Bridging Skeletal Motion and Wearable Sensing: A Survey of LLM-driven Human Motion Computing

Anonymous ACL submission

Abstract

Human motion computing is experiencing a paradigm shift from traditional deep learning approaches to large language model (LLM)-driven frameworks. Following the PRISMA guideline, this paper systematically analyzed 101 core papers and 21 multimodal datasets. To bridge the gap between skeletal motion sequences and wearable sensor data, we propose a full-stack classification framework encompassing motion representation, modeling, and agent interaction. We examine how LLMs enhance motion perception, understanding, generation, and planning, and summarize emerging application paradigms in advanced domains such as healthcare, sports science, and embodied intelligence. Finally, we proposed future directions for bottlenecks such as reasoning delay, interpretability, and multi-modal unification, aiming to drive the leap of human motion computing from single sequence fitting to general motion intelligent systems with physical feedback and long-term planning capabilities.

1 Introduction

Human motion computing serves as the core foundation for embodied intelligence, human-computer interaction, and medical rehabilitation (Sui et al., 2025; Jo et al., 2025), encompassing multimodal data such as skeletal sequences, inertial measurement units (IMU), and video (Ordóñez and Roggen, 2016; Driess et al., 2023; Arava et al., 2025). During the deep learning wave of the past decade, DeepConvLSTM (Ordóñez and Roggen, 2016) established a foundational architecture for sensor-based recognition (Ordóñez and Roggen, 2016), while ST-GCN (Yan et al., 2018) enabled the efficient modeling of skeletal topological features (Yan et al., 2018). These task-specific models have achieved significant progress on closed-loop benchmarks such as UCI-HAR and PAMAP2 (Reiss and Stricker, 2012). However, traditional methods still face three major bottlenecks in real-world scenarios: **Semantic Gap**: A lack of a unified semantic alignment mechanism between heterogeneous modalities (Li et al., 2025d; Zhang et al., 2025b). **Generalization Limits**: Affected by domain shifts, model performance frequently degrades in unseen domains (Cai et al., 2025). **Cognitive Deficiency**: Existing discriminative models are restricted to predefined labels, lacking common-sense reasoning and the capacity to process complex open-domain instructions (Li et al., 2025d; Jo et al., 2025; Zheng and Arakawa, 2025).

Traditional models exhibit limited generalization and logical representation capabilities in complex scenarios. In contrast, Large Language Models (LLMs), with their vast world knowledge, logical reasoning, and generalization abilities (Brown et al., 2020; Grattafiori et al., 2024; Wei et al.), are poised to drive human motion computing research from traditional temporal signal processing toward a new paradigm of cross-modal semantic translation and advanced cognitive reasoning, showing immense developmental potential.

Specifically within the domain of human motion, several representative works integrating LLMs with motion data have emerged. (Zhang et al., 2023a) and (Jiang et al., 2023) have significantly enhanced the modeling depth and generation fidelity for complex behaviors by employing discrete representations or modeling motion as a specialized language. On the perception front, (Li et al., 2025d) and (Zhang et al., 2025b) have established a new paradigm for wearable sensing foundation models through the deep alignment of heterogeneous signals and natural language. The technical trajectory leveraging LLMs as an engine has effectively overcome the limitations of task-specific models, becoming the mainstream choice for reshaping research paradigms in current human motion computing.

In recent years, although research on Large Language Model (LLM)-driven human motion has

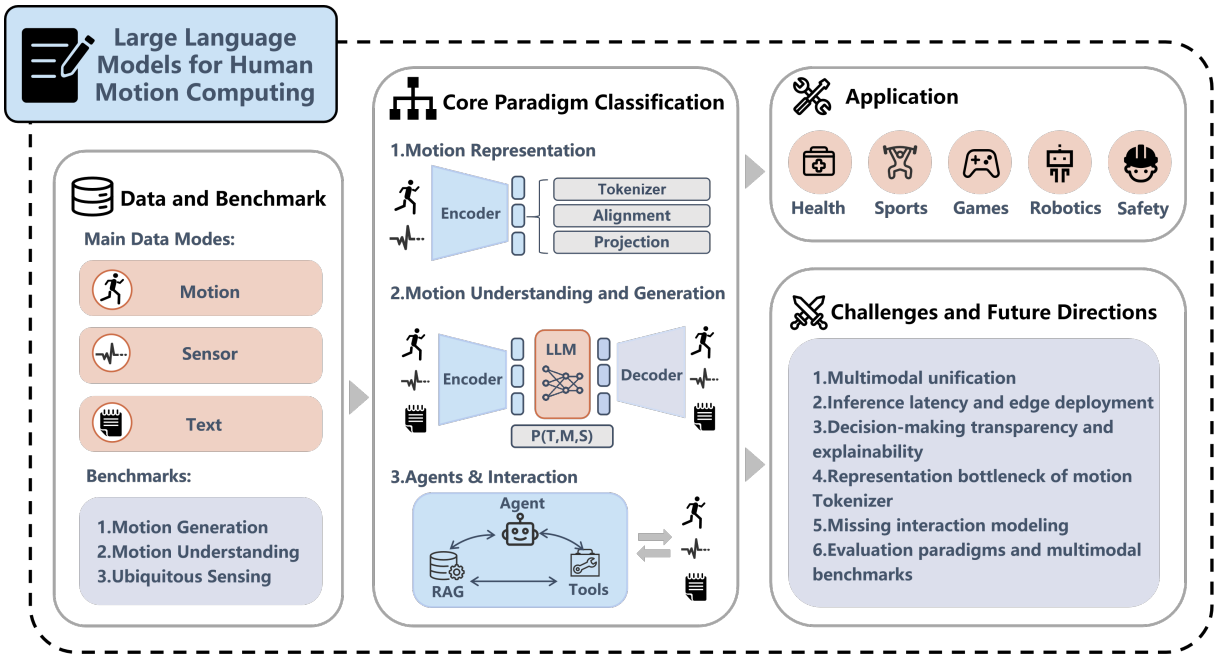

Figure 1: The main framework of this article: It covers the evolution path of core technologies from multimodal data and benchmarks to motion representation, understanding generation, and agent interaction, and sorts out the application scenarios and future technical challenges in this field

continued to emerge, there remains a lack of a unified framework to integrate the full-link applications of LLMs across motion perception, understanding, generation, and planning. This survey adopted a literature screening strategy guided by PRISMA(Page et al., 2021). After searching and screening mainstream databases, a core research literature collection for systematic analysis was constructed. Aimed at answering the following core research questions (RQs) :

- **RQ1**: What are the primary multimodal datasets and evaluation metrics employed in the current landscape of LLM-driven human motion computing?
- **RQ2**: What are the prevailing strategies for aligning heterogeneous motion modalities with language spaces in existing frameworks?
- **RQ3**: Within the proposed three-tiered taxonomy, what are the dominant technical paradigms for motion understanding and generation?
- **RQ4**: In what fields can LLM-based human motion computing be currently applied?
- **RQ5**: What are the key technical challenges and future evolution directions still faced in realizing a universal motion intelligent system?

By systematically addressing these questions, this paper aims to construct a clear technical roadmap and reveal the evolutionary logic of this field.

Although intersection research between Large Language Models (LLMs) and human motion computing has made significant progress (Ferrara, 2024; Jo et al., 2025; Zheng and Arakawa, 2025), existing surveys often focus on specific technical branches or single task types (Zhu et al., 2024; Obludzyner et al., 2024; Cai et al., 2025). Distinct from existing surveys that focus on specific technical branches , this work provides a unified, PRISMA-backed synthesis of LLM-driven human motion computing, covering the full technical stack from representation to agent interaction.

The main contributions of this article are as follows:

- **Methodological Rigor:** We provide a PRISMA-backed systematic survey, identifying 101 core records and 21 high-impact datasets to ensure transparency and reproducibility .
- **Systematic Taxonomy:** We propose a three-tiered taxonomy (Representation, Modeling, Agent Interaction) that bridges research silos between skeleton and wearable sensor communities within a unified framework .

- **Technical Roadmap:** An in-depth analysis of technical paradigms and multi-field applications was conducted, and key bottlenecks such as reasoning delay and modal unification were identified, providing a future path towards universal motion intelligent systems.

As it can be seen from the comparison in Table 1 of the appendix, this paper has made innovative contributions in multiple aspects.

The remaining chapters of this review are arranged as follows: Chapter 2: PRISMA Literature Screening Process Chapter 3: Datasets and Benchmarking (RQ1) Chapter 4-5: Architecture Analysis and LLM Role Analysis (RQ2, RQ3) Chapters 6-7: Applications, Challenges and Future Trends (RQ4, RQ5).

2 Survey Methodology

To ensure the systematicness and reproducibility of the review, this article follows the PRISMA guidelines (Page et al., 2021). We searched four major databases: ACM Digital Library, IEEE Xplore, Scopus, and ACL Anthology. Systematic retrieval was conducted based on core logical expressions ("Large Language Model" OR "LLM" OR "GPT") AND ("Human Motion" OR "IMU" OR "Skeleton"). As shown in the PRISMA flowchart (see Appendix), our multi-stage screening resulted in a final collection of 101 relevant records. From this collection, we curated 50 representative core studies for technical analysis (Table 3) based on their methodological novelty and 21 high-frequency datasets (Table 2) based on their prevalence in the literature. In addition to the records managed by PRISMA, we supplement a limited number of late-breaking works (e.g., 2025 arXiv preprints) to illustrate the latest research context. These supplementary citations are only for supplementation and are not included in the PRISMA statistics to maintain the rigor of the method.

3 Dataset and Benchmarks

We curated 21 core datasets underpinning human motion research and categorized them into three major benchmark systems: Generation, Understanding, and Sensing. These multimodal data resources—spanning skeletal, visual, sensor, and linguistic modalities—collectively constitute the foundation for evaluating multi-dimensional model capabilities, including semantic alignment, physical consistency, and generalization,, aiming to answer RQ1.

3.1 Benchmarks for Motion Generation Tasks

This benchmark primarily evaluates the semantic consistency (e.g., R-Precision) and generation quality (e.g., FID) of Text-to-Motion models.

Semantic-driven and Whole-body Representation: (Guo et al.), (Plappert et al., 2016), and (Lin et al., 2024) serve as the infrastructure for this domain. The former two established the standardized alignment specifications for "text-skeleton" pairs, while Motion-X further extends data dimensionality to include SMPL-X whole-body meshes (encompassing facial expressions and finger movements) alongside multi-level semantic annotations.

Stylization and Interaction Generation: Addressing fine-grained control, (Ghorbani et al., 2022), (Liu et al., 2022), and (Wang et al., b) have established generation specifications for stylized transfer, emotional conversational gestures, and Human-Scene Interaction (HSI), respectively. These datasets specifically assess model expressiveness under zero-shot control and physical constraints.

3.2 Benchmarks for Motion Understanding Tasks

This benchmark focuses on evaluating the parsing accuracy of motion semantics and the reasoning capabilities regarding physical environments.

Semantic Annotation and Parametric Models: (Mahmood et al., 2019) and (Punnakkal et al., 2021) integrate parametric human models with frame-level linguistic labels, serving as core resources for training long-sequence motion segmentation and fine-grained semantic description models.

Visual Recognition and Scene Perception: (Shahroudy et al., 2016) and (Liu et al., 2017) provide large-scale samples for classification and motion detection. Meanwhile, (Grauman et al., 2022) and (Hassan et al., 2019) introduce egocentric perspectives and static scene geometric constraints, facilitating the evaluation of hand-object interaction analysis in complex environments and the physical consistency of monocular human pose estimation.

3.3 Benchmarks for Ubiquitous Sensing

Covering IMU and physiological signal data, this benchmark aims to verify model robustness (e.g., F1-Score) in real-world (In-the-wild) settings.

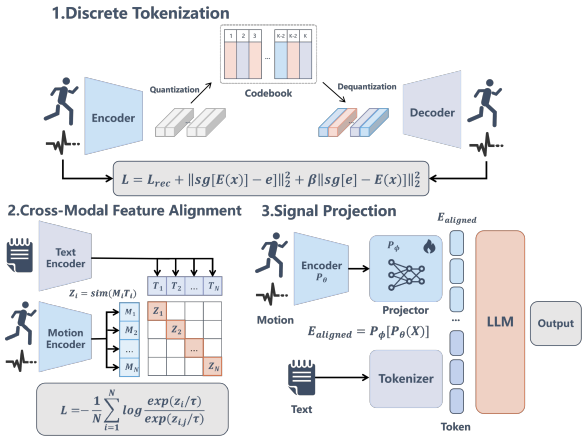

Figure 2: The framework 4.1:Methodologies for motion representation and latent alignment.(1)Discrete Tokenization via VQ-VAE mapping to codebook **e**; (2)Cross-modal Alignment using similarity matrix $\mathbf{Z} \in \mathbb{R}^{N\times N}$; (3)Signal Projection mapping heterogeneous embeddings into LLM semantic space.

Basic Recognition and Heterogeneity Challenges: (Garcia-Gonzalez et al., 2020) and (Reiss and Stricker, 2012) laid the foundation for multi-position sensor recognition. (Stisen et al., 2015), (Sztyler and Stuckenschmidt, 2016), (Shoaib et al., 2014), and (Malekzadeh et al., 2019) focus on addressing heterogeneous generalization metrics, including device model discrepancies, dynamic changes in wearing positions, and privacy protection.

In-the-wild Environments and Physiological Computing: (Chan et al., 2024), (Schmidt et al., 2018), and (Al Dahdah et al., 2015) extend research scenarios to all-weather wild environments and health monitoring, providing benchmarks for stress detection and long-term behavioral analysis combined with physiological signals such as ECG and EDA.

4 Method

This chapter systematically constructs a panoramic view of the evolution of human motion behavior computing technology centered on large language models (LLMS) from three core dimensions: modal representation of human motion, core task modeling, and agent interaction, aiming to answer RQ2 and RQ3.

4.1 Motion Representation

4.1.1 Discrete Tokenization

Unlike the method of directly extracting features in a continuous space(Xu et al., 2021; Zhu et al., 2023),motion discretization compresses continuous features into LLM-compatible tokens. The mainstream VQ-VAE paradigm maps motions to a discrete codebook e by minimizing(Zhang et al., 2023a; Jiang et al., 2023, 2024):

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \|\text{sg}[E(x)] - e\|_2^2 + \beta\|\text{sg}[e] - E(x)\|_2^2 \tag{1}$$

To enhance quality, subsequent works introduced Residual VQ (RVQ) for high-frequency details (Guo et al., 2023; Yao et al., 2023), adopted part-decoupled strategies for local precision (Wang et al., 2024b; Huang et al., 2025a), or optimized for noise and streaming via jitter reduction and causal convolutions (Shan et al., 2025; Jiang et al.).

4.1.2 Cross-Modal Feature Alignment

Mainstream motion representation learning adopts a cross-modal paradigm based on CLIP, mapping motion M and text T to a shared space through an encoder. By comparing and maximizing the correlation of positive samples with $Z_{i,i} = Mi * Ti$, and suppressing the interference of negative samples:

$$\mathcal{L}_{cl} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{e^{Z_{i,i}/\tau}}{\sum_{j=1}^{N} e^{Z_{i,j}/\tau}} + \log\frac{e^{Z_{i,i}/\tau}}{\sum_{j=1}^{N} e^{Z_{j,i}/\tau}}\right) \tag{2}$$

Recent works optimize this objective via three strategies: (1)Semantic Augmentation: utilizing LLMs to enrich sparse labels for stronger textual supervision (Chen et al., 2025b; Chivereanu et al., 2024; Zhang et al., c); (2)Feature Alignment: projecting motion features into CLIP space or employing joint pre-training to enable zero-shot capabilities (Fritsch et al., 2025; Li et al., 2025e; Maldonado et al., 2025; Ao et al., 2023; Moon et al., 2023; Tevet et al., 2022; Yuan et al., 2022); and (3)Cross-Domain Synthesis: leveraging physical simulation or knowledge transfer to align low-resource modalities (Yang et al., 2024; Zhang et al., c).

4.1.3 Signal Projection

LLM-based motion modality alignment is essentially a cross-modal manifold mapping process that maps sensor signals X_{motion} into embeddings

$$\mathbf{E}_{aligned} = \mathcal{P}_\phi(\mathcal{F}_\theta(X)) \tag{3}$$

via an encoder $\mathcal{F}_\theta$ and a projector $\mathcal{P}_\phi$, aiming to endow LLMs with the capability to understand and generate physical motion. Current mainstream technical paradigms include: Direct projection utilizing

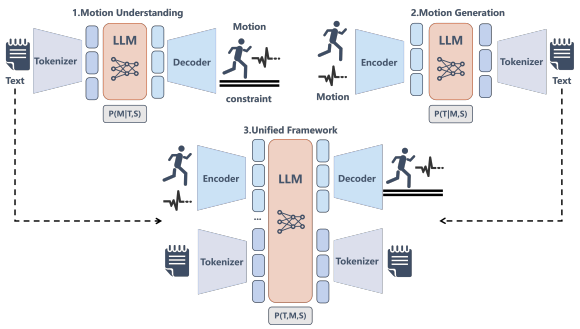

Figure 3: The framework of 4.2 demonstrates the classification of modeling paradigms. Above: Task-specific branches for understanding $P(T|M,S)$ and generating $P(M|T,S)$. The unified framework models the joint distribution $P(T,M,S)$ through the shared token space.

Linear/MLP layers (Imran et al., 2025), (Hong et al., 2025b); Semantic distillation leveraging teacher models (Chen et al., 2024b), (Zhang et al., 2025b); Natural language-based description alignment (Li et al., 2025d), (Qiu et al., 2025); Representation enhancement via LLM-driven data (Wei et al., 2025a); And attention-based fusion by treating motion as a second modality (Zhu et al., 2025a). Ultimately, these methods aim to empower LLMs with physical motion understanding and generation capabilities.

4.2 Motion understanding and generation

4.2.1 Motion Understanding

While traditional methods focus on fitting motion M to Category Y $f : M \rightarrow Y$, the new LLM-based paradigm aims to realize the transition from signal perception to semantic cognition by modeling a conditional distribution $P(T|M,S)$of language about motion and scene S. Studies such as (Ji et al., 2024; Li et al., 2025c; Su et al., 2025; Yan et al., 2024) have verified the zero-shot potential of LLMs in understanding physical motion. Subsequently, the underlying alignment between heterogeneous signals and natural language has been established through the construction of multimodal foundation models (Zhang et al., 2025b; Li et al., 2025d; Chen et al., 2025a; Qu et al., 2024; Yazdian et al., 2025). Furthermore, (Arrotta et al., 2024; Hong et al., 2025b; Yan et al., 2025; Chen et al., 2024a; Wei et al., 2025b; Hota et al., 2025) have effectively tackled challenges in cross-domain generalization and motion segmentation via semantic injection. Finally, (Yu et al., 2025; Ouyang and Srivastava, 2024; Mo et al., 2025; Wang et al., 2024a; Zhu et al., 2025c) have extended the capability boundary to long-term logical reasoning and expert-level interactive feedback, completing the closed loop of general embodied intelligent systems.

4.2.2 Motion Generation

The motion generation based on LLM aims to model the motion $P(M|T,S)$ under the joint constraints of semantic $\Psi(T)$ and physical Scene S, and generate the motion flow by using the physical common sense of LLM. First of all,to address the challenges of instruction ambiguity and motion editing, (Leng et al., 2023; Kalakonda et al., 2023; Leng et al., 2024; Liu et al., 2024b; Harithas and Sridhar, 2025; Jiang et al., 2023) utilize LLMs for text enhancement, operator design, and other techniques to achieve interactive alignment; (Wang et al., 2024b; Zhu et al., 2025a; Qazi et al., 2024; Qing et al., 2023; Sun et al., 2024) break the modality barrier through discrete or continuous representations to realize unified generation of "motion as language"; for complex spatiotemporal constraints, (Zhang et al., 2023a; Wang et al., 2025a,d,c,b; Zhang et al., 2025c) introduce fine-grained control to improve generation accuracy; for data-scarce scenarios, (Athanasiou et al., 2024; Chen et al., 2024c; Fan et al., 2025; Goel et al., 2024; Gui et al., 2025; Huang et al., 2025b) leverage the physical common sense and code simulation of multimodal large models to achieve zero-shot generation without motion data.

4.2.3 Unified Framework

Different from traditional methods that optimize conditional probabilities $P(Y|X)$, unified motion-language models aim to model the joint distribution $P(T,M,S)$ of text T, motion M, and scene S, achieving multi-task unification through learning cross-modal general representations. The MotionGPT series (Jiang et al., 2023; Wang et al., 2024b; Zhu et al., 2025a) and (Li et al., 2025e) pioneered the establishment of bidirectional translation and feature alignment representations for motion-language. (Zhou et al., 2023), (Wu et al., 2025), and (Li et al., 2025b) utilized chain-of-thought or hybrid scripts to bridge multi-granular semantics from high-level planning to microscopic control. Advanced models further break task barriers: (Wang et al., 2025b) integrates generation and editing, (Li et al., 2025a) reformulates perception as controlled generation, and (Wang et al., a) finally incorporates physical scene interactions, propelling the field towards embodied generalist intelligence.

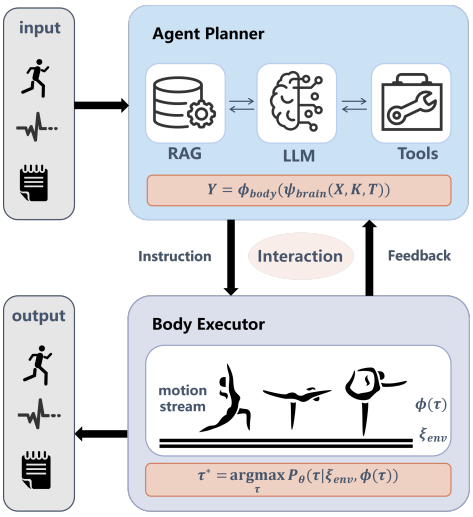

Figure 4: The framework of 4.3 presents a hierarchical Agent architecture. Through the closed-loop interaction between Agent Planner and Body Executor, it realizes decoupling and collaboration from high-level cognitive decision-making to low-level physical motion execution.

4.3 Agents & Interaction

4.3.1 Knowledge & Retrieval

Knowledge-driven human motion processing can be expressed as

$$Y \sim P(Y \mid \mathcal{E}(X), \mathcal{R}(\mathcal{E}(X), \mathcal{M}); \mathcal{K}_{LLM}) \quad (4)$$

that is, performing deep reasoning on the semanticized input $\mathcal{E}(X)$ based on knowledge $\mathcal{K}_{LLM}$ and retrieval $\mathcal{R}$. At the perception and understanding level, Large Language Models (LLMs) enable zero-shot reasoning and spatiotemporal analysis [(Arrotta et al., 2024), (Ouyang and Srivastava, 2024), (Li et al., 2025c), (Su et al., 2025)], and guide low-level skeleton representation learning [(Xu et al., 2025a), (Ye et al., 2025)]. At the generation and synthesis level, Retrieval-Augmented Generation (RAG) significantly enhances the physical realism and semantic consistency of generated motions by retrieving from external sample libraries [(Zhang et al., 2023b), (Kalakonda et al., 2025), (Zhu et al., 2025b), (Zhang et al., 2024), (Zhang et al., a)]. At the interactive application level, a closed loop of personalized linguistic feedback has been realized in the fields of rehabilitation, sports, and robotics [(Wang et al., 2024a), (Zhu et al., 2025c), (Graule and Isler, 2024), (Tian et al., 2025a)], promoting the transformation of the field towards a "knowledge + data" dual-driven paradigm.

4.3.2 Agent Planning

Large Language Models (LLMs) are evolving from low-level executors to high-level decision-makers in human motion computing. This paradigm shift decouples the traditional end-to-end mapping into a composite of a high-level cognitive function Φ and a low-level executive function Ψ:

$$Y = \Psi_{\text{body}}\left(\Phi_{\text{brain}}(X, \mathcal{K}, \mathcal{T})\right) \quad (5)$$

First, acting as a Hierarchical Task & Motion Planner, LLMs employ progressive generation strategies (Jin et al., 2025; Fu et al., 2024; Zhang et al., b) to resolve long-sequence generation issues, while integrating environmental geometry and fine-grained planning (Qian et al., 2025; Graule and Isler, 2024; Li et al., 2025b) to ensure physical and spatiotemporal consistency. Second, as an Explicit Logical Reasoner, LLMs replace black-box fitting with explicit reasoning via mathematical operators (Tang et al., 2025), coding tools (Quan et al., 2025), knowledge retrieval (Li et al., 2025c), and intent analysis (Gao et al., 2025), thereby achieving interpretable zero-shot sensing. Finally, functioning as a Multi-Agent Orchestrator, LLMs coordinate multi-agent collaboration (Zhao et al., 2025; Zhang et al., 2025a) and unify multimodal architectures with closed-loop toolchains (Zhou et al., 2023; Jiang et al., 2024; Wu et al., 2024; Miao et al.). This evolution marks a significant transition toward cognitive motion intelligence.

4.3.3 Embodied Interaction

Under the LLM framework, interaction generation is reformulated as a problem of maximizing the joint probability of multi-modal trajectories under scene embeddings $\mathcal{E}_{\text{env}}$ and physical constraints $\Phi(\tau)$:

$$\tau^* = \arg\max_{\tau} P_\theta(\tau \mid \mathcal{E}_{\text{env}}, \Phi(\tau)). \quad (6)$$

Based on this paradigm, (Wang et al., a), (Zhou et al., 2023), and (Harithas and Sridhar, 2025) established a unified multi-modal token space.(Cen et al., 2024), (Qian et al., 2025), (Graule and Isler, 2024), (Takeyama et al., 2025) focused on the processing of $\mathcal{E}_{\text{env}}$, achieving precise navigation and intention prediction. (Huang et al., 2025a), (Xu et al.), (Liu et al., 2024a), (Wang et al., c), and Fan et al. (Fan et al., 2025) tackled the challenges of fine-grained interaction by optimizing the constraints $\Phi(\tau)$. Meanwhile, (Yao et al., 2023), (Shan et al.,

2025), (Tang et al., 2025), (Hong et al., 2025a) successfully applied this model to physical simulation and perception applications.

5 Application

This chapter explores how LLM-based motion modeling can empower various vertical application scenarios, aiming to answer RQ4.

5.1 Medical Care and Health

Healthcare scenarios are transitioning from numerical recording to semantic interpretation. Arava et al. (2025) and Pulse-PPG (Saha et al., 2025) establish foundational frameworks for wearables and physiological signals. DailyLLM (Tian et al., 2025b) automates health log generation, while Ubi-Physio and Myo-Trainer (Wang et al., 2024a; He et al., 2025) deliver real-time rehabilitation guidance via biomechanical analysis. To overcome data bottlenecks, SHADE-AD and Dobhal et al. (2024) demonstrate the potential of LLMs in synthesizing high-quality behavioral data.

5.2 Sports Training and Coaching

The sports domain leverages LLMs to achieve an evolution from simple motion recognition to expert-level coaching. (Tian et al., 2025a)(Zhu et al., 2025c; Tang et al., 2025), establishes a closed-loop assessment system that transforms motion sequences into interpretable guidance. (He et al., 2025; Ma et al., 2025b) further enhance the analytical depth through muscle-aware sensing and domain-specific knowledge injection. Additionally,(Ma et al., 2025a; Hirano et al., 2025) provide critical multi-modal data support for developing instructional models in this field.

5.3 Games and Entertainment

LLMs empower digital humans with superior narrative expressiveness and stylistic control. Story-to-Motion and MotionScript (Qing et al., 2023; Yazdian et al., 2025) achieve direct mapping from long-form text to fine-grained motion sequences. Zeng et al. (2024) and InterDreamer (Xu et al.) introduce breakthroughs in affective style editing and complex human-object interaction generation. Finally, GameMLD (Fu et al., 2025b) provides large-scale data support for such stylized modeling.

5.4 Embodied Intelligence and Robotics

Embodied intelligence anchors motion semantics within physical space. GG-LLM and HOIGPT (Graule and Isler, 2024; Huang et al., 2025a) realize geometrically-guided activity forecasting and hand-object interaction modeling. MotionGlot and Walk-the-Talk (Harithas and Sridhar, 2025; Ramesh and Flohr, 2024) demonstrate potential in cross-embodied motion generation and pedestrian trajectory simulation, while Liu et al. (2024b) enhances open-set control through programmable logic.

5.5 Safety and Surveillance

The safety sector is evolving into an autonomous warning system equipped with semantic reasoning. REMONI and Zhang et al. (2025a) offer automated monitoring solutions for telehealth and industrial environments. LLM4HAR and Ray et al. (2025) focus on enhancing open-vocabulary recognition capabilities for edge devices. Lastly, SensorBench and SHADE-AD (Quan et al., 2025; Fu et al., 2025a) ensure system reliability through standardized benchmarks and high-fidelity data augmentation.

6 Challenges and Future Directions

6.1 Multimodal unification

Existing research often treats skeleton, visual motion, and sensor signals as independent modalities, where the lack of a unified semantic space leads to difficulties in cross-modal generalization (Cai et al., 2025). Currently, IMU2CLIP and UniMTS (Moon et al., 2023; Zhang et al., c) achieve cross-modal alignment through language mediation, while MotionLLM and MuJo (Chen et al., 2025a; Fritsch et al., 2025) attempt to construct multimodal joint feature spaces. Future research should focus on a shared latent space centered on "motion concepts" and utilize generative technologies to supplement scarce alignment resources, aiming to construct universal motion-language large models.

6.2 Inference latency and edge deployment

A structural contradiction exists between the autoregressive generation paradigm of large models and the requirements for real-time feedback (Hong et al., 2025b; Tian et al., 2025b). LLM4HAR and DailyLLM demonstrate that partial training and context-aware frameworks can enable model compression for edge devices (Hong et al., 2025b; Tian et al., 2025b), while Myo-Trainer and SensorLLM (He et al., 2025; Li et al., 2025d) verify the feasibility of real-time feedback in task-specific scenarios. Future research should explore pruning, quantiza-

tion, and low-rank decomposition techniques tailored for motion features, and integrate knowledge distillation to construct low-latency, high-response embodied architectures.

6.3 Decision-making transparency and explainability

The opacity of decision-making mechanisms limits the application of models in high-trust scenarios such as healthcare and health monitoring (Li et al., 2025c; Ji et al., 2024). HARGPT and Xu et al. (Ji et al., 2024; Xu et al., 2025b) have revealed the perception bottlenecks of models in zero-shot scenarios, whereas ZARA, ContextGPT, and SportsGPT (Li et al., 2025c; Arrotta et al., 2024; Tian et al., 2025a) attempt to break through "black-box" constraints through agent-based reasoning and logic injection. Future research should systematically introduce mature semantic probing and neural circuit dissection methods from the NLP field to evaluate the encoding quality of motion concepts and construct goal-driven logical reasoning systems.

6.4 Representation bottleneck of motion Tokenizer

Current tokenizers face bottlenecks such as the loss of dynamic information and challenges in cross-model reuse. MotionGPT, T2M-GPT, and MoConVQ (Jiang et al., 2023; Zhang et al., 2023a; Yao et al., 2023) have laid the foundation for discretized representations and physical control, while MoMask and Causal Motion Tokenizer (Guo et al., 2023; Jiang et al.) balance high-level semantics with fine-grained details. Future developments should evolve toward the deep decoupling of semantics, style, and dynamics, constructing efficient, universal cross-task interfaces through multi-scale residual quantization and structured event tokens.

6.5 Missing interaction modeling

Current motion large models generally face limitations due to the weak coupling between motion modeling and environmental geometry or object states, making it difficult to simulate physical contact constraints and object affordances, and lacking explicit perception of 3D spatial dimensions. To address this bottleneck, Cen et al. (Cen et al., 2024) and HOIGPT (Huang et al., 2025a) have improved the quality of 3D interactive motion generation through unified interaction tokenization and spatial reasoning, while GG-LLM (Graule and Isler, 2024) and TR-LLM (Takeyama et al., 2025) have corrected the spatial awareness deficiencies of LLMs by introducing geometric grounding and trajectory data. Future research will focus on embedding motion generation into embodied world models with physical feedback capabilities, bridging the gap from offline sequence fitting to the generation of executable behaviors in real-world scenarios by constructing joint representation frameworks for scene and contact awareness.

6.6 Evaluation paradigms and multimodal benchmarks

Current metrics rely excessively on static datasets, making it difficult to reflect generalization performance and causal logic consistency. CAPTURE-24 (Chan et al., 2024) provides large-scale, in-the-wild data, while LLaSA and MotionLLM (Imran et al., 2025; Chen et al., 2025a) expand the dimensions of reasoning and behavioral evaluation. SensorBench, HHAR, and relevant reviews (Quan et al., 2025; Stisen et al., 2015; Zheng and Arakawa, 2025) reveal the degradation of models in complex and heterogeneous scenarios. In the future, there is an urgent need to construct a unified framework centered on "task execution," integrating physical constraints and long-term rationality metrics to drive the advancement of evaluation systems for embodied intelligence.

7 Conclusion

This survey has constructed a full-stack framework covering perception, understanding, generation and agency through the PRISMA process, innovatively achieving the deep integration of skeletal motion and wearable sensing modalities. Research shows that motion computing is evolving from low-level signal processing to cross-modal semantic translation and advanced cognitive reasoning. LLM has become the core engine driving human motion computing systems and has been applied in many fields. However, there are still bottlenecks and challenges. This study lays a theoretical and empirical foundation for realizing a general motion intelligence system with physical feedback and long-term planning capabilities.

Limitations

Although this review has a series of innovative points and contributions, there are still several limitations. Firstly, literature retrieval mainly focuses on LLM-related research from 2023 to the present, which may lead to insufficient sorting out of classic human motion calculation models before the LLM era. Secondly, to ensure domain focus, pure robotics or computer vision research without human prior knowledge was excluded during the screening process, which to some extent limited the comprehensive coverage of the broader interdisciplinary field of embodied intelligence. At the same time, in the face of the rapid iteration of LLM technology, this review may have a certain lag for the latest preprint achievements that emerge after 2025. Finally, although the full-stack classification method was proposed, there is still a certain degree of subjectivity when defining the attribution of some cross-task models.

References

Marine Al Dahdah, Annabel Desgrées Du Loû, and Cécile Méadel. 2015. Mobile health and maternal care: A winning combination for healthcare in the developing world? *Health Policy and Technology*, 4(3):225–231.

Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Transactions on Graphics*, 42(4):1–18.

Poojareddy Arava, Aditi Singh, Abul Ehtesham, and Saket Kumar. 2025. Large Language Models in Wearable Devices: Applications, Challenges, and Future Directions. In *2025 IEEE World AI IoT Congress (AIIoT)*, pages 1003–1009, Seattle, WA, USA. IEEE.

Luca Arrotta, Claudio Bettini, Gabriele Civitarese, and Michele Fiori. 2024. ContextGPT: Infusing LLMs Knowledge into Neuro-Symbolic Activity Recognition Models. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 881–884. ArXiv:2403.06586 [cs].

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. 2024. MotionFix: Text-Driven 3D Human Motion Editing. *arXiv preprint*. ArXiv:2408.00712 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Yize Cai, Baoshen Guo, Flora Salim, and Zhiqing Hong. 2025. Towards Generalizable Human Activity Recognition: A Survey. *arXiv preprint*. ArXiv:2508.12213 [eess].

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. 2024. Generating Human Motion in 3D Scenes from Text Descriptions. *arXiv preprint*. ArXiv:2405.07784 [cs].

Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. 2024. CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data*, 11(1):1135.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2025a. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Qixiu Chen, Yingan Liu, Peng Huang, and Jiani Huang. 2024a. Linguistic-Driven Partial Semantic Relevance Learning for Skeleton-Based Action Recognition. *Sensors*, 24(15):4860.

Wenqiang Chen, Jiaxuan Cheng, Leyao Wang, Wei Zhao, and Wojciech Matusik. 2024b. Sensor2Text: Enabling Natural Language Interactions for Daily Activity Tracking Using Wearable Sensors. *arXiv preprint*. ArXiv:2410.20034 [cs].

Wenshuo Chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, and Chen Chen. 2024c. SATO: Stable Text-to-Motion Framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6989–6997. ArXiv:2405.01461 [cs].

Yukun Chen, Liang Wan, Zihuan Qiu, Mingzhou He, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. 2025b. DPM-CLIP: Zero-Shot Multimodal Egocentric Activity Recognition based on Dual-Prediction Mechanism. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 1786–1791, Anchorage, AK, USA. IEEE.

Radu Chivereanu, Adrian Cosma, Andy Catruna, Razvan Rughinis, and Emilian Radoi. 2024. Aligning Actions and Walking to LLM-Generated Textual Descriptions. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, Istanbul, Turkiye. IEEE.

Umang Dobhal, Christina Garcia, and Sozo Inoue. 2024. Synthetic Skeleton Data Generation using Large Language Model for Nurse Activity Recognition. In *Companion of the 2024 on ACM International Joint*

Conference on Pervasive and Ubiquitous Computing, pages 493–499, Melbourne VIC Australia. ACM. TLDR: The proposed approach using LLM with prompts containing the explanation of the task with the description of the datasets to generate synthetic data has improved the overall ES classification performance.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint*. ArXiv:2303.03378 [cs].

Siyuan Fan, Longlin Sun, Bo Peng, Bo Du, and Xiantao Cai. 2025. Fine-Grained Body Part Control in Text-Driven Motion Synthesis with Interactive Intention. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Nantes, France. IEEE.

Emilio Ferrara. 2024. Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling: A Survey of Early Trends, Datasets, and Challenges. *arXiv preprint*. ArXiv:2407.07196 [cs].

Stefan Gerd Fritsch, Cennet Oguz, Vitor Fortes Rey, Lala Ray, Maximilian Kiefer-Emmanouilidis, and Paul Lukowicz. 2025. MuJo: Multimodal Joint Feature Space Learning for Human Activity Recognition. In *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–12, Washington DC, DC, USA. IEEE.

Heming Fu, Hongkai Chen, Shan Lin, and Guoliang Xing. 2025a. SHADE-AD: An LLM-Based Framework for Synthesizing Activity Data of Alzheimer's Patients. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 290–296, UC Irvine Student Center. Irvine CA USA. ACM.

Jiajun Fu, Yuxing Long, Xiaojie Wang, and Jianqin Yin. 2024. LLM-Driven "Coach-Athlete" Pretraining Framework for Complex Text-To-Motion Generation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Yokohama, Japan. IEEE.

Yiyu Fu, Ziming Cheng, Yihao Liao, Jiangfeiyang Wang, Ruomei Wang, Guanghui Yue, Chenlei Lv, and Baoquan Zhao. 2025b. GameMLD: A Game-Sourced Motion-Language Dataset for Stylized Motion Generation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Nantes, France. IEEE.

Nan Gao, Yihua Bao, Dongdong Weng, Jiayi Zhao, Jia Li, Yan Zhou, and Pengfei Wan. 2025. SARGes: Semantically Aligned Reliable Gesture Generation via Intent Chain. In *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents*, pages 13–21, Dublin Ireland. ACM.

Daniel Garcia-Gonzalez, Daniel Rivero, Enrique Fernandez-Blanco, and Miguel R. Luaces. 2020. A Public Domain Dataset for Real-Life Human Activity Recognition Using Smartphone Sensors. *Sensors*, 20(8):2200.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *arXiv preprint*. ArXiv:2209.07556 [cs].

Purvi Goel, Kuan-Chieh Wang, C. Karen Liu, and Kayvon Fatahalian. 2024. Iterative Motion Editing with Natural Language. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–9. ArXiv:2312.11538 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Moritz A. Graule and Volkan Isler. 2024. GG-LLM: Geometrically Grounding Large Language Models for Zero-shot Human Activity Forecasting in Human-Aware Task Planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 568–574, Yokohama, Japan. IEEE.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, New Orleans, LA, USA. IEEE.

Wenxuan Gui, Ruiyan Wang, Rong Xie, and Li Song. 2025. Motion Dictionary-Based Approach for Highly Generalizable Text-To-Motion Generation. In *2025 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–7, Dublin, Ireland. IEEE.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. *arXiv preprint*. ArXiv:2312.00063 [cs].

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating Diverse and Natural 3D Human Motions From Text.

Sudarshan Harithas and Srinath Sridhar. 2025. MotionGlot: A Multi-Embodied Motion Generation Model. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16109–16116, Atlanta, GA, USA. IEEE.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. 2019. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. *arXiv preprint*. ArXiv:1908.06963 [cs].

Yuting He, Xinyan Wang, Mu Yuan, Bufang Yang, Siyang Jiang, Yihua Huang, Doris S. F. Yu, Guoliang Xing, and Hongkai Chen. 2025. Myo-Trainer: A Vision-based Muscle-Aware Motion Feedback System for In-Home Resistance Training. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*, pages 742–757, Kerry Hotel, Hong Kong Hong Kong China. ACM.

Toshihiro Hirano, Yuki Tabei, Yichen Peng, Chen-Chieh Liao, Erwin Wu, and Hideki Koike. 2025. SkiTechCoach: A Multimodal Alpine Skiing Dataset with 3D Body Pose, Sole Pressure, and Expert Coaching. In *Proceedings of the 8th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 39–46, Dublin Ireland. ACM.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. 2025a. EgoLM: Multi-Modal Language Model of Egocentric Motions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5344–5354, Nashville, TN, USA. IEEE.

Zhiqing Hong, Yiwei Song, Zelong Li, Anlan Yu, Shuxin Zhong, Yi Ding, Tian He, and Desheng Zhang. 2025b. LLM4HAR: Generalizable On-device Human Activity Recognition with Pretrained LLMs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 4511–4521, Toronto ON Canada. ACM.

Aritra Hota, Soumyajit Chatterjee, and Sandip Chakraborty. 2025. Evaluating Large Language Models as Virtual Annotators for Time-series Physical Sensing Data. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–25. ArXiv:2403.01133 [cs].

Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J. Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, and Hao Tang. 2025a. HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7136–7146, Nashville, TN, USA. IEEE.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. 2025b. CoMo: Controllable Motion Generation Through Language Guided Pose Code Editing. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15087, pages 180–196. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2025. LLaSA: A Sensor-Aware LLM for Natural Language Reasoning of Human Activity from IMU Data. *arXiv preprint*. ArXiv:2406.14498 [cs].

Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? *arXiv preprint*. ArXiv:2403.02727 [cs].

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint*. ArXiv:2306.14795 [cs].

Biao Jiang, Xin Chen, Ailing Zeng, Xinru Sun, Fukun Yin, Xianfang Zeng, Xuanyang Zhang, Gang Yu, and Tao Chen. Causal Motion Tokenizer for Streaming Motion Generation.

Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang YU, and Jiayuan Fan. 2024. MotionChain: Conversational Motion Controllers via Multimodal Prompts. *arXiv preprint*. ArXiv:2404.01700 [cs] TLDR: This work presents MotionChain, a conversational human motion controller to generate continuous and long-term human motion through multimodal prompts, coupled with a Vision-Motion-aware Language model, which comprehends each instruction in multi-turn conversation and generates human motions followed by these prompts.

Chuhao Jin, Haosen Li, Bingzi Zhang, Che Liu, Xiting Wang, Ruihua Song, Wenbing Huang, Ying Qin, Fuzheng Zhang, and Di Zhang. 2025. PlanMoGPT: Flow-Enhanced Progressive Planning for Text to Motion Synthesis. *arXiv preprint*. ArXiv:2506.17912 [cs].

Kanghyun Jo, Jehwan Choi, Kwanho Kim, Seongmin Kim, Duy-Linh Nguyen, Xuan-Thuy Vo, Adri Priadana, and Tien-Dat Tran. 2025. Artificial Behavior Intelligence: Technology, Challenges, and Future Directions. In *2025 17th International Conference on Human System Interaction (HSI)*, pages 01–10, Ulsan, Korea, Republic of. IEEE.

Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2023. Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 31–36, Brisbane, Australia. IEEE.

Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2025. Morag - Multi-Fusion Retrieval Augmented Generation for Human Motion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4564–4573, Tucson, AZ, USA. IEEE.

Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. 2024. IMUGPT 2.0: Language-Based Cross Modality Transfer for Sensor-Based Human Activity Recognition. *arXiv preprint*. ArXiv:2402.01049 [cs].

Zikang Leng, Hyeokhyen Kwon, and Thomas Plötz. 2023. Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition. *arXiv preprint*. ArXiv:2305.03187 [cs].

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. 2025a. GENMO: A GENeralist Model for Human MOtion. *arXiv preprint*. ArXiv:2505.01425 [cs].

Ronghui Li, Lingxiao Han, Shi Shu, Yueyao Liu, Yukang Lin, Yue Ma, Jie Guo, Ziwei Liu, and Xiu Li. 2025b. A Motion is Worth a Hybrid Sentence: Taming Language Model for Unified Motion Generation by Fine-grained Planning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1404–1413, Dublin Ireland. ACM. TLDR: This work introduced a novel motion-to-text generation method that combines atomic motion operators with GPT-4o, resulting in 68.2 million fine-grained textual descriptions across diverse modalities, and is the first to support multimodal condition joint driven motion generation tasks.

Zechen Li, Baiyu Chen, Hao Xue, and Flora D. Salim. 2025c. ZARA: Zero-shot Motion Time-Series Analysis via Knowledge and Retrieval Driven LLM Agents. *arXiv preprint*. ArXiv:2508.04038 [cs].

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. 2025d. SensorLLM: Aligning Large Language Models with Motion Sensors for Human Activity Recognition. *arXiv preprint*. ArXiv:2410.10624 [cs].

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T. Yang. 2025e. LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioning. *arXiv preprint*. ArXiv:2410.07093 [cs].

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2024. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *arXiv preprint*. ArXiv:2307.00818 [cs].

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*, pages 1–8, Mountain View California USA. ACM. TLDR: A new large scale benchmark (PKU-MMD) for continuous skeleton-based human action understanding and cover a wide range of complex human activities with well annotated information to benefit future researches on action detection for the community.

Frank Liu, Yuean Wang, and Hai Ci. 2024a. Object-Driven Human Motion Generation from Images. In *Proceedings of the 1st Workshop on Efficiency, Security, and Generalization of Multimedia Foundation Models*, pages 1–5, Melbourne VIC Australia. ACM.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-modal Dataset for Conversational Gestures Synthesis. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13667, pages 612–630. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. 2024b. Programmable Motion Generation for Open-Set Motion Control Tasks. *arXiv preprint*. ArXiv:2405.19283 [cs].

Ji Ma, Jiale Wu, Haoyu Wang, Yanze Zhang, Xiao Xie, Zheng Zhou, Hui Zhang, Jiachen Wang, and Yingcai Wu. 2025a. T3Set: A Multimodal Dataset with Targeted Suggestions for LLM-based Virtual Coach in Table Tennis Training.

Xinnan Ma, Yaochen Li, Limeng Zhao, ChenXu Zhou, and Yuncheng Xu. 2025b. SCALE-Pose: Skeletal Correction and Language Knowledge-assisted for 3D Human Pose Estimation. In Zhouchen Lin, Ming-Ming Cheng, Ran He, Kurban Ubul, Wushouer Silamu, Hongbin Zha, Jie Zhou, and Cheng-Lin Liu, editors, *Pattern Recognition and Computer Vision*, volume 15041, pages 578–592. Springer Nature Singapore, Singapore. Series Title: Lecture Notes in Computer Science.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. *arXiv preprint*. ArXiv:1904.03278 [cs].

Gabriel Maldonado, Armin Danesh Pazho, Ghazal Alinezhad Noghre, Vinit Katariya, and Hamed Tabkhi. 2025. MoCLIP: Motion-Aware Fine-Tuning and Distillation of CLIP for Human Motion Generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2922–2932, Nashville, TN, USA. IEEE. TLDR: MoCLIP is introduced, a fine-tuned CLIP model with an additional motion encoding head, trained on motion sequences using contrastive learning and tethering loss, establishing it as a robust framework for enhancing motion generation.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile Sensor Data Anonymization. In *Proceedings of the International Conference on Internet of Things Design and*

Implementation, pages 49–58. ArXiv:1810.11546 [cs].

Honglei Miao, Fan Ma, Ruijie Quan, Kun Zhan, and Yi Yang. Autonomous LLM-Enhanced Adversarial Attack for Text-to-Motion.

Jiawei Mo, Yixuan Chen, Rifen Lin, Yongkang Ni, Feng Liang, Min Zeng, Xiping Hu, and Min Li. 2025. MoChat: Joints-Grouped Spatio-Temporal Grounding Multimodal Large Language Model for Multi-Turn Motion Comprehension and Description. *IEEE Journal of Biomedical and Health Informatics*, pages 1–15.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253, Singapore. Association for Computational Linguistics. TLDR: Several new IMU-based Wearable AI applications such as motion-based media search, or an LM-based multi-modal reasoning with motion sensor data – all using text as the grounding platform are introduced.

Andrés Obludzyner, Fabio Zaldivar, and Oscar E. Ramos. 2024. A Survey on Large Language Models for Motion Generation from Text. In *2024 IEEE XXXI International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–8, Lima, Peru. IEEE.

Francisco Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, 16(1):115.

Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. *arXiv preprint*. ArXiv:2403.19857 [cs].

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71.

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data*, 4(4):236–252. ArXiv:1607.03827 [cs].

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. 2021. BABEL: Bodies, Action and Behavior with English Labels. *arXiv preprint*. ArXiv:2106.09696 [cs].

Muhammad Hasham Qazi, Sandesh Kumar, and Farhan Khan. 2024. LLM-Guided Speech Processing for 3D Human Motion Generation. In *2024 26th International Multi-Topic Conference (INMIC)*, pages 1–6, Karachi, Pakistan. IEEE.

Siyi Qian, Jian Fang, Yuzhou Mao, Yayun Zou, Wentao Zhang, and Haiwei Xue. 2025. Human Motion Generation in 3D Scenes from Open-Ended Textual Instructions with MLLM Planning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9404–9413, Dublin Ireland. ACM.

Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. 2023. Story-to-Motion: Synthesizing Infinite and Controllable Character Animation from Long Text. *arXiv preprint*. ArXiv:2311.07446 [cs].

Minghui Qiu, Cekai Weng, Mingming Fan, and Kaishun Wu. 2025. Towards Customizable Foundation Models for Human Activity Recognition with Wearable Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–29.

Haoxuan Qu, Yujun Cai, and Jun Liu. 2024. LLMs are Good Action Recognizers. *arXiv preprint*. ArXiv:2404.00532 [cs].

Pengrui Quan, Xiaomin Ouyang, Jeya Vikranth Jeyakumar, Ziqi Wang, Yang Xing, and Mani Srivastava. 2025. SensorBench: Benchmarking LLMs in Coding-Based Sensor Processing. *arXiv preprint*. ArXiv:2410.10741 [cs].

Mohan Ramesh and Fabian B. Flohr. 2024. Walk-the-Talk: LLM driven pedestrian motion generation. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 3057–3062, Jeju Island, Korea, Republic of. IEEE.

Lala Shakti Swarup Ray, Bo Zhou, Sungho Suh, and Paul Lukowicz. 2025. Initial Findings on Sensor based Open Vocabulary Activity Recognition via Text Embedding Inversion. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 685–688, Washington DC, DC, USA. IEEE.

Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109, Newcastle, United Kingdom. IEEE.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. 2025. Pulse-PPG: An Open-Source Field-Trained PPG Foundation Model for Wearable Applications across Lab and Field Settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the*

20th ACM International Conference on Multimodal Interaction, pages 400–408, Boulder CO USA. ACM. TLDR: This work introduces WESAD, a new publicly available dataset for wearable stress and affect detection that bridges the gap between previous lab studies on stress and emotions, by containing three different affective states (neutral, stress, amusement).

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *arXiv preprint*. ArXiv:1604.02808 [cs].

Ziwei Shan, Yaoyu He, Chengfeng Zhao, Jiashen Du, Jingyan Zhang, Qixuan Zhang, Jingyi Yu, and Lan Xu. 2025. Mojito: LLM-Aided Motion Instructor with Jitter-Reduced Inertial Tokens. *arXiv preprint*. ArXiv:2502.16175 [cs].

Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga. 2014. Fusion of Smartphone Motion Sensors for Physical Activity Recognition. *Sensors*, 14(6):10146–10176.

Milyun Ni'ma Shoumi and Sozo Inoue. Leveraging the Large Language Model for Activity Recognition: A Comprehensive Review.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart Devices are Different: Assessing and MitigatingMobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 127–140, Seoul South Korea. ACM.

Jie Su, Fengtong Ge, Zhenyu Wen, Taotao Li, Yang Bai, Yejian Zhou, and Xiaoqin Zhang. 2025. *IMUZero:* Zero-Shot Human Activity Recognition by Language-Based Cross Modality Fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):1–28.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. 2025. A Survey on Human Interaction Motion Generation. *arXiv preprint*. ArXiv:2503.12763 [cs] TLDR: This survey presents, for the first time, a comprehensive overview of the literature in human interaction motion generation and systematically review existing solutions and datasets across three primary interaction tasks – human-human, human-object, and human-scene interactions.

Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. 2024. LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–9. ArXiv:2405.03485 [cs] TLDR: LGTM utilizes a diffusion-based architecture and aims to address the challenge of accurately translating textual descriptions into semantically coherent human motion in computer animation, and proposes a two-stage pipeline to overcome this challenge.

Timo Sztyler and Heiner Stuckenschmidt. 2016. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9, Sydney, Australia. IEEE.

Kojiro Takeyama, Yimeng Liu, and Misha Sra. 2025. TR-LLM: Integrating Trajectory Data for Scene-Aware LLM-Based Human Action Prediction. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 18708–18715, Hangzhou, China. IEEE.

Zhenhui Tang, Jiahao Li, Ping Guo, Bowen Tian, Qingjun Xing, Xuyang Xing, and Peng Wang. 2025. FitnessAgent: A Unified Agent Framework for Open-Set and Personalized Fitness Evaluation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12437–12444, Atlanta, GA, USA. IEEE.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. *arXiv preprint*. ArXiv:2203.08063 [cs] TLDR: Although CLIP has never seen the motion domain, MotionCLIP offers unprecedented text-to-motion abilities, allowing out-of-domain actions, disentangled editing, and abstract language specification.

Wenbo Tian, Ruting Lin, Hongxian Zheng, Yaodong Yang, Geng Wu, Zihao Zhang, and Zhang Zhang. 2025a. SportsGPT: An LLM-driven Framework for Interpretable Sports Motion Assessment and Training Guidance. *arXiv preprint*. ArXiv:2512.14121 [cs].

Ye Tian, Xiaoyuan Ren, Zihao Wang, Onat Gungor, Xiaofan Yu, and Tajana Rosing. 2025b. DailyLLM: Context-Aware Activity Log Generation Using Multi-Modal Sensors and LLMs. In *2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pages 372–380, Chicago, IL, USA. IEEE.

Chongyang Wang, Yuan Feng, Lingxiao Zhong, Siyi Zhu, Chi Zhang, Siqi Zheng, Chen Liang, Yuntao Wang, Chengqi He, Chun Yu, and Yuanchun Shi. 2024a. UbiPhysio: Support Daily Functioning, Fitness, and Rehabilitation with Action Understanding and Feedback in Natural Language. *arXiv preprint*. ArXiv:2308.10526 [cs].

Fuming Wang, Wenlong Wang, Dahua Gao, Xunliang Huang, Xiaodan Song, Haoyuan Sun, and Cheng Peng. 2025a. A LLM-guided hybrid Mamba-Transformer architecture for part-to-whole motion synthesis. *Computer Vision and Image Understanding*, 262:104549.

Ruoyu Wang, Yangfan He, Tengjiao Sun, Xiang Li, and Tianyu Shi. 2025b. UniTMGE: Uniform Text-Motion Generation and Editing Model via Diffusion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6104–6114, Tucson, AZ, USA. IEEE.

Wenlong Wang, Dahua Gao, and Xinyu Liu. 2025c. LLM-Guided Dual-Branch Diffusion Model for Fine-Grained Motion Synthesis. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Hyderabad, India. IEEE.

Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. 2025d. Fg-T2M++: LLMs-Augmented Fine-Grained Text Driven Human Motion Generation. *arXiv preprint*. ArXiv:2502.05534 [cs].

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. 2024b. MotionGPT-2: A General-Purpose Motion-Language Model for Motion Generation and Understanding. *arXiv preprint*. ArXiv:2410.21747 [cs] TLDR: The proposed MotionGPT-2 is highly adaptable to the challenging 3D holistic motion generation task, enabled by the innovative motion discretization framework, Part-Aware VQVAE, which ensures fine-grained representations of body and hand movements.

Yuan Wang, Yali Li, Xiang Li, and Shengjin Wang. a. HSI-GPT: A General-Purpose Large Scene-Motion-Language Model for Human Scene Interaction.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. b. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. c. InterControl: Zero-shot Human Interaction Generation by Controlling Every Joint.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models.

Qingxin Wei, Jiaming Huang, Yi Gao, and Wei Dong. 2025a. One Model to Fit Them All: Universal IMU-based Human Activity Recognition with LLM-assisted Cross-dataset Representation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–22. TLDR: OneHAR is proposed, an LLM-assisted universal IMU-based HAR system designed to achieve "one model to fit them all" — just one model that can adapt to diverse HAR datasets without any dataset-specific operation.

Ran Wei, Hui Jie Zhang, Chang Cao, Fang Zhang, Jun Ling Gao, Xiao Tian Li, and Lei Geng. 2025b. Large Language Model-Based Spatio-Temporal Semantic Enhancement for Skeleton Action Understanding. *IET Computer Vision*, 19(1):e70041.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. 2025. MG-MotionLLM: A Unified Framework for Motion Comprehension and Generation across Multiple Granularities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27849–27858, Nashville, TN, USA. IEEE. TLDR: This work introduces a comprehensive multi-granularity training scheme by incorporating a set of novel auxiliary tasks, such as localizing temporal boundaries of motion segments via detailed text as well as motion detailed captioning, to facilitate mutual reinforcement for motion-text modeling across various levels of granularity.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2024. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. *arXiv preprint*. ArXiv:2405.17013 [cs].

Haojun Xu, Yan Gao, Zheng Hui, Jie Li, and Xinbo Gao. 2025a. Language Knowledge-Assisted Representation Learning for Skeleton-Based Action Recognition. *IEEE Transactions on Multimedia*, 27:5784–5799.

Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233, Coimbra Portugal. ACM. TLDR: This work presents LIMU-BERT, a novel representation learning model that can make use of unlabeled IMU data and extract generalized rather than task-specific features and significantly outperform existing approaches in two typical IMU sensing applications.

Lilin Xu, Kaiyuan Hou, and Xiaofan Jiang. 2025b. Exploring the Capabilities of LLMs for IMU-based Fine-grained Human Activity Understanding. In *Proceedings of the 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things*, pages 13–18, Irvine CA USA. ACM.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction.

Hua Yan, Heng Tan, Yi Ding, Pengfei Zhou, Vinod Namboodiri, and Yu Yang. 2025. Large Language Model-Guided Semantic Alignment for Human Activity Recognition. *arXiv preprint*. ArXiv:2410.00003 [cs].

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Tingbing Yan, Wenzheng Zeng, Yang Xiao, Xingyu Tong, Bo Tan, Zhiwen Fang, Zhiguo Cao, and Joey Tianyi Zhou. 2024. CrossGLG: LLM Guides One-shot Skeleton-based 3D Action Recognition in a Cross-level Manner. *arXiv preprint*. ArXiv:2403.10082 [cs].

Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. 2024. OmniMotionGPT: Animal Motion Generation with Limited Data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1249–1259, Seattle, WA, USA. IEEE. TLDR: This work designs a model architecture that imitates Generative Pretraining Transformer (GPT), utilizing prior knowledge learned from human data to the animal domain, and introduces AnimalML3D, the first text-animal motion dataset with 1240 animation sequences spanning 36 different ani-mal identities.

Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. 2023. MoConVQ: Unified Physics-Based Motion Control via Scalable Discrete Representations. *arXiv preprint*. ArXiv:2310.10198 [cs].

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, and Angelica Lim. 2025. MotionScript: Natural Language Descriptions for Expressive 3D Human Motions. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21574–21581, Hangzhou, China. IEEE.

Qilang Ye, Yu Zhou, Lian He, Jie Zhang, Xuanming Guo, Jiayu Zhang, Mingkui Tan, Weicheng Xie, Yue Sun, Tao Tan, Xiaochen Yuan, Ghada Khoriba, and Zitong Yu. 2025. SUGAR: Learning Skeleton Representation with Visual-Motion Knowledge for Action Recognition. *arXiv preprint*. ArXiv:2511.10091 [cs].

Xiaofan Yu, Lanxiang Hu, Benjamin Reichman, Dylan Chu, Rushil Chandrupatla, Xiyuan Zhang, Larry Heck, and Tajana S. Rosing. 2025. SensorChat: Answering Qualitative and Quantitative Questions during Long-term Multimodal Sensor Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35.

Lin Yuan, Zhen He, Qiang Wang, Leiyang Xu, and Xiang Ma. 2022. SkeletonCLIP: Recognizing Skeleton-based Human Actions with Text Prompts. In *2022 8th International Conference on Systems and Informatics (ICSAI)*, pages 1–6, Kunming, China. IEEE.

Jie Zeng, Yoshiki Takahashi, Yukiko I. Nakano, Tatsuya Sakato, and Hannes Högni Vilhjálmsson. 2024. Modifying Gesture Style with Impression Words. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 1–9, GLASGOW United Kingdom. ACM.

Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. a. SGAR: Structural Generative Augmentation for 3D Human Motion Retrieval.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023a. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. *arXiv preprint*. ArXiv:2301.06052 [cs].

Jiawen Zhang, Shuai Han, Yiwei Weng, Yinong Hu, Mingyu Zhang, Mingzhou Cheng, and Guozong Zhang. 2025a. Large language model-driven framework for inertial measurement unit-based worker activity recognition. *Automation in Construction*, 180:106594.

Jusheng Zhang, Jinzhou Tang, Sidi Liu, Mingyan Li, Sheng Zhang, Jian Wang, and Keze Wang. b. From Motion to Behavior: Hierarchical Modeling of Humanoid Generative Behavior Control.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023b. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv preprint*. ArXiv:2304.01116 [cs].

Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K Gupta, and Jingbo Shang. c. UniMTS: Unified Pre-training for Motion Time Series.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. 2025b. SensorLM: Learning the Language of Wearable Sensors. *arXiv preprint*. ArXiv:2506.09108 [cs].

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis. *ACM Transactions on Graphics*, 43(4):1–17. ArXiv:2405.09814 [cs].

Zhikai Zhang, Yitang Li, Haofeng Huang, Mingxian Lin, and Li Yi. 2025c. FreeMotion: MoCap-Free Human Motion Synthesis with Multimodal Large Language Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15081, pages 403–421. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Yilin Zhao, Yihong Chen, Rui Tang, and Wenjie Zhao. 2025. MultiagentsCR: A Multi-Agent Collaborative Reasoning Framework Based on LLM for Human Activity Recognition. In *2025 10th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pages 90–95, Nanchong, China. IEEE.

Wenqi Zheng and Yutaka Arakawa. 2025. A Review of Wearables-Based Activities of Daily Living Recognition With LLMs: Overview, Progress, and Trends. *IEEE Access*, 13:143813–143830.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. 2023. AvatarGPT: All-in-One Framework for Motion Understanding, Planning, Generation and Beyond. *arXiv preprint*. ArXiv:2311.16468 [cs].

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. 2025a. MotionGPT3: Human Motion as a Second Modality. *arXiv preprint*. ArXiv:2506.24086 [cs].

Chenhui Zhu, Yilu Wu, Shuai Wang, Gangshan Wu, and Limin Wang. 2025b. MotionRAG: Motion Retrieval-Augmented Image-to-Video Generation. *arXiv preprint*. ArXiv:2509.26391 [cs].

Man Zhu, Pengfei Huang, Xiaolong Xu, Houpeng He, and Lijie Zhang. 2025c. BoxingPro: An IoT-LLM Framework for Automated Boxing Coaching via Wearable Sensor Data Fusion. *Electronics*, 14(21):4155.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. *arXiv preprint*. ArXiv:2210.06551 [cs].

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2024. Human Motion Generation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449.

A Appendix

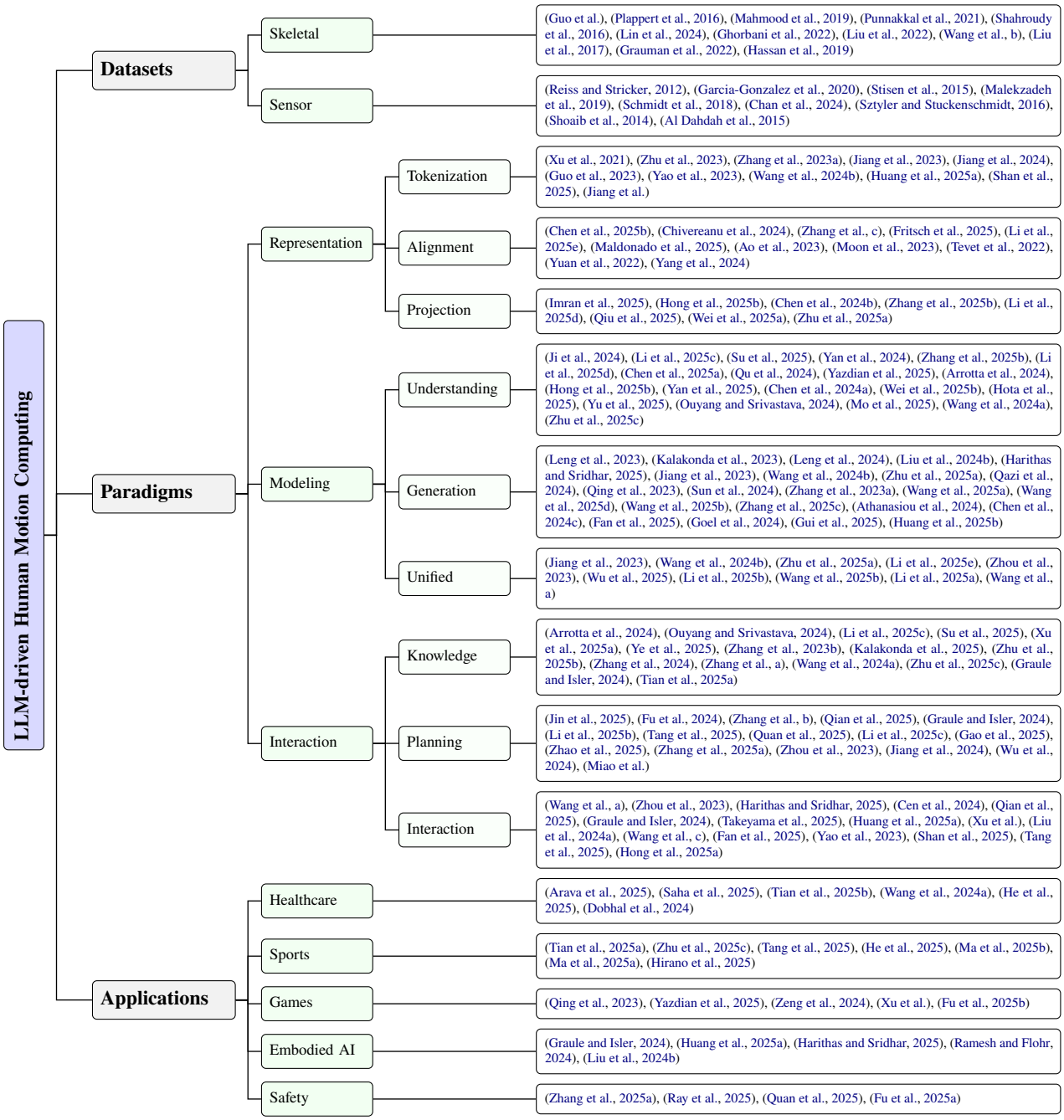

Figure 5: Taxonomy of LLM-based Human Motion Computing.

Table 1: This article reviews a horizontal comparison with existing related reviews. The unique value of this work lies in its proposed full-stack classification system and the innovative and in-depth integration of the two major research camps of skeletal movement and wearable sensors under the guidance of the PRISMA methodology.

Core Literature (Author/Year)	**PRISMA Method**	**Full-stack Framework**	**Skeleton Coverage**	**Sensor Coverage**	**Agent Planning**	**Scenario Summary**	**Future Planning**
Zhu et al. (Zhu et al., 2024)	–	–	✓	–	–	Partial	Partial
Obludzyner (Obludzyner et al., 2024)	–	–	✓	–	–	–	Partial
Sui et al. (Sui et al., 2025)	–	–	✓	–	–	Partial	Partial
Ferrara (Ferrara, 2024)	–	–	–	✓	–	Medical	✓
Arava et al. (Arava et al., 2025)	–	–	–	✓	–	Fitness/Health	✓
Zheng et al. (Zheng and Arakawa, 2025)	–	–	–	✓	–	Daily Life	✓
Jo et al. (Jo et al., 2025)	–	–	–	–	–	Driving/Surv.	✓
Shoumi (Shoumi and Inoue)	✓(SLR)	–	✓	✓	–	Partial	✓
Cai et al. (Cai et al., 2025)	–	–	–	✓	–	Indust./Health	✓
Ours (This Survey)	✓	✓**(Unique)**	✓	✓	✓**(Unique)**	✓	✓

Figure 6: PRISMA-based literature screening process. We queried four databases (ACM, IEEE, Scopus, ACL) using the logic ("LLM" OR "GPT") AND ("Human Motion" OR "IMU" OR "Skeleton") for records published since 2023. From 1,469 initial records (searched Nov 2025), 1,372 remained after deduplication. We applied four inclusion criteria: (IC1) keyword intersection in titles/abstracts; (IC2) removal of superficial mentions; (IC3) exclusion of out-of-scope domains (e.g., pure computer vision, WiFi sensing); and (IC4) focusing on core human motion modalities (Skeleton, IMU, etc.) where LLMs play a central methodological role. This resulted in 101 included records, further refined into 50 core studies and 21 datasets.

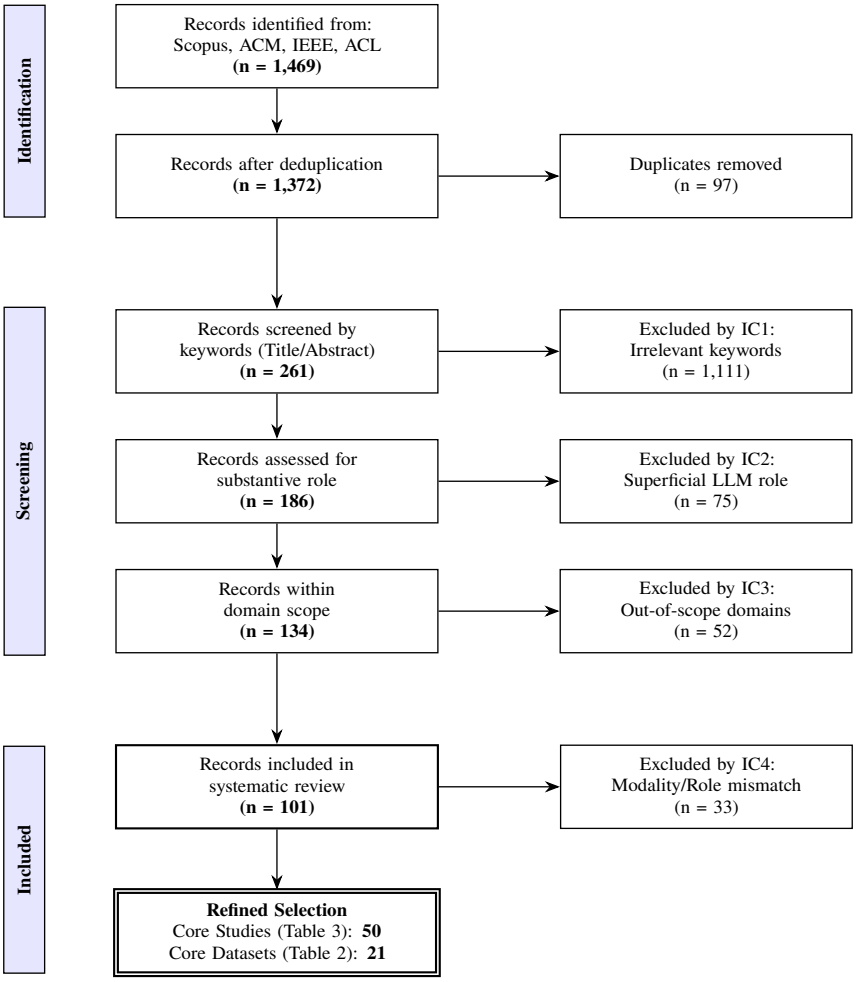

Table 2: The features of 21 core multimodal datasets were systematically sorted out, covering skeletal motion and sensor modalities. The scale, task type, evaluation metrics and semantic granularity were statistically analyzed, providing fundamental support for the construction of an evaluation system for large-scale human motion models.

Dataset	**Modality**	**Scale**	**Semantic Granularity**	**Key Tasks**	**Metrics**
HumanML3D (Guo et al.)	Skeleton, Text	14.6k seq	Fine-grained scripts of actions/styles	T2M Generation, Captioning	FID, R-Prec
KIT-ML (Plappert et al., 2016)	Skeleton, Text	3.9k seq	Holistic natural language for locomotion	Motion Generation, Retrieval	FID, R-Prec
BABEL (Punnakkal et al., 2021)	Skeleton, Text	43.5h	Hierarchical sequence and frame labels	Recognition, Segmentation	Acc, F1
AMASS (Mahmood et al., 2019)	Skeleton	42h	Kinematic pose parameters (unlabeled)	Tokenizer Pretrain, Repr.	Recon. Err
NTU-120 (Shahroudy et al., 2016)	Skeleton, Video.	114k clips	Discrete labels for 120 action classes	Recognition, ZSL	CS Acc.
Motion-X (Lin et al., 2024)	Skeleton., Text, Vid.	144h	Whole-body scripts and pose labels	Expressive Gen, Mesh Recovery	FID, R-Prec
BEAT (Liu et al., 2022)	Skeleton., Aud., Text	76h	Semantic/emotional relevance labels	Conversational Gesture Gen	SRGR, FID
Ego4D (Grauman et al., 2022)	Video., Aud., IMU	3,670h	Dense narratives and NL queries	Motion Prediction, Interaction	mAP, Acc
HUMANISE (Wang et al., b)	Skeleton., Text, 3D	19.6k seq and 643 scenes	Template interaction and object desc.	Scene-aware Gen, HSI	Phys. Real
PROX (Hassan et al., 2019)	Skeleton., Video., 3D	20 subjects	Proximal and contact constraints	Pose Estimation, Interaction	Joint Err
PKU-MMD (Liu et al., 2017)	Skeleton., Video., Dep.	1,076 seq	Temporal boundaries with 51 categories	Action Detection, Recog.	mAP
ZeroEGGS (Ghorbani et al., 2022)	Skeleton	19 styles	Stylized example-based descriptions	Zero-shot Gesture Gen	Style Score
UCI-HAR (Garcia-Gonzalez et al., 2020)	IMU	19 persons	Discrete labels for daily activities	Zero-shot HAR, Sig-to-Txt	Acc, F1
PAMAP2 (Reiss and Stricker, 2012)	IMU, Phys.	9 subjects	Labels defined by physical intensity	Intensity Class., HAR	F1-Score
HHAR (Stisen et al., 2015)	IMU	36 devices	Device/workload heterogeneous labels	Cross-dataset Learning	Accuracy
MotionSense (Malekzadeh et al., 2019)	IMU	24 users	Time-series states and user attributes	Signal Decoding, Attribution	Accuracy
RealWorld (Sztyler and Stuckenschmidt, 2016)	IMU	15 users	Position-aware activity/location labels	Localization, HAR	F-Measure
Shoaib (Shoaib et al., 2014)	IMU	10 users	Standardized labels for fused sensor	Multi-sensor Fusion, Recog.	Accuracy
MHEALTH (Al Dahdah et al., 2015)	IMU, Phys.	10 subjects	Health-related activities/vital signs	Health Monitoring, HAR	Accuracy
WESAD (Schmidt et al., 2018)	IMU, Phys.	15 subjects	Contextual affective and stress labels	Affect Detection, Reasoning	Macro F1
CAPTURE-24 (Chan et al., 2024)	IMU, Video.	3,883h	Annotated free-living behavioral context	Behavioral Tracking	Macro F1

Table 3: Systematic taxonomy and benchmarking of 50 core studies. Studies are categorized into Representation (P1), Modeling (P2), and Agent Interaction (P3) layers. ✓/✗ denote inclusion/exclusion. Abbreviations: [Modality] Skl.: Skeleton, Vid.: Video, Txt.: Text, Obj.: 3D Object, Rob.: Robot; [Task] Under.: Understanding, Gen.: Generation, Edit: Editing, Plan: Planning.

Model / Study	**Modality**	**Task**	**P1**	**P2**	**P3**	**Year**	**Base Model**	**Code**
MotionFix (Athanasiou et al., 2024)	Text, Skl.	Edit	✓	✓	✗	2024	T5	Yes
Cen et al. (Cen et al., 2024)	Text, Skl, 3D	Gen	✓	✓	✓	2024	GPT-4	Yes
Sensor2Text (Chen et al., 2024b)	IMU, Text	Under	✓	✓	✗	2024	Llama-2	No
MotionLLM (Chen et al., 2025a)	Text, Skl, Vid	Under	✓	✓	✗	2025	Llama-2	Yes
ContextGPT (Arrotta et al., 2024)	IMU, Text	Under	✗	✓	✓	2024	GPT-4	Yes
FitnessAgent (Tang et al., 2025)	Sensor, Text	Plan/Under	✗	✓	✓	2025	GPT-4	Yes
GENMO (Li et al., 2025a)	Text, Skl, Vid	Gen/Under	✓	✓	✗	2025	Llama-3	Yes
GG-LLM (Graule and Isler, 2024)	Text, Skl, 3D	Plan/Under	✗	✓	✓	2024	GPT-4	Yes
HAR-FM (Qiu et al., 2025)	IMU, Sensor	Under	✓	✗	✗	2024	BERT	Yes
HARGPT (Ji et al., 2024)	IMU, Text	Under	✓	✓	✗	2024	GPT-4	No
EgoLM (Hong et al., 2025a)	Vid, Skl, Text	Under	✓	✓	✓	2025	Llama-2	Yes
LLM4HAR (Hong et al., 2025b)	IMU, Text	Under	✓	✓	✗	2025	Llama-2	Yes
HSI-GPT (Wang et al., a)	Skl, 3D, Text	Gen	✓	✓	✓	2024	Llama-2	Yes
HOIGPT (Huang et al., 2025a)	Skl, Obj, Text	Gen	✓	✓	✓	2025	Llama-2	Yes
IMUGPT (Leng et al., 2023)	IMU, Text	Gen	✓	✓	✗	2023	GPT-3.5/4	No
Action-GPT (Kalakonda et al., 2023)	Text, Skl.	Gen	✓	✓	✗	2023	GPT-3	Yes
MoRAG (Kalakonda et al., 2025)	Text, Skl.	Gen	✓	✓	✓	2025	Llama-2	Yes
LanHAR (Yan et al., 2025)	IMU, Text	Under	✓	✓	✗	2025	Llama-2	Yes
IMUGPT 2.0 (Leng et al., 2024)	IMU, Text	Under	✓	✓	✗	2024	GPT-4	No
MotionUPG (Li et al., 2025b)	Text, Skl.	Gen	✓	✓	✓	2025	Llama-2	Yes
LIMU-BERT (Xu et al., 2021)	IMU	Under	✓	✗	✗	2021	BERT	Yes
LLaSA (Imran et al., 2025)	IMU, Text	Under	✓	✓	✗	2025	Llama-2	Yes
LLMSense (Ouyang and Srivastava, 2024)	Sensor, Text	Plan/Under	✗	✓	✓	2024	GPT-4	No
MoCLIP (Maldonado et al., 2025)	Skl, Text	Gen	✓	✓	✗	2025	CLIP	No
Mojito (Shan et al., 2025)	IMU, Text	Plan/Under	✓	✓	✓	2025	Llama-2	No
IMU2CLIP (Moon et al., 2023)	IMU, Vid, Text	Under	✓	✗	✗	2023	CLIP	Yes
Motion-Agent (Wu et al., 2024)	Skl, Text	Gen/Edit	✓	✓	✓	2023	GPT-4	Yes
MotionBERT (Zhu et al., 2023)	Skeleton	Under	✓	✗	✗	2023	Custom	Yes
MotionChain (Jiang et al., 2024)	Skl, Text, Img	Plan/Gen	✓	✓	✓	2024	Llama-2	Yes
MotionCLIP (Tevet et al., 2022)	Skl, Text	Gen	✓	✗	✗	2021	CLIP	Yes
MotionGlot (Harithas and Sridhar, 2025)	Text, Skl, Rob	Gen/Under	✓	✓	✓	2025	Llama-2	Yes
MotionGPT (Jiang et al., 2023)	Text, Skl.	Gen/Under	✓	✓	✗	2023	Flan-T5	Yes
MotionGPT-2 (Wang et al., 2024b)	Text, Skl, Hand	Gen/Under	✓	✓	✗	2024	Llama-2	Yes
MotionGPT-3 (Zhu et al., 2025a)	Text, Skl.	Gen/Under	✓	✓	✗	2025	Llama-3	Yes
MotionRAG (Zhu et al., 2025b)	Img, Vid, Mot	Gen	✓	✓	✗	2025	Custom	Yes
PlanMoGPT (Jin et al., 2025)	Text, Skl.	Gen	✓	✓	✗	2025	Llama-2	No
Qian et al. (Qian et al., 2025)	Skl, 3D, Text	Plan/Gen	✓	✓	✓	2025	GPT-4o	No
LLM-AR (Qu et al., 2024)	Skl, Text	Under	✓	✓	✗	2024	Llama-2	No
ReMoDiffuse (Zhang et al., 2023b)	Text, Skl.	Gen	✓	✓	✗	2023	CLIP	Yes
SensorLLM (Li et al., 2025d)	IMU, Text	Under	✓	✓	✗	2025	GPT-4o	No
SensorLM (Zhang et al., 2025b)	IMU, Sensor	Under	✓	✓	✗	2025	CoCa / CLIP	No
SGAR (Zhang et al., a)	Skl, Text	Under	✓	✓	✗	2025	Llama-2	No
T2M-GPT (Zhang et al., 2023a)	Text, Skl.	Gen	✓	✓	✗	2023	GPT	Yes
UbiPhysio (Wang et al., 2024a)	Sensor, Text	Plan/Under	✗	✓	✓	2024	GPT-4	No
oneHAR (Wei et al., 2025a)	IMU, Text	Under	✓	✓	✗	2025	Llama-3	No
MG-MotionLLM (Wu et al., 2025)	Text, Skl.	Gen/Under	✓	✓	✗	2025	Llama-3	Yes
ZARA (Li et al., 2025c)	IMU, Text	Under	✗	✓	✓	2025	GPT-4o	No
UniMTS (Zhang et al., c)	IMU, Text	Under	✓	✓	✗	2024	BERT	Yes
AvatarGPT (Zhou et al., 2023)	Text, Skl.	Plan/Gen	✓	✓	✓	2023	Llama-2	Yes
GBC (Zhang et al., b)	Text, Skl.	Plan/Gen	✗	✓	✓	2025	Llama-3	No