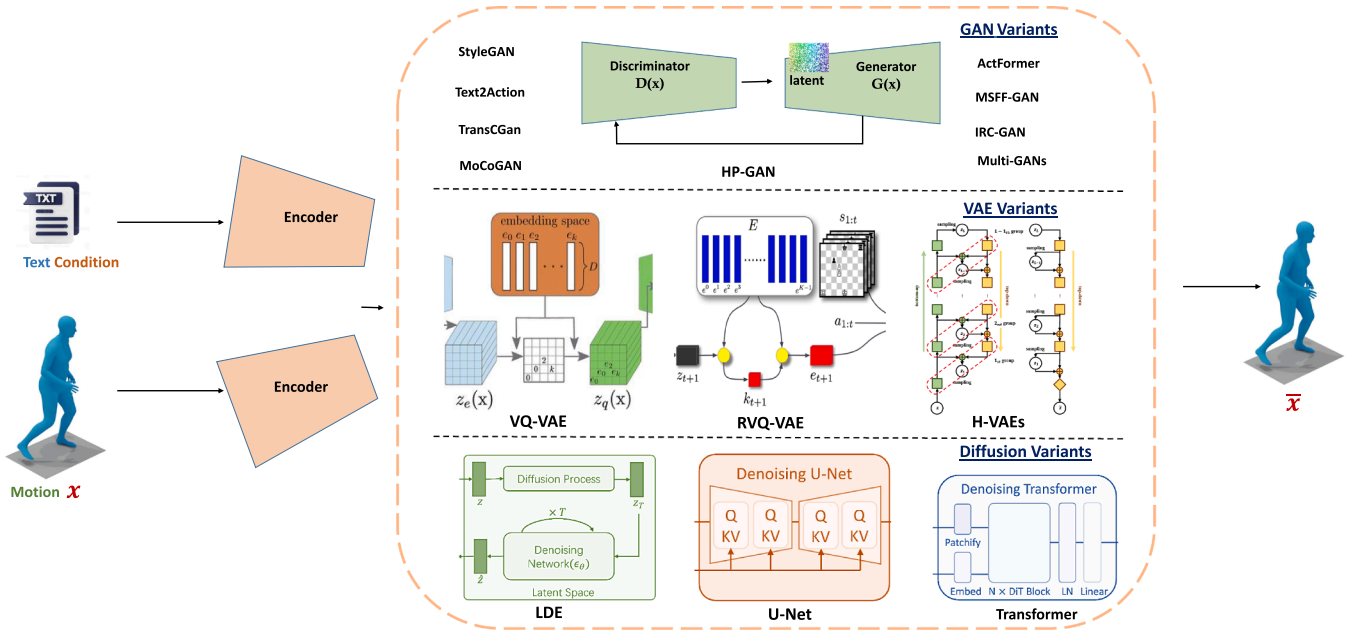


Fig. 6. An illustration of multimodal diffusion GAN and VAE variants in the architecture. The figure illustrates how each model variant is integrated into the architecture: the diffusion GAN enables generative learning across multiple modalities, and the VAE variants facilitate effective latent space modeling for multimodal data synthesis and transformation.

where x_t represents the motion data at step t, and β_t controls the amount of noise added at each step. As the number of steps T approaches infinity, the motion data becomes indistinguishable from Gaussian noise. However, the key to generating realistic motion lies in the reverse process. Given the noisy motion data, a neural network p_θ learns to predict the previous, less noisy step, gradually recovering the clean motion:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{13}$$

This process generates human motion by iteratively refining noisy sequences, and the model is trained using an objective similar to that of Variational Autoencoders (VAEs) [87], which minimizes the evidence lower bound (ELBO).

When applied to human motion, diffusion models can generate smooth, continuous sequences such as walking or running by training on large motion datasets. They provide high-quality results and stable training compared to other generative models, such as GANs, but are computationally expensive due to the extended sequence of reverse steps required.

To mitigate computational cost, Latent Diffusion Models (**LDMs**) [11] offer a more efficient approach by performing the diffusion process in a compressed latent space rather than directly on motion sequences. This reduces the dimensionality of the data and speeds up the generation process while maintaining motion realism. This latent space can be compressed using an encoder, similar to VQGAN [96], enabling conditional generation in which additional inputs, such as text or semantic information, can guide motion generation.

In essence, diffusion models are a powerful tool for generating high-fidelity human motion, though optimizing their efficiency remains a key focus of research.

4.1. Text conditioned motion generation

Text is a powerful and intuitive medium that enables the specification of human actions, making it a compelling modality for motion sequence synthesis compared to other inputs, such as predefined motion categories or voice commands. Text-to-motion generation leverages the descriptive nature of language, capturing complex aspects of human movement, including action types, velocities, directions, and destinations. This flexibility enables rich, nuanced control over motion synthesis, allowing for diverse applications in animation, robotics, and healthcare. The task can be divided into two main categories: action-to-motion, where the text directly maps to predefined actions, and text-to-motion, which aims to dynamically generate realistic motion sequences from free-form language descriptions, as shown in Table 3.

4.1.1. Word-based action to motion

The action-to-motion task [129–131] focuses on generating human motion sequences based on predefined action categories, such as “Walk” and “Jump”. These actions are typically represented using simple encoding methods, such as one-hot vectors, making it easier to translate specific actions into motion. This method [132–134] offers a direct and efficient approach to motion generation, providing more control and predictability compared to text-to-motion tasks, which deal with the complexities of interpreting natural language. However, action-to-motion generation is more rigid, often limited to individual actions or short sequences. It struggles to handle more complex motions involving multiple actions, detailed trajectories, or nuanced body movements. To overcome these limitations, researchers have introduced the concept of using localized action signals to guide the generation of more complex, global motion sequences, allowing for smoother transitions and more accurate representations of human movement [107], actformer [94].

4.1.2. Text-description to motion

The text-to-motion task involves generating human motion sequences directly from natural language descriptions, tapping into the rich expressiveness of language [63]. Unlike action-to-motion, which relies on a fixed set of predefined labels, text-to-motion allows for the creation of a wide range of dynamic and complex motions based on diverse textual inputs. This approach enables more flexible and nuanced motion generation but also presents challenges in translating detailed linguistic descriptions into accurate physical movements, requiring a deep understanding of both language and human motion dynamics. We will discuss each text-to-motion generative model, along with its available variants, one by one, as shown in Fig. 5.

Table 3
A comprehensive comparison of technical work in multimodal LLM in motion understanding and generation. Showing the modalities being used, representations, backbone, and datasets. It also highlights the tasks and subtasks catered to in the research, with open-source availability.

Method	Link	Year	Input	GenAI Model	Represen-tation	Backbone	Datasets	Tasks	Open source	Venue
Text2Motion	[125]	2018	Text + Motion	GAN	2D, 3D	Enc-Dec	MSR-VTT	Text-2-Motion, Motion-2-Text	✗	ICRA
HP-GAN	[122]	2018	Text + Motion	WGAN	3D	Enc-Dec	NTU , Human3.6M	Motion Prediction	✗	CVPR Wksp.
GAN model	[122]	2021	Text + Motion	GAN, LSTM, BERT	2D	Enc-Dec	KIT-ML	Quality Motion Gen.	✓	ICCV
ActFormer	[109]	2023	Text	GAN, Transformer	2D, 3D	Enc-Denoiser	GTA, NTU, BABEL	Multi-Person Gen.	✓	ICCV
Style-GAN	[126]	2024	Text + Motion	GAN	3D	Autoreg. Enc-Dec	HumanML3D, KIT-ML	Motion Generation	✗	ICCGV
LS-GAN	[100]	2024	Text + Motion	WGAN-GP	3D	Enc-Dec	HumanAct12, HumanML3D	Motion Synthesis	✗	arXiv
MSFF-GAN	[123]	2024	Motion	MSFF-GAN	Keypts, 3D	Enc-Dec	Human Video	Motion Transfer	✗	CCDC
MotVAE	[124]	2023	Text + Motion	VQ-VAE, Trans.	3D	Enc-Dec	KIT-ML, HumanML3D	Motion Generation	✓	CVPR
VAE DLS	[89]	2024	Text + Motion	VQ-VAEs	Keypts, 2D	Enc-Dec	KIT-ML	Motion Reconstruction	✗	ICMEW
Tevet	[99]	2023	Text + Motion	Diffusion, Trans.	2D, 3D	Enc-Dec	HumanML3D, KIT-ML	Interp., Editing, Recog.	✓	ECCV
FineMoGen	[114]	2023	Text+ Motion	LLM, Diffusion	3D	Enc-Dec	HuMMan, KIT-ML	Motion Gen., Editing	✓	NeurIPS
ReMoDiffuse	[127]	2023	Text+ Motion	Retrieval Diffusion	3D	Enc-Dec	HumanML3D, KIT-ML	Hybrid Retrieval	✓	ICCV
MotionDiffuse	[5]	2024	Text + Motion	Diffusion, Trans.	2D, 3D	Text Enc, Diffuse	HumanML3D, KIT-ML, BABEL	Text-to-Motion	✓	TPAMI
MotionFix	[128]	2024	Text+ Motion	Diffusion	3D	Trans.-Denoiser	HumanML3D, KIT-ML, PoseFix	Motion Gen., Editing	✓	SIG- GRAPH

4.2. T2M using GAN with G-variants

Now, after discussing the architecture of GAN in Section 4.1 and illustrating it in Fig. 4. We explore text-to-motion generation using GANs with G-variants, which involves generating human motion sequences from text by leveraging **GANs** with various generator architectures. These **G-variants** enable the system to produce more realistic and diverse motions by improving the quality of generated sequences. This approach addresses challenges such as motion realism and fine-tuning complex actions based on textual descriptions, as illustrated in Fig. 5.

The researchers proposed Text2Motion [135], in which they utilized a GAN architecture to generate diverse motion from high-level natural language descriptions. For example, GAN body pose utilizes a CNN for detection and generates high-quality human body poses through a GAN. Another study generates a motion dataset from a textual description using a GAN model [124]. One of the research areas is decomposing motion [7] to generate video content. The authors proposed ActFormer [94], which employs a GAN and transformer to generate action-conditioned motion based on text. It helps by adapting the diverse motion representations in single-person motion generation tasks. They also introduced synthetic data for multi-person behavior. In terms of prediction, HP-GAN [19] proposed a probabilistic 3D human motion work by modifying the Wasserstein GAN (WGAN) model. They introduced the novel loss function and predicted the next motions non-deterministically. Style-GAN [96] generates an improved human motion generation by interpolating the intermediate latent variables. In TransCGan [126], the researcher used the transformer architecture with a conditional GAN to generate human motions. MSFF-GAN [125] proposed the multiscale feature fusion GAN for motion transfer by using global-local perceptual loss and pose consistency loss.

4.3. T2M using VAE with V-variants

After discussing the architecture of VAE in Section 4.2 and illustrating it in Fig. 4, we can see that text-to-motion generation is achieved using Variational Autoencoders (**VAEs**) [100,122]. It is the most widely used GenAI model that transforms natural-language descriptions into human motion sequences by learning a compressed latent representation of motion data. VAEs [123] capture the complex variability in human motion and generate outputs by sampling from this learned space, conditioned on text. This enables the model to produce realistic, diverse motions with smooth transitions while reducing noise and enhancing overall quality [131]. Advanced VAE variants, such as VQ-VAE [81,83,87–89] and the most recent RVQ-VAE [20,84,90–93,136–138], further enhance the process by introducing discrete latent spaces. **VQ-VAE** produces sharper, more coherent outputs. For instance, AttT2M [79] proposed a body-part and global-local two-stage attention mechanism that uses a discrete latent space. In contrast, **RVQ-VAE** employs multiple levels of quantization to achieve more detailed and fine-tuned motions. Other variants, such as Conditional VAEs (**CVAEs**) [11] and Hierarchical VAEs (**HVAEs**) [32,101], enhance control over motion generation by incorporating features such as text conditioning and multi-level latent representations, thereby enabling more intricate and precise motion generation. HumanTOMATO [139] uses the hierarchical VAE to generate a vivid and well-aligned whole-body motion generation through textual descriptions.

4.4. T2M using diffusion with D-variants

In recent years, diffusion models have demonstrated remarkable success in text-to-image generation, prompting researchers to extend these models to text-to-motion and video synthesis. By adapting attention mechanisms, researchers have achieved text-to-motion generation without introducing additional parameters. The pioneering work L2Pose [140] addresses challenges in multimodal integration by constructing a joint embedding space using shorter, more tractable sequences. MotionDiffuse [5] represents a seminal contribution in this area, drawing significant attention for enabling high-quality text-to-motion generation using a probabilistic diffusion framework. This model employs a cross-modality linear transformer that incrementally integrates text into the motion sequence, thereby mitigating the dominance of text features within the joint embedding space and its associated high-dimensional latent representations. Given the strong correlation between natural language commands and human body movements, this approach effectively controls individual body parts.

A related effort added by FLAME [141], with a focus on high-fidelity motion generation that aligns well with the given text prompt. The combination of these architectures yields an efficient and robust model for generating diverse, high-quality motion. Another researcher [142] uses a classifier-free diffusion model with BERT embedding and achieves comparable results in complex NLP scenarios. The authors demonstrated zero-shot motion generation using unseen text guidance. Guided, con-

trollable motion synthesis [143] using the diffusion model offers a new direction for incorporating spatial constraints into the generation process. They injected sparse key-frame signals, which are susceptible to being missed during the reverse denoising step. Another area of interest is representation learning in latent space, a technique that offers efficient modeling. The Motion Latent Diffusion (**MLD**) [11] generates reasonable human motion sequences that are compatible with text descriptions or action classes. While comparing this MLD model to Motion Diffusion Model (**MDM**), there is very little need for computational overhead to generate vivid motions, and it is also faster than MDM and more effective, as shown in the architecture in Fig. 5. These controllable motion generation methods are generally classified into two types: classifier-based and classifier-free guidance, as outlined in previous studies.

MotionDiffuse is grounded in the Denoising Diffusion Probabilistic Model (**DDPM**) [101], which excels at producing diverse outputs while accommodating additional constraints throughout the denoising process. Building on this foundation, Tevet [77] introduces a transformer-based diffusion model that adapts classifier-free guidance for the human motion domain. Similarly, FLAME [141] focuses on high-fidelity motion generation aligned with textual prompts, thereby contributing to the development of robust models capable of synthesizing diverse, high-quality human motions.

Additional work includes a classifier-free diffusion approach that integrates BERT embeddings, achieving competitive performance in complex natural language processing tasks, including zero-shot motion generation from unseen text prompts [142]. Furthermore, research in guided and controllable motion synthesis has proposed mechanisms for injecting spatial constraints during generation, such as sparse keyframe signals, which are particularly vulnerable to omission during the reverse denoising phase [143].

An emerging area of interest involves representation learning in the latent space for more efficient motion modeling. The MLD model [11] is capable of generating coherent motion sequences corresponding to text descriptions or action classes, with lower computational overhead and faster inference compared to MDM models, as illustrated in Fig. 5. These controllable generation methods are generally categorized into classifier-based and classifier-free guidance strategies, as outlined in earlier studies.

Some approaches concentrate on generating motion from fine-grained textual descriptions. For instance, FineMoGen [113] introduces a spatiotemporal fine-grained motion generation method supported by the HuMMan-MoGen dataset, which also facilitates motion editing a promising advancement in human motion manipulation. The issue of ensuring consistency between text and motion is addressed by ReMoDiffuse [127], which combines retrieval-based techniques with diffusion and transformer architectures and introduces a novel generalizability metric. More recently, MotionFix [128] introduced a text-annotated dataset for unconstrained human motion editing. By training a conditional diffusion model on this dataset, the approach outperforms existing state-of-the-art methods. Likewise, GUESS [144] proposes a cascaded latent denoising diffusion model that incrementally enriches motion generation by adaptively inferring joint conditions. Through semantic graph modeling, hierarchical information is retrieved to support fine-grained generation, with semantic disentanglement performed across motions, actions, and specific details.

Broadly, existing research employs a two-pronged intervention strategy to enhance motion generation: one at the textual level to improve descriptive quality, and another at the motion level through joint-space optimization. The overarching objective across these efforts is to establish a generation pipeline that is efficient, controllable, high-quality, and diverse, while minimizing both training and inference time.

5. Unified human motion model

In the preceding sections, we explored the capabilities of multimodal LLMs and multi-model diffusion models in addressing various aspects of human motion planning, understanding, and generation. In this section, we present a unified framework that integrates these approaches, aligning with the broader trend toward Artificial General Intelligence (**AGI**). This unified approach envisions a single model that can plan, understand, and generate human motion seamlessly. We examine this framework from perspectives such as model architecture and probabilistic methodologies, and discuss when and why unification is helpful in practical scenarios.

5.1. Multimodal probabilistic modelling

Autoregressive models, as demonstrated in multimodal LLMs, excel in text understanding and generation due to their autoregressive nature. Similarly, diffusion models have proven indispensable for generating high-quality motion, as outlined in the prior sections. Recent advances propose combining these two paradigms, leveraging their respective strengths to create a unified framework for understanding and generating human motion. For instance, by aligning motion data and text through a shared latent space, these models enable the synthesis of realistic motion sequences guided by textual descriptions. This unification helps reduce model redundancy, improves generalization across tasks, and ensures consistency between generation and comprehension.

5.2. Autoregressive models in motion generation

Although diffusion models dominate in generating high-quality visual motion, autoregressive models also hold significant promise for text-conditioned motion generation. MotionLLaMA [145], MotionGPT-2 [116], MotionLLM [64], and InstructMotion [103] employ autoregressive approaches by discretizing motion data into tokens through Vector Quantized GANs or VAEs. These tokens, combined with text tokens, are processed by multimodal LLMs to generate sequences. For decoding, the tokens are mapped back into motion data using VQGAN or VAE, enabling the simultaneous understanding and generation of human motion.

Despite significant advancements, several challenges remain. Autoregressive models rely heavily on visual tokenizers that compress pixel-level motion information, often at the cost of losing fine-grained details critical for realistic motion generation. Additionally, the causal attention mechanism inherent to autoregressive models complicates token sequencing, frequently necessitating large datasets to achieve generalization and diversity, as shown in Fig. 8a. BAMM [145] addresses these issues to some extent by introducing a novel framework that captures rich bidirectional dependencies and learns a probabilistic mapping between text inputs and motion outputs. Here, unification can help by combining the fine-grained generation of diffusion models with the semantic richness of autoregressive models.

5.3. Unify AR and diffusion model capability

To address these limitations, a unified framework combining the strengths of diffusion and autoregressive (AR) models is proposed, as shown in Fig. 7a. While diffusion models excel at generating high-quality and temporally consistent motion, they often lack explicit controllability compared to AR models, which generate sequences step-by-step based on prior context. Integrating these paradigms enables improved balance between motion realism and semantic control. Existing approaches often use multimodal LLMs as controllers, with diffusion models handling motion generation; however, such designs may falter when text conditions fail to capture fine-grained motion details. A unified approach helps bridge the gap between text-driven guidance and high-fidelity motion generation, improving robustness across diverse scenarios.

A more advanced unification strategy [146] involves training both diffusion models and multimodal LLMs within a shared, high-dimensional embedding space, as shown in Fig. 8a. Compared to loosely

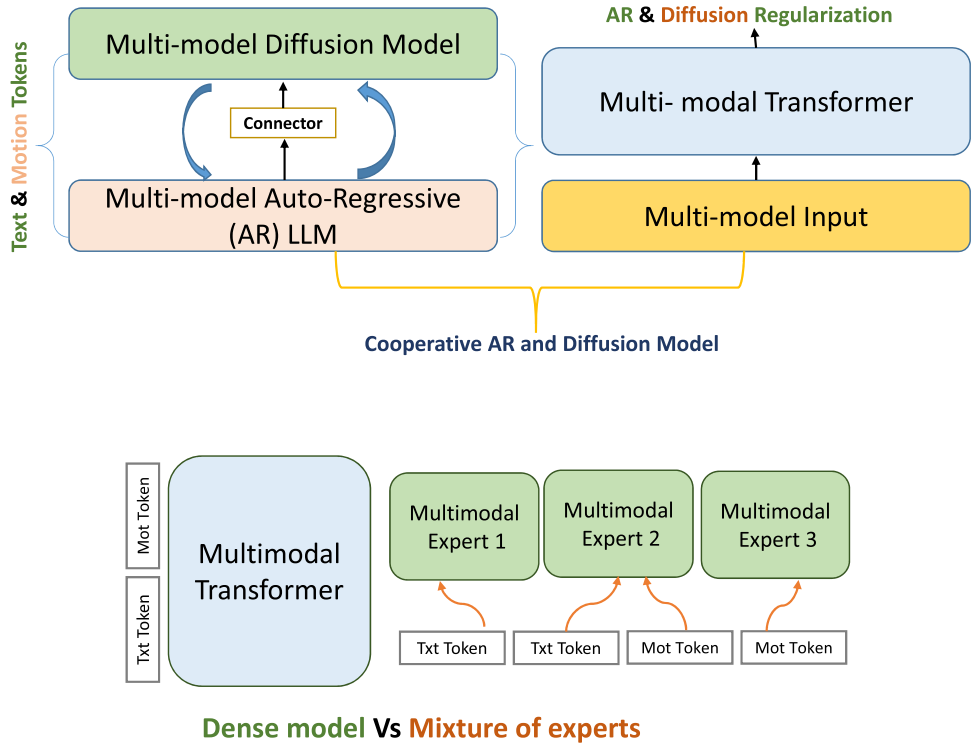


Fig. 7. An illustration of **a)** the cooperative capabilities of an AR and Diffusion Model in a single unified framework, and **b)** the Dense versus Mixture of Experts alignment methodologies.

coupled pipelines, this shared representation enables tighter alignment between modalities, facilitating both motion generation and understanding from multimodal embeddings. For example, transformer-based multimodal frameworks such as AMD [147] and its improved variant AAMD [148] jointly process text and motion using full attention mechanisms, resulting in better cross-modal coherence than earlier sequential or modular designs. Similarly, MotionVerse [149] introduces a Large Motion Model (LMM) with ArtAttention, combining a multimodal transformer with a diffusion backbone (Fig. 8d), and demonstrates improved generalizability across diverse tasks and datasets. In comparison, methods such as M2D2M [85] and DART [150] focus more on enhancing long-text understanding and motion controllability, highlighting a trade-off between scalability to complex language inputs and fine-grained motion synthesis. Collectively, these unified frameworks indicate that tightly coupled multimodal architectures enable more effective joint generation and reasoning, although challenges such as alignment consistency and computational complexity still persist.

5.4. Unified model architecture

The unified model architecture is designed to address multiple tasks, including comprehension, planning, understanding, reasoning, captioning, and motion generation. This architecture diverges from single-task approaches by incorporating joint pixel-level and semantic information, which enables simultaneous improvements in accuracy, generalization, and diversity across tasks, thereby overcoming limitations in current theoretical concepts and visual motion generation techniques.

5.4.1. Multimodal input architecture

As shown in Fig. 7, text is tokenized using standard methods, while motion data requires more specialized processing. Compared to approaches that rely solely on visual encoders for pixel-level inputs, a hybrid design that combines pixel-level and semantic encoders enables richer representations by capturing both low-level motion details and high-level contextual semantics. This dual-token strategy supports adaptive token selection, which is particularly beneficial in tasks requiring both fine-grained motion synthesis and semantic alignment. For example, in text-to-motion generation, semantic tokens guide overall action intent, while pixel-level tokens preserve detailed joint dynamics, resulting in more coherent outputs. Incorporating a mixture-of-experts mechanism further enhances this process by selectively activating relevant experts based on input modality or task complexity, improving efficiency compared to static encoding schemes. This unified design therefore enables more flexible and computationally efficient multimodal processing, although it introduces additional complexity in expert coordination and routing.

5.4.2. Multimodal transformer architecture

The multimodal transformer captures complex relationships between text and motion modalities [93,103,151]. As shown in Fig. 7b, two configurations are explored. The first uses a unified dense transformer that jointly models text and motion tokens, enabling shared representation learning across modalities. The second adopts a mixture-of-experts (MoE) design, where specialized experts focus on distinct functions such as semantic alignment, motion synthesis, and temporal refinement. A lightweight gating mechanism dynamically routes tokens to relevant experts, enabling conditional computation and improving efficiency compared to fully activated dense models. This comparison highlights a trade-off between simplicity in unified models and scalability/specialization in MoE-based designs.

Overall, this section highlights a unified multimodal framework for HMUG that integrates transformer backbones with diffusion and LLM-based paradigms, as illustrated in Fig. 8 and Table 4. Compared to task-specific or modality-isolated architectures, such unified designs demonstrate improved cross-modal alignment and computational efficiency by enabling shared representation learning across both understanding and generation tasks. This consolidation reduces architectural redundancy while enhancing generalization across diverse motion synthesis and reasoning scenarios.

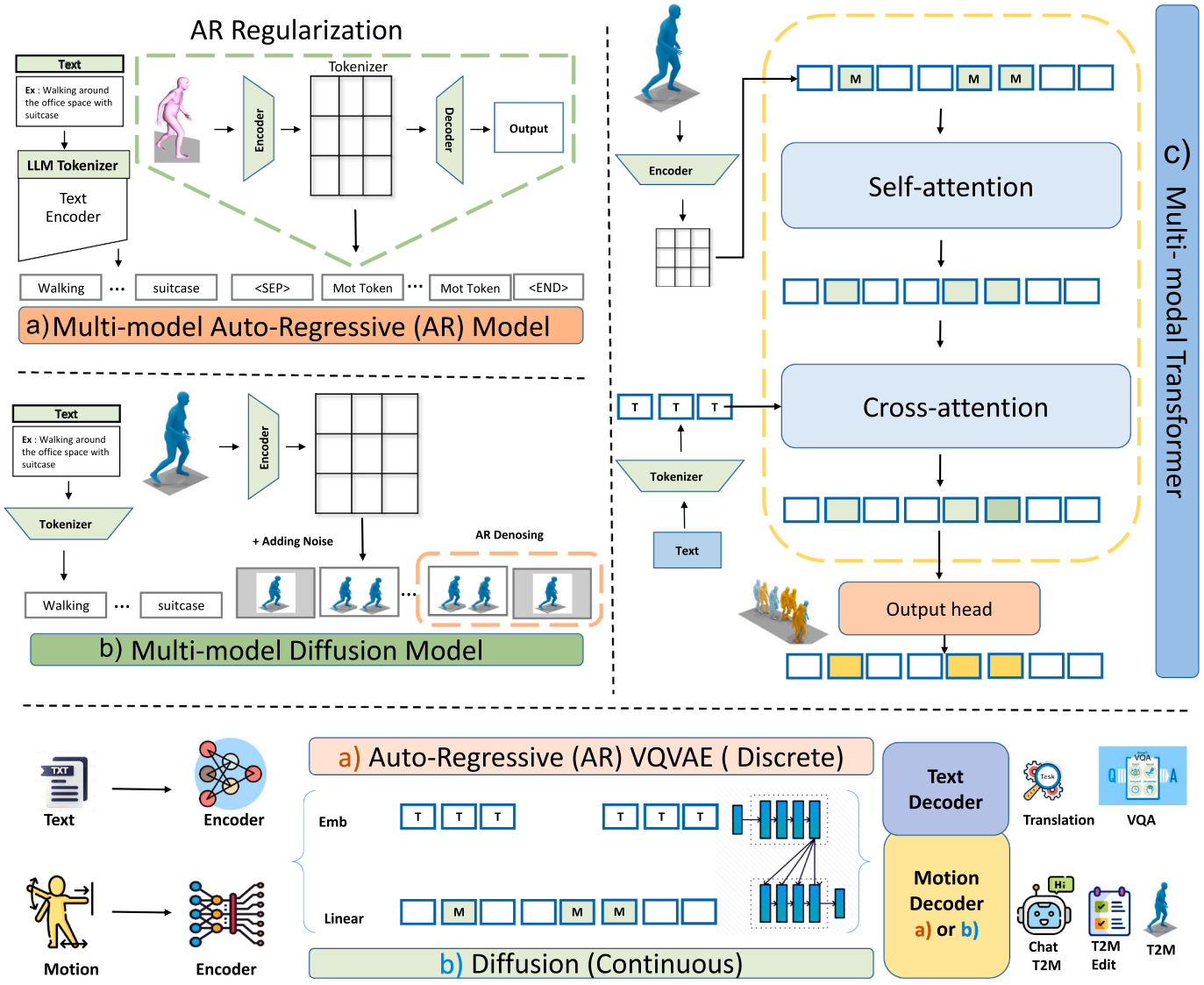


Fig. 8. A detailed illustration of **a)** multimodal LLM, **b)** multimodal Diffusion, **c)** multimodal Transformer, and unified approaches with an encoder-decoder backbone architecture.

6. Datasets and evaluation metrics

Following an in-depth discussion of multimodal **HMUG** architectures and their use cases, it is crucial to emphasize the role of datasets and evaluation metrics in implementing GenAI models. This section explores commonly used text-to-motion datasets and evaluation metrics, structured into two subsections: Datasets and Evaluation Metrics. Furthermore, many multimodal large models utilize these foundational datasets to construct task-specific datasets for downstream applications, such as motion captioning, conversational motion modeling, reasoning, and VQA, thereby enhancing their generalization capabilities. Table 6 provides a summary of the key properties of these datasets, along with prominent state-of-the-art evaluation criteria, including newly developed datasets specifically designed for text-to-motion generation.

6.1. Benchmark datasets

To evaluate text-to-motion models, researchers commonly rely on benchmark datasets that provide paired motion and textual descriptions, motion-only sequences, and unified datasets that integrate multiple modalities. These datasets offer standardized, high-quality data for both training and quantitative evaluation of multimodal generative models.

6.1.1. Text and motion pairs data

KIT-ML. KIT Motion Language (KIT-ML) [155] is a well-structured paired dataset designed for the integration of motion and text modalities. The motion data are captured using advanced optical marker-based systems, ensuring high precision in motion representation. Each motion sequence is paired with detailed language annotations that describe the associated movement. These annotations provide rich textual descriptions, making the dataset highly suitable for tasks such as motion understanding, generation, and text-to-motion mapping. KITML plays a crucial role in bridging the gap between natural language processing and motion analysis, enabling researchers to effectively develop and evaluate multimodal GenAI models.

BABEL. The Bodies, Action, and Behavior with English Labels (BABEL) dataset [133] extends the comprehensive AMASS [156] motion dataset by providing rich text annotations for motion sequences. These

Table 4
Summary of related works on Unified text to human motion understanding, generation, and other subtasks. This table presents key details of recent studies, including their input modalities, methods, backbone models, motion representations, datasets, tasks, and applications, open-source availability, and publication dates.

Paper Title	Link	Method	Venue	Input Modality	Backbone Model	Motion Representation	Dataset	Tasks Applications	Open Source	Online Date
A Unified 3D Human Motion Synthesis Model via Conditional VAE	[152]	Unified 3D, Human-Motion	ICCV	Text, Motion	Conditional Variational Auto-Encoder	2D, 3D	Human 3.6M, CMU Mocap	Human pose series generation	✗	28 Feb 2022
Pretrained Diffusion Models for Unified Human Motion Synthesis	[151]	Mo-Fusion	CVPR	Text, Motion, Music	Transformer, DDPM, DistilBERT	3D Rot.	KIT, AMASS, AIST++, LaFAN1, BABEL	Interactive motion editing and generation	✓	06 Dec 2022
UDE: A Unified Driving Engine for HMG	[153]	UDE Engine	CVPR	Text, Audio, Motion	VQ-VAE, Diffusion Decoder	3D Rot.	Human ML3D, AIST++	Motion quantization and generation	✗	24 June 2023
A Unified Framework for Multimodal, Multi-Part Human Motion Synthesis	[154]	Unified Motion	ArXiv	Text, Motion, Speech, Music	Hierarchical VQ-VAE, HuBERT, Transformer, ED	3D Rot.	Human ML3D, AIST++, BEAT, MANO	Multi part human motion generation	✓	28 No. 2023
MotionLLM: Multimodal Motion Language Learning with LLMs	[64]	Motion LLM	ICLR	Text, Motion	RVQ, LLaMA, LoRA, Gemma, AR Learning	3D Rot.	AIST++, Human ML3D	Single and multi-human motion generation	✓	27 May 2024
AAMDM: AR Motion Diffusion Model	[148]	AAMDM	CVPR	Text, Motion	AR, Motion Diffusion Model	3D Rot.	LaFAN1, AAMDM	Human motion generation and polishing	✗	21 Aug 2024
LMM for Unified Multimodal Motion Generation	[149]	LMM, MotionVerse, ArtAttention	ECCV	Text, Audio, Motion	Transformer, Diffusion Model	3D Rot.	Human ML3D, AMASS, 3DPW	Motion Generation, Prediction	✓	26 Oct 2024
MotionGPT2: A General Purpose Motion Language Model for Motion Generation and Understanding	[116]	Unified Motion GPT	ArXiv	Text, Motion	VQ-VAE, LLM, LoRA, T5, ED	3D Rot.	Human ML3D, KIT-ML, MotionX	Human body-controlled motion generation	✗	29 Oct 2024
MotionLLaMA: A Unified Framework for Motion Synthesis and Comprehension	[145]	Motion LLaMA	ArXiv	Text, Audio, Motion	RVQ, LLaMA, LoRA, AR Learning	3D Rot.	AIST++, Human ML3D, InterGen	Human motion synthesis and comprehension	✓	26 No. 2024
M2D2M: Multi-Motion Generation from Text with Discrete Diffusion Models	[85]	DDM, VQ-VAE	ECCV	Text, Motion	Transformer, Diffusion, ED	3D Rot.	Human ML3D, KIT-ML	Single and multimodal human motion generation	✗	05 Dec 2024

annotations are available at two levels of granularity: sequence-level labels, which describe entire motion sequences, and frame-level labels, which provide insights into specific motion frames. BABEL encompasses over 28,000 motion sequences and 63,000 individual frames, spanning 250 diverse motion categories. This dual-level labeling makes BABEL an invaluable resource for tasks such as motion segmentation, classification, and text-to-motion generation, enabling more nuanced understanding and generation of human motion in GenAI models.

HumanML3D. Human Motion Language 3D (HumanML3D) [157], a widely utilized dataset, is created by combining the HumanAct12 [158] and AMASS [156] datasets. It enriches motion sequences by associating each with three distinct text descriptions, enabling a deeper connection between textual inputs and 3D human motion. The dataset spans a diverse array of activities, including daily routines, sports, acrobatics, and artistic expressions. This diversity makes HumanML3D a valuable resource for advancing research in text-to-motion generation, offering robust benchmarks for aligning multimodal representations of text and motion.

6.1.2. Motion only data

UESTC. UESTC [153] is a comprehensive dataset that captures motion data across three modalities: RGB videos, depth images, and skeleton sequences, all collected using the Microsoft Kinect V2 sensor. It is organized into 40 distinct action categories, with 15 categories that can be performed in both standing and sitting positions and 25 categories exclusive to standing. The multimodal nature of UESTC makes it a valuable resource for tasks involving motion analysis, understanding, and generation, as it provides diverse representations of human actions in varying positions and contexts. Its detailed categorization supports the development and evaluation of advanced multimodal models for human motion.

NTU-RGB+D 120. The NTU-RGB+D 120 dataset [159] is an extended version of the widely-used NTU-RGB+D dataset [160], created to support more comprehensive research in **HMUG**. This extension introduces 60 additional action classes, bringing the total to 120 distinct categories that span a diverse range of activities, including daily routines and health-related tasks. Furthermore, it adds 57,600 new RGB+D video samples, significantly increasing the dataset's size and variety. Captured from multiple camera views and featuring subjects from diverse demographics, NTU-RGB+D 120 serves as a benchmark dataset for tasks such as action recognition, motion generation, and multimodal reasoning, providing a robust foundation for generalization and real-world applications.

HumanAct12. The HumanAct12 dataset [158], derived from the PHSPD [161], offers a curated collection of 3D motion clips, specifically designed to capture a diverse range of human actions. This dataset is structured around 12 main motion classes, such as walking, running, sitting, and warming up, which are further divided into 34 detailed sub-

classes to represent nuanced variations of these behaviors. Each motion clip is carefully segmented, yielding high-quality, detailed motion data suitable for tasks such as motion understanding, generation, and action recognition. HumanAct12 is particularly valuable for applications focused on everyday human activities, serving as a benchmark for evaluating the capabilities of GenAI and multimodal models in understanding and replicating realistic human motion.

6.1.3. Unified data

Several foundational datasets have also been extensively used in combination for text-to-motion research, each contributing unique strengths to advance the field. **Human3.6M** [162] is a large-scale motion capture dataset that features diverse human activities and accompanying annotations, making it instrumental in understanding motion dynamics. **CMU MoCap** [163] provides a rich repository of skeletal motion data, spanning a diverse range of activities and offering a robust foundation for motion modeling. **AMASS** [156] unifies over 15 motion capture datasets into a common parametric representation, enabling the creation of high-quality 3D motion sequences. **HuMMan** [164], on the other hand, is a more recent dataset designed to bridge the gap between **HMUG** by providing multimodal data including RGB, depth, and 3D motion. Together, these datasets provide a robust foundation for building large-scale, diverse, and multimodal text-to-motion benchmarks that support downstream tasks such as motion captioning, reasoning, and generative modeling.

To achieve generalization and enhance the diversity of generated motions, several works have utilized the fusion of multiple datasets. **HUMANISE** [165], for instance, integrates scene, object, and textual data, with a total of 643 scenes. The dataset offers a rich combination of visual contexts and text annotations, enabling models to understand and generate motion sequences informed by complex interactions between objects, scenes, and associated text. In addition, **TED-Gesture** [166] contains paired motion and text data, with a focus on gesture-based motion generation. The inclusion of gestures and corresponding text descriptions enables models to learn the mapping between textual prompts and specific movements, supporting tasks such as text-to-gesture motion generation. Building upon this, **TED-Gesture++** [167] and **PATS** [168] extend the fusion by introducing speech along with text and motion. These datasets enable the training of models to understand the impact of speech and text on motion, thereby enhancing the realism of generated gestures and motions in applications such as conversational AI and gesture recognition. **BEAT** [169] further complicates the relationship by incorporating emotional context along with text and speech, enabling models to capture how emotions can influence motion, such as body language or facial expressions. This multimodal approach, using datasets with diverse combinations of motion, text, speech, and emotion, enables large multimodal models to generalize more effectively and produce more sophisticated, contextually aware motion sequences, leading to improved performance on text-to-motion generation tasks.

6.2. Evaluation metrics

After discussing the datasets, it is essential to highlight the role of evaluation metrics, which are crucial for benchmarking the performance and quality of generative models in **HMUG**. These metrics serve to standardize comparisons across methodologies, enabling the field to progress. Assessing synthesized human motion is uniquely challenging due to its subjective nature, the intricacies of high-level conditional signals, and the one-to-many mapping between input and output. Broadly, evaluation focuses on aspects such as **fidelity**, **realism**, **diversity**, **coherence**, and **retrieval** performance. Metrics such as the Fréchet Inception Distance (**FID**) are widely used to assess distributional similarity between generated and real datasets. Table 5 presents the various evaluation metrics used in human motion research, along with their interpretation, hyperparameters, bounds, and categories.

To assess structural accuracy and motion quality, metrics such as Mean Per Joint Position Error (**MPJPE**) are used. Retrieval metrics, such as R-Precision, are used to evaluate the model's ability to accurately retrieve data. Perceptual evaluations, such as text-to-motion accuracy, provide insights into the alignment between generated motion and conditional input signals.

Diversity metrics, including Coverage, Recall, and Multimodality Score (**MS**), assess the variety and novelty of generated samples, ensuring the model captures diverse modes of human motion. Despite the diverse array of metrics available, establishing a unified evaluation framework for human motion synthesis remains a critical challenge [58]. For clarity, the evaluation metrics can be categorized based on their primary purposes in assessing generative models, such as fidelity, diversity, accuracy, and motion quality.

6.2.1. Fidelity metrics

These metrics assess how closely the generated motion matches the real-world data, focusing on realism and naturalness:

Fréchet Inception Distance (FID). FID is a widely used metric for evaluating the quality of generative models. It measures the similarity between the real and generated data distributions by calculating the Fréchet Distance between their feature representations in a latent space [12,78,89,144]. These features are typically extracted using a pretrained neural network.

Proposed by Heusel et al. [170], **FID** improves upon the Inception Score [12] by directly comparing the distributions of real and generated samples. It assumes that both distributions are multivariate Gaussian and computes their means and covariances. The FID score quantifies the effort, or "energy", required to transform one distribution into the other, with lower FID indicating higher fidelity in the generated samples. Mathematically, FID is computed using:

$$\mathrm{FID}(P_1, P_2)^2 = \mathrm{trace}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}) + \sum_{i=1}^{f}(\mu_{1,i} - \mu_{2,i})^2, \tag{14}$$

where μ_1, μ_2 are the means and Σ_1, Σ_2 are the covariance matrices of the feature representations of real and generated samples, respectively.

FID captures both the fidelity (how realistic the generated samples are) and diversity (how well they cover the real data distribution). However, it has limitations: a perfect FID score of 0 can be achieved if the model merely replicates the real data without producing novel samples. This underscores the importance of interpreting FID in conjunction with other metrics or qualitative evaluations.

Precision (R-Precision). It evaluates how well the generated samples conform to the real data distribution, with a focus on fidelity. In the context of **HMUG**, it measures the proportion of generated motions that are close to the real motions in the feature space [79,81,84,91,100,136,139]. Higher precision indicates that the AI-generated motions are realistic and align well with the characteristics of the real data.

Mathematically, precision is often computed using a neighborhood-based approach, in which a set of real samples and generated samples is compared in the latent space. For a given generated sample, if it falls within the neighborhood of any real sample (e.g., using k-nearest neighbors), it is considered precise. The metric is formally expressed as:

$$\mathrm{Precision} = \frac{\text{Gen samples in real neighborhood}}{\text{Total gen samples}}. \tag{15}$$

In motion generation, precision helps determine whether the model produces high-quality, plausible motions that are consistent with real human motion. However, it does not assess diversity, meaning a high precision score could still occur if the model generates highly realistic but repetitive motions.

6.2.2. Diversity metrics

These metrics assess the variety and novelty of the generated samples, ensuring that models do not suffer from mode collapse. In the context of **HMUG**, diversity metrics are crucial for evaluating whether the

Table 5
Summary of datasets and evaluation metrics used in various top-tier research with the backbone of Autoregressive LLMs, diffusion, and unified models for multimodal human motion understanding, planning, and generation, published in top-tier venues.

Metric	Category	Hyper-Parameters	Bounds	Interpretation	Used in Research
FID	Fidelity	None	$0 \leq \text{FID} < \infty$	Lower is better; compares feature distributions with real data	[12,78,89,144,170]
Precision	Fidelity	T Nearest points	$0 \leq \text{Precision} \leq 1$	Closer to 1 indicates better fidelity	[79,81,84,91,100,136,139]
Coverage	Diversity	T Nearest points	$0 \leq \text{Coverage} \leq 1$	Higher indicate better diversity	[5,64,65,78,89,91,124,135,136,142,144]
Recall	Diversity	T Nearest points	$0 \leq \text{Recall} \leq 1$	Higher indicate better diversity	[17,101,127,139]
MM Score	Diversity	T2M Accuracy	$0 \leq \text{MMS} \leq 1$	Higher indicate better multimodality representation	[79,81,82,84,86,91,100,102,134,136,139,171,172]
MM Distrib.	Diversity	T2m Consistency	Depends on stat measures	Measures distribution similarity across modalities	[17,82,84,86,102,134,150,153,169,171–177]
MPJPE	Accuracy	None	$0 \leq \text{MPJPE} < \infty$	Lower indicate better accuracy in joint position estimation	[83,92,122,141,147,148]

generated motions capture the full range of natural variations present in real human motion [5,64,65,78,89].

Coverage. It measures how well the generated motions represent the diversity of the real motion data. It evaluates whether all major patterns or variations in real motion are reflected in the generated samples [91,124,135,136,142,144]. Mathematically, coverage is often computed using a neighborhood-based approach in the latent space, where the proportion of real samples that are covered by the generated samples is calculated:

$$\text{Coverage} = \frac{\text{Real samples in generated neighborhood}}{\text{Total real samples}}. \tag{16}$$

Higher coverage implies that the generative model captures a wider variety of real motion, ensuring that all significant motion patterns are adequately represented.

Recall. This metric quantifies the percentage of real motion samples that fall within the generated data distribution. In the context of motion generation, it measures the model's ability to recreate the diverse features of real human motion. Recall is calculated as:

$$\text{Recall} = \frac{\text{Real samples near generated}}{\text{Total real samples}}. \tag{17}$$

Closeness is typically measured using a distance threshold in a latent feature space (e.g., L2 distance or cosine similarity). A high recall score reflects the model's ability to reproduce the diversity observed in real motion data. This metric is also used to evaluate a model's understanding and retrieval of human motion [17,127].

Multimodality Score (MMS). The MMS evaluates the extent of distinct modes or patterns present in the generated motions [79,81,82,84,86,91,100,102,134,136,139,171,172]. It ensures that the model generates a variety of motion types, styles, or trajectories rather than repetitive or overly similar ones. Mathematically, MMS is computed as:

$$\text{MMS} = \frac{1}{K}\sum_{k=1}^{K} \frac{\text{Unique Modes in Generated Motion}\,(G_k)}{\text{Modes in Real Motion Data}\,(R_k)}, \tag{18}$$

where K represents the number of motion categories or clusters, such as action types like walking, running, or dancing. For each motion category k, G_k denotes the unique modes in the generated motion, referring to the distinct motion patterns produced. Similarly, R_k refers to the modes in real motion data, capturing the number of distinct motion patterns observed in the real-world dataset for the k^{th} category.

Distinct modes are identified based on clustering in the latent space or feature embeddings of the generated motions. Higher MMS indicates that the generative model effectively captures the diversity of motion categories, which is crucial for applications that require varied, human-like motions. The detailed metrics comparison along with their interpretation is shown in Table 5.

6.2.3. Accuracy and consistency metrics

These metrics assess the degree to which the generated human motion aligns with the conditional inputs, such as text descriptions, audio, or other modalities. They are essential for assessing the quality of text-to-motion generation and ensuring that generated motions are both accurate and consistent with the input conditions.

Condition Consistency (Text-Motion Accuracy). Condition consistency measures the accuracy with which the generated motions match the corresponding textual descriptions. For example, if a text prompt describes "a person walking slowly", the generated motion should visually and dynamically resemble slow walking. This metric is often computed by comparing the embeddings of the text and motion in a shared feature space, such as those learned by a multimodal model [17,18,75,79–81,84,91,100,101,107,124,128,135,136,139,142,144,150,167] . Higher consistency scores indicate that the model effectively understands and translates textual inputs into motion.

Text-Motion Consistency. Text-motion consistency evaluates the coherence between the generated motions and their corresponding textual descriptions over the entire sequence. While condition consistency focuses on matching the input description to generated motions, text-motion consistency ensures that the generated motions remain semantically aligned with the text throughout the motion sequence. This can be assessed by calculating a similarity score between text and motion embeddings [17,82,84,86,102,134,150,153,169,171–177], ensuring the generated motion reflects the text in a temporally consistent manner.

R-Precision. It quantifies the ranking accuracy of retrieved motion samples against a text query [127]. It measures whether the generated motion is ranked among the top R relevant motions for a given textual prompt in a dataset. For example, if $R = 5$, R-Precision checks if the correct motion appears in the top 5 most relevant results. Mathematically, it can be expressed as:

$$\text{R-Precision} = \frac{\text{Relevant motions in top-}R}{R}. \tag{19}$$

R-Precision is particularly useful in retrieval-based evaluation settings, where the goal is to assess the alignment between the textual query and the retrieved motion samples. Higher R-Precision indicates better alignment and ranking accuracy [17], indicating that the model generates or retrieves motions most relevant to the input query.

Table 6
Comprehensive quantitative comparisons based on two important datasets: HumanML3D and KIT-ML. The table also describes the backbone architecture, tasks, and subtasks. It highlights results including FID, Precision (Prec), Diversity (Div), and MM metrics.

Types	Method	Data Representation	Backbone	Tasks	Open-source	Venue	HumanML3D [157]				KIT-ML [155]			
							FID	Prec	Div	MM	FID	Prec	Div	MM
Multimodal LLM	MotionLLM [64]	3D	Enc-Dec	Motion Generation, Captioning, Multi-turn Editing, Reasoning, Composition	✓	ICLR	0.230	0.515	9.908	2.967	0.438	0.439	11.151	2.872
	AvatarGPT [74]	3D Latent	Autoregressive	Motion Generation, Understanding, Summarization	✓	CVPR	0.168	0.510	9.624	–	0.376	–	–	–
	MotionGPT-3 [104]	2D, 3D	Enc-Denoiser	Retrieval, Generation	✓	CVPR	0.567	0.411	9.006	3.775	0.597	0.376	10.540	3.394
	MotionGPT-2 [116]	3D	Enc-Dec	Motion Generation, Captioning, Prediction	✗	ArXiv	0.232	0.492	9.528	3.096	0.510	0.366	10.350	2.328
	T2M-GPT [78]	3D Latent	Autoregressive Enc-Dec	Reconstruction, Generation	✓	CVPR	0.116	0.491	9.761	1.856	0.514	0.416	10.920	1.570
	MotionChain [117]	3D	Enc-Dec	Generation, Reasoning, Translation	✗	ECCV	0.248	0.504	9.470	1.715	–	–	–	–
	WalkLLM [146]	Keypts, 2D	Enc-Dec	Pedestrian Motion Generation	✗	IVS	0.197	–	8.157	–	–	–	–	–
	FineMoGen [113]	3D	Enc-Dec	Generation, Editing	✓	NeurIPS	0.151	0.504	9.263	2.696	0.178	0.432	10.850	1.877
	AlertMotion [75]	3D	Enc-Dec	Adversarial Similarity, Naturality	✗	ArXiv	4.170	5.400	5.680	–	–	–	–	–
	TAAT [107]	3D	Enc-Dec	Adversarial Similarity, Naturality	✗	ArXiv	0.461	0.329	10.038	2.929	0.488	0.225	8.552	2.957
Multimodal Diffusion	ActFormer [94]	2D, 3D	Enc-Denoiser	Single/Multi-person Gen.	✓	ICCV	0.130	2.550	–	–	–	–	–	–
	LS-GAN [100]	3D	Enc-Dec	Motion Synthesis	✗	ArXiv	0.482	0.391	9.249	3.501	–	–	–	–
	VQ-VAE Mot [124]	Keypts, 2D	Enc-Dec	Motion Recon.	✗	ICMEW	0.017	0.505	9.532	3.005	0.251	0.417	10.750	3.143
	VQ-VAE DLS [89]	3D	Enc-Dec	Motion Gen.	✓	CVPR	0.141	0.492	9.722	1.831	0.514	0.416	10.921	1.570
	MotionDiffuse [5]	2D, 3D	Enc-Denoiser	Text-to-Motion	✓	TPAMI	0.630	0.491	9.410	1.553	0.514	0.416	10.921	1.570
	Tevet [77]	2D, 3D	Enc-Dec	Interp., Editing, Recog.	✓	ECCV	4.569	0.645	7.688	1.264	0.497	0.396	10.847	1.907
	ReMoDiffuse [127]	3D	Enc-Dec	Hybrid Retrieval	✓	ICCV	0.103	0.510	9.018	1.795	0.155	0.427	10.800	1.239
Unified Model	Unify Motion [152]	3D Rot.	Enc-Dec	Multi-Part Generation	✓	ArXiv	1.167	0.548	14.720	2.010	–	–	–	–
	Unify MoGPT [151]	3D Rot.	Enc-Dec	Body-Controlled Gener.	✗	ArXiv	0.191	0.496	9.860	2.137	0.614	0.427	11.256	2.357
	MotionLLaMA [145]	3D Rot.	Autoregressive	Synthesis and Comprehen.	✓	ArXiv	0.1133	0.4926	20.343	1.0637	–	–	–	–
	UDE Engine [153]	3D Rot.	Enc-Dec	Motion Refinement	✗	ECCV	2.670	8.210	6.750	2.340	-	–	–	–

Motion Quality Metrics. These metrics evaluate the realism, smoothness, and plausibility of the generated human motion sequences. They focus on ensuring that the generated motions not only match the intended task or context but also appear natural and fluid, closely resembling human-like movements.

Motion Quality. Motion quality encompasses both subjective and objective evaluations of the naturalness and fluidity of generated motion sequences. Subjective evaluations rely on human judges rating the plausibility [178] of the motion, whereas objective evaluations analyze specific motion attributes, such as velocity, acceleration, and joint coherence [85]. High-quality motions exhibit smooth transitions, realistic joint movements, and adherence to biomechanical principles.

Mean Per Joint Position Error (MPJPE). It is an objective metric that quantifies the average error between the joint positions of the generated and ground-truth motions across all frames. It measures how closely the generated motion replicates the true joint positions in space, indicating motion accuracy. The formula for MPJPE is:

$$\text{MPJPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J}\sum_{j=1}^{J}\left\|p_{t,j}^{\text{gen}} - p_{t,j}^{\text{gt}}\right\|_2, \tag{20}$$

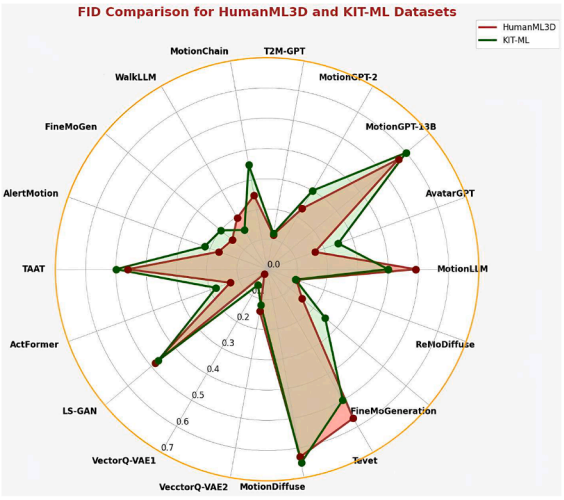
where T represents the number of frames in the motion sequence, and J denotes the number of joints in the skeleton. The term $p_{t,j}^{\text{gen}}$ refers to the position of the jth joint at frame t in the generated motion, while $p_{t,j}^{\text{gt}}$ denotes the position of the jth joint at frame t in the ground-truth motion.

A lower MPJPE value indicates higher fidelity [83,92,122,141,147,148] and closer alignment [139] between the generated motion and the ground-truth motion.

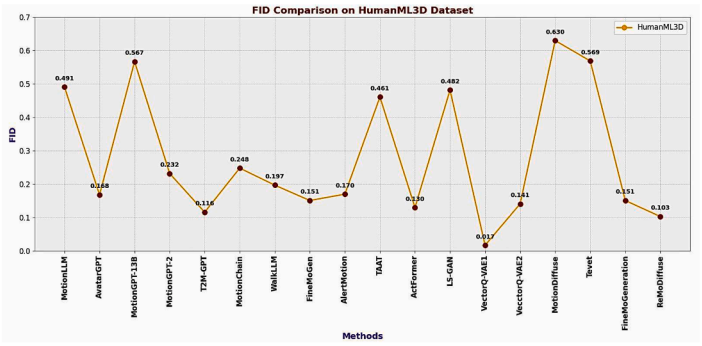
6.3. Evaluation trade-offs and need for unified metrics

Human motion generation models are commonly evaluated along multiple, often competing, dimensions, including motion realism, diversity, and computational efficiency. As recent methodologies increasingly span multimodal LLM-based, diffusion-based, and unified architectural paradigms, understanding the inherent trade-offs among these criteria becomes essential for fair and meaningful comparison. This subsection analyzes these trade-offs using standardized quantitative results across benchmark datasets.

Diversity-Fidelity-Efficiency Trade-offs Table 6 reveals a consistent trade-off between motion fidelity, diversity, and computational efficiency across multimodal LLM-based, diffusion-based, and unified human motion generation paradigms. Fidelity-oriented metrics such as FID and Precision generally favor models with strong structural priors or au-



(a) FID comparison on HumanML3D and KIT-ML.



(b) FID-based comparison of different methodologies on HumanML3D.

Fig. 9. Comparative analysis of text-to-motion generation methods based on the **FID** metric across benchmark datasets.

toregressive decoding strategies, whereas diversity metrics highlight the strengths of stochastic generative mechanisms.

Multimodal LLM-based approaches, including T2M-GPT [78], FineMoGen [113], and MotionChain [117], typically achieve competitive FID and Precision scores on both HumanML3D and KIT-ML, indicating strong semantic alignment between textual descriptions and generated motions. However, their diversity scores remain comparatively moderate, reflecting the tendency of encoder-decoder or autoregressive pipelines to prioritize semantic correctness over motion variability. From an efficiency perspective, these models benefit from single-pass or token-level decoding, making them relatively suitable for interactive and real-time scenarios.

In contrast, diffusion-based methods such as MotionDiffuse [5], Tevet [77], and ReMoDiffuse [127] demonstrate stronger diversity performance, particularly on KIT-ML, where stochastic denoising enables broader exploration of the motion space. This increase in diversity often comes at the expense of higher FID or reduced Precision, indicating a trade-off between variability and fine-grained realism. Moreover, the iterative nature of diffusion inference introduces higher computational overhead, limiting applicability in latency-sensitive deployment settings.

Unified architectures aim to balance these competing objectives. Models such as MotionLLaMA [145] and Unify Motion [152] exhibit substantially higher diversity scores, indicating strong capacity for multi-part and multimodal motion synthesis. However, this expressiveness may be accompanied by degraded fidelity or increased architectural complexity. While unified frameworks conceptually enable shared representations and cross-task generalization, their practical efficiency depends heavily on backbone design and inference optimization strategies.

These trends are further illustrated in Fig. 9a, which compares FID-based performance across HumanML3D and KIT-ML, and Fig. 9b, which provides a detailed FID analysis on the HumanML3D dataset. Together with Table 6, these results demonstrate that no single modeling paradigm simultaneously optimizes fidelity, diversity, and computational efficiency.

Importance of Unified Evaluation Metrics The observed trade-offs across modeling paradigms highlight the limitations of relying on isolated evaluation metrics. Current practices often adopt inconsistent metric definitions, preprocessing pipelines, or evaluation protocols, making cross-model and cross-dataset comparisons difficult. As emphasized in [58], unified evaluation metrics are essential to ensure consistency, reproducibility, and fairness across diverse motion generation tasks and representations. Furthermore, [179] advocates for evaluation criteria that better align with human perceptual judgments in order to holistically assess motion realism, naturalness, and semantic coherence.

Given the diversity-fidelity-efficiency trade-offs evident in Table 6, Fig. 9a, and b, a unified evaluation framework that jointly accounts for motion quality, diversity, semantic alignment, and computational cost is urgently needed. Such a framework would enable more transparent benchmarking, facilitate fair comparison between competing paradigms, and support the development of deployment-ready human motion generation systems.

7. Potential applications

In recent years, advancements in GenAI, particularly Multimodal LLMs and Diffusion Models, have transformed the landscape of **HMUG**. With the emergence of the metaverse, these cutting-edge technologies have become integral to creating realistic, dynamic virtual humans in metaverse space, seamlessly integrating motion, text, and other modalities. Their impact is most pronounced in augmented reality (**AR**), virtual reality (**VR**), and the **metaverse**, where they enable immersive, interactive digital experiences. The scope of applications extends beyond entertainment to domains such as healthcare, education, and training simulations, underscoring the transformative potential of multimodal human motion generation. Leading technology companies, including **Nvidia, Meta, Google, Microsoft, Samsung, Epic Games, Tencent, and Digital Domain**, are spearheading initiatives such as virtual avatars, digital humans, and “MetaHuman” platforms, further accelerating innovation in this field. Below are key application domains where multimodal human motion generation could play a transformative role:

7.1. Healthcare and rehabilitation

Therapeutic exercises can be enhanced by automatically generating tailored human motion sequences for physiotherapy and rehabilitation [180–183], addressing individual patient needs. Virtual simulations for gait analysis and correction provide a means to assess gait impairments and propose targeted exercises, facilitating personalized treatment [184–189]. Additionally, AI-driven assistive technologies are being developed to support patients with motor disabilities by providing

real-time motion predictions and training, further improving rehabilitation and daily functioning [190,191].

7.2. Virtual reality (VR) and augmented reality (AR)

VR and AR applications are significantly enhanced through the generation of realistic human movements that adapt seamlessly to user interactions, creating immersive experiences [192–194]. In gaming, this advancement enables non-player characters (NPCs) to exhibit more natural movements, thereby increasing player engagement [195,196]. Furthermore, realistic human motion simulations are utilized in training scenarios across various professional domains, such as medicine, sports, and military operations, to provide effective and practical skill development [196].

7.3. Entertainment and fashion industry

In the entertainment and fashion industries, (AI-powered human motion generation has revolutionized the integration of creativity and realism into digital content. In film and animation production [197], AI automates the generation of realistic human movements, thereby minimizing reliance on manual animation techniques and traditional motion capture, while enabling the creation of lifelike characters and immersive storylines. Game development benefits from diverse motion styles [198–200], which enhance realism and engagement in gaming experiences. Similarly, in fashion [201–203], virtual models driven by AI-generated motions deliver dynamic, runway-like presentations in virtual try-on applications [204,205] and digital fashion showcases, offering consumers an interactive, personalized experience that redefines retail and marketing strategies.

7.4. Robotics and human-robot interaction (HRI)

Robotics and Human-Robot Interaction (HRI) significantly benefit from advancements in multimodal human motion generation. By learning from human motion patterns, these technologies enhance robotic motion planning and execution, enabling humanoid robots to perform tasks more efficiently and naturally [206–210]. Additionally, they facilitate seamless collaboration between humans and robots by predicting and generating human-like movements, fostering more intuitive and effective interactions in both industrial and personal settings.

7.5. Security and surveillance

Advances in human motion generation are critical to enhancing security and surveillance systems. By analyzing and simulating human motion, these technologies enable accurate prediction of behavior and the detection of abnormal activities in real time [211,212]. Additionally, they facilitate forensic analysis by reconstructing human motion [213–215] from available evidence, thereby aiding investigations and improving the overall efficacy of security protocols.

7.6. Autonomous vehicles

Human motion generation is integral to improving the safety and efficiency of autonomous vehicles by enabling accurate prediction and simulation of pedestrian movements. This technology enables vehicles to anticipate human behavior in complex scenarios, thereby improving decision-making and enhancing collision avoidance. For instance, generative models have been used to simulate diverse pedestrian trajectories, enhancing the training of autonomous systems in real-world environments [118,216,217]. These advancements underscore the importance of human motion understanding in developing reliable and adaptive autonomous driving solutions.

Multimodal human motion generation is a transformative technology with applications spanning healthcare, entertainment, robotics, education, and beyond. By integrating advancements in machine learning, GenAI, and multimodal processing, this technology is paving the way for innovative solutions to complex challenges across industries. As the field advances, its potential to transform our interactions with digital and physical environments will continue to grow.

8. Challenges and future work

Despite significant progress in multimodal GenAI for human motion understanding and generation, several fundamental challenges remain in achieving efficient, scalable, and temporally coherent models across diverse modalities. Existing paradigms, including diffusion-based, autoregressive/LLM-based, and transformer-unified frameworks, exhibit complementary strengths but also introduce distinct limitations that hinder real-world deployment. Key challenges include computational inefficiency (e.g., inference time and FLOPs), modality-specific representation constraints, and difficulty in capturing long-range temporal dependencies. Addressing these limitations is essential for enabling real-time, high-fidelity, and cross-modal motion generation systems. This section outlines key research directions in **HMUG**, focusing on improvements in model efficiency, architectural design, and evaluation protocols.

Multimodal Architecture Level Challenges Multimodal LLMs with autoregressive probabilistic modeling, as shown in Fig. 8, have become increasingly dominant in visual understanding and generation. However, their effectiveness is constrained by challenges in visual tokenization, representation learning, and cross-task generalization. Existing tokenization approaches, such as VQGAN [96], face scalability issues with large codebooks, leading to inefficient multi-scale compression and increased inference latency. Although VAE-based variants, including VQ-VAE [81,87,88], RVQ-VAE [20,84,93], and VQ-Diffuse [85], partially mitigate these limitations, trade-offs between discrete and continuous representations persist. While continuous embeddings support smoother motion transitions, they often introduce training instability and slower convergence.

Furthermore, autoregressive architectures exhibit limited inductive bias for visual data, making them less effective without hierarchical tokenization or structured priors [8,10,147,150]. In contrast, multimodal LLMs show stronger generalization across tasks, whereas visual autoregressive models remain less adaptable to downstream motion understanding tasks. In addition, capturing long-range dependencies in human motion remains challenging, the development of more motivating more efficient causal or sparse attention mechanisms. Early fusion strategies, although effective for alignment, are computationally expensive for video-scale motion modeling, limiting real-time applicability. Alternative alignment architectures (Figs. 4 and 7) offer improved efficiency but still require further refinement for scalable deployment. Future research should focus on hierarchical visual tokenization, improved continuous-discrete hybrid objectives, structured priors for motion modeling, and task-adaptive pretraining strategies to enhance efficiency, generalization, and robustness in multimodal autoregressive frameworks.

Multimodal diffusion architectures [25], as shown in Figs. 6 and 8, achieve strong performance in human motion generation while mitigating mode collapse. However, their effectiveness is limited by high computational cost arising from iterative denoising, leading to increased training FLOPs and inference latency, which restricts real-time deployment. Additional challenges include scalability to long-sequence motion generation due to accumulated computational overhead and complex spatiotemporal dependencies, as well as difficulties in effectively integrating heterogeneous modalities such as text, vision, audio, and motion while preserving realism and structural consistency.

Future research should focus on improving the efficiency of latent diffusion models [11,98], particularly through reduced sampling steps and optimized inference strategies. Complementary directions include

model compression techniques such as pruning and distillation, as well as approximate inference to enable faster generation. While alternative generative paradigms such as GAN-based models offer faster single-pass generation, they generally lag behind diffusion models in stability and fidelity, suggesting potential value in hybrid architectures. More broadly, unifying diffusion, GAN, and VAE paradigms (Fig. 6) through hierarchical tokenization, physics-informed priors, and progressive generation strategies remains a promising direction for scalable multimodal motion synthesis.

A unified framework [58,145,151,153] offers significant advantages by leveraging the complementary strengths of the two architectures, as depicted in Fig. 8. Multimodal LLMs excel at understanding and generating structured sequences across text and motion, while diffusion models generate high-quality outputs with superior diversity and realism. Ensuring synchronization across different modalities, such as motion-text correspondence, remains difficult due to inconsistencies in data representations. Some works [85,148,149] employ the connector strategy, utilizing a dense or a mixture of expert models to address the synchronization issues illustrated in Fig. 7. The advanced strategy involves training both models in a shared high-dimensional embedding space. Large-scale multimodal transformers, as shown in Fig. 8c, help both models align and share a common space, enhancing comprehension and generation capabilities [35,93,176]. However, they also suffer from gradient instability and mode collapse, making training optimization a key research priority. The high computational cost is quantified and compared, and future work is proposed to explore lightweight and efficient transformer variants for real-time applications. Generating long-form video and motion sequences is another significant challenge, as most models are limited to short-term outputs. Future research should focus on refining the above strategies and cross-attention mechanisms, as well as adaptive training strategies such as reinforcement learning, and developing lightweight architectures to reduce inference costs. Dynamic conditioning methods and edge AI deployment are proposed as concrete solutions to improve efficiency.

Multimodal Datasets and Benchmarks The development of multimodal datasets and benchmarks for human motion generation faces several significant challenges. One of the main issues is the lack of large-scale, unified datasets that can effectively support cross-validation and evaluation across different tasks [17]. Current datasets often focus on either understanding or generation tasks, but do not integrate both aspects in a unified manner. This fragmented approach hinders the comprehensive assessment of models that simultaneously handle multiple tasks [132,155,159]. Additionally, data scarcity, particularly in human motion video generation, remains a significant challenge due to privacy concerns, Erroneous data quality, inconsistent human annotations and descriptions, and the high costs of data collection. Future work should focus on creating diverse, high-quality datasets that cover a wide range of human activities and include rich annotations across multiple modalities. Furthermore, designing unified benchmark datasets that combine understanding and generation tasks is critical for more accurate model evaluation, greater diversity, and improvements in this domain. Expanding data sources and addressing privacy concerns will be key to developing robust models that can generalize well across various applications.

Scene grounding and Fine-Grained Body Part Modeling Text-conditioned human motion generation has achieved significant progress; however, most existing methods remain largely scene-agnostic, generating motions without explicit awareness of physical environments. Compared to scene-aware approaches that incorporate environmental constraints, scene-agnostic models often produce physically implausible interactions. In practical settings, scene grounding is critical for ensuring physical plausibility, including collision avoidance, object interaction modeling, and stable foot-ground contact. Although recent works [26,218–222] introduce contact-aware losses, scene constraints, or physics-based post-processing, these components are typically heuristic and weakly integrated into end-to-end learning frameworks, limiting systematic evaluation and generalization. As a result, generated motions may preserve semantic consistency while violating physical constraints, restricting their applicability in robotics, AR/VR, and embodied interaction tasks.

In parallel, fine-grained body part modeling particularly hands and facial expressions remains under-explored despite its importance for expressive and semantically aligned motion generation [223–225]. Compared to full-body-focused methods, approaches incorporating hand- and face-aware modeling (e.g., SMPL-X-based pipelines [48]) achieve more expressive and realistic motion, but are less frequently adopted in large-scale text-to-motion frameworks due to data and annotation limitations. This limitation is further exacerbated by the lack of dedicated datasets and evaluation protocols targeting fine-grained motion fidelity and semantic alignment. Overall, advancing scene-aware motion generation alongside fine-grained modeling of hands and facial expressions is essential for physically realistic and expressive human motion synthesis in deployment-oriented applications.

Evaluation Metrics Recent efforts have advanced evaluation metrics, enabling more systematic assessment of models across domains [46,151]. However, compared to well-established benchmarks in vision and language domains, evaluation in human motion generation remains significantly more fragmented, with inconsistent metric definitions and reporting protocols [58]. While existing metrics such as FID-based and diversity-based measures provide useful insights, they often capture only isolated aspects of motion quality, leading to incomplete or non-comparable evaluations across studies. More robust and comprehensive evaluation protocols are therefore required to improve accuracy, reliability, and cross-paper comparability. In particular, fine-grained evaluation of body parts such as hands and facial expressions remains largely underexplored, despite its importance for perceptual realism and semantic alignment [46,151,153].

To this end, a unified evaluation metric [38] that integrates multiple criteria into a single coherent framework is urgently needed. Such a framework would enable consistent and fair comparisons across methods, akin to unified benchmarking practices in the language and vision domains, thereby improving reproducibility and transparency. Overall, existing efforts provide a foundational basis, but further standardization is needed to enable reliable benchmarking of human motion generation.

Applications. Human motion applications, particularly in lightweight multimodal GenAI, face several challenges that require innovative solutions. One major hurdle is the lack of photorealism, necessitating more natural, lifelike human motion [197,202], including accurate hand and body movements. Additionally, expanding the duration and refining the control of motion [190] generation, from short clips to longer, more dynamic videos with multi-person interactions, remains a significant challenge. Future advancements should prioritize improving the generation of long-duration human videos while also enhancing the semantic depth and interactivity of motion. Another critical direction is the integration of these models with emerging technologies, such as AR/VR [192,194] and robotics [118,216,217,226], which enables more immersive and interactive experiences. Furthermore, as AI continues to evolve, there is a need to adapt multimodal GenAI systems [25] to function effectively in dynamic environments, ensuring that human motion generation remains efficient, controllable, and real-time. These advancements have broad applicability across diverse fields, including fashion, health, autonomous vehicles, and education, where lifelike, controlled motion generation can enhance user experiences, improve training, and offer innovative solutions.

Ethical and Adversarial Effects Ethics and adversarial impacts pose significant challenges [107] in the development and deployment of motion models, especially with the rise of digital humanoids and virtual humans [108,109,113,114]. Ensuring fairness and ethical considerations is paramount, particularly regarding privacy and the responsible use of human-generated data. Providing privacy-free human image data can alleviate concerns about personal data protection, but it also raises issues regarding consent and the potential for misuse. Adversarial effects,

such as the manipulation or generation of misleading or harmful content [75], remain a significant concern, particularly when models are deployed in sensitive domains such as healthcare, education, or entertainment.

To address these challenges, future work should focus on creating robust ethical frameworks that govern the creation, use, and distribution of motion models. These frameworks should prioritize fairness, transparency, and accountability, while also ensuring that the generated content does not perpetuate bias or harm. Additionally, advancements in adversarial training and model robustness are needed to prevent malicious exploitation of these technologies. Moving forward, the field must strike a balance between innovation and ethical responsibility, ensuring that generative models are developed and used in ways that benefit society while minimizing potential risks.

9. Conclusion

In this review paper, we explored recent advancements in data representation, MLLMs, autoregressive models, multimodal diffusion models, and emerging unified frameworks that integrate these technologies. We also examined the role of multimodal transformers in enhancing HMUG and their broader applications across diverse fields. By providing a comprehensive overview of the architectures and key contributions in this area, we aim to offer valuable insights into the current landscape and highlight opportunities for future research.

Despite significant progress, several challenges persist, including improving photorealism, extending motion durations, refining control mechanisms, developing unified evaluation metrics, and addressing ethical concerns. Future work should focus on addressing these issues, enhancing model scalability and robustness, and exploring integration with emerging technologies, such as augmented and virtual reality (AR/VR) and robotics. These advancements have the potential to transform industries such as fashion, healthcare, autonomous systems, and education, enabling more interactive, efficient, and context-aware applications.

CRediT authorship contribution statement

Muhammad Islam: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation, Conceptualization; **Tao Huang:** Writing – review & editing, Supervision, Project administration; **Euijoon Ahn:** Writing – review & editing, Supervision, Project administration; **Usman Naseem:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Fei, et al., Towards artificial general intelligence via a multimodal foundation model 2021, https://doi.org/10.48550/ARXIV.2110.14378

[2] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, & A. Ramesh, Video generation models as world simulators openai SORA. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[3] OpenAI, et al., GPT-4 Technical report, 2023, https://doi.org/10.48550/ARXIV.2303.08774

[4] OpenAI, DALL·E 3. Available online: https://openai.com/index/dall-e-3/. Accessed: 18 April 2025.

[5] M. Zhang, Z. Cai, L. Pan, F. Hog, X. Guo, L. Yang, Z. Liu Motiondiffuse: Text-driven human motion generation with diffusion model IEEE transactions on pattern analysis and machine intelligence, 46 (6), 4115–4128 2024, https://doi.org/10.1109/TPAMI.2024.3355414.

[6] R. Kazmierczak, E. Berthier, G. Frehse, & G. Franchi, (2025). Explainability for vision foundation models: a survey. Inf. Fusion, https://doi.org/10.1016/j.inffus.2025.103184

[7] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, Mocogan: Decomposing motion and content for video generation 2018, 1526–1535 https://doi.org/10.1109/CVPR.2018.00165

[8] 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. 44 (11), pp. 7327-7347, 2022, https://doi.org/10.1109/TPAMI.2021.3116668

[9] W. Yin, H. Yin, D. Kragic, and M. Bjorkman, Graph-based normalizing flow for human motion generation and reconstruction, 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) Vancouver, BC, Canada: IEEE, 2021, pp. 641-648, https://doi.org/10.1109/RO-MAN50785.2021.9515316

[10] E. Pinyoanuntapong, M. U. Saleem, P. Wang, M. Lee, S. Das, and C. Chen, BAMM: Bidirectional Autoregressive Motion Model, In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) Computer Vision - ECCV 2024. ECCV 2024, 2025, Lecture Notes in Computer Science, 15073, Springer, Cham. https://doi.org/10.1007/978-3-031-72633-0_10

[11] X. Chen, et al., Executing your Commands via Motion Diffusion in Latent Space, 2023, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18000–18010, https://doi.org/10.1109/CVPR52729.2023.01726

[12] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc. Red Hook, NY, USA, Article 574 (2020), 6840–6851 https://doi.org/10.5555/3495724.3496298.

[13] Y. Huang, W. Wan, Y. Yang, C. Callison-Burch, M. Yatskar, L. Liu, Como: controllable motion generation through language guided pose code editing, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2024, pp. 180–196.

[14] W. Zhou, Z. Dou, Z. Cao, Z. Liao, J. Wang, W. Wang, L. Liu, EMDM: efficient motion diffusion model for fast and high-quality motion generation, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2024, pp. 18–38.

[15] K.S. Kalyan, A survey of GPT-3 family large language models including ChatGPT and GPT-4, Nat. Lang. Process. J. 6 (2024) 100048. https://doi.org/10.1016/j.nlp.2023.100048

[16] A. Buscemi, D. Proverbio, ChatGPT vs Gemini vs LLaMA on multilingual sentiment analysis, 2024, Accessed: May 06, 2024. [Online]. Available: http://arxiv.org/abs/2402.01715

[17] N. Messina, J. Sedmidubsky, F. Falchi, T. Rebok, Text-to-motion retrieval: towards joint understanding of human motion data and natural language, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2420–2425. https://doi.org/10.1145/3539618.3592069

[18] J. Ni, G.H. Abrego, N. Constant, J. Ma, K. Hall, D. Cer, Y. Yang, Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models In Findings of the association for computational linguistics: ACL (2022), 1864–1874 https://doi.org/10.18653/v1/2022.findings-acl.146.

[19] E. Barsoum, J. Kender, Z. Liu, Hp-gan: Probabilistic 3d human motion prediction via gan, In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, (2018) 1418–1427. https://doi.org/10.1109/CVPRW.2018.00191.

[20] M. Mezghanni, M. Boulkenafed, M. Ovsjanikov, RIVQ-VAE: discrete rotation-invariant 3D representation learning, in: 2024 International Conference on 3D Vision (3DV), IEEE, Davos, Switzerland, 2024, pp. 1382–1391. https://doi.org/10.1109/3DV62453.2024.00129

[21] F. Nazarieh, Z. Feng, M. Awais, W. Wang, J. Kittler, A survey of cross-modal visual content generation, IEEE Trans. Circuits Syst. Video Technol. 34 (8) (2024) 6814–6832. https://doi.org/10.1109/TCSVT.2024.3351601

[22] Z. Jia, Z. Zhang, L. Wang, T. Tan, Human image generation: a comprehensive survey, ACM Comput. Surv. 56 (11) (2024) 1–39. https://doi.org/10.1145/3665869

[23] K. Lyu, H. Chen, Z. Liu, B. Zhang, R. Wang, 3D human motion prediction: a survey, Neurocomputing 489 (2022) 345–365. https://doi.org/10.1016/j.neucom.2022.02.045

[24] W. Zhu, et al., Human motion generation: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 46 (4) (2024) 2430–2449. https://doi.org/10.1109/TPAMI.2023.3330935

[25] X. Wang, Y. Zhou, B. Huang, H. Chen, W. Zhu, Multi-Modal Generative AI: Multi-Modal LLMs, Diffusions, and the Unification, in IEEE Transactions on Circuits and Systems for Video Technology, 36 4 2026, pp. 5621–5641 https://doi.org/10.1109/TCSVT.2025.3635224.

[26] K. Sui, A. Ghosh, I. Hwang, B. Zhou, J. Wang, C. Guo, A survey on human interaction motion generation, International Journal of Computer Vision, 134 (3), (2026) 113 https://doi.org/10.1007/s11263-025-02582-5.

[27] Z. Shi, et al., Deep generative models on 3D representations: a survey (2022). arXiv:2210.15663

[28] X. Guo, J. Choi, Human motion prediction via learning local structure representations and temporal dependencies, in: Proc. AAAI Conf. Artif. Intell., 33, 2019, pp. 2580–2587. https://doi.org/10.1609/aaai.v33i01.33012580

[29] Y. Wang, X. Wang, P. Jiang, F. Wang, RNN-based human motion prediction via differential sequence representation, in: 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, Singapore, 2019, pp. 138–143. https://doi.org/10.1109/CCIS48116.2019.9073734

[30] J. Martinez, M.J. Black, J. Romero, On human motion prediction using recurrent neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017, pp. 4674–4683. https://doi.org/10.1109/CVPR.2017.497

[31] K. Fragkiadaki, S. Levine, P. Felsen, J. Malik, Recurrent network models for human dynamics, In Proceedings of the IEEE international conference on computer vision 2015, 4346–4354 https://doi.org/10.5555/2919332.2919834.

[32] Y. Li, et al., Efficient convolutional hierarchical autoencoder for human motion prediction, Vis. Comput. 35 (7–8), 2019. https://doi.org/10.1007/s00371-019-01692-9

[33] H. Zhou, C. Guo, H. Zhang, Y. Wang, Learning Multiscale Correlations for Human Motion Prediction, IEEE International Conference on Development and Learning (ICDL), Beijing, China, 2021, 1–7 https://doi.org/10.1109/ICDL49984.2021.9515609.

[34] Z. Liu, K. Lyu, S. Wu, H. Chen, Y. Hao, S. Ji, Aggregated multi-GANs for controlled 3D human motion prediction, in: Proc. AAAI Conf. Artif. Intell., 35 (3), 2021, pp. 2225–2232. https://doi.org/10.1609/aaai.v35i3.16321

[35] M. Zanfir, A. Zanfir, E.G. Bazavan, W.T. Freeman, R. Sukthankar, C. Sminchisescu, THUNDR: Transformer-based 3D HUmaN Reconstruction with Markers (2021). IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 12951-12960 https://doi.org/10.1109/ICCV48922.2021.01273.

[36] Y. Zhang, M.J. Black, S. Tang, We are more than our joints: predicting how 3D bodies move, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 3371–3381. https://doi.org/10.1109/CVPR46437.2021.00338

[37] X. Ma, J. Su, C. Wang, W. Zhu, Y. Wang, 3d human mesh estimation from virtual markers, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 534–543) arXiv:2303.11726 https://doi.org/10.1109/CVPR52729.2023.00059.

[38] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: a skinned multi-person linear model, ACM Trans. Graph. 34 (6) (2015) 1–16. https://doi.org/10.1145/2816795.2818013

[39] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A. Osman, D. Tzionas, M.J. Black, Expressive body capture: 3d hands, face, and body from a single image, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 10975–10985 arXiv:1904.05866 https://doi.org/10.1109/CVPR.2019.01123.

[40] A.A.A. Osman, T. Bolkart, M.J. Black, STAR: Sparse Trained Articulated Human Body Regressor, In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision - ECCV 2020. ECCV, 2020 Lecture Notes in Computer Science, 12351, Springer, Cham https://doi.org/10.1007/978-3-030-58539-6_36

[41] H. Xu, E.G. Bazavan, A. Zanfir, W.T. Freeman, R. Sukthankar, C. Sminchisescu, GHUM & GHUML: generative 3D human shape and articulated pose models, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, pp. 6183–6192. https://doi.org/10.1109/CVPR42600.2020.00622

[42] D. Pavllo, D. Grangier, M. Auli, QuaterNet: a quaternion-based recurrent model for human motion (2018). arXiv:1805.06485.

[43] Z. Liu, et al., Towards natural and accurate future motion prediction of humans and animals, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 9996–10004. https://doi.org/10.1109/CVPR.2019.01024

[44] P. Su, Z. Liu, S. Wu, L. Zhu, Y. Yin, X. Shen, Motion prediction via joint dependency modeling in phase space, in: Proceedings of the 29th ACM International Conference on Multimedia, ACM, Virtual Event China, 2021, pp. 713–721. https://doi.org/10.1145/3474085.3475237

[45] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Multiscale spatio-temporal graph neural networks for 3D skeleton-based motion prediction, IEEE Trans. Image Process. 30 (2021) 7760–7775. https://doi.org/10.1109/TIP.2021.3108708

[46] Computer vision - ECCV 2020: 16th european conference, glasgow, UK, August 23–28, 2020, proceedings, Part XIV, in: Lecture Notes in Computer Science Ser, No. v. 12359, Springer International Publishing AG, Cham, 2020.

[47] A. Jain, A.R. Zamir, S. Savarese, A. Saxena, Structural-RNN: deep learning on spatio-temporal graphs, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 5308–5317. https://doi.org/10.1109/CVPR.2016.573

[48] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, pp. 211–220. https://doi.org/10.1109/CVPR42600.2020.00029

[49] W. Mao, M. Liu, M. Salzmann, H. Li, Learning trajectory dependencies for human motion prediction, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 9488–9496. https://doi.org/10.1109/ICCV.2019.00958

[50] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, H. Liu, TrajectoryCNN: a new spatio-temporal feature learning network for human motion prediction, IEEE Trans. Circuits Syst. Video Technol. 31 (6) (2021) 2133–2146. https://doi.org/10.1109/TCSVT.2020.3021409

[51] A. Jain, A.R. Zamir, S. Savarese, A. Saxena, Structural-RNN: deep learning on spatio-temporal graphs, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 5308–5317. https://doi.org/10.1109/CVPR.2016.573

[52] Q. Cui, H. Sun, Towards accurate 3D human motion prediction from incomplete observations, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 4799–4808. https://doi.org/10.1109/CVPR46437.2021.00477

[53] V. Ferrari (Ed.), Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IV, in: Lecture Notes in Computer Science Ser, no. v. 11208, Springer International Publishing AG, Cham, 2018.

[54] J. Sedmidubsky, P. Elias, P. Budikova, P. Zezula, Content-based management of human motion data: survey and challenges, IEEE Access 9 (2021) 64241–64255.

[55] H. Xue, et al., Human motion video generation: a survey, IEEE Trans. Pattern Anal.Mach. Intell. 47 (11) (2025) 10709–10730.

[56] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, A. Dantcheva, Synthetic data in human analysis: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 46 (7) (2024) 4957–4976.

[57] F. Liao, X. Zou, W. Wong, (2024). Appearance and pose-guided human generation: a survey, ACM computing surveys, 56 (5) 1–35.

[58] A. Ismail-Fawaz, M. Devanne, S. Berretti, J. Weber, G. Forestier, Establishing a unified evaluation framework for human motion generation: A comparative analysis of metrics, Computer Vision and Image Understanding, 254, 104337 bibinfonotearXiv:2405.07680 https://doi.org/10.1016/j.cviu.2025.104337 (2025).

[59] J. Xiong, G. Liu, L. Huang, C. Wu, T. Wu, Y. Mu, N. Wong, (2024). Autoregressive models in vision: a survey. arXiv:2411.05902.

[60] M. Grohe, Word2vec, node2vec, graph2vec, X2vec: towards a theory of vector embeddings of structured data, in: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, ACM, Portland OR USA, 2020, pp. 1–16. https://doi.org/10.1145/3375395.3387641

[61] J. Pennington, R. Socher, C. Manning, GloVe: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

[62] M.V. Koroteev, BERT: a review of applications in natural language processing and understanding (2021). arXiv:2103.11943.

[63] Y. Tian, H. Zhang, Y. Liu, L. Wang, Recovering 3D human mesh from monocular images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (12) (2023) 15406–15425. https://doi.org/10.1109/TPAMI.2023.3298850

[64] Q. Wu, Y. Zhao, Y. Wang, Y.-W. Tai, C.-K. Tang, MotionLLM: multimodal motion-language learning with large language models, 2024, arXiv:arXiv:2405.17013. Accessed: Sep. 29, 2024. [Online]. Available:.

[65] Y. Zhang, Y. Wang, Y. Wang, Z. Li, J. Song, Z. Wang, Y. Wang, Y. Yang, Y. Li, MotionGPT: finetuned LLMs are general-purpose motion generators, in: Proc. AAAI Conf. Artif. Intell., 38, 2024, pp. 7368–7376. https://doi.org/10.1609/aaai.v38i7.28567

[66] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1325–1339. https://doi.org/10.1109/TPAMI.2013.248

[67] M. Loper, N. Mahmood, M.J. Black, MoSh: motion and shape capture from sparse markers, ACM Trans. Graph. 33 (6) (2014) 1–13. https://doi.org/10.1145/2661229.2661273

[68] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose (2019). 7718–7727 arXiv:1905.05754 https://doi.org/10.1109/ICCV.2019.00781.

[69] H. Tu, C. Wang, W. Zeng, VoxelPose: towards multi-camera 3D human pose estimation in wild environment, (2020), 97–212 https://doi.org/10.1007/978-3-030-58452-8_12

[70] H. Ye, W. Zhu, C. Wang, R. Wu, & Y. Wang, Faster voxelpose: real-time 3D human pose estimation by orthographic projection, 142–159 arXiv:2207.10955, (2022). https://doi.org/10.1007/978-3-031-20068-7

[71] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields (2016) 7291–7299. arXiv:1611.08050 https://doi.org/10.1109/CVPR.2017.143.

[72] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training (2018) 7753–7762. arXiv:1811.11742. https://doi.org/10.1109/CVPR.2019.00794

[73] S. Shen, L.-J. Li, H. Tan, M. Bansal, A. Rohrbach, K.W. Chang, K. Keutzer, How much can CLIP benefit vision-and-language tasks?, 2021, arXiv:2107.06383.

[74] Z. Zhou, Y. Wan, B. Wang, AvatarGPT: all-in-One framework for motion understanding, planning, generation and beyond, arXiv:2311.16468, (2023) 1357–1366. https://doi.org/10.1109/CVPR52733.2024.00135

[75] H. Miao, F. Ma, R. Quan, K. Zhan, Y. Yang, Autonomous LLM-enhanced adversarial attack for text-to-motion, 39 (6), pp. 6144-6152 arXiv:2408.00352 (2024). Accessed: Sep. 22, 2024. [Online], https://doi.org/10.1609/aaai.v39i6.32657.

[76] M. Ramesh, F.B. Flohr, Walk-the-talk: LLM driven pedestrian motion generation, in: 2024 IEEE Intelligent Vehicles Symposium (IV), IEEE, Jeju Island, Korea, Republic of, 2024, pp. 3057–3062. https://doi.org/10.1109/IV55156.2024.10588860

[77] G. Tevet, B. Gordon, A. Hertz, A.H. Bermano, D. Cohen-Or, MotionCLIP: exposing human motion generation to CLIP space, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, 13682 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2022, pp. 358–374. https://doi.org/10.1007/978-3-031-20047-2_21

[78] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, MotionGPT: human motion as a foreign language, 2023, 36 20067–20079 bibinfonoteAccessed: Sep. 22, 2024. [Online]. Available at: https://doi.org/10.5555/3666122.3667002.

[79] C. Zhong, L. Hu, Z. Zhang, S. Xia, AttT2M: text-driven human motion generation with multi-perspective attention mechanism, 2023, 509–519 arXiv:2309.00796. Accessed: Sep. 22, 2024. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.00053.

[80] J. Zhang, et al., Generating human motion from textual descriptions with discrete representations, 2023, 14730–14740 arXiv:2301.06052. Accessed: Sep. 29, 2024. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.01415.

[81] J. Zhang, et al., Generating human motion from textual descriptions with discrete representations, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, 2023, pp. 14730–14740. https:

//doi.org/10.1109/CVPR52729.2023.01415
[82] J. Lin, J. Chang, L. Liu, G. Li, L. Lin, Q. Tian, C.-W. Chen, 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.
[83] H. Kong, K. Gong, D. Lian, M.B. Mi, X. Wang, Priority-Centric Human Motion Generation in Discrete Latent Space, 2023, 14760–14770 https://doi.org/10.1109/ICCV51070.2023.01360
[84] C. Wang, T2M-HiFiGPT: generating high quality human motion from textual descriptions with residual discrete representations (2023). arXiv:2312.10628.
[85] S. Chi, Others, M2D2M: multi-motion generation from text with discrete diffusion models, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, 15072 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2025, pp. 18–36. https://doi.org/10.1007/978-3-031-72630-9_2
[86] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, Y. Shan, T2M-GPT: generating human motion from textual descriptions with discrete representations, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 14730–14740.
[87] L. Chen, Z. Niu, Q. Liu, J. Wang, J. Xue, K. Lu, Anatomically-informed vector quantization variational auto-encoder for text to motion generation, in: 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, Niagara Falls, ON, Canada, 2024, pp. 1–6. https://doi.org/10.1109/ICMEW63481.2024.10645445
[88] S. Ma, Q. Cao, J. Zhang, D. Tao, Contact-aware human motion generation from textual descriptions, 2024, arXiv:2403.15709.
[89] Z. Zhou, B. Wang, UDE: a unified driving engine for human motion generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, 2023, pp. 5632–5641. https://doi.org/10.1109/CVPR52729.2023.00545
[90] Q. Wu, Y. Zhao, Y. Wang, X. Liu, Y.-W. Tai, C.-K. Tang, Motion-agent: a conversational framework for human motion generation with LLMs, 2024, arXiv:2405.17013.
[91] H. Yang, et al., Unimumo: Unified text, music, and motion generation, arXiv:2410.04534 (2024), 39 (24), 25615–25623. https://doi.org/10.1609/aaai.v39i24.34752
[92] C. Guo, Y. Mu, M.G. Javed, S. Wang, L. Cheng, MoMask: generative masked modeling of 3D human motions, (2023) 1900–1910. https://doi.org/10.1109/CVPR52733.2024.00186
[93] Z. Zhang, et al., InfiniMotion: mamba boosts memory in transformer for arbitrary long motion generation, 2024, Accessed: Sep. 22, 2024. [Online]. Available: http://arxiv.org/abs/2407.10061.
[94] L. Xu, et al., ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation (2022) 2228–2238. arXiv:2203.07706 https://doi.org/10.1109/ICCV51070.2023.00212.
[95] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan (2019) 8110-8119. https://doi.org/10.1109/CVPR42600.2020.00813
[96] K. Yamamoto, M. Murakami, Human motion generation with StyleGAN, in: J. Li (Ed.), Seventh International Conference on Computer Graphics and Virtuality (ICCGV 2024), SPIE, Hangzhou, China, 2024, p. 6. https://doi.org/10.1117/12.3029447
[97] Y. Jiang, S. Yang, T.L. Koh, W. Wu, C.C. Loy, Z. Liu, Text2performer: text-driven human video generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22747–22757.
[98] Z. Huang, Y. Yu, L. Yang, C. Qin, B. Zheng, X. Zheng, W. Yang, Motion-aware latent diffusion models for video frame interpolation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 1043–1052.
[99] L.-H. Chen, et al., Motionllm: Understanding human behaviors from human motions and videos, 2024, https://doi.org/10.1109/TPAMI.2025.3627546
[100] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, Z. Liu, AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars, 2022, arXiv:2205.08535.
[101] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents (2022). arXiv:2204.06125.
[102] G. Tevet, B. Gordon, A. Hertz, A.H. Bermano, D. Cohen-Or, Motionclip: exposing human motion generation to clip space, in: Proc. Eur. Conf. Comput. Vis., 2022, pp. 358–374.
[103] Y. Mao, X. Liu, W. Zhou, Z. Lu, H. Li, Learning generalizable human motion generator with reinforcement learning, 2024, arXiv:2405.15541. Accessed: Sep. 29, 2024. [Online]. Available:
[104] Ribeiro Gomes, et al., MotionGPT: human motion synthesis with improved diversity and realism via GPT-3 prompting, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2024, pp. 5058–5068. https://doi.org/10.1109/WACV57701.2024.00499
[105] P.J. Yazdian, E. Liu, L. Cheng, A. Lim, Motionscript: Natural language descriptions for expressive 3d human motions, 2023, 21574–21581 Accessed: Sep. 29, 2024. [Online]. Available: https://doi.org/10.1109/IROS60139.2025.11247385.
[106] K. Li, Y. Feng, Motion generation from fine-grained textual descriptions, 2024, pp. 11625-11641 Accessed: Sep. 22, 2024. [Online]. https://aclanthology.org/2024.lrec-main.1016/.
[107] R. Wang, C. Ma, G. Li, Z. Wang, You Think, You ACT: The New Task of Arbitrary Text to Motion Generation, (2024) pp. 12012–12022. https://doi.org/10.48550/ARXIV.2404.14745
[108] K. Deng, T. Fei, X. Huang, Y. Peng, IRC-GAN: introspective recurrent convolutional GAN for text-to-video generation, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 2019, pp. 2216–2222. https://doi.org/10.24963/ijcai.2019/307
[109] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, (2015). https://doi.org/10.48550/ARXIV.1511.06434
[110] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, (2017). https://doi.org/10.48550/ARXIV.1710.10196
[111] T. Huang, J. Liu, X. Zhou, D.C. Nguyen, M.R. Azghadi, Y. Xia, S. Sun, V2X cooperative perception for autonomous driving: recent advances and challenges, (2023) 113 (5), pp. 443–477. arXiv:2310.03525 https://doi.org/10.1109/JPROC.2025.3600903.
[112] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 4396–4405. https://doi.org/10.1109/CVPR.2019.00453
[113] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, Z. Liu, FineMoGen: fine-grained spatio-temporal motion generation and editing,
[114] Y. Pan, Z. Qiu, T. Yao, H. Li, T. Mei, To create what you tell: Generating videos from captions, (2018) 1789–1798. https://doi.org/10.1145/3123266.3127905
[115] P. Susarla, A. Mukherjee, M.L. Cañellas, C.A. Casado, X. Wu, O. Silvén, M.B. López, Non-contact multimodal indoor human monitoring systems: a survey, Inf. Fusion 110 (2024) 102457.
[116] Y.e.a. Wang, MotionGPT-2: a general-purpose motion-language model for motion generation and understanding, 2024, arXiv:2410.21747.
[117] B. Jiang, et al., Motionchain: Conversational motion controllers via multimodal prompts, 2024, 54–74 https://doi.org/10.1007/978-3-031-73347-5_4
[118] M. Ramesh, F.B. Flohr, Walk-the-talk: LLM-driven pedestrian motion generation, in: 2024 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2024, pp. 3057–3062.
[119] J. Liu, W. Dai, C. Wang, Y. Cheng, Y. Tang, X. Tong, Plan, posture and go: towards open-world text-to-motion generation, (2023). arXiv:2312.14828
[120] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Commun. ACM 63 (11) (2020) 139–144. https://doi.org/10.1145/3422622
[121] A. Amballa, G. Akkinapalli, V. Muralikrishnan, LS-GAN: human motion synthesis with latent-space GANs, 2024, 326–335 arXiv:2501.01449 https://doi.org/10.1109/WACVW65960.2025.00039
[122] Z. Geng, C. Han, Z. Hayder, J. Liu, M. Shah, A. Mian, Text-guided 3D human motion generation with keyframe-based parallel skip transformer, 2024, Accessed: Sep. 22, 2024. [Online]. Available: http://arxiv.org/abs/2405.15439.
[123] D.P. Kingma, M. Welling, Auto-encoding variational bayes, (2013). https://doi.org/10.48550/ARXIV.1312.6114
[124] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, P. Slusallek, Synthesis of compositional animations from textual descriptions, 2021, 1376–1386 https://doi.org/10.1109/ICCV48922.2021.00143
[125] Y. Xue, L. Chai, MSFF-GAN: a refined method for human video motion transfer, in: 2024 36th Chinese Control and Decision Conference (CCDC), IEEE, Xi'an, China, 2024, pp. 4416–4421. https://doi.org/10.1109/CCDC62350.2024.10587557
[126] W. Yu, TransCGan-based human motion generator, in: Y. Zhong (Ed.), Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022), SPIE, Chongqing, China, 2023, p. 188. https://doi.org/10.1117/12.2668277
[127] M. Zhang, et al., ReMoDiffuse: retrieval-augmented motion diffusion model (2023), pp. 364–373. arXiv:2304.01116 https://doi.org/10.1109/ICCV51070.2023.00040.
[128] N. Athanasiou, A. Ceske, M. Diomataris, M.J. Black, G. Varol, MotionFix: text-driven 3D human motion editing, (2024). Accessed: Sep. 22, 2024. [Online]. Available: http://arxiv.org/abs/2408.00712.
[129] Q. Lu, Y. Zhang, M. Lu, V. Roychowdhury, Action-conditioned on-demand motion generation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 2249–2257.
[130] P. Jin, H. Li, Z. Cheng, K. Li, R. Yu, C. Liu, J. Chen, Local action-guided motion diffusion model for text-to-motion generation, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2024, pp. 392–409.
[131] C. Guo, et al., Action2Motion: conditioned generation of 3D human motions, (2020). https://doi.org/10.48550/ARXIV.2007.15240
[132] Y. Ji, F. Xu, Y. Yang, F. Shen, H.T. Shen, W.S. Zheng, A large-scale RGB-D database for arbitrary-view human action recognition, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 1510–1518.
[133] A.R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. QuirosRamirez, M.J. Black, Babel: bodies, action and behavior with English labels, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 722–731.
[134] T. Lee, G. Moon, K.M. Lee, MultiAct: long-term 3D human motion generation from multiple action labels, in: Proc. Assoc. Advance. Artif. Intell., 2023, pp. 1231–1239.
[135] H. Ahn, T. Ha, Y. Choi, H. Yoo, S. Oh, Text2Action: generative adversarial synthesis from language to action (2017). arXiv:1710.05298.
[136] S.R. Hosseyni, A.A. Rahmani, S.J. Seyedmohammadi, S. Seyedin, A. Mohammadi, BAD: bidirectional auto-regressive diffusion for text-to-motion generation (2024). arXiv:2409.10847.
[137] E. Pinyoanuntapong, et al., ControlMM: controllable masked motion generation, 2024, arXiv:2410.10780.
[138] B. Chen, Y. Li, Y.-X. Ding, T. Shao, K. Zhou, Enabling synergistic full-body control in prompt-based co-speech motion generation, 2024, arXiv:2410.00464.
[139] S. Lu, et al., HumanTOMATO: text-aligned whole-body motion generation, 2023, Accessed: Sep. 22, 2024. [Online]. Available: http://arxiv.org/abs/2310.12978.
[140] C. Ahuja, L.-P. Morency, Language2Pose: natural language grounded pose forecasting, 2019, arXiv:1907.01108.

[141] J. Kim, J. Kim, S. Choi, FLAME: free-form language-based motion synthesis & editing, 2022, arXiv:2209.00349.

[142] Z. Ren, Z. Pan, X. Zhou, L. Kang, Diffusion motion: generate text-guided 3D human motion by diffusion model, 2022, arXiv:2210.12315.

[143] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, S. Tang, Guided motion diffusion for controllable human motion synthesis, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, 2023, pp. 2151–2162. https://doi.org/10.1109/ICCV51070.2023.00205

[144] X. Gao, Y. Yang, Z. Xie, S. Du, Z. Sun, Y. Wu, GUESS: GradUally enriching synthesis for text-driven human motion generation, IEEE Trans. Vis. Comput. Graph. (2024) 1–13. https://doi.org/10.1109/TVCG.2024.3352002

[145] Z. Ling, B. Han, S. Li, H. Shen, J. Cheng, C. Zou, MotionLLaMA: a unified framework for motion synthesis and comprehension, 2024, arXiv:2411.17335.

[146] Z. Kiang, W. Chai, Z. Zhou, C.Y. Yang, H.W. Huang, J.N. Hwang, PackDiT: joint human motion and text generation via mutual prompting, 2025, arXiv:2501.16551

[147] B. Han, H. Peng, M. Dong, Y. Ren, Y. Shen, C. Xu, AMD: autoregressive motion diffusion, in: Proc. AAAI Conf. Artif. Intell., 38, 2024, pp. 2022–2030. https://doi.org/10.1609/aaai.v38i3.27973

[148] T. Li, C. Qiao, G. Ren, K. Yin, S. Ha, AAMDM: accelerated auto-Regressive motion diffusion model, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 1813–1823. https://doi.org/10.1109/CVPR52733.2024.00178

[149] M. Zhang, et al., Large motion model for unified multi-modal motion generation, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, 15071 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2025, pp. 397–421. https://doi.org/10.1007/978-3-031-72624-8_23

[150] K. Zhao, G. Li, S. Tang, DART: a diffusion-based autoregressive motion model for real-time text-driven motion control (2024). https://doi.org/10.48550/ARXIV.2410.05260

[151] J. Ma, S. Bai, C. Zhou, Pretrained diffusion models for unified human motion synthesis, 2022, https://doi.org/10.48550/arXiv.2212.02837

[152] Y. Cai, et al., A unified 3D human motion synthesis model via conditional variational auto-encoder, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 11625–11635.

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

[154] Z. Zhou, Y. Wan, B. Wang, A unified framework for multimodal, multi-part human motion synthesis (2023). arXiv:2311.16471

[155] M. Plappert, C. Mandery, T. Asfour, The kit motion-language dataset, Big Data 4 (4) (2016) 236–252.

[156] N. Mahmood, N. Ghorbani, N.F. Troje, G. Pons-Moll, M.J. Black, AMASS: archive of motion capture as surface shapes, in: Proc. Int. Conf. Comput. Vis. (ICCV), 2019, pp. 5441–5450.

[157] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng, Generating diverse and natural 3d human motions from text, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 5152–5161.

[158] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, L. Cheng, Action2motion: conditioned generation of 3d human motions, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2021–2029. https://doi.org/10.1145/3394171.3413869

[159] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2020) 2684–2701.

[160] A. Shahroudy, J. Liu, T.-T. Ng, 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.

[161] S. Zou, X. Zuo, Y. Qian, S. Wang, C. Xu, M. Gong, L. Cheng, 3D human shape reconstruction from a polarization image, in: Proc. Eur. Conf. Comput. Vis., 2020, pp. 351–368.

[162] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 2014.

[163] CMU Graphics Lab, CMU Graphics Lab Motion Capture Database. Available online: https://mocap.cs.cmu.edu/. Accessed: 26 March 2025.

[164] Z. Cai, D. Ren, A. Zeng, Z. Lin, T. Yu, W. Wang, X. Fan, Y. Gao, Y. Yu, L. Pan, F. Hong, M. Zhang, C.C. Loy, L. Yang, Z. Liu, Humman: multi-modal 4D human dataset for versatile sensing and modeling, in: Proc. Eur. Conf. Comput. Vis., 2022.

[165] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, S. Huang, HUMANISE: language-conditioned human motion generation in 3D scenes, in: Proc. Adv. Neural Inform. Process. Syst., 2022.

[166] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, G. Lee, Robots learn social skills: end-to-end learning of co-speech gesture generation for humanoid robots, in: Int. Conf. on Robot. and Automa., 2019, pp. 4303–4309.

[167] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, G. Lee, Speech gesture generation from the trimodal context of text, audio, and speaker identity, ACM Trans. Graph. 39 (6) (2020).

[168] C. Ahuja, D.W. Lee, Y.I. Nakano, L.-P. Morency, Style transfer for co-speech gesture animation: a multi-speaker conditional mixture approach, in: Proc. Eur. Conf. Comput. Vis., 2020.

[169] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, B. Zheng, Beat: a large-scale semantic and emotional multimodal dataset for conversational gestures synthesis, in: Proc. Eur. Conf. Comput. Vis., 2022.

[170] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two-time-scale update rule converge to a local Nash equilibrium, in: Advances in Neural Information Processing Systems 30, 2017.

[171] C. Xin, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, J. Yu, G. Yu, Executing your commands via motion diffusion in latent space, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 18000–18010.

[172] C. Guo, X. Zuo, S. Wang, 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.

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

[174] R. Dabral, M.H. Mughal, V. Golyanik, C. Theobalt, Mofusion: a framework for denoising-diffusion-based motion synthesis, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 9760–9770.

[175] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, A.H. Bermano, Human motion diffusion model, in: Proc. Int. Conf. Learn. Represent., 2023.

[176] M. Petrovich, M.J. Black, G. Varol, Action-conditioned 3D human motion synthesis with transformer VAE, in: Proc. Int. Conf. Comput. Vis., 2021.

[177] Q. Lu, Y. Zhang, M. Lu, V. Roychowdhury, Action-conditioned on-demand motion generation, in: Proceedings of the ACM International Conference on Multimedia, 2022, pp. 2249–2257.

[178] H. Wang, W. Zhu, L. Miao, Y. Xu, F. Gao, Q. Tian, (2024), Y. Wang, Aligning Human Motion Generation with Human Perceptions, arXiv:2407.02272

[179] J. Voas, Y. Wang, Q. Huang, R. Mooney, What is the best automated metric for text-to-motion generation?, in: SIGGRAPH Asia 2023 Conference Papers, 2023, pp. 1–11.

[180] A. Qazi, A. Iqbal, ExerAIde: AI-assisted multimodal diagnosis for enhanced sports performance and personalised rehabilitation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3430–3438.

[181] D. Ekambaram, V. Ponnusamy, AI-assisted physical therapy for post-injury rehabilitation: current state of the art, IEIE Trans. Smart Process. Comput. 12 (3) (2023) 234–242.

[182] Emerging role of artificial intelligence and robotics in physiotherapy: past, present, and future perspective.

[183] J. Kaur, Futurecare: AI robots revolutionizing health and healing, in: Revolutionizing the Healthcare Sector with AI, IGI Global, 2024, pp. 311–340.

[184] B.J. Darter, J.M. Wilken, Gait training with virtual reality-based real-time feedback: improving gait performance following transfemoral amputation, Phys. Ther., 91 (9), 1385–1394

[185] D.T. Lai, R.K. Begg, M. Palaniswami, Computational intelligence in gait research: a perspective on current applications and future challenges, IEEE Trans. Inf. Technol. Biomed. 13 (5) (2009) 687–702.

[186] C.A. Cifuentes, C. Rodriguez, A. Frizera-Neto, T.F. Bastos-Filho, R. Carelli, Multimodal human–robot interaction for walker-assisted gait, IEEE Syst. J. 10 (3) (2014) 933–943.

[187] C. Cui, G.B. Bian, Z.G. Hou, J. Zhao, H. Zhou, A multimodal framework based on integration of cortical and muscular activities for decoding human intentions about lower limb motions, IEEE Trans. Biomed. Circuits Syst. 11 (4) (2017) 889–899.

[188] A. Vakanski, H.P. Jun, D. Paul, R. Baker, A data set of human body movements for physical rehabilitation exercises, Data 3 (1) (2018) 2.

[189] L. Sun, Y. Wang, W. Qin, A language-directed virtual human motion generation approach based on musculoskeletal models, Comput. Animat. Virtual Worlds 35 (3) (2024) e2257.

[190] K.D. Katyal, M.S. Johannes, T.G. McGee, A.J. Harris, R.S. Armiger, A.H. Firpi, B.A. Wester, HARMONIE: a multimodal control framework for human assistive robotics, in: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2013, pp. 1274–1278.

[191] Y. Liu, X. Cao, T. Chen, Y. Jiang, J. You, M. Wu, J. Chen, (2025). A survey of embodied AI in healthcare: techniques, applications, and opportunities, arXiv:2501.07468

[192] M. Park, Y. Cho, G. Na, J. Kim, Application of virtual avatar using motion capture in immersive virtual environment, Int. J. Hum.–Comput. Interact. 40 (20) (2024) 6344–6358.

[193] S. Ma, 3D Human Motion Recovery and Generation: From Noisy Pose to Language Description, Ph.D. thesis, Doctoral dissertation, 2024.

[194] Z. Fan, P. Dai, Z. Su, X. Gao, Z. Lv, J. Zhang, Y. Zhang, (2024). Emhi: a multimodal egocentric human motion dataset with hmd and body-worn imus, arXiv:2408.17168

[195] H. Armanto, H.A. Rosyid, Improved non-player character (NPC) behavior using evolutionary algorithm—a systematic review, Entertain. Comput. 52 (2024) 100875.

[196] N. Macwan, A.J. Hude, H. Iyer, H. Jeong, S. Guo, High-fidelity worker motion simulation with generative AI, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68, SAGE Publications, Sage CA: Los Angeles, CA, 2024, pp. 1540–1541.

[197] J. Zhao, D. Weng, Q. Du, Z. Tian, Motion generation review: exploring deep learning for lifelike animation with manifold, 2024, arXiv:2412.10458).

[198] H. Yi, J. Thies, M.J. Black, X.B. Peng, D. Rempe, Generating human interaction motions in scenes with text control, in: European Conference on Computer Vision, Springer, Cham, 2025, pp. 246–263.

[199] W. Menapace, A. Siarohin, S. Lathuilière, P. Achlioptas, V. Golyanik, S. Tulyakov, E. Ricci, Promptable game models: text-guided game simulation via masked diffusion models, ACM Trans. Graph. 43 (2) (2024) 1–16.

[200] H. Yang, C. Li, Z. Wu, G. Li, J. Wang, J. Yu, L. Xu, SMGDiff: soccer motion generation using diffusion probabilistic models (2024) (pp. 11807–11817). arXiv:2411.16216

[201] J. Wang, Y. Liu, Z. Dou, Z. Yu, Y. Liang, C. Lin, W. Wang, Disentangled clothed avatar generation from text descriptions, in: European Conference on Computer Vision, Springer, Cham, 2025, pp. 381–401.

[202] Y. Huang, H. Yi, Y. Xiu, T. Liao, J. Tang, D. Cai, J. Thies, Tech: text-guided reconstruction of lifelike clothed humans, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 1531–1542.
[203] Y. Cho, L.S.S. Ray, K.S.P. Thota, S. Suh, P. Lukowicz, ClothFit: cloth-human-attribute guided virtual try-on network using 3D simulated dataset, in: 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 3484–3488.
[204] X. Han, Z. Wu, Z. Wu, R. Yu, L.S. Davis, Viton: an image-based virtual try-on network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7543–7552.
[205] Z. Jia, Z. Zhang, L. Wang, T. Tan, Human image generation: a comprehensive survey, ACM Comput. Surv. 56 (11) (2024) 1–39.
[206] S. Kim, C. Kim, B. You, S. Oh, Stable whole-body motion generation for humanoid robots to imitate human motions, in: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2009, pp. 2518–2524.
[207] L.Y. Gui, K. Zhang, Y.X. Wang, X. Liang, J.M. Moura, M. Veloso, Teaching robots to predict human motion, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 562–567.
[208] J. Bütepage, H. Kjellström, D. Kragic, Anticipating many futures: online human motion prediction and generation for human-robot interaction, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 4563–4570.
[209] M.S. Yasar, M.M. Islam, T. Iqbal, PoseTron: enabling close-proximity human-robot collaboration through multi-human motion prediction, in: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, 2024, pp. 830–839.
[210] J. Cha, J. Kim, J.S. Yoon, S. Baek, Text2HOI: text-guided 3D motion generation for hand-Object interaction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1577–1585.
[211] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, L. Zhang, Grounded SAM: assembling open-world models for diverse visual tasks, (2024) arXiv:2401.14159.
[212] N. Kaur, S. Rani, S. Kaur, Real-time video surveillance-based human fall detection system using a hybrid Haar cascade classifier, Multimed. Tools Appl. 83 (2024) (71599–71617).
[213] D. Rangelov, J. Knotter, R. Miltchev, 3D reconstruction in crime scenes investigation: impacts, benefits, and limitations, in: Intelligent Systems Conference, Springer Nature Switzerland, Cham, 2024, pp. 46–64.
[214] K. Maksymowicz, A. Kuzan, W. Tunikowski, 3D reconstruction of events: search for a spatial correlation between injuries and the geometry of the body discovery site, Forensic Sci. Int. 357 (2024) 111970.
[215] V. de Vette, K. Hutchinson, W. Mugge, A. Loeve, J.P. van Zandwijk, Applicability of the madymo pedestrian model for forensic fall analysis, Forensic Sci. Int. 112068 (2024).
[216] H. Yi, J. Thies, M.J. Black, X.B. Peng, D. Rempe, Generating human interaction motions in scenes with text control, in: European Conference on Computer Vision, Springer, Cham, 2025, pp. 246–263.
[217] H. Yi, J. Thies, M.J. Black, X.B. Peng, D. Rempe, Generating human interaction motions in scenes with text control, in: European Conference on Computer Vision, Springer, Cham, 2025, pp. 246–263.
[218] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, B. Dai, Towards diverse and natural scene-aware 3d human motion synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20460–20469.
[219] A. Mir, X. Puig, A. Kanazawa, G. Pons-Moll, Generating continual human motion in diverse 3d scenes, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 903–913.
[220] F.B. Mueller, J. Tanke, J. Gall, Massively multi-person 3d human motion forecasting with scene context, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2024, pp. 130–147.
[221] J. Li, T. Huang, Q. Zhu, T.T. Wong, Physics-based scene layout generation from human motion, in: ACM SIGGRAPH 2024 Conference Papers, 2024, pp. 1–10.
[222] C. Diller, A. Dai, CG-Hoi: contact-guided 3d human-object interaction generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19888–19901.
[223] P. Lin, HandDiffuse: generative controllers for two-hand interactions via diffusion models, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39(5), 2025, pp. 5280–5288.
[224] Y. Bian, A. Zeng, X. Ju, X. Liu, Z. Zhang, W. Liu, Q. Xu, MotionCraft: crafting whole-body motion with plug-and-play multimodal controls, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39(2), 2025, pp. 1880–1888.
[225] Y. Ma, J. Qi, C. Ji, P. Zhang, B. Zhang, Z. Deng, L. Bo, Exploring timeline control for facial motion generation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 1940–1950.
[226] L.Y. Gui, K. Zhang, Y.X. Wang, X. Liang, J.M.F. Moura, M. Veloso, Teaching robots to predict human motion, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 562–567. https://doi.org/10.1109/IROS.2018.8594452