

UbiPhysio: Support Daily Functioning, Fitness, and Rehabilitation with Action Understanding and Feedback in Natural Language

CHONGYANG WANG¹, YUAN FENG², LINGXIAO ZHONG¹, and SIYI ZHU²,

¹Tsinghua University, China ²West China Hospital, Sichuan University, China

CHI ZHANG, SIQI ZHENG, CHEN LIANG, and YUNTAO WANG, Tsinghua University, China

CHENGQI HE, West China Hospital, Sichuan University, China

CHUN YU and YUANCHUN SHI, Tsinghua University, China

We introduce UbiPhysio, a milestone framework that delivers fine-grained action description and feedback in natural language to support people's daily functioning, fitness, and rehabilitation activities. This expert-like capability assists users in properly executing actions and maintaining engagement in remote fitness and rehabilitation programs. Specifically, the proposed UbiPhysio framework comprises a fine-grained action descriptor and a knowledge retrieval-enhanced feedback module. The action descriptor translates action data, represented by a set of biomechanical movement features we designed based on clinical priors, into textual descriptions of action types and potential movement patterns. Building on physiotherapeutic domain knowledge, the feedback module provides clear and engaging expert feedback. We evaluated UbiPhysio's performance through extensive experiments with data from 104 diverse participants, collected in a home-like setting during 25 types of everyday activities and exercises. We assessed the quality of the language output under different tuning strategies using standard benchmarks. We conducted a user study to gather insights from clinical experts and potential users on our framework. Our initial tests show promise for deploying UbiPhysio in real-life settings without specialized devices.

CCS Concepts: • Human-centered computing → Ubiquitous and mobile computing; Human computer interaction (HCI); • Applied computing → Life and medical sciences; • Computing methodologies → Natural language generation.

Additional Key Words and Phrases: action understanding, feedback generation, rehabilitation, fitness, activities of daily life

ACM Reference Format:

Chongyang Wang¹, Yuan Feng², Lingxiao Zhong¹, Siyi Zhu², Chi Zhang, Sisi Zheng, Chen Liang, Yuntao Wang, Chengqi He, Chun Yu, and Yuanchun Shi. 2023. UbiPhysio: Support Daily Functioning, Fitness, and Rehabilitation with Action Understanding and Feedback in Natural Language. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 0 (2023), 27 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 INTRODUCTION

People's domestic activities and exercise routines provide vital insights for health providers to understand their difficulties, problems, and behavioral changes [11]. Recently, home-based physiotherapy concerning everyday functioning and regular exercise has gained increasing importance, given the benefits they provide for enhancing individuals' quality of life [48], maintaining functional independence [14], combating chronic diseases [36], and

Authors' addresses: Chongyang Wang¹, wangchongyang@tsinghua.edu.cn; Yuan Feng²; Lingxiao Zhong¹; Siyi Zhu²,

¹Tsinghua University, China ²West China Hospital, Sichuan University, China; Chi Zhang; Sisi Zheng; Chen Liang; Yuntao Wang, Tsinghua University, China; Chengqi He, West China Hospital, Sichuan University, China; Chun Yu; Yuanchun Shi, Tsinghua University, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

2474-9567/2023/0-ART0 \$15.00

<https://doi.org/XXXXXX.XXXXXXX>

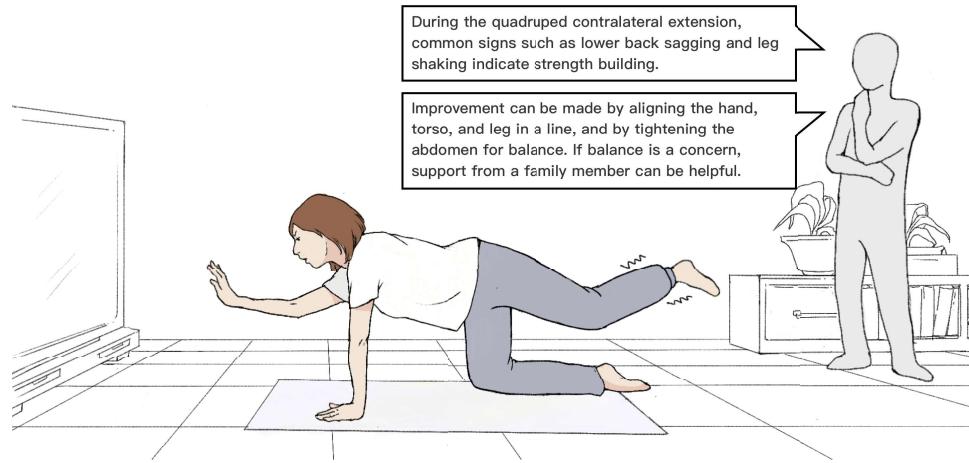


Fig. 1. By accurately understanding the detailed movement patterns of a person, a physio is able to provide guidance on how to make the exercise more effective and suggestions that could help them achieve the goal properly. Inspired by such vivid and expert interactions, this work aims to narrow the gap between ubiquitous technology and the basic skill of a physio in evaluating one's action and generating detailed feedback.

promoting recovery from illness or injury outside hospital settings [44]. However, without guidance, individuals can easily develop improper habitual behaviors that pose long-term musculoskeletal risks. For instance, bending to pick up a heavy bag could significantly harm the lower back, whereas squatting and then lifting the bag is a safer approach. Moreover, performing physical exercises without professional supervision can result in ineffective training or even injury due to incorrect postures and movements. Another concern is the lack of timely feedback in domestic settings, which can discourage those with chronic diseases from adhering to self-organized rehabilitation programs, especially given the high cost of clinical services [20, 45]. Recent advances in action captioning [18, 24, 31, 71, 73] and knowledge-driven natural language generation (NLG) [8, 13, 21, 53] present an opportunity to create a system that can perceive, understand, and offer expert-like feedback on home-based daily functional activities and exercises. Compared to existing works, our approach delves into fine-grained action descriptions that encompass both the type of action and potential movement patterns that could inform intervention. Furthermore, we introduce a simple yet efficient retrieval-based method that leverages the common-sense knowledge of a large language model (LLM) to generate professional feedback.

As illustrated in Fig 1, we receive motivations from the real life scenario, where a physio provides direct feedback on conducting an action properly and making the exercise doable and effective, grounded in a fine-grained understanding of the user's performance. In light of this, we ask: **how can technology support people's daily functioning and exercises with expert-like action understanding and feedback?** We introduce UbiPhysio, the first framework of its kind designed to function as a virtual physiotherapist, facilitating daily functional activities and supporting rehabilitation exercises. UbiPhysio accepts action data as input and connects it to natural language through self-supervised tokenization. In contrast to previous approaches, we propose employing biomechanical movement features and action-conditioned instruction tuning to deliver detailed action descriptions. Ultimately, leveraging the common-sense knowledge of a LLM, the framework generates vivid and professional feedback using a retrieval-enhanced prompting strategy.

Our experiment leverages data collected from 104 participants, capturing full-body action data while performing 25 different types of daily functioning and exercise activity. We established the experiment with an aim to reveal

the performance of participants at their homes. We achieved this by converting the lab space into a furnished, home-like environment, using comfortable suits instead of bandages to attach IMU sensors, and implementing a continuous, less intrusive experimental session. To evaluate the efficacy of UbiPhysio, we carried out rigorous cross-validation tests on unseen participants, measuring the accuracy of the system's language output with standard objective benchmarking tools. Additionally, we conducted subjective evaluations with our participants and clinical partners. The results strongly suggest that UbiPhysio has the potential to be successfully integrated into real-life scenarios, paving the way for future advancements in ubiquitous fitness and rehabilitation technology. The main contributions of this work are summarized below.

- We introduce UbiPhysio, a novel framework capable of describing human action in terms of its type and patterns-of-interest. It can also generate personalized and professional feedback to improve user engagement and performance.
- To enhance the framework, we implement several tuning strategies, including the integration of comprehensive biomechanical action features for precise action quantification and the use of action-conditioned instruction tuning to optimize language output.
- We evaluate the framework with data collected from 104 diverse participants in a furnished, home-like environment. To further demonstrate the framework's potential for real-world application, we conduct a user study and deployment test, which demonstrate promising results.

2 RELATED WORK

This section provides a literature review on relevant studies on action-centric applications, as well as multimodal language modeling with action data.

2.1 Analyzing Actions for Healthcare, Assistance, and Fitness

Recognizing the type or category of an action serves as the foundation of many ubiquitous systems and applications. However, human perception of actions extends beyond mere categorization, and are capable of inferring the emotional and physical status, needs, and difficulties, based on detailed action patterns [3, 4, 15]. Therein, the difference between a layperson and a domain expert lies in their ability to **accurately capture such patterns and respond with proper language**. This capability is crucial for applications where technology is expected to play the role of a specialist in relevant domains [68].

This work differs significantly from many previous studies focusing on activities of daily living (ADLs) [5, 26, 46], assisted fitness [10, 19, 27, 32], and remote rehabilitation [23, 57]. Most of these studies aim to identify discrete activity classes and provide general insights, such as revealing behavioral changes (e.g., a decrease in exercise) through longitudinal observations, detecting critical or abnormal activities (e.g., falls or prolonged bed rest) [12], and assessing health risks by tracking specific activities like drinking and eating [7, 50]. Some studies move on to evaluate the quality of an action using quantitative kinematic features to predict the development of chronic diseases [33, 55, 56]. They monitor the execution of exercises by calculating metrics (e.g., speed, range of movement, variations) against *gold standards* [10, 27, 32, 57], and detecting the presence of abnormal movement behavior [61] to estimate the difficulty one may have in specific activities.

Aside from the studies above on supporting people's healthcare, assistive and fitness needs with diverse ubiquitous technologies, we also found the following studies that move deeper into action understanding, focusing on tasks that require more specific insights or feedback. Guo et al. [19] highlights the importance of reminding the user to maintain proper posture during workouts. However, the feedback provided in their system is based on the comparison of low-level features like repetition consistency and temporal differences between normal and expert users, which is indirect for this purpose.

Wei et al. [67] proposed a system aimed at facilitating remote training for individuals with Parkinson's disease. This system is capable of action understanding and feedback generation for three actions (referred to as *tasks* in their study), namely squat, forward lunge, and backward lunge. By analyzing angular features computed from the pose estimation by Kinect [1], the system can identify errors in execution (i.e. violation of criteria set by physios such as '*keep the back knee straight*'). Based on this analysis, the system then provides recommendations, deciding whether the user should repeat the exercise, adjust the difficulty level (for instance, by reducing the rotation angle or introducing additional support), or progress to the next level. In a separate study, Wang et al. [64] only used a smartwatch to evaluate the quality of three upper-arm exercises, namely shoulder abduction, external rotation, and forward flexion, and provide feedback on the number of repetitions, range of movements, and stability of each exercise. Lee et al. [37] proposed a more comprehensive set of features to measure the performance of stroke patients performing three upper limb exercises. By facilitating a collaborative approach between the AI system and physiotherapists in determining which features to examine, they showed improved evaluation results. These studies rely on what we can categorize as *rule-based* approaches, where the features designed for action evaluation are derived from a convergence of quantitative computational methods (informed by full-body pose or data from a wearable device attached to a specific body part) and qualitative analysis of the specific action by domain experts. A major disadvantage of these approaches is the lack of generalizability of the features, particularly when confronted with more complex movements with diverse patterns [39]. Also, the feedback provided in these systems mainly evaluates the outcome or quality of the action, and falls short of guiding a person to act more properly.

2.2 Modeling Actions with Natural Language Description

Modeling actions with natural language description is essential for understanding and interpreting the semantics of human actions. Early investigations [28, 29, 49, 65, 69] on such a modeling process mainly focused on human motions with large body movements (e.g., waving the hand), general motion categories (e.g., walking around) and explicit descriptions (e.g., counts of a specific gesture pattern). The user's 3D pose series, represented in the positions and attitudes of a set of body joints (e.g., 24 body key points under the SMPL model [42]), was taken as the main input of motion-language models. Although Plappert et al. [49] and Yamada et al. [69] demonstrated the feasibility of establishing a bidirectional mapping between human whole-body motion and natural language, the target pose sequences were usually deterministic or with limited variety, probably due to insufficient motion feature descriptions and the hard-supervised training process. To enrich the description space for capturing the dynamics and kinetic features of human motion, Holden et al. [28, 29] utilized rigid body dynamic features such as joint velocity, orientation, and foot contact state for better modeling of human motion in temporal and spatial domains. Guo et al. [17] introduced a temporal variational autoencoder framework, guided by a learned distribution function, to construct a probabilistic mapping between motion sequences and textual descriptions.

Although the abovementioned work managed to bridge general motions and semantics, the problem become far more challenging given a finer granularity (e.g., mining certain micro-expressions) or specific attention patterns (e.g., telling the difference of different walking patterns) for the description target [59, 73]. This challenge originates from the inherent modality discrepancy between text and motion series, which would result in substantial ambiguity and information loss in the interpretation process. From a higher perspective, human motions are far more complex than they appear to be. For example, a "walking" motion series could potentially indicate the subject's mental state (e.g., casually or hurriedly) and physical state (e.g., any disease-related features from clinical observations). Aiming at these challenges, previous work have investigated different methods and models for enriching motion representation in different levels of hidden semantic spaces. For example, VQ-VAE (Vector Quantized Variational Autoencoder) [59] presented an elegant solution of discretizing the hidden distributions into parameterized "codebooks" to increase the capacity of hidden semantics. Guo et al. [18] proposed the use of

compact and discrete motion tokens to offer a more flexible and adaptable array of motion representation for generating text descriptions and the inverse process based on VQ-VAE [59], which achieved modeling explicit connections between atomic actions and fine-grained semantics. Kim et al. [35] present a joint representation model of human motion and language with contrastive learning by leveraging both unpaired and paired motion and language datasets. HUMANISE [66] investigated the representation of scene-aware and goal-oriented human motions. Taking advantages of the success of large language models (LLMs), Zhang et al. [71] introduced a LLM-based framework (e.g., a transformer architecture with 12 layers), leveraging the semantic expressiveness of LLMs, to yield high-quality discrete representations of human motion given language instructions as input. Later on, other language models like T5 [9, 51] and Llama [58] are also adopted to promote the accuracy in building action-language interactions with a similar pipeline [31, 73].

In our work, we took the first step to improve the capability of VQ-VAE for representing fine-grained motion patterns and a language model for action-conditioned action description in our UbiPhysio framework. Specifically, our work aimed at uncovering implicit motion features related to a variety of daily functioning, fitness, and rehabilitation actions, where we collect feedbacks and comments from clinical experts' observations, so that to provide accurate and professional training guidance.

3 METHOD

This section details our proposed framework, providing a comprehensive roadmap from the initial preprocessing of action data input to the final generation of detailed feedback.

3.1 Biomechanical Features-Driven Action Tokenization

In our proposed framework, we begin by discretizing the continuous action data into succinct tokens. This process facilitates the learning of the language model by establishing a mapping between these discrete action tokens and corresponding language tokens.

3.1.1 Biomechanical Features. We employ a discretization process driven by low-level features extracted from comprehensive full-body action data. As demonstrated in prior research [18, 31, 71, 73], this includes the position, orientation, and velocity of each joint, as well as specific global features pertinent to the root joint. Given a skeleton of 24 joints, as shown in Fig 2, a total of 287 features can be extracted. Such a methodology enables the generation of descriptive phrases that characterize a person's actions, for example: "*The person is standing up from a chair and walks forward in a circle*", and "*The person is bending down to pick up a thing and carry it forward with both hands*". However, our preliminary experiments, which replicated their pipelines, revealed limitations in producing accurate and consistent descriptions of diverse movement patterns in daily functioning and exercises. For instance, the model struggled to generate the description like "*This individual is practicing body side bend, but with reduced mobility and a tendency to bend the spine and knees*". To enhance the model's ability to capture detailed movement patterns - as emphasized by our clinical partners - we have developed an additional set of features, which we detail in the following.

- **Bilateral balance:**

$$\begin{aligned} \text{feats}_{BB, upper} &= \mathcal{D}(\mathbf{J}_{left\,forearm}, \mathbf{J}_{spine2}) - \mathcal{D}(\mathbf{J}_{right\,forearm}, \mathbf{J}_{spine2}) \\ \text{feats}_{BB, lower} &= \mathcal{D}(\mathbf{J}_{left\,foot}, \mathbf{J}_{hip}) - \mathcal{D}(\mathbf{J}_{right\,foot}, \mathbf{J}_{hip}) + \mathcal{D}(\mathbf{J}_{left\,leg}, \mathbf{J}_{hip}) - \mathcal{D}(\mathbf{J}_{right\,leg}, \mathbf{J}_{hip}) \end{aligned} \quad (1)$$

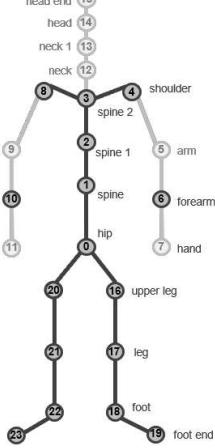


Fig. 2. The skeleton of the action data used in this study. Darker joints stand for those used for the computation of biomechanical features.

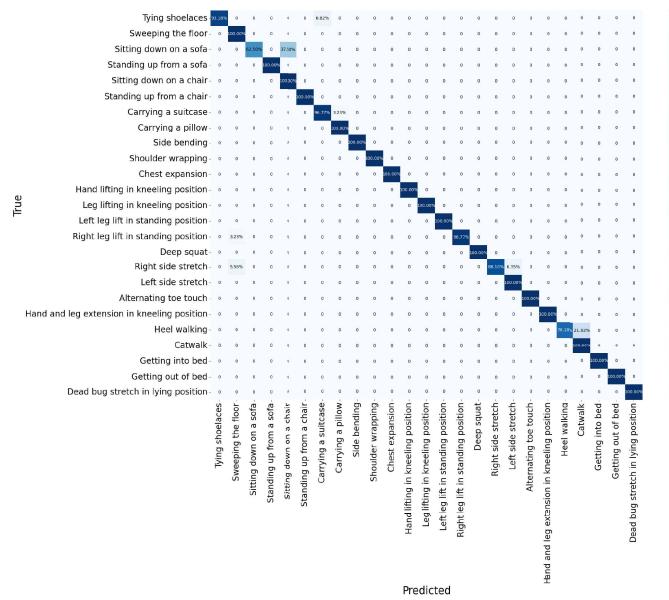


Fig. 3. The confusion matrix for the action classification result of our action recognition module.

• Inter-joint angle:

$$\text{feats}_{IJA} = [\angle(\mathbf{J}_{\text{left shoulder}}, \mathbf{J}_{\text{spine2}}, \mathbf{J}_{\text{spine1}}), \angle(\mathbf{J}_{\text{right shoulder}}, \mathbf{J}_{\text{spine2}}, \mathbf{J}_{\text{spine1}}), \angle(\mathbf{J}_{\text{spine2}}, \mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{spine}}), \angle(\mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{spine}}, \mathbf{J}_{\text{hip}}), \angle(\mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{left upperleg}}), \angle(\mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{right upperleg}}), \angle(\mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{left leg}}), \angle(\mathbf{J}_{\text{spine1}}, \mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{right leg}}), \angle(\mathbf{J}_{\text{left upperleg}}, \mathbf{J}_{\text{left leg}}, \mathbf{J}_{\text{left foot}}), \angle(\mathbf{J}_{\text{right upperleg}}, \mathbf{J}_{\text{right leg}}, \mathbf{J}_{\text{right foot}}), \angle(\mathbf{J}_{\text{left shoulder}}, \mathbf{J}_{\text{spine2}}, \mathbf{J}_{\text{hip}}), \angle(\mathbf{J}_{\text{right shoulder}}, \mathbf{J}_{\text{spine2}}, \mathbf{J}_{\text{hip}}), \angle(\mathbf{J}_{\text{left shoulder}}, \mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{right shoulder}}), \angle(\mathbf{J}_{\text{left leg}}, \mathbf{J}_{\text{left foot}}, \mathbf{J}_{\text{left footend}}), \angle(\mathbf{J}_{\text{right leg}}, \mathbf{J}_{\text{right foot}}, \mathbf{J}_{\text{right footend}})] \quad (2)$$

• Inter-line angle:

$$\text{feats}_{ILA} = \angle(\mathbf{J}_{\text{left shoulder}}, \mathbf{J}_{\text{right shoulder}}, \mathbf{J}_{\text{left upperleg}}, \mathbf{J}_{\text{right upperleg}}) \quad (3)$$

• Inter-joint distance and ratio:

$$\text{feats}_{IJR} = [\mathcal{D}(\mathbf{J}_{\text{left shoulder}}, \mathbf{J}_{\text{hip}}), \mathcal{D}(\mathbf{J}_{\text{right shoulder}}, \mathbf{J}_{\text{hip}}), \mathcal{D}(\mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{left foot}}), \mathcal{D}(\mathbf{J}_{\text{hip}}, \mathbf{J}_{\text{right foot}}), \mathcal{D}(\mathbf{J}_{\text{left forearm}}, \mathbf{J}_{\text{left leg}}), \mathcal{D}(\mathbf{J}_{\text{left forearm}}, \mathbf{J}_{\text{right leg}}), \mathcal{D}(\mathbf{J}_{\text{right forearm}}, \mathbf{J}_{\text{left leg}}), \mathcal{D}(\mathbf{J}_{\text{right forearm}}, \mathbf{J}_{\text{right leg}})] \quad (4)$$

$$\text{feats}_{IJR} = [\frac{\mathcal{D}(\mathbf{J}_{\text{left forearm}}, \mathbf{J}_{\text{right leg}})}{\mathcal{D}(\mathbf{J}_{\text{left forearm}}, \mathbf{J}_{\text{left leg}})}, \frac{\mathcal{D}(\mathbf{J}_{\text{right forearm}}, \mathbf{J}_{\text{left leg}})}{\mathcal{D}(\mathbf{J}_{\text{right forearm}}, \mathbf{J}_{\text{right leg}})}] \quad (5)$$

where \mathbf{J}_* refers to corresponding joints. $\mathcal{D}(*, *)$ represents the Euclidean distance, taking 3D joint positions as input, with $\mathcal{D}(a, b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2}$. The inter-joint (with three joints) and inter-line (with four joints) angle is computed as $\angle(a, b, c) = \arctan\left(\frac{\|\vec{ab} \times \vec{bc}\|}{\vec{ab} \cdot \vec{bc}}\right)$, and $\angle(a, b, c, d) = \arctan\left(\frac{\|\vec{ab} \times \vec{cd}\|}{\vec{ab} \cdot \vec{cd}}\right)$, respectively. It should be noted that the above features are computed for each timestep, presenting potential opportunities for the real-time deployment of our system. The total number of features computed here is 28.

We provide such details here as not only to support the replication of this work, but also to inform researchers working on rule-based applications about the features that could be considered for action pattern analysis.

3.1.2 Action Tokenization. By computing the features per timestep, for each action sequence of duration T , we have $\mathbf{X} = \{x_t\}_{t=1}^T$ and $x_i \in \mathbb{R}^{d_{action}}$, where d_{action} is the dimension of features, which is 315 for a pose data input of 24 joints. A VQ-VAE model [59] is further used to learn the discrete action tokens, using a codebook of length K , denoted as $\mathcal{B}\{b_k\}_{k=1}^K$ and $b_k \in \mathbb{R}^{d_{code}}$. Similar to other encoder-decoder architecture, the encoder of VQ-VAE first encodes the input action features into latent vectors with 1D convolution as $\mathcal{E}(\mathbf{X}) = \mathbf{H}$, with $\mathbf{H} = \{h_n\}_{n=1}^N$, $N = T/l$, and l denotes the downsampling rate of the convolutional encoder. Then, the quantization through the codebook \mathcal{B} is carried out to search for the most close counterpart vector b_k for each encoded latent vector h_n as:

$$z_n = \underset{b_k \in \mathcal{B}}{\operatorname{argmin}} \|h_n - b_k\|_2, \quad (6)$$

where the sequence of indexes of all the searched b_k in the codebook \mathcal{B} , denoted as $\mathbf{Q} = [q_1, \dots, q_N]$ with $q_i \in [1, 2, \dots, q_N]$, becomes the sequence of tokens for the input action, which is used to fine-tune a language model as described later. Thereon, the decoder reconstructs the input action data with $\mathbf{X}_r = \mathcal{D}(\mathbf{Z})$. The VQ-VAE is trained in a self-supervised way given the loss computed against the input action data as:

$$\mathcal{L}_{vqvae} = \mathcal{L}_{recon} + \|\mathbf{H} - sg[\mathbf{Z}]\|_2 + \beta \|sg[\mathbf{H}] - \mathbf{Z}\|_2, \quad (7)$$

where the three parts of the total loss stand for the reconstruction loss of the VAE, and the embedding and commitment losses of the quantization process, $sg[\cdot]$ denotes stop-gradient operation, β is a weighting hyperparameter. Motivated by the velocity regularization used in [71], the reconstruction loss is added with the regularization from reconstructing the biomechanical features as:

$$\mathcal{L}_{recon} = |\mathbf{X} - \mathbf{X}_r| + \alpha |\mathbf{X}[\text{bio}] - \mathbf{X}[\text{bio}]_r|, \quad (8)$$

where $\mathbf{X}[\text{bio}]$ stands for getting out the biomechanical features from the original (or reconstructed) data, α is a balancing hyperparameter. The L1 smooth loss is used in our experiment according to existing works [31, 71, 73].

3.1.3 Human Action Recognition Module. During our experiments, we found it effective to insert the action type information into the instruction for tuning language models to improve its performance on describing the detailed movement patterns. A similar finding is also reported in [61], where the inclusion of activity type information was found to be advantageous in improving the detection of specific movement behaviors. Furthermore, the action type information can also be used as a query to extract background knowledge from our pre-defined knowledge base, enhancing the feedback of a language model.

By utilizing the proposed action features extracted from each action instance as input, we trained a 1D convolutional neural network to classify the different action types, i.e., $y = \text{ConvNet}(\mathbf{X})$, $y \in \mathbf{Y}_C$, where y is the predicted action label, C is the number of action classes. For the dataset we collected, which is described in detail in the next section, the network is able to achieve an average macro F1 score of 0.9636 with a standard deviation of 0.0067, after five different runs of validation (each run divides participants with 85% for training, 5% for validation, and 10% for testing). The confusion matrix is shown in Fig 3. The mistakes are mostly about misclassifying sitting on the chair vs. on the sofa, and heel-walking vs. heel-to-toe walking. Please kindly refer to Appendix section for details about the convolutional backbones used in this module and the VQ-VAE model.

3.2 Action-Conditioned Description with a Language Model

An overview of our proposed Ubiphysio framework is shown in Fig 4. After the step 1 of extracting discrete action tokens, namely the sequence of indexes of codebook representations that found most close to our encoded motion features, we conduct the fine-tuning of a language model for action description in step 2. Following the

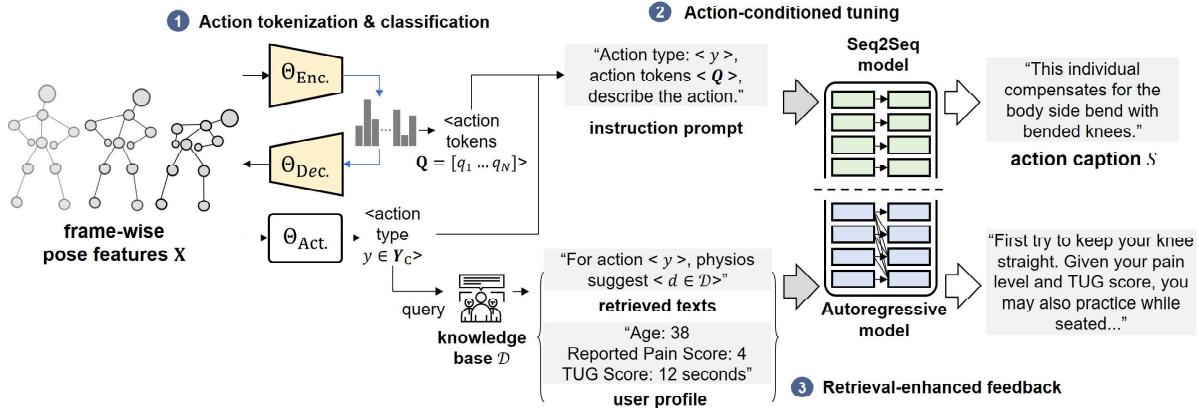


Fig. 4. With 3D body poses as input, we first calculate a set of movement features per timestep, which are designed to characterize the local dynamics of each body part and movement patterns. Then, a self-supervised feature quantization step is conducted to acquire discrete tokens for each action instance. A retrieval-enhanced large language model is further adopted to convert such tokens into descriptive texts, and generate vivid feedback and suggestion.

instruction structure adopted in [31, 73], we formalize the instruction prompt for each action instance using a mix of texts and action tokens as follows:

- Task prompt \mathcal{T} : *{I want you to act as an action interpreter. Given the type of human action and tokens representing the action, please generate a natural language description of the action.}*
- Condition input: *{The action you need to describe is as following, type: [y], tokens: [Q].}*

The action type information y is particularly added here to make action-conditioned input to the language model, given our experimental results that are reported later. We will also demonstrate that such a piece of information help the model better understand the movement patterns hidden in the action tokens. Thereon, for each action instance, the complete instruction tuning pair we prepare for a language model includes, the complete input prompt $\{\mathcal{T}, y, Q\}$, and the label of action description S_{gt} . It should be noted that the categorical label y is translated into texts here. The optimization of the language model (LM) is as follows:

$$\mathcal{L}_{lm} = - \sum_{\theta}^L \log p_{\theta}(S_{gt} | \{\mathcal{T}, y, Q\}), \quad (9)$$

where L denotes the output length of the language model.

3.3 Retrieval-Enhanced Feedback Generation

The common sense knowledge of existing large language models (LLMs) come from their pretraining against a huge amount of textual data. By providing the LLM (taking GPT-4¹ for example) with the detailed action description S , it could yet give some generic feedbacks as follows:

- Prompt: *{You are acting as an expert of sport rehabilitation, you will support a person (age: 32, reported pain score: 4 (0-10), Time-Up-and-Go score: 12 seconds) during the following session. From now, you will receive detailed action descriptions about their performances, to which you should provide instant feedback to help this person conduct each exercise properly. Action description: [This individual compensates for the body side bend with bended knees].}*

¹The OpenAI ChatGPT (<https://chat.openai.com/>) at its August 3rd, 2023 version.

- Original response: *{I noticed that you're bending your knees to compensate for the side bend. It's essential to maintain straight knees to get the most benefit from the exercise and to target the right muscles. Focus on keeping your knees extended and try to bend from your waist. This will ensure proper alignment and maximize the benefits of the exercise. Let's try that again and keep those knees straight!}*

As is shown, the LLM is able to suggest a direct way of correcting the action and help the user to maintain a full engagement of the exercise. However, if such a pattern persists, the language model would continue asking the person to try to stretch the knee, instead of suggesting an option that require less of this person, during our multi-round conversation with it. Whereas, a physio at this moment would suggest, "*let's try this exercise again while seating on a chair*". We believe such a piece of knowledge from the domain expert will strengthen the language model to better act vividly and professionally. By providing this information to GPT-4, its response is improved as follows:

- Knowledge-enhanced response: *{I understand that you might be facing some challenges with this exercise. To ensure your comfort and safety, consider doing the training while seated. This way, you can get a better grip on the movement without feeling too strained or in pain. We can take it step by step, aiming for gradual progress while staying comfortable.}*

Here, in a collaboration with our clinical partners, we collect a rehabilitation knowledge base $\mathcal{D} = d$ detailing the low-demand and high-demand feedbacks they would consider for different movement patterns per each of the actions that included in this work. The low-demand ones refer to the suggestion that require less of the participant's capacity, it usually proposes an alternative way to perform the action properly, e.g., carry out side-bend against a wall if trunk rotation or hip lateral shift are found. Whereas, the high-demand suggestion point to the way of correcting the movement patterns in a straightforward manner. We organize the data along an index corresponding to different action types. Thus, the knowledge for action type y is retrieved as $d \in \mathcal{D} : I(d) = y$, where $I(\cdot)$ denotes the index function.

4 EXPERIMENT

This section presents the experimental details, including the collection of a large-scale action-language dataset and the implementation details of our framework.



Fig. 5. The furnished, home-like environment used for our data collection session, where we made our best to reduce the gap between a lab and the living space of the user.



Fig. 6. Our self-made motion capture suit. By stitching Velcro pads onto the anatomical points of the suit that suggested by the sensor manufacturer for full-body motion capturing, the wireless IMU sensors with Velcro pads are directly attached onto it. This design largely reduces the discomfort of the user, and helps them perform more naturally during the experiment.

4.1 Data Collection in a Home-Like Space

Targeting the future deployment of our proposed Ubiphysio framework at home, we converted a lab room into a modern-style living space, as shown in Fig 5. This space is equipped with essential furniture and equipment that facilitate daily activities and home exercises. These include a television (TV), chair, sofa, soft carpet, bed, cushions, and green plants. Although not visible in the figure, other items such as a broom, yoga pad, and a suitcase loaded with 2.5Kg are also used during the experimental sessions, particularly for daily activities like sweeping the floor and carrying heavy objects. Instead of using stand-alone cameras that may cause discomfort to the participant, we choose to use mobile phones to record the experimental session, which are small and user-friendly. Three mobile phones are put at the top of the TV (facing the participant at 0°), the corner of the space (at 65°), and the left border of the space (at 90°, where Fig 5 is captured). The visual data is collected to aid the annotation from experts and open opportunities for visual-based pose reconstruction that may speed up the deployment of our framework. This experiment is approved by the Institutional Review Board (IRB) of the University.

4.1.1 IMU-Integrated Suit. To collect the action data, we use the wireless IMU sensors from Noitom [2] together with our custom-made suit (as shown in Fig 6) that allows direct attachment of the sensor without the need for bandages. The raw IMU data collected by the sensor was processed by manufacturer-provided software to compute pose and 3D coordinates, which constituted the *raw* data utilized in this study. In the last part of the paper, we also demonstrate the feasibility of using a visual-based technique for acquiring similar data, which could facilitate future low-cost deployment of our proposed framework. The sensor operates at 60Hz and is equipped with an internal battery and a 2.4G Hz Wi-Fi communication module. Based on our experience, it typically lasts for a continuous usage of approximately four hours. The suit is available in different sizes (e.g., S, M, L, XL, XXL) and is thin, breathable, and stretchable. It received positive feedback from our participants, with comments such as "*It wears just like my yoga suit*". To maintain hygiene, we prepared three suits of each size, ensuring that each participant wore a clean suit every day. It is noteworthy that for comprehensive full-body motion capture, we additionally utilized five soft, breathable bandages to attach the sensors to the participants' feet, hands, and head. A total of 17 sensors were used, with 12 attached directly to the suit.

4.1.2 Daily Activities and Exercises. The collected data could be categorized into two groups: one encompassing activities of daily life, and the other comprises exercises that are commonly incorporated in fitness and rehabilitation programs. A comprehensive overview of all actions is shown in Fig 7. It is important to note that our participant demographic included both average healthy individuals and individuals suffering from chronic lower back pain, spanning a diverse age range from 18 to 60 years old. These actions were specifically chosen due to their ability to engage the muscle groups of the back [16, 22, 34]. Healthy participants were included to simulate scenarios of daily fitness support, while participants with chronic lower back pain were recruited to emulate

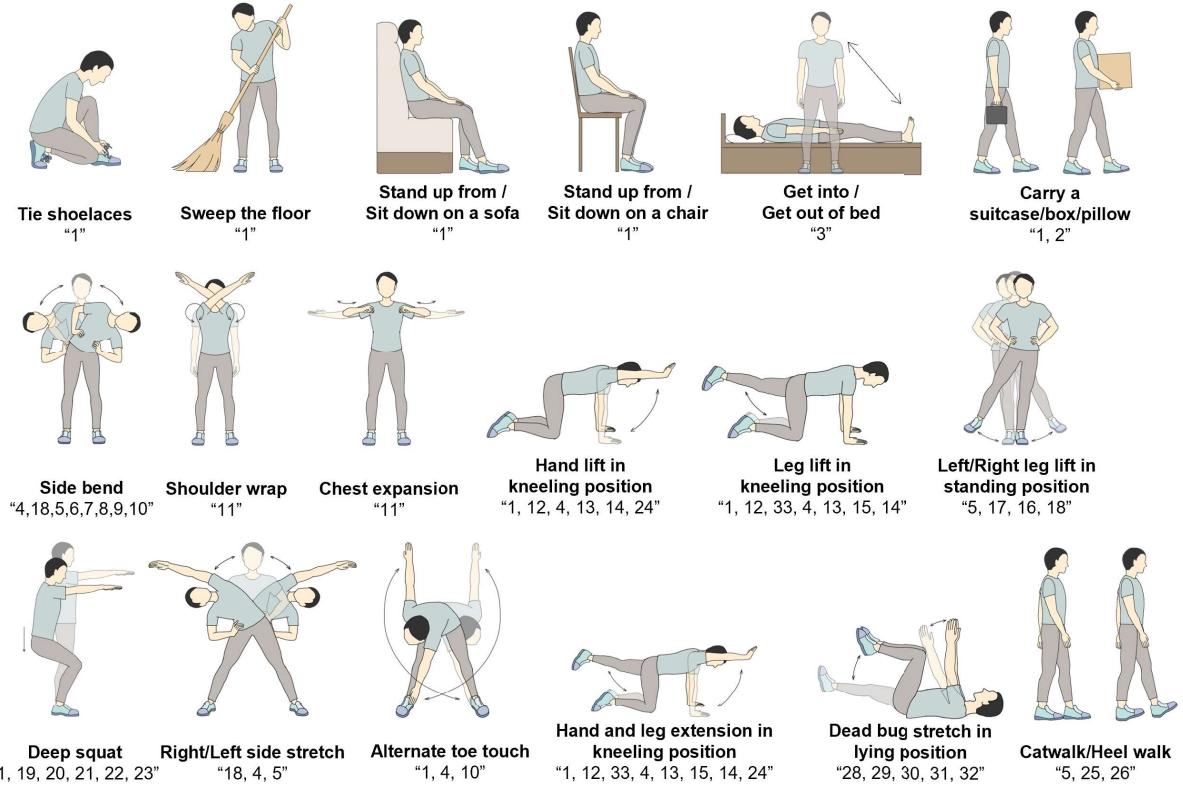


Fig. 7. Our experiment considers actions that cover daily functioning, fitness, and rehabilitation activities. Numbers in quotations represent the possible movement patterns, each of which may exist alone or in a combination with others.

scenarios of home rehabilitation assistance. The only difference of our system in dealing with these two groups of participants lies in the generation of feedback. The language model functions as a coach for healthy individuals and as a physiotherapist for chronic pain sufferers.

4.1.3 A Continuous Session with Small Interventions. Each participant took a single data collection session, consisting of three parts. First, as we assisted the participant in wearing the suit, we briefed them on the key aspects of the experiment. This included the designated area for floor sweeping, the suitcase and pillow to be carried, and the path along which walking actions should be performed. We also allowed participants to familiarize themselves with all the actions, offering practice if necessary. Fig 8 illustrates the spatial distribution of these actions within the experimental space. Subsequently, participants were instructed to perform each action independently. A visual-acoustic reminder was displayed on a television screen, indicating the action to be performed and counting the remaining repetitions. Although the counting was manually operated by the experimenter, participants were informed that *a real-time automatic action recognition system is in use*. This Wizard-of-Oz design aimed to minimize the experimenter's intervention in participants' behaviors. Except for obvious errors, participants were assured that the experimenter would not interfere with their actions, and there was no need to seek confirmation of correctness. Only one experimenter remain in the room during the formal data collection stage. In the second and third parts of the experiment, participant carry out daily functioning and

exercise actions respectively. Each action was followed by a rest period, and participants could request longer intervals if needed. The average duration of a complete data collection session was approximately 25 minutes. For this study, we left out the transitions between different actions, primarily comprised of standing still, casual walking, and self-massage. The targeted actions themselves are very dynamic, with positions spanning the entire experimental space. By extracting each single action instance (i.e., a single execution) and segmenting the longer actions (i.e., sweeping the floor and chest fly) into 15-second non-overlapping windows, we obtain a total of 8780 action instances from 104 participants. Each segment last from 2 to 20 seconds. Although we initially recruited 119 participants, 15 were excluded due to identifiable flaws in their action data upon visual inspection.

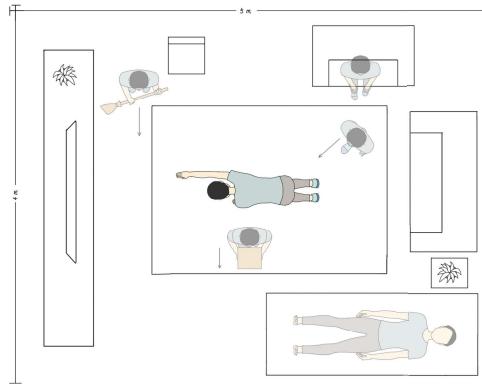


Fig. 8. The spatial distribution of each action performed in our home-like space.

4.1.4 Data Annotation. The annotation of action types was conducted concurrently with each session, with the experimenter marking the onset and offset timestamps as each action was performed by the participant. A post-hoc inspection is carried out to correct potential errors. Possible movement patterns are listed below each action in Fig 7 with the reference table shown in Table 1. These patterns were predefined by a committee comprising our clinical partners, leveraging their extensive experience and professional knowledge. These movement patterns typically serve as evidence for physios to provide feedback and support during clinical rehabilitation sessions, as well as in telemedicine scenarios [6]. They offer more specific insight into the improper movement behaviors of

Table 1. The reference table for the movement patterns included in this work.

Index	Patterns	Index	Patterns	Index	Patterns
1	Lumbar Flexion/Extension	12	Lumbar Hyperextension/Swayback	23	Uneven Bilateral Loading
2	Lumbar Hyperextension after Lifting Heavy Objects	13	Trunk Deviation from Midline/Trunk Lateral Shift	24	Thoracic Hyperextension Compensating for Shoulder Joint Movement
3	Getting Up Directly from / Lying Down Directly into a Supine Position	14	Hip Hyperflexion	25	Insufficient Ankle Dorsiflexion
4	Trunk Rotation	15	Upper Chest Depression	26	Trunk Anterior Lean
5	Hip Lateral Shift	16	Hip Tilt	27	Incorrect Walking Pattern
6	Spinal Extension	17	Lumbar Lateral Flexion	28	Lumbar Lift off the Bed
7	Cervical Lateral Flexion Compensation	18	Trunk Flexion	29	Same-side Hand and Foot Movement
8	On tiptoes/Pelvic Tilt	19	Excessive Anterior Knee Displacement	30	Head Not Touching the Ground
9	No Trunk Activity	20	Upright Trunk Squat	31	Thigh Not Perpendicular to the Floor
10	Knee Flexion Compensation	21	Excessive Hip and Knee Flexion Angle/Too Deep Squat	32	Calf Not Parallel to the Bed Surface/Calf Dangling
11	Lumbar Hyperextension/Pelvic Anterior Tilt	22	Shallow Squat	33	Hip hyperextension leading to lumbar hyperextension

different body parts compared to the metrics used in some previous studies [61–63]. Consequently, this allows for the generation of comprehensive and direct feedback. The extensive set of patterns suggests that traditional machine learning methods might require downsampling these patterns into a few discrete classes to achieve acceptable performance. As such, modeling the action and its patterns using natural language proves to be a more flexible approach.

By accessing synchronized video data collected from three different viewing angles during the experiment, a committee of clinical physios annotated each individual action execution. This is achieved by selecting one or more patterns from the potential candidates pre-defined for each action, as reported in Table 1. Consequently, multiple movement patterns could be marked at the same time for a single action instance. The counting of patterns per action is shown in Fig 9. Based on the annotations (represented by indices in Fig 7), we employ GPT-4 to generate a natural language description for each action instance. To facilitate this, we provide examples that lay out the basic structure of the description, which should cover both the action type and movement patterns. Three different descriptions are generated for each action instance. The average length of the description is 15 words, with a maximum of 25 words. Examples of the generated description are as follows:

- Action type: <shoulder wrap>, movement patterns: <11-lumbar hyperextension>, description: "*When executing a shoulder wrap, this individual exhibits poor core stability and their waist is overextended.*".
- Action type: <hand lift in kneeling position>, movement patterns: <4-trunk rotation, 13-trunk lateral shift>, output: "*In the kneeling hand forward exercise, this individual moves their trunk laterally.*".
- Action type: <leg lift in kneeling position>, movement patterns: <12-Lumbar hyperextension, 14-Hip hyperextension>, output: "*This individual overextends their hip, causing their back to arch excessively when performing the kneeling leg backward exercise.*".

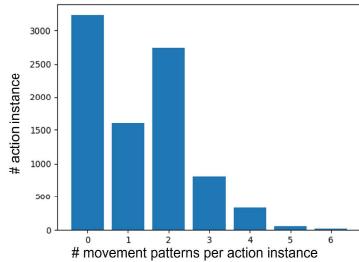


Fig. 9. The number of different movement patterns per each action instance.

4.2 Implementations

During the feature extraction process, the raw 3D pose data is normalized through skeleton normalization and rotation to face the Z+ direction. After feature extraction, the dimensionality of the action data is 315. The first 287 features are derived from methodologies used in previous studies [18, 31, 71, 73] and the last 28 features are the ones proposed as above in this work to help capture the diverse movement patterns. For all experiments, the data collected from the 104 participants was split into training, validation, and test sets, comprising 85%, 5%, and 10% of the data, respectively. It was ensured that there was no subject overlap among these sets. We report the metrics scores obtained on the test set for each method. For experiments that have multiple runs, we additionally report the 95% confidence interval to illustrate the variability in performance. For language modeling, we evaluate various open-sourced foundational language models, including T5 (in its small, base, large, and 3B variants) [52], Llama 2 7B[58], and ChatGLM2 6B[13, 70]. We use an extensive set of benchmark metrics to evaluate the output of each language model against the annotated description, including Bleu [47], Rouge [40], Cider [60],

and BertScore [72]. The duration of a complete training on a single RTX 4090 graphics card of VQ-VAE is 18 hrs, and are as follows for the action-language modeling: T5 small (6 hrs), T5 base (4.5 hrs), T5 large (3.7 hrs), T5 3B (19.5 hrs), Llama 2 7B (5 hrs), and ChatGLM2 6B (16 hrs). It should be noted that, for smaller models, the training could have met the GPU inefficiency problem, leading to suboptimal utilization of the high-performance graphics card and longer training time. Please kindly refer to Appendix section for more details of the hyperparameters used for each model.

5 RESULTS

The objective evaluation of our framework in terms of its language output is reported in this section, where we look at performances given different foundational language models and tuning strategies that proposed in this work. Additionally, we conducted a user study with participants and physios to collect the subjective preference and opinions about the output of our framework. Finally, we look into the practical deployment of our system in real-world scenarios without specialized devices, highlighting the potential and limitations of our framework.

5.1 Quantitative Evaluation of Action Description with Language Models

We first compare the performance of open-sourced foundation models on our action description task with benchmark metrics for natural language generation tasks. The action tokens provided here belong to two groups, the first is acquired by training the VQ-VAE with the 287 features that widely adopted in recent action-language modeling studies [18, 31, 71, 73], while the other is acquired by adding our proposed biomechanical features that provide the input features to VQ-VAE in a dimension of 315. A window length of 64 is applied during the training of VQ-VAE. The action label is removed from the tuning instructions, and the description label comprises full descriptions of the action type and movement patterns. Results of a standalone cross validation experiment are reported in Table 2.

The results confirm the effectiveness of our proposed biomechanical features (Dim.=315) in assisting the VQ-VAE model, and consequently the language models, in better capturing the diverse movement patterns hidden behind each action. The use of action tokens driven by these features led to improvements across all metrics for different models. It is worth noting that these features are computed for each timestep, providing a significant opportunity for real-time operation of the framework.

As the results indicate, the T5-large model demonstrates the most promising performance in this action description task, despite its smaller size compared to large language models (LLMs) such as Llama 2 7B and ChatGLM2 6B. The Llama 2 7B model's fine-tuning on our data for this task was unsuccessful, despite our

Table 2. The comparison of different foundational language models on our action description task with different feature sets as input. Params.= the number of trainable parameters, Dim. = the dimension of the input action data, Iter. = the total number of training iterations. The best results under different feature inputs are highlighted in bold.

Model	Params.	Tuning	Dim.	Iter.	Bleu@1↑	Bleu@4↑	Rogue 1↑	Rogue 2↑	Rogue L↑	Cider↑	BertScore↑
T5 small	80M	Full-params.	287	50K	63.40	41.94	58.71	44.08	55.45	1.44	56.80
T5 base	248M	Full-params.	287	50K	61.56	39.04	57.12	41.60	53.21	1.31	53.47
T5 large	783M	Full-params.	287	50K	66.38	44.75	62.37	48.02	60.01	1.49	58.52
T5 3B	3B	LoRA	287	100K	65.37	43.29	62.55	48.23	59.86	1.42	56.04
ChatGLM2 6B	6B	P-tuning v2	287	10k	38.67	13.56	35.35	17.76	32.93	0.13	21.86
T5 small	80M	Full-params.	315	50K	64.09	42.89	60.64	46.24	57.74	1.44	56.70
T5 base	248M	Full-params.	315	50K	65.09	42.86	60.93	46.61	57.68	1.43	56.29
T5 large	783M	Full-params.	315	50K	68.21	47.64	64.74	50.95	61.39	1.62	59.84
T5 3B	3B	LoRA	315	100K	66.70	45.67	62.64	49.35	60.47	1.51	57.69
ChatGLM2 6B	6B	P-tuning v2	315	10K	57.54	33.78	53.46	37.63	49.90	1.07	48.94

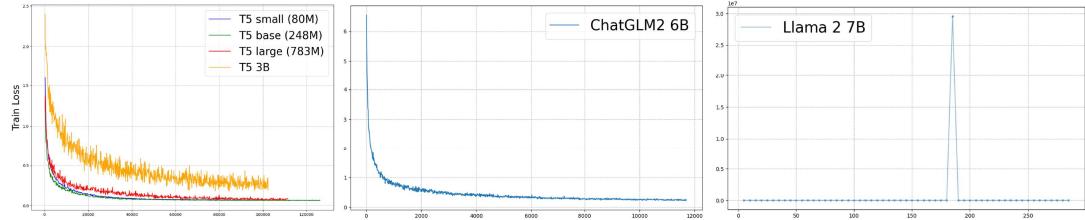


Fig. 10. The training curves of models adopted in the action description experiment. While the training losses of most models reduce normally in this task, the Llama 2 7B model is not training well after our several tries with different hyperparameters.

attempts to modify various hyperparameters of the LoRa framework, learning rate, and input/output lengths. A representative training curve of this model is depicted in Fig 10, where the training loss does not decrease as expected. The performance disparities among the models may stem from their differing architectures and training objectives. The task at hand resembles language translation, converting discrete action tokens (a form of 'second language') into formal language. This task might be more effectively tackled by sequence-to-sequence (Seq2Seq) models, such as the T5 family. For LLMs like Llama, the 'translation task' becomes more challenging due to their decoder-only architecture and autoregressive objective, as predicting the next token based on previous ones is not particularly beneficial in this context. However, ChatGLM2 6B does produce some meaningful results, which could be owed to its blank-filling objective during pre-training, facilitating fine-tuning in a Seq2Seq manner.

5.1.1 Impact of Window Length. For action tokenization with the VQ-VAE model, a sliding window is used to accommodate action feature inputs of different lengths. To evaluate the impact of this variable window length, we conducted an experiment using the T5 large model. This model was chosen due to its superior performance when compared to other language models in prior evaluations. The experiment was conducted across five different cross-validation folds, and the results are presented as an average performance, accompanied by a 95% confidence interval. The 'description label' in this context refers to comprehensive descriptions of both the type of action and the associated movement patterns. The detailed results are reported in Table 3. A key finding from this experiment is that the performance of language modeling does not appear to be sensitive to the length of the sliding window employed during action tokenization. This could be attributed to the fact that the tokens are extracted from complete action sequences, thereby potentially neutralizing the impact of varying window lengths.

Table 3. The ablation experiment on different window lengths applied during action tokenization, sampling rate, and action-conditioned instruction tuning. Action. = the action type information as a condition added to the instruction. The relatively higher performances in each comparison are marked in bold.

Model	Window length	Sampling rate	Instruction	Bleu@1↑	Bleu@4↑	Rogue 1↑	Rogue 2↑	Rogue L↑	Cider↑	BertScore↑
T5 large	64	60Hz	Tokens	65.78±1.95	46.20±2.43	64.50±1.81	50.12±2.44	62.11±1.97	1.58±0.09	59.22±2.38
	128	60Hz	Tokens	65.63±0.45	45.95±0.56	63.79±0.53	49.69±0.62	61.32±0.60	1.58±0.03	58.95±0.74
	192	60Hz	Tokens	66.20±0.68	46.99±1.27	64.96±0.64	50.95±1.21	62.45±0.78	1.62±0.07	60.44±1.23
T5 large	96	30Hz	Tokens	65.79±1.12	46.20±1.69	64.20±1.30	50.15±1.77	61.99±1.52	1.59±0.06	59.42±1.28
T5 large	64	60Hz	Action.+Tokens	68.97±0.76	50.06±1.10	67.82±0.99	54.23±1.34	65.75±1.06	1.74±0.04	62.78±0.78
	128	60Hz	Action.+Tokens	67.57±0.89	48.37±1.28	66.58±1.04	52.62±1.30	64.23±1.13	1.68±0.05	61.40±0.71
	192	60Hz	Action.+Tokens	68.59±1.16	49.92±1.46	67.74±1.20	53.93±1.58	65.42±1.25	1.73±0.06	62.67±1.60
T5 large (patterns only)	192	60Hz	Action.+Tokens	60.25±0.89	47.18±1.17	60.04±1.06	49.51±1.22	59.44±1.00	1.71±0.11	55.50±1.02
	192	60Hz	Tokens	59.33±1.40	45.46±2.02	58.90±1.52	47.79±1.92	58.18±1.53	1.63±0.1	55.43±1.39

In a secondary experiment, we downsampled the raw action data from 60Hz to 30Hz and tested the model with a window length of 96 (equivalent to the 192 used for data at 60Hz). The results showed a slight decrease in performance, suggesting that a higher data frequency might enhance the sufficiency of action descriptions. However, this reduction in performance was not significant. Therefore, the decision to lower the frequency could be considered as a viable strategy for reducing hardware costs and computational loads without substantially compromising model performance.

5.1.2 Impact of Action-Conditioned Instruction Tuning. As reported in Table 3, integrating the action type information, denoted as y , into the instruction significantly enhances the language model’s description performance. This enhancement is two-fold. Firstly, the model becomes more proficient at describing the correct type of action. Secondly, it exhibits an improved capacity to capture diverse movement patterns. This strategy is particularly beneficial for systems that engage directly with users in a home environment. It serves as a reliability measure for the language model’s output, especially considering the high accuracy achieved in classifying different actions using an efficient classifier, as documented in Section 3.1.3. To further validate whether this strategy indeed improves the model’s ability to capture movement patterns, we conducted an additional experiment. In this experiment, the model was trained exclusively with labels that contained only pattern information. The results, presented in the last two rows of Table 3, confirm our hypothesis. The enhancement brought about by the integration of action type information is not confined to the accurate description of the action type alone. It also extends to the precise depiction of movement patterns, demonstrating that the improvement is comprehensive and not overshadowed by the correct identification of action types.

5.2 User Study

We conduct a user study with the participants and physios to evaluate the output of the language model in terms of the quality of its generated action description and feedback. For action description, the T5 large model with window length of 64 and action conditioned instruction tuning is used. For feedback generation, we utilize GPT-4 with the following prompt: “*You are now a physiotherapist to support the daily functioning and exercise of a person. In the following, you will receive the <knowledge> specifying the pre-defined low-demand and high-demand feedbacks for the action, <user profile> indicating the age, self-reported pain score (0-10), and Timed-Up-and-Go (TUG) score, <action description> detailing the action type and movement patterns of the person. You need to return the <instant feedback> in a vivid tongue*”.

For healthy participants, the role is transferred to a fitness coach.

We create a questionnaire that includes both the model-generated outputs and the actual action descriptions (ground truth). Similarly, feedbacks generated without our retrieved domain knowledge are created. These are included and randomly organized in the questionnaire presented to users, in a 50% by 50% ratio, in order to eliminate any potential bias. The questionnaire, as shown in Fig 11, is composed of multiple sections, each presents a video of an action instance, paired with its action description and feedback.

We reach out to 15 participants that are randomly selected from the testing folds of our previous evaluation, and invite them to complete the questionnaire, where each of them is rating given their own

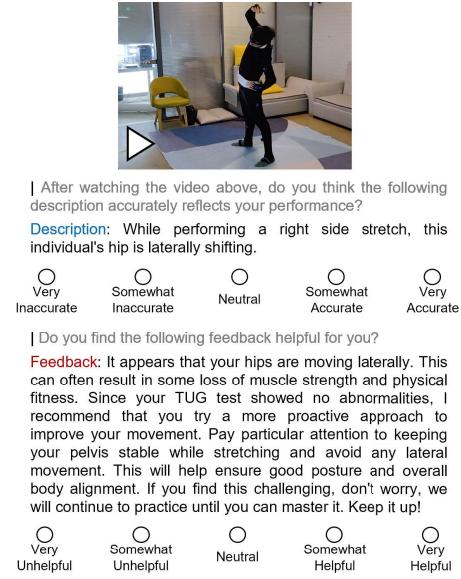


Fig. 11. The adopted questionnaire in our user study, rated by participants shown in the video and physios.

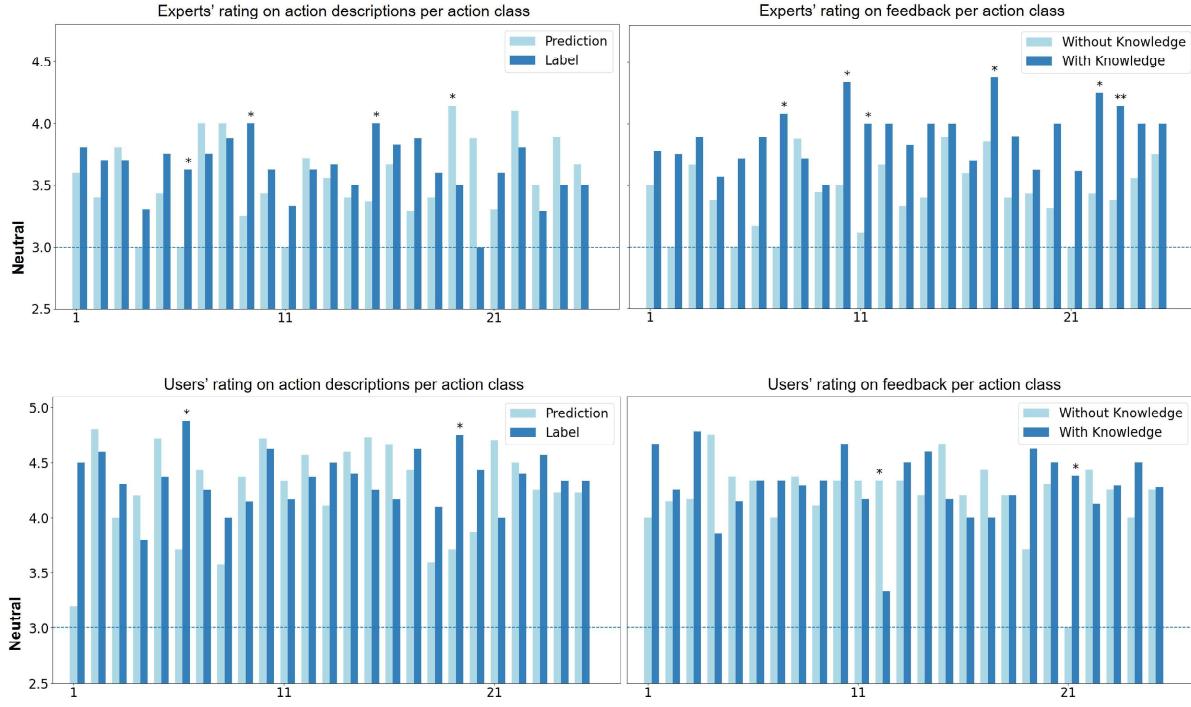


Fig. 12. Results of the user study on rating the outputs of our framework per each action type by clinical experts and normal participants. Significant results are marked with * for $\alpha = 0.05$ and ** for $\alpha = 0.01$.

videos. In addition to participants, three clinical physiotherapists are also recruited to provide a professional evaluation of the model outputs. Participants and physiotherapists are asked to rate action description and feedback of each section using a standard 5-point rating scale. This approach allowed us to assess the quality and effectiveness of the language model in a structured, rigorous manner. The average results of the study per each group of the people are reported in Fig 12.

For action descriptions, both participants and physiotherapists evaluated the model outputs of most actions as comparable to or even better than the ground truth. However, a subset of the actions was rated with a difference that exceeded the significance level, indicating room for improvement. When describing the action of standing up from a chair (Action #6), the model consistently generated outputs such as '*this individual is rising from a chair, curving their back forward*'. It appears that distinguishing between normal patterns and improper back curvature is challenging for the model due to their high similarities. In the case of the left-leg lift in a standing position (Action #15), the model frequently predicted the users' performances as unstable - a description criticized by physiotherapists but accepted by participants, who often found this action difficult to execute while maintaining balance. More interestingly, for the action of alternating toe touch (Action #19), while physiotherapists agreed with the model's prediction of the performances as '*rounding their back excessively*', some participants believe they performed the action correctly. In a clinic setting, physiotherapists would actively engage with the patient to determine if such a discrepancy is due to a lack of capacity (i.e., the performance is the best the patient can do) or cognitive bias (i.e., the patient is unaware of the harm in performing the action in that particular manner). Based on this interaction, they would come to a consensus and decide whether to adjust the action's difficulty or

work on improving the patient's self-awareness. **We believe such an interaction could also happen between our framework and a user.** We would like to leave this to its future development.

For generated feedback, there was a noticeable divergence in the assessment provided by the physiotherapists and the participants. The physiotherapists rated the retrieval-enhanced feedbacks, those generated with domain knowledge, as significantly superior to the plain ones generated without such knowledge. This highlights the importance of domain knowledge in improving the quality of generated feedback. In contrast, the participants exhibited a more neutral stance, indicating no strong preference between the two types of feedback. This discrepancy in ratings may due to the lack of professional knowledge in our participants. Consequently, they may perceive the feedback provided by GPT-4, based on its common sense knowledge, as sufficiently satisfactory. For the action of hand lift in a kneeling position (Action #12), we observed that the feedback generated by the model without domain knowledge tended to be more fluent and natural. When domain knowledge was incorporated, the feedback became somewhat rigid and unnatural. This could explain why participants found the former feedback significantly more helpful.

In general, the participants and physios are all excited about the achievement made by our proposed framework. During the user study, we also collect some opinions from both participants and experts. Most of them underscored the need for further refinement of the model's language use and tone to enhance the acceptability and effectiveness of the conversation. For example, the model tends to use the language such as '*one is carefully conducting an action*', which is criticized by our physiotherapists due to its ambiguity in action description. Such language might be interpreted in multiple ways. Some may perceive it as indicating that the individual is performing the action slowly yet with balance, while others might interpret it as a sign of the difficulties the person is encountering during the action.

5.3 Preliminary Evaluation towards Real-Life Deployment

Here, we provide an initial examination of the feasibility of implementing our framework for practical use in home settings. The considerations primarily revolve around two aspects: the system's inference time, spanning from action data preprocessing to feedback generation, and the extra devices needed to run this framework.

5.3.1 Inference Time. From the collection of 3D pose data to the generation of action description and finally the feedback, the complete inference pipeline is shown in Fig 13. Given our environment of RTX 4090 and a CPU with 2.30 GHz frequency, the time spends on data preprocessing and action tokenization takes 1.06ms and 1.28ms per each frame in an action data sequence, respectively. While the actual operation of the algorithm will use the multi-thread processing capacity of CPU and GPU, it takes nearly 2.34ms to process a single frame in a naive manner to transform action data into tokens, equivalent to processing 427 frames per second. Given the action data in an arbitrary duration, the length of action tokens range from 0 to 512 (with codebook size set to 512), while the processing time for action description stays almost the same. For T5 large, our machine needs 0.5s to provide a complete description per action instance. Meanwhile, the generation of feedback using the GPT-4

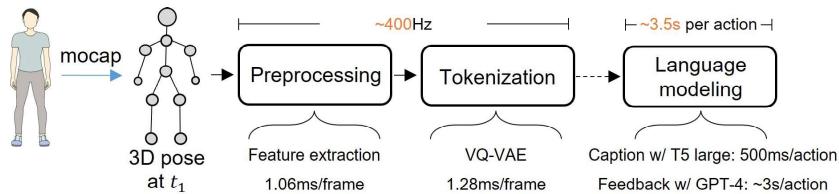


Fig. 13. The inference pipeline of the proposed framework, inference time is computed based on a machine with RTX 4090.

API costs approximately 3 seconds before returning responses. In short, given an action lasting 10 seconds at a streaming frequency of 60Hz, our system requires 4.91s to generate feedback for the user in a real-time setting, using a machine with RTX 4090. This duration is comparable to the time it takes a physio to make a suggestion in a clinical setting, according to our clinical partners. However, if instant feedback for each action instance is not imperative, the longer latency associated with more commonly used computing devices may also be acceptable.

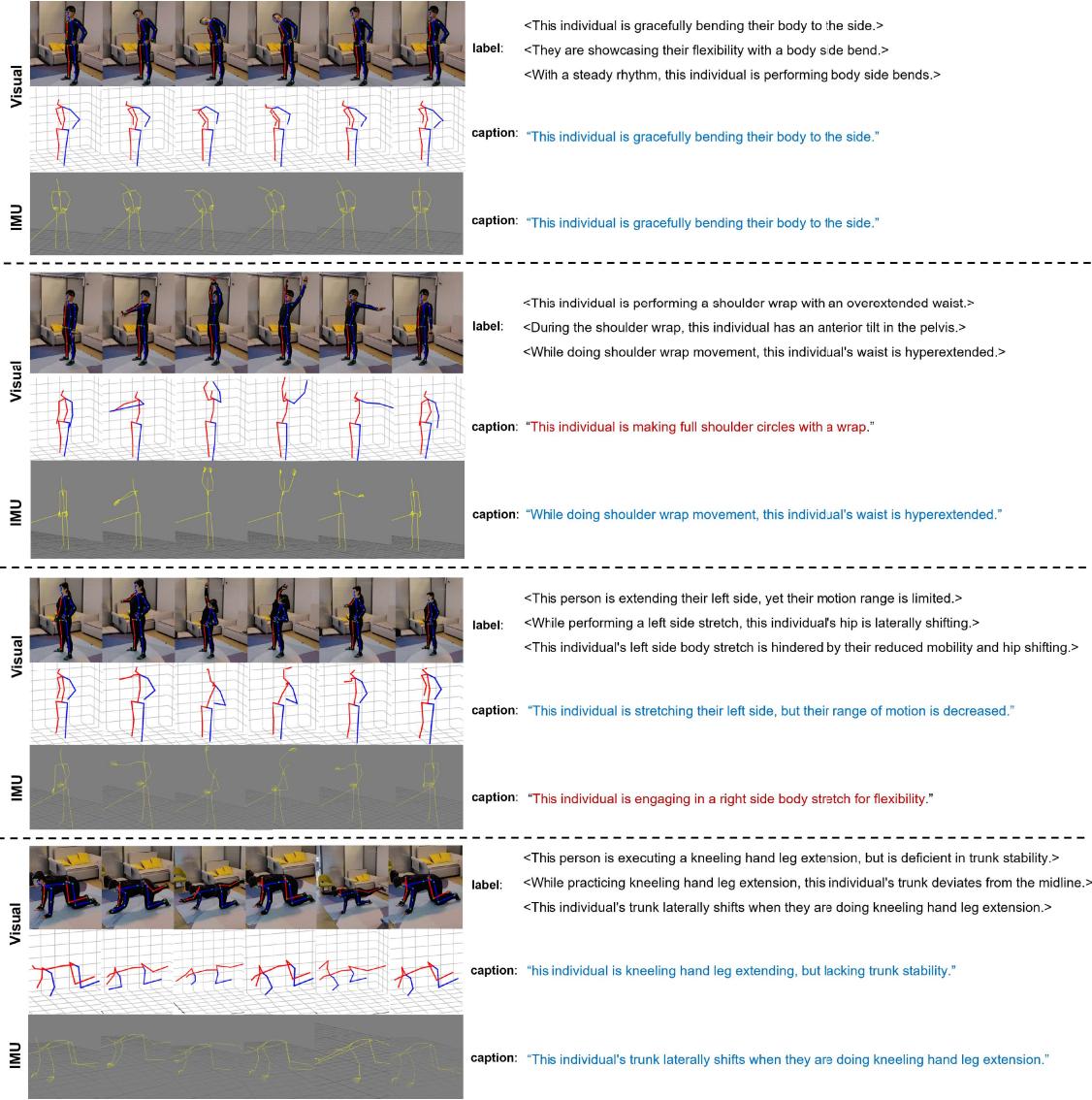


Fig. 14. Visualizations of the results acquired on action data collected from a visual system and IMUs. The model's outputs are marked in blue for correct predictions and red for incorrect ones.

5.3.2 Visual Motion Capture as an Alternative to IMUs. In our experiment, in order to improve the comfort of users when wearing the IMUs, we designed a thin, stretchable, and breathable suit with wireless sensors directly attached. Although our participants generally found the suit comfortable, requiring users to have such equipment for home use is costly. This cost is not only financial but also practical, due to issues with maintaining accurate data capture over extended periods, typically beyond 10 minutes. This latter problem led to the exclusion of data from 15 participants due to its poor quality. Therefore, we suggest the exploration of visual motion capture as a potential alternative to IMUs, which could mitigate these issues while maintaining data fidelity.

Here, we run an extra experiment with the most recent technology of acquiring 3D human poses from a video. We use the plug-and-play PoseFormer V2 [74], applying it to videos collected at 65° from two randomly selected participants from the test folds. Given the common 17-joint skeleton data returned by this kind of method, we simulate the joint data of head end, both hands, and feet using their nearest joint and the distance d in-between, e.g., using the known coordinate of left forearm A , the coordinate of left hand B that is missing in the visual poses is simulated as $(\hat{x}, \hat{y}, \hat{z}) = (x, y, z) + \frac{\vec{AB}}{|AB|} \times d$. The joint of neck 1 is acquired by taking an average of the joints of neck and head. Such that, we transform the 17-joint visual pose data into the skeleton used in this work, and put into the inference pipeline shown above. It is important to note that the model used in this experiment was trained solely on data collected from the IMUs. The preprocessing conducted on the visual poses remains the same as those applied to IMU data, including Z-normalization (with their respective distributions), skeleton normalization, and alignment towards the Z+ direction. Results with visualization examples are shown in Fig 14.

Although the quality of vision-captured poses could get largely affected by occlusions given a single view angle, the results are promising and comparable to those using data from IMUs. For movement patterns that describe the behavior of localized body parts, such as waist hyperextension during shoulder wrap, more accurate tracking is required. Unfortunately, the data from the visual system often fails to provide this level of detail. As for other movements, vision-captured poses offer sufficient information for our framework, leading to satisfactory results. We posit that by further refining the visual system—for instance, by fine-tuning it with our IMU data or integrating it with signals from a limited set of IMUs—we can enhance performance. This could pave the way for the system’s real-world deployment, eliminating the need for users to purchase extra devices.

6 OPPORTUNITIES AND FUTURE DEVELOPMENT

As an initial attempt, the proposed framework is shown to understand the fine-grained action performances of a user, provide accurate and professional language outputs, and deliver helpful feedback. During our experiment, we further found the following opportunities that may guide the future development of this work.

6.1 From One-Shot Feedback to the Interactive Functioning

The current framework provides action description for each action instance, but it has not yet been tested in the context of continuous user interaction. For one thing, the system needs to segment the data stream into meaningful action instances, or one should verify by putting continuous data into the pipeline the language output is still useable. For another, in real-life scenarios, physios would typically respond to the performance of the user consistently, providing not only feedback, but also encouragements and other forms of verbal support. Particularly, in cases where there is a discrepancy in the description of movement patterns, further interaction with the user would be initiated to discern their potential physical limitations and cognitive biases. We are optimistic that the day when an artificial intelligence agent can deliver such a comprehensive service to users is not far off. This belief is reinforced by the rapid advancements observed in Large Language Models (LLMs) in recent months, as well as the paradigm shift among academics and industry professionals, i.e., the latest technology being increasingly applied to previously unresolved tasks, gaining widespread acceptance and popularity.

6.2 Enriching Feedback Generation via Real-time Physiological Signal Monitoring

In our domiciliary scenarios involving daily functioning, fitness, and rehabilitation activities, the user's physiological signals, such as heart rate, breath rate, and heart rate variability (HRV), are essential measurements complementary to the exterior human motion features. These physiological signals could serve as direct indicators of the user's internal state, such as physical load and fatigue level. Therefore, combining physiological signals with motion capture data could help the smart agent to establish more comprehensive user profiles, which, in turn, can enhance the generation of appropriate intervention and rehabilitation suggestions.

As an initial exploration, We re-examined the auxiliary audio data collected from our main experiment sessions (with a wireless collar clip microphone) to figure out whether behaviors indicating the physiological state (e.g., heavy breaths) can be detected. We manually labeled 3737 audio clips from 40 users with one of the three labels - "heavy breath, sigh, or moan", "talking or friction noise", and "background or microphone noise" - by reviewing the corresponding video clips. We preprocessed these audio samples into Mel spectrograms and trained a ResNet [25] image classifier with 80% audio samples. Evaluation on the rest 20% audio samples achieved an average recognition accuracy of 66.7% and a Marco F-1 score of 60.7%. These results suggest that the signal-to-noise ratio (SNR) of breaths is not sufficient for robust recognition in our current hardware settings (e.g., breath typically has a lower amplitude and is easily overwhelmed by friction noise and background noise). To improve the recognition accuracy and robustness, one possible solution is to optimize the hardware form (e.g., using the inner microphone from an ANC earbud [38]).

In future research, we aim to incorporate alternative sensor channels, such as PPG, RPPG, Wi-Fi, and mm wave radar, into our UbiPhysio framework to enhance the sensing capability for both motional signals and physiological signals. In this context, how these signals jointly influence the intervention and rehabilitation strategy would be an interesting research question for further investigation.

7 CONCLUSION

In this paper, we proposed UbiPhysio, a pioneering framework designed to offer fine-grained action descriptions, in terms of the action type and movement patterns, and feedback for daily functioning, fitness, and rehabilitation activities. We have assessed the framework with extensive experiments involving 104 diverse participants who engaged in 25 types of everyday activities and exercises in a home-like setting. The quality of language output, evaluated using standard benchmarks under various tuning strategies, showed promising results. Additionally, we carried out a user study involving clinical experts and non-professional users, whose results further affirmed the framework's suitability and usability. We also explored the feasibility of deploying our framework in real-world settings using vision-captured motion data, instead of relying on the data returned by many IMUs. With these promising results, we believe that UbiPhysio can significantly enhance the effectiveness of remote fitness and rehabilitation programs, improve user engagement, and ultimately contribute to better health outcomes.

REFERENCES

- [1] Kinect. 2023. www.xbox.com/en-US/kinect
- [2] Noitom. 2023. www.noitom.com/perception-neuron-series
- [3] Hillel Aviezer, Yaakov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.
- [4] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current directions in psychological science* 20, 5 (2011), 286–290.
- [5] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
- [6] Scott A Biely, Sheri P Silfies, Susan S Smith, and Gregory E Hicks. 2014. Clinical observation of standing trunk movements: What do the aberrant movement patterns tell us? *Journal of orthopaedic & sports physical therapy* 44, 4 (2014), 262–272.

- [7] Lora E Burke, Jing Wang, and Mary Ann Sevick. 2011. Self-monitoring in weight loss: a systematic review of the literature. *Journal of the American Dietetic Association* 111, 1 (2011), 92–102.
- [8] Jiangjie Chen and Yanghua Xiao. 2022. Harnessing Knowledge and Reasoning for Human-Like Natural Language Generation: A Brief Review. *arXiv preprint arXiv:2212.03747* (2022).
- [9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
- [10] Christopher Clarke, Doga Cavdir, Patrick Chiu, Laurent Denoue, and Don Kimber. 2020. Reactive video: adaptive video playback based on user motion for supporting physical activity. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 196–208.
- [11] Diane J Cook, Juan C Augusto, and Vikramaditya R Jakkula. 2009. Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and mobile computing* 5, 4 (2009), 277–298.
- [12] Diane J Cook, Aaron S Crandall, Brian L Thomas, and Narayanan C Krishnan. 2012. CASAS: A smart home in a box. *Computer* 46, 7 (2012), 62–69.
- [13] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhang Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 320–335.
- [14] Carol Ewing Garber, Bryan Blissmer, Michael R Deschenes, Barry A Franklin, Michael J Lamonte, I-Min Lee, David C Nieman, and David P Swain. 2011. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. (2011).
- [15] Martin A Giese and Tomaso Poggio. 2003. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience* 4, 3 (2003), 179–192.
- [16] Sahar Modares Gorji, Hadi Mohammadi Nia Samakosh, Peter Watt, Paulo Henrique Marchetti, and Rafael Oliveira. 2022. Pain neuroscience education and motor control exercises versus core stability exercises on pain, disability, and balance in women with chronic low back pain. *International Journal of Environmental Research and Public Health* 19, 5 (2022), 2694.
- [17] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.
- [18] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *ECCV*.
- [19] Xiaonan Guo, Jian Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2018. Device-free personalized fitness assistant using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.
- [20] Héctor Gutiérrez-Espinoza, Felipe Araya-Quintanilla, Cristian Olguín-Huerta, Iván Valdés-Orrego, and Oscar Sepúlveda-Osses. 2022. Effectiveness of supervised physiotherapy versus home exercise in subjects with rotator cuff disorders treated surgically: A systematic review and meta-analysis. *Physiotherapy Research International* 27, 2 (2022), e1942.
- [21] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.
- [22] Mark H Halliday, Evangelos Pappas, Mark J Hancock, Helen A Clare, Rafael Z Pinto, Gavin Robertson, and Paulo H Ferreira. 2019. A randomized clinical trial comparing the McKenzie method and motor control exercises in people with chronic low back pain and a directional preference: 1-year follow-up. *Physiotherapy* 105, 4 (2019), 442–445.
- [23] Shane Halloran, Lin Tang, Yu Guan, Jian Qing Shi, and Janet Eyre. 2019. Remote monitoring of stroke patients' rehabilitation using wearable accelerometers. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 72–77.
- [24] Harish Haresamudram, Irfan Essa, and Thomas Ploetz. 2023. Towards Learning Discrete Representations via Self-Supervision for Wearables-Based Human Activity Recognition. *arXiv preprint arXiv:2306.01108* (2023).
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [26] Shruthi K Hiremath, Yasutaka Nishimura, Sonia Chernova, and Thomas Plötz. 2022. Bootstrapping Human Activity Recognition Systems for Smart Homes from Scratch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.
- [27] Thuong N Hoang, Martin Reinoso, Frank Vetere, and Egemen Tanin. 2016. Onebody: remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 1–10.
- [28] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned Motion Matching. *ACM Trans. Graph.* 39, 4, Article 53 (aug 2020), 13 pages. <https://doi.org/10.1145/3386569.3392440>
- [29] Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4, Article 42 (jul 2017), 13 pages. <https://doi.org/10.1145/3072959.3073663>
- [30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

- [31] 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* (2023).
- [32] Hye-Young Jo, Laurenz Seidel, Michel Pahud, Mike Sinclair, and Andrea Bianchi. 2023. FlowAR: How Different Augmented Reality Visualizations of Online Fitness Videos Support Flow for At-Home Yoga Exercises. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [33] Balasundaram Kadirvelu, Constantinos Gavriel, Sathiji Nageshwaran, Jackson Ping Kei Chan, Suran Nethisinghe, Stavros Athanasopoulos, Valeria Ricotti, Thomas Voit, Paola Giunti, Richard Festerstein, et al. 2023. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. *Nature Medicine* 29, 1 (2023), 86–94.
- [34] Beomryong Kim and Jongeon Yim. 2020. Core stability and hip exercises improve physical function and activity in patients with non-specific low back pain: a randomized controlled trial. *The Tohoku journal of experimental medicine* 251, 3 (2020), 193–206.
- [35] Jihoon Kim, Youngjae Yu, Seungyoun Shin, Taehyun Byun, and Sungjoon Choi. 2022. Learning Joint Representation of Human Motion and Language. *arXiv:2210.15187* [cs.CV]
- [36] Lawla LF Law, Fiona Barnett, Matthew K Yau, and Marion A Gray. 2014. Effects of functional tasks exercise on older adults with cognitive impairment at risk of Alzheimer's disease: a randomised controlled trial. *Age and ageing* 43, 6 (2014), 813–820.
- [37] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
- [38] Zisu Li, Chen Liang, Yuntao Wang, Yue Qin, Chun Yu, Yukang Yan, Mingming Fan, and Yuanchun Shi. 2023. Enabling Voice-Accompanying Hand-to-Face Gesture Recognition with Cross-Device Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 313, 17 pages. <https://doi.org/10.1145/3544548.3581008>
- [39] Yalin Liao, Aleksandar Vakanski, Min Xian, David Paul, and Russell Baker. 2020. A review of computational approaches for evaluation of rehabilitation exercises. *Computers in biology and medicine* 119 (2020), 103687.
- [40] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [41] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 61–68.
- [42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2023. SMPL: A Skinned Multi-Person Linear Model (1 ed.). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3596711.3596800>
- [43] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [44] Mel E Major, Daniela Dettling-Ihnfeldt, Stephan PJ Ramaekers, Raoul HH Engelbert, and Marike van der Schaaf. 2021. Feasibility of a home-based interdisciplinary rehabilitation program for patients with Post-Intensive Care Syndrome: the REACH study. *Critical Care* 25, 1 (2021), 1–15.
- [45] Sionadh Mairi McLean, Maria Burton, Lesley Bradley, and Chris Littlewood. 2010. Interventions for enhancing adherence with physiotherapy: a systematic review. *Manual therapy* 15, 6 (2010), 514–521.
- [46] Alessandra Moschetti, Laura Fiorini, Dario Esposito, Paolo Dario, and Filippo Cavallo. 2017. Daily activity recognition with inertial ring and bracelet: An unsupervised approach. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3250–3255.
- [47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [48] Lukasz Piwek, David A Ellis, Sally Andrews, and Adam Joinson. 2016. The rise of consumer health wearables: promises and barriers. *PLoS medicine* 13, 2 (2016), e1001953.
- [49] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems* (Nov 2018), 13–26. <https://doi.org/10.1016/j.robot.2018.07.006>
- [50] Thomas Plötz, Paula Moynihan, Cuong Pham, and Patrick Olivier. 2011. Activity recognition and healthier food preparation. *Activity Recognition in Pervasive Intelligent Environments* (2011), 313–329.
- [51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
- [52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. <http://jmlr.org/papers/v21/20-074.html>
- [53] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083* (2023).

- [54] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).
- [55] Valeria Ricotti, Balasundaram Kadirvelu, Victoria Selby, Richard Festenstein, Eugenio Mercuri, Thomas Voit, and A Aldo Faisal. 2023. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. *Nature medicine* 29, 1 (2023), 95–103.
- [56] Ann-Kathrin Schalkamp, Kathryn J Peall, Neil A Harrison, and Cynthia Sandor. 2023. Wearable movement-tracking data identify Parkinson’s disease years before clinical diagnosis. *Nature Medicine* (2023), 1–9.
- [57] Chuan-Jun Su, Chang-Yu Chiang, and Jing-Yan Huang. 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Applied Soft Computing* 22 (2014), 652–666.
- [58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [59] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
- [60] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.
- [61] Chongyang Wang, Yuan Gao, Akhil Mathur, Amanda C De C. Williams, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2021. Leveraging activity recognition to enable protective behavior detection in continuous data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.
- [62] Chongyang Wang, Temitayo A Olugbade, Akhil Mathur, Amanda C De C. Williams, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2019. Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 225–230.
- [63] Chongyang Wang, Temitayo A Olugbade, Akhil Mathur, Amanda C DE C Williams, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2021. Chronic pain protective behavior detection with deep learning. *ACM Transactions on Computing for Healthcare* 2, 3 (2021), 1–24.
- [64] Hanchen David Wang and Meiyi Ma. 2023. PhysiQ: Off-site Quality Assessment of Exercise in Physical Therapy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–25.
- [65] Liang Wang, Weiming Hu, and Tieniu Tan. 2003. Recent developments in human motion analysis. *Pattern recognition* 36, 3 (2003), 585–601.
- [66] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 14959–14971. https://proceedings.neurips.cc/paper_files/paper/2022/file/6030db5195150ac86d942186f4abdad8-Paper-Conference.pdf
- [67] Wenchuan Wei, Carter McElroy, and Sujit Dey. 2019. Towards on-demand virtual physical therapist: Machine learning-based patient action understanding, assessment and task recommendation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 9 (2019), 1824–1835.
- [68] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit+: Personalized fitness monitoring via audio signals on smart speakers. *IEEE Transactions on Mobile Computing* (2021).
- [69] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. 2018. Paired Recurrent Autoencoders for Bidirectional Translation Between Robot Actions and Linguistic Descriptions. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3441–3448. <https://doi.org/10.1109/LRA.2018.2852838>
- [70] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).
- [71] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [72] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [73] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2023. MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. *arXiv preprint arXiv:2306.10900* (2023).
- [74] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8877–8886.

A IMPLEMENTATION DETAILS

For VQ-VAE, the size of the codebook \mathcal{B} is 512×512 , the downsampling rate is set to $l = 4$. Based on empirical evidence, β is set to 1.0 for the commitment loss, and α is set to 0.5 for the regularization loss using biomechanical features. Typical VQ-VAE training techniques such as codebook reset and exponential moving average (EMA) [54] are used, where the exponential moving constant is set as $\lambda = 0.99$. During training, the batch size is set to 128, the default AdamW optimizer [43] is used for optimization, the initial learning rate is set to 2e-4 and is decreased by multiplying 0.1 after the first 100K steps, and the total number of steps is 150K. A non-overlapped sliding window is applied to the action feature inputs of different lengths. The window length is set to 64 for most experiments, although we also conducted an ablation experiment on this hyperparameter, which is reported in the experimental section.

For all the language models, AdamW [43] is used as the optimizer, and the pre-trained weights are all loaded before fine-tuning. For T5 models, the initial learning rate is set to 1e-4, and the batch size is 16. Particularly, T5 3B is fine-tuned using LoRA[30], with $\alpha_{lora} = 16$, $dropout_{lora} = 0.1$, $rank_{lora} = 64$. For Llama 2 7B, the initial learning rate is set to 3e-3 with a cosine learning rate scheduler, and the batch size is 64, 8-bit quantization with single precision is used. LoRA is also used for fine-tuning Llama 2 7B, with $\alpha_{lora} = 16$, $dropout_{lora} = 0.1$, $rank_{lora} = 64$. For ChatGLM2 6B, the initial learning rate is set to 5e-3, and the batch size is 16, the single precision is used without quantization. Here, P-tuning v2 [41] is used for its fine-tuning, with soft prompt length set to 128.

The architecture of the two convolutional backbones used in VQ-VAE and action recognition module are reported in the following tables, respectively.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 4. The architecture of the convolutional backbone used in VQ-VAE.

Components	Architecture
Encoder	<ul style="list-style-type: none"> (0): Conv1d(315, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (1): ReLU() (2): $2 \times$ Sequential(<ul style="list-style-type: none"> (0): Conv1d(512, 512, kernel_size=(4,), stride=(2,), padding=(1,)) (1): Resnet1D(<ul style="list-style-type: none"> (model): Sequential(<ul style="list-style-type: none"> (0): 3 \times ResConv1DBlock(<ul style="list-style-type: none"> (norm1): Identity() (norm2): Identity() (activation1): ReLU() (activation2): ReLU()) (conv1): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(9,), dilation=(9,)) (conv2): Conv1d(512, 512, kernel_size=(1,), stride=(1,))))) (4): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)))
Codebook	nn.Parameter((512, 512), requires_grad=False)
Decoder	<ul style="list-style-type: none"> (0): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (1): ReLU() (2): $2 \times$ Sequential(<ul style="list-style-type: none"> (0): Resnet1D(<ul style="list-style-type: none"> (0): 3 \times ResConv1DBlock(<ul style="list-style-type: none"> (norm1): Identity() (norm2): Identity() (activation1): ReLU() (activation2): ReLU()) (conv1): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(9,), dilation=(9,)) (conv2): Conv1d(512, 512, kernel_size=(1,), stride=(1,)))) (1): Upsample(scale_factor=2.0, mode='nearest') (2): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (4): Conv1d(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (5): ReLU() (6): Conv1d(512, 315, kernel_size=(3,), stride=(1,), padding=(1,))

Table 5. The of the convolutional backbone used in the action recognition module.

Components	Architecture
Action Type Classifier	<ul style="list-style-type: none"> (0): Conv1d(315, 64, kernel_size=(7,), stride=(2,), padding=(3,)) (1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True) (2): ReLU() (3): MaxPool1d(kernel_size=3, stride=2, padding=1, dilation=1, ceil_mode=False) (4): 2 × Sequential(<ul style="list-style-type: none"> (0): 2 × ResidualBlock(<ul style="list-style-type: none"> (1): Conv1d(64, 64, kernel_size=(3,), stride=(1,), padding=(1,)) (2): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True) (3): ReLU() (4): Conv1d(64, 64, kernel_size=(3,), stride=(1,), padding=(1,)) (5): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True) (6): Dropout(p=0.5, inplace=False) (7): Identity())) (5): AdaptiveAvgPool1d(output_size=1) (6): Linear(in_features=128, out_features=25, bias=True)