PosePrompt: Soft-Prompted Large Language Models for Natural Language Description of Human Motion from 2D Pose

Liyun Gong
School of Agri-food Technology and Manufacturing
University of Lincoln
Lincoln, UK
Email: lgong@lincoln.ac.uk

Oluwapelumi Alagbe
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: 29404843@students.lincoln.ac.uk

Sheldon McCall
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: ShMccall@lincoln.ac.uk

Philp Williams
Great Northern Physiotherapy
Lincoln, UK
Email: phil@greatnorthernphysiotherapy.co.uk

Huizhi Liang
School of Computing
Newcastle University
Newcastle upon Tyne, UK
Email: Huizhi.Liang@newcastle.ac.uk

Miao Yu
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: myu@lincoln.ac.uu

***Abstract*—Natural language description of human motion enables interpretable motion understanding beyond categorical action labels. While motion–language models predominantly rely on 3D motion capture or parametric body models, such as A Skinned Multi-Person Linear Model (SMPL), these representations require specialised hardware or complex reconstruction pipelines. In contrast, 2D human pose estimation is widely accessible, yet remains underexplored for motion description. We propose PosePrompt, a parameter-efficient framework that generates natural language descriptions directly from 2D pose sequences using large language models. Spatiotemporal pose dynamics are encoded into compact motion features and injected into a pre-trained language model via soft prompt learning through multi-head attention layers, avoiding full model fine-tuning. This approach leverages strong linguistic priors while remaining effective in low-data settings. Experiments on a curated pose–text dataset demonstrate that PosePrompt produces accurate and coherent motion descriptions from 2D poses extracted from video sequences, with lightweight training requirements.**

Keywords- 2D poses; LLM; softprompt; multi-head attention.

I. INTRODUCTION

Understanding and describing human motion is a fundamental problem in computer vision, with applications spanning activity recognition, healthcare monitoring, human–computer interaction, and sports analysis. Natural language descriptions of motion provide semantic, human-interpretable explanations of movement patterns, enabling richer interaction and more advanced downstream reasoning beyond discrete action labels.

Recent advances in vision–language models and large language models (LLMs) have demonstrated strong capabilities in image and video understanding, including general video captioning and human motion description [1,2]. However, most existing LLM-based human description approaches rely on parametric 3D body models, such as SMPL, which require specialised motion-capture systems as well as heavy fine-tuning of LLMs, which incur substantial computational overhead, limiting their scalability and practical deployment.

To address these challenges, we propose **PosePrompt**, a parameter-efficient framework that adapts pre-trained LLMs to generate natural-language descriptions directly from 2D human pose sequences via soft prompt learning. Rather than relying on 3D motion representations such as SMPL, PosePrompt exploits 2D pose sequences, which capture the essential kinematic structure of human motion and have been widely adopted in action-recognition research. A lightweight pose encoder transforms the 2D skeletal sequences into compact spatiotemporal representations and produces a set of learnable continuous soft-prompt embeddings. These embeddings are injected into a pre-trained LLM to guide text generation. Crucially, this design avoids fine-tuning the entire language model, enabling efficient adaptation while preserving the strong linguistic priors of the pre-trained LLM.

We evaluate PosePrompt on a curated dataset of short human motion clips paired with textual descriptions, constructed from multiple public action datasets. Experimental results demonstrate that the proposed lightweight PosePrompt framework can generate accurate and semantically meaningful text descriptions using purely 2D

pose inputs, highlighting the potential of soft-prompted LLMs as an effective and scalable solution for interpretable human motion understanding from 2D skeletal data.

In Section 2, related work on pose-based action recognition and motion description is reviewed. In Section 3, the proposed PosePrompt methodology is described, including pose extraction, preprocessing, the pose encoder, and soft prompt generation. In Section 4, the experimental setup and results are presented. In Section 5, the conclusions and future work are provided.

II. RELATED WORKS

There exists a substantial body of work exploiting 2D and 3D human pose representations for motion understanding, particularly in the context of action recognition and, more recently, motion description.

Pose-Based Action Recognition:

Skeleton- and pose-based action recognition has been extensively studied as an alternative to RGB video, owing to its robustness to appearance variations, background clutter, and illumination changes. Early deep-learning approaches primarily employed recurrent architectures, such as long short-term memory (LSTM) networks, to model temporal dynamics in pose sequences for action classification [3]. These methods treat pose sequences as time series and capture motion patterns through sequential modeling of joint coordinates. Subsequent work introduced spatiotemporal graph convolutional networks (ST-GCN) [4] and their variants, which represent the human skeleton as a graph with joints as nodes and bones as edges. By explicitly modeling both spatial relationships between joints and their temporal evolution, these approaches more effectively capture structured motion dependencies than earlier time-series-based methods as in [3].

Beyond graph-based formulations, alternative representations have also been explored. Some approaches convert pose sequences into heatmap volumes or pseudo-images and apply 2D or 3D convolutional neural networks [5], achieving competitive performance while avoiding explicit graph construction. More recently, transformer-based architectures have been introduced for pose-based action recognition [5], leveraging self-attention mechanisms to model long-range temporal dependencies and complex inter-joint interactions across entire motion sequences. Despite these advances, most existing pose-based methods focus on predicting discrete action labels, offering limited semantic interpretability.

From Motion Labels to Motion Description:

Compared with action recognition and text-to-motion generation, motion-to-text description remains relatively underexplored. Early studies in this area predominantly adopted sequence-to-sequence learning paradigms. For instance, [6] employed an LSTM-based seq2seq framework to map 3D human motion sequences to natural language descriptions, while subsequent work incorporated generative adversarial learning into seq2seq models to better capture the diversity and realism of motion–language mappings [7].

More recently, driven by advances in LLMs, increasing attention has been paid to exploiting LLMs for motion interpretation using 3D motion representations, particularly those derived from the Archive of Motion Capture As Surface Shapes (AMASS) dataset and parameterized by SMPL or SMPL-X body models [8]. In particular, several recent works [1][2][9] explicitly treat human motion as a foreign language by discretizing continuous motion sequences into motion tokens and extending the vocabulary of language models. This formulation enables motion and text to be represented within a unified token space, allowing bidirectional translation between motion and natural language through end-to-end fine-tuning of LLMs.

While these LLM-based approaches have substantially advanced motion–language understanding, they typically rely on accurate 3D motion capture data or reliable estimation of SMPL parameters. Such requirements often involve specialised hardware, controlled capture environments, or computationally expensive reconstruction and training pipelines. Consequently, their applicability in unconstrained real-world settings—where only monocular RGB video is available—remains limited. This limitation motivates the exploration of motion-to-text description directly from more accessible representations, such as 2D pose sequences, without the need for heavy fine-tuning of large language models.

III. METHODOLOGY

Given a sequence of 2D human poses extracted and pre-processed from a video clip, the goal of this work is to generate a natural-language description that provides a semantic summary of the observed human motion. We address this task by leveraging a large language model enhanced through soft prompt learning, enabling effective adaptation to pose-based motion understanding, as illustrated in Figure 1. The proposed methodology is described below:

i. 2D pose extraction and pre-processing:

We employ the Detectron2 library [10] to extract 2D human poses from the original video sequences. Utilizing this library, we obtain the 2D skeleton, from which we extract the (x, y) coordinates of 12 key body points (including left/right shoulders, elbows, wrists, hips, knees, and ankles) for each video frame. After key point extraction by Detectron2, several preprocessing tasks are executed. Initially, we designate the midpoint of the shoulders as the origin and adjust the position of each key point accordingly to derive relative key point positions. Following this, we standardize the key point positions using the length of the trunk, defined as the distance between the midpoint of the shoulders and that of the hips, as a normalization factor. Upon completion of the preprocessing steps, an individual key point $\boldsymbol{p}_i$ is transformed according to the following formula:

$$\widehat{\boldsymbol{p}}_i = \frac{\boldsymbol{p}_i - \boldsymbol{p}_{mid_{shoulder}}}{L} \quad (1)$$

where $\boldsymbol{p}_{mid\_shoulder}$ represents the position of the middle of the left and right shoulders, and L represents the trunk length.

Through the utilization of the aforementioned preprocessing methods, we ascertain that the extracted 2D poses maintain invariance concerning scale and position.

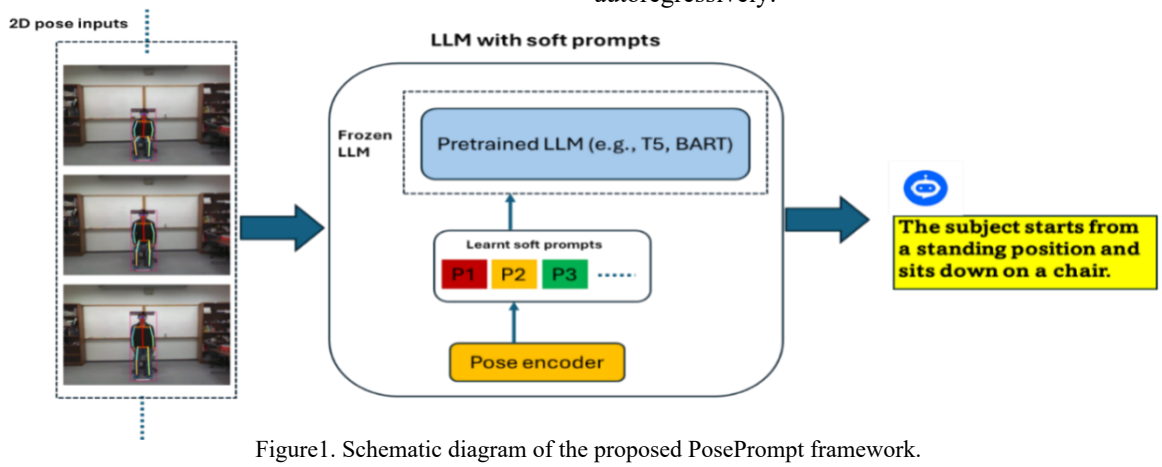


Figure1. Schematic diagram of the proposed PosePrompt framework.

Each extracted and processed 2D pose in a video sequence is concatenated and represented as

$$\mathbf{P} = \{\mathbf{p}_t\}_{t=1}^{T}, \mathbf{p}_t \in \mathbb{R}^{J\times 2}, \quad (2)$$

where $\boldsymbol{p}_t$ denotes the pre-processed 2D pose at time step t, consisting of $J = 12$ body joints. T represents the total number of frames in the sequence. The resulting pose sequence $\boldsymbol{P}$ serves as the input to the proposed PosePrompt framework as below.

ii. PosePrompt for motion to text

The proposed **PosePrompt** framework consists of three main components: i. Pose Encoder ii. Soft Prompt Generator and iii, Pre-trained Language Model, as shown in Figure 1.

Instead of training a motion-to-text model from scratch, we leverage a pre-trained large language model (LLM) and adapt it to pose-based motion interpretation using soft prompt learning. To capture motion dynamics from 2D poses, we employ a lightweight pose encoder f_{pose} that maps the input pose sequence $\mathbf{P}$ to a fixed-length feature representation:

$$\mathbf{z} = f_{\text{pose}}(\mathbf{P}), \mathbf{z} \in \mathbb{R}^{d}. \quad (3)$$

The pose encoder operates on normalized joint coordinates to model motion dynamics. In practice, f_{pose} can be implemented using a transformer-based encoder with positional embeddings [11], which can capture both short-term joint movements and longer-range temporal dependencies.

Given the pose feature $\mathbf{z}$, we generate a sequence of K soft prompt vectors:

$$\mathbf{S} = g(\mathbf{z}) \in \mathbb{R}^{K\times d_{\text{LM}}}, \quad (4)$$

where $g(\cdot)$ is a prompt projection network (e.g., a linear or MLP layer), and d_{LM} is the embedding dimension of the LLM. These soft prompts could be prepended to the input token embeddings of the language model, together with the textual tokens:

$$[\mathbf{S}; \mathbf{E}(x_1), \ldots, \mathbf{E}(x_n)], \quad (5)$$

where $\mathbf{E}(\cdot)$ denotes the token embedding function and x_i are textual tokens.Conditioned on the soft prompts, the pre-trained LLM could generate a natural language description autoregressively:

$$p(\mathbf{Y} \mid \mathbf{P}) = \prod_{l=1}^{L} p(y_l \mid y_{<l}, \mathbf{S}). \quad (6)$$

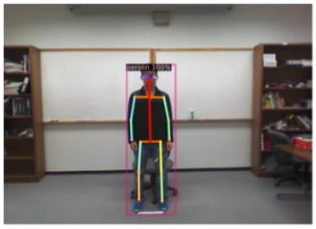
where $\mathbf{Y}$ represents the output texts.

Both the pose encoder and the prompt projection network are trained using a standard cross-entropy loss between the generated tokens and the corresponding ground-truth motion descriptions. This training objective enables the model to learn appropriate soft prompts from input 2D pose sequences for text generation, while keeping the large language model frozen. As a result, the proposed approach achieves efficient adaptation while preserving the rich linguistic knowledge of the pre-trained model.

IV. EXPERIMENTAL RESULTS

We evaluated the proposed PosePrompt framework for motion-to-text generation using a dataset composed of samples from three public datasets: Weizmann, Intelligent Sensing Lab Dataset (ISLD), and Unbiasing through Textual Descriptions (UTD) [12]. In total, the dataset contains 530 video samples, each paired with a corresponding natural-language description. The data are split into 70% for training and 30% for testing.

The pose encoder adopts a multi-head attention architecture consisting of three attention layers, each with four heads and an output feature dimension of 256. The number of soft prompt vectors (denoted as K in Equation (4)) is set to 40 for each input sequence. Two pre-trained (LLM) backbones are investigated in this work: a lightweight T5-small model and a relatively larger Bidirectional and Auto-Regressive Transformer (BART) -base model [13]. The pose



Output text: The subject starts in a seated position and stands up to a full upright posture.

Output text: A person is bending down and reaching with one hand to pick up an object.

Output text: The subject stands upright, starts with their hands close to the chest, and pushes them forward.

Output text: The person walks forward at a steady pace.

Figure 2. Output text results from video clip samples

encoder and prompt projection networks are trained using the AdamW optimizer [14], with a learning rate of $2 \times 10^{-4}$and a weight decay of 0.01 with 1000 epochs.

After training, the PosePrompt framework is evaluated on the different test datasets. Qualitative results are illustrated in Figure 2, which shows example motion descriptions generated from the extracted 2D pose sequences. Furthermore, we conduct both subjective and objective evaluations across the entire test set. Subjective evaluation is performed by human comparison between the ground-truth descriptions to evaluate the text generation accuracy, while objective evaluation is conducted using the Bilingual Evaluation Understudy (BLEU) metric [15], as reported in Table I. The results indicate that the T5-based PosePrompt model achieves higher motion-to-text generation accuracy than its BERT-based counterpart, attaining over 90% accuracy of the generated text descriptions.

TABLE I. POSEPROMPT RESULTS BASED ON DIFFERENT LLM MODELS

	T5-small	***BART-base***
No. of parameters	~60M	~139M
Subjective Predict Accuracy	91.19%	88.68%
BLEU	92.01	88.19

V. CONCLUSIONS AND FUTURE WORK

We proposed PosePrompt, an efficient framework for generating natural language descriptions of human motion directly from 2D pose sequences using soft-prompted large language models. By conditioning a frozen LLM on pose-derived motion features, our approach avoids full model fine-tuning while effectively bridging skeletal representations and language generation. Experiments on a curated pose–text dataset demonstrate that PosePrompt could achieve accurate and coherent motion descriptions based on 2D poses. These results highlight the potential of soft-prompted LLMs for interpretable human motion understanding from widely available and privacy-preserving 2D pose data. Future work will extend PosePrompt to a broader range of large language model architectures and investigate its application to domain-specific tasks, such as gait analysis and healthcare assessment.

ACKNOWLEDGMENT

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778602 ULTRACEPT.

REFERENCES

[1] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "MotionGPT: human motion as a foreign language*", NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, no. 880, Pages 20067 – 20079.

[2] Y. Wang et al., "MotionGPT-2: A General-Purpose Motion-Language Model for Motion Generation and Understanding," Online. Available from: https://arxiv.org/html/2410.21747v1

[3] J. Liu, G. Wang, P. Hu, L. Duan, and A. Kot, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[4] S. Yan, Y. Xiong, and D. Lin," Spatial-temporal graph convolutional networks for skeleton-based action recognition," *AAAI'18*, no. 912, pp. 7444-7452, 2018.

[5] M. Liu et al., "3D Skeleton-Based Action Recognition: A Review," [Online]. Available from: https://arxiv.org/abs/2506.00915

[6] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robotics and Autonomous Systems,* vol. 109, pp. 13-26,2018.

[7] Y. Goutsu, and T. Inamura, "Linguistic Descriptions of Human Motion with Generative Adversarial Seq2Seq Learning," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021.

[8] N. Mahmood, N. Ghorbani, N. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of Motion Capture As Surface Shapes," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019.

[9] C. Guo, S. Wang, and L. Cheng, "TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts," *European Conference on Computer Vision,* Tel Aviv, Israel, 2022.

[10] Y. Wu, A. Kirillov, F. Massa, WY. Lo, and R. Girshick, Detectron2, [Online]. Available from: https://github.com/facebookresearch/detectron2

[11] A. Vaswani, et al., Attention is All you Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[12] MPOSE2021, [Online]. Available from: https://github.com/PIC4SeR/MPOSE2021_Dataset
[13] Pretrained models, [Online]. Available from: https://huggingface.co/transformers/v4.6.0/pretrained_models.html
[14] I. Loshcgilov and F. Hutter, "Decoupled Weight Decay Regularization," *International Conference on Learning Representations* (ICLR 2019), LA, USA, 2019.
[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu , BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL), Philadelphia, 2002.