Towards Closed-Loop Embodied Empathy Evolution: Probing LLM-Centric Lifelong Empathic Motion Generation in Unseen Scenarios

Jiawen Wang[1], **Jingjing Wang**[1*], **Tianyang Chen**[1], **Min Zhang**[2], **Guodong Zhou**[1]

[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Harbin Institute of Technology, Shenzhen, China
jwwang1031@stu.suda.edu.cn, djingwang@suda.edu.cn, 20244227061@stu.suda.edu.cn, {mzhang, gdzhou}@suda.edu.cn

Abstract

In the literature, existing human-centric emotional motion generation methods primarily focus on boosting performance within a single scale-fixed dataset, largely neglecting the flexible and scale-increasing motion scenarios (e.g., sports, dance), whereas effectively learning these newly emerging scenarios can significantly enhance the model's real-world generalization ability. Inspired by this, this paper proposes a new **L**LM-Centric **L**ifelong **E**mpathic **M**otion **G**eneration (L^2-EMG) task, which aims to equip LLMs with the capability to continually acquire emotional motion generation knowledge across different unseen scenarios, potentially contributing to building a closed-loop and self-evolving embodied agent equipped with both empathy and intelligence. Further, this paper poses two key challenges in the L^2-EMG task, i.e., the emotion decoupling challenge and the scenario adapting challenge. To this end, this paper proposes an **E**motion-Transferable and **S**cenario-Adapted **Mi**xture **o**f **E**xperts (ES-MoE) approach which designs a causal-guided emotion decoupling block and a scenario-adapted expert constructing block to address the two challenges, respectively. Especially, this paper constructs multiple L^2-EMG datasets to validate the effectiveness of the ES-MoE approach. Extensive evaluations show that ES-MoE outperforms advanced baselines.

Introduction

Human motion generation technology focuses on generating either 3D static motions (i.e., poses (Feng et al. 2023)) or 3D sequential motions (i.e., motions (Guo et al. 2020; Mao et al. 2019; Tevet et al. 2023)) based on condition signals, which mainly include text (Guo et al. 2022; Tevet et al. 2023; Zhang et al. 2023a) and speech (Tseng et al. 2023; Li et al. 2022), and has broad applications in virtual reality, the metaverse, and game development. Recently, some works (Yu et al. 2024b; Chen et al. 2024) begin to consider introducing emotional information into the motion generation process due to its potential applications in the field of empathic robots and emotional virtual avatars. However, due to the reliance on specialized equipment for motion data collection and the complexity of emotion annotation processes, existing human motion generation datasets struggle to achieve rapid dynamic updates that are synchronized with real-world scenarios. This lag in data updates causes deployed models to continuously encounter unseen motion scenarios, leading to issues with model generalization decay. Furthermore, even when datasets are updated in a timely manner, considering limited storage and computational budgets, storing all historical data and repeatedly retraining models remains expensive and almost unfeasible.

With these in mind, this paper proposes a **L**LM-Centric **L**ifelong **E**mpathic **M**otion **G**eneration (L^2-EMG) task, which enables Large Language Models (LLMs) to continually learn emotional motion generation abilities across different unseen scenarios. It can powerfully contribute to developing an empathetic and intelligent embodied agent with closed-loop self-evolving. Specifically, the L^2-EMG task sequentially trains LLM on datasets from different unseen motion scenarios, continually learning the emotional generation for new scenarios while preventing the forgetting of motion generation knowledge learned from previous scenarios. As illustrated in Figure 1, for scenarios *"Daily Life"* and *"Sports"*, the model can accurately generate emotional motions that align with the descriptions *"walk"* and *"jog"*, while consistently expressing their shared emotion label *"Sad"*. In this paper, we believe that this new task at least faces two key challenges, which are illustrated as follows.

On the one hand, ensuring the sustainable transfer of emotional representation commonality across scenarios during lifelong learning is challenging, namely the emotion decoupling challenge. The diversity of emotions is not limited to the singularity of a scenario; that is, the motion in each individual scenario contains a variety of emotions. Thus, the model needs to effectively decouple emotional representations with cross-scenario transferability while learning the motion generation style of a specific scenario. As illustrated in Figure 1, two motions from different scenarios—*"Shows"* (training) and *"Sports"*(unseen)—both express the *"Sad"* emotion. We expect the model to capture invariant emotional features (e.g., reduced limb movement and lowered head) from scenario *"Shows"* and transfer them to scenario *"Sports"*. Therefore, this paper believes that a well-behaved approach should be able to decouple emotional representation commonality during cross-scenario lifelong learning and transfer it to new scenarios.

On the other hand, ensuring that the uniqueness of each motion scenario is not forgotten during cross-scenario life-

*Corresponding Author: Jingjing Wang

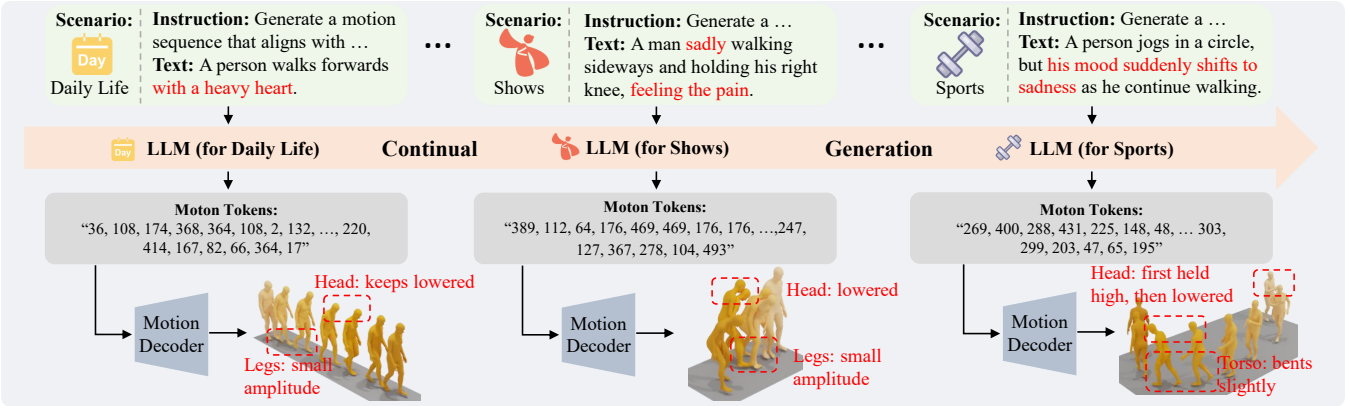

Figure 1: Examples of three scenarios to illustrate our L^2-EMG task for continual generation. Red words indicate emotional information in the input, while red boxes highlight the expected emotional motion expressions for the shared emotion *Sad*.

long learning is also challenging, namely the scenario adapting challenge. Human motion types are diverse and can be categorized into various scenarios, so the model needs to learn the personalized expressions of motion in each scenario while forgetting as little as possible when learning new scenarios. For example, Figure 1 shows motions from three different scenarios. In the second scenario, the *"Shows"* scenario expresses emotions in a more exaggerated manner, while in the third scenario, the *"Sports"* scenario conveys emotions through more professional and skill-oriented movements. Therefore, this paper believes that a better-behaved approach should be able to capture the uniqueness of each scenario during cross-scenario lifelong learning and avoid catastrophic forgetting when learning new scenarios.

To address the above challenges, this paper proposes an **E**motion-Transferable and **S**cenario-Adapted **M**ixture **o**f **E**xperts (ES-MoE) approach, which enables cross-scenario lifelong learning and emotion-enriched human motion generation. Specifically, inspired by recent causal decoupling works (Wang et al. 2022; Liu et al. 2023a), ES-MoE designs a causal-guided emotion decoupling block to decouple and highlight the common emotional representations shared across different motion scenarios. Then, ES-MoE designs a scenario-adapted expert constructing block based on the MoE (Jacobs et al. 1991; Yu et al. 2024a) architecture to learn scenario-specific expressions and facilitate efficient knowledge transfer. Furthermore, we construct multiple L^2-EMG datasets to validate the effectiveness of ES-MoE. Comprehensive experiments demonstrate that ES-MoE achieves significant improvements compared to the advanced lifelong learning baselines.

Related Work

Human Motion Generation. Human motion generation produces diverse and realistic 3D motions from various controls, such as text (Zhang et al. 2023a; Guo et al. 2023), audio (Tseng et al. 2023; Li et al. 2022), pose (Liu et al. 2023b; Mao et al. 2019), and trajectories. Earlier works (Ghosh et al. 2021; Guo et al. 2022; Tevet et al. 2022) build shared latent spaces to learn input–motion relations, while recent diffusion-based approaches (Chen et al. 2023; Zhang et al. 2023b) (e.g., MotionDiffuse (Zhang et al. 2024a), MDM (Tevet et al. 2023), EDGE (Tseng et al. 2023)) further improve motion quality. Other works (Zhang et al. 2023a, 2024b; Jiang et al. 2023; Yang et al. 2024) explore discrete motion representations via VQ-VAE (Oord 2017). Recent efforts incorporate emotion understanding (Yu et al. 2024b; Chen et al. 2024) for emotion-controllable generation. Unlike them, we propose a new L^2-EMG task to enhance lifelong learning in T2M models across expanding motion scenarios, enabling intelligent and empathetic embodied agents.

Lifelong Learning. Lifelong (continual) learning improves future-task generalization while mitigating catastrophic forgetting. Existing methods include parameter regularization (Lopez-Paz and Ranzato 2017; Isele and Cosgun 2018; Zhang et al. 2025), replay (Kirkpatrick et al. 2016; Isele and Cosgun 2018), and model expansion (Gao et al. 2024; Zhao et al. 2024). Regularization preserves past knowledge (e.g., EWC (Kirkpatrick et al. 2016)); replay rehearses stored samples; and expansion methods (Wang et al. 2023a; Zhao et al. 2024) allocate task-specific parameters, with SAPT (Zhao et al. 2024) enabling selective sharing. MoE-based architectures (Jacobs et al. 1991; Yu et al. 2024a) have recently shown strong potential in continual learning. Building on these insights, we introduce ES-MoE for the L^2-EMG task, which incorporates a causal-guided emotion decoupling block and a scenario-adapted expert construction block to continually learn emotional motion generation across diverse scenarios.

Approach

In this paper, we propose the L^2-EMG task, which aims to enable the model to retain and utilize the emotional motion generation abilities learned from different scenarios and generate natural emotional motions that are appropriate for the new scenario, even when facing entirely new, unseen motion scenarios. Our L^2-EMG task is formulated as follows: $\{S_1, S_2, \ldots, S_n\}$ represent n different motion scenarios, and $\{D_1, D_2, \ldots, D_n\}$ represent the corresponding datasets for these scenarios, where each dataset $D_i =$

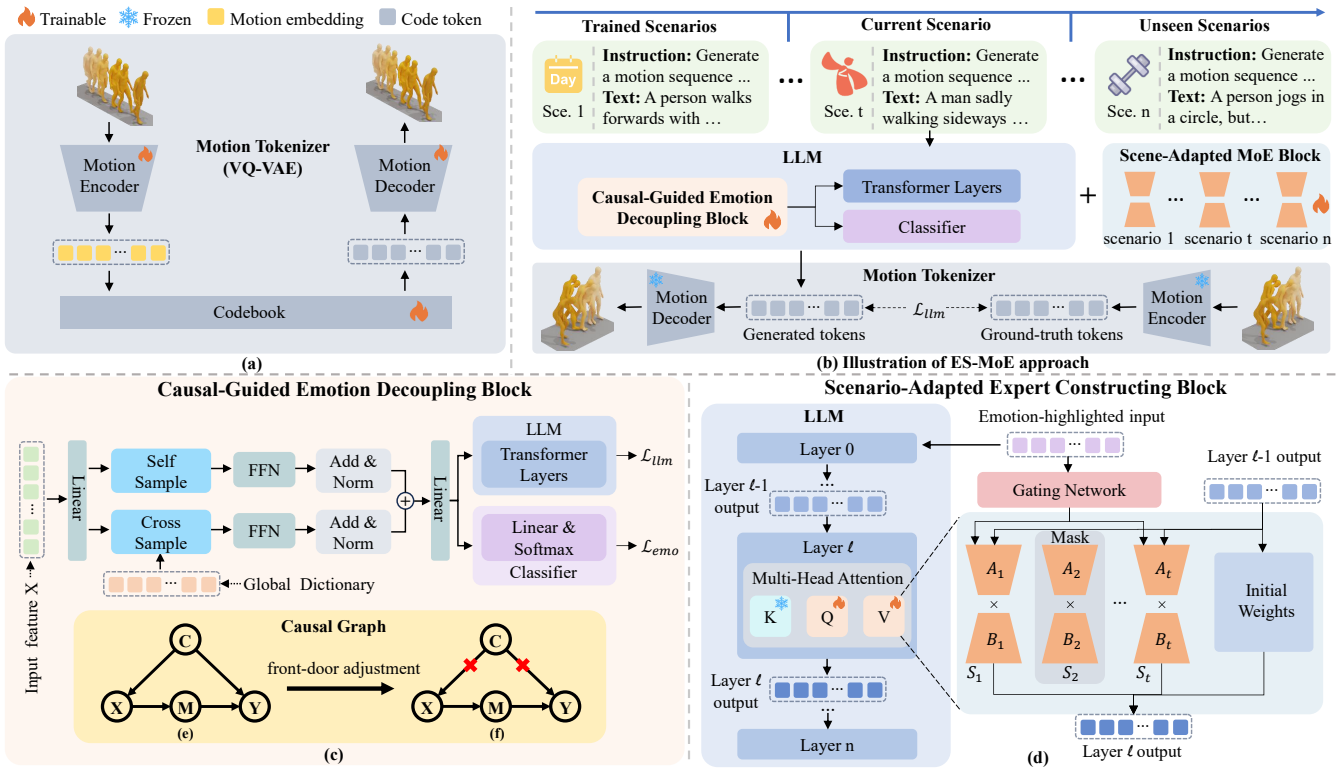

Figure 2: ES-MoE overview: (a) Motion Tokenizer training phase; (b) overall architecture of our approach; (c)/(d) causal-guided emotion decoupling block and scenario-adapted expert constructing block, respectively; (e)/(f) causal intervention graph.

$\{(s^i, t^i_j, \mathbf{mt}^i_j)\}_{j=1}^{N_i}$ has the size of N_i and contains scenario-specific instructions s^i, text descriptions t^i_j of motions, and output motion tokens $\mathbf{mt}^i_j$. We train the model sequentially on the $\{D_1, D_2, \ldots, D_n\}$ datasets. At each time step i, the model only has access to the D_i dataset, with the goal that the model learns to generate motions for the scenario at time i without forgetting the scenarios learned before time i.

To address the L^2-EMG task, we propose the ES-MoE framework (Figure 2). It contains a Motion Tokenizer that encodes motion sequences into motion tokens, a causal-guided emotion decoupling block that handles emotion decoupling, and a scenario-adapted expert constructing block that enables effective scenario adaptation.

Motion Tokenizer

To convert 3D human motion data into tokens, we use a VQ-VAE (Vector Quantized Variational AutoEncoder) (Oord 2017) as the motion tokenizer. As shown in Figure 2 (a), the motion tokenizer consists of two main components: the motion encoder $\mathcal{E}$, which encodes the motion sequence into discrete tokens, and the motion decoder $\mathcal{D}$, which decodes the motion tokens back into the motion sequence.

Specifically, the input to the motion tokenizer is a human motion sequence $\mathbf{mo}$, and the output is the reconstructed motion sequence $\mathbf{mo}'$. At first, motion encoder $\mathcal{E}$ encodes the human motion sequence $\mathbf{mo}$ into the motion feature embedding $\mathbf{e}$. Next, a quantization operation $\mathrm{Quan}(\cdot)$ is applied to transform the motion feature embedding $\mathbf{e}$ into $\mathbf{z}$, a sequence of code vectors, from the learnable codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$, where k is the size of the codebook and each code vector in the codebook is associated with a discrete token. The quantization operation refers to finding the code vectors in the codebook that are most similar to the motion feature embedding $\mathbf{e}$, which can be mathematically written as:$\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_i, ..., \mathbf{z}_{d_\mathbf{e}}\}, \mathbf{z}_i = \mathrm{Quan}(\mathbf{e}_i) = \arg\min \|\mathbf{e}_i - \mathbf{c}_j\|_2$,

where $i \in (1, d_\mathbf{e})$ and $\mathbf{c}_j \in \mathcal{C}$. $d_\mathbf{e}$ represents the number of columns in the motion embedding $\mathbf{e}$. $\mathbf{e}_i$ and $\mathbf{z}_i$ represent the row vectors of $\mathbf{e}$ and its quantized output, respectively. Finally, the decoder $\mathcal{D}$ can decode quantized code tokens into reconstructed motion sequences $\mathbf{mo}'$.

Causal-Guided Emotion Decoupling Block

In this paper, we leverage the causal intervention technique (Pearl et al. 2018) and design a causal-guided emotion decoupling block to decouple the emotional representation commonality in the motion and further emphasize the commonality of emotion in lifelong learning across different scenarios. This block consists of two parts: the causal intervention graph aiming to decouple emotion via the front-door adjustment strategy (Pearl et al. 2018), and the deconfounded causal attention that implements this strategy through attention mechanisms in the L^2-EMG task, as described below:

Causal Intervention Graph for Emotion Decoupling. First, to accurately model the causal relation in the de-

coupling process, we construct a causal intervention graph as shown in Figure 2 (c). Here, X represents the model's input features, including motion and emotion information. M denotes the decoupled feature representation. Y denotes the motion's emotion category, and C represents emotion-irrelevant confounding factors, such as shallow motion semantics. $\mathrm{X} \rightarrow \mathrm{M} \rightarrow \mathrm{Y}$ is a front-door path from X to Y, representing the causal effect of the input features X on the emotion category Y. $\mathrm{X} \leftarrow \mathrm{C} \rightarrow \mathrm{Y}$ is a back-door path from X to Y, representing the causal effect of the confounding factors C on the features X and emotion category Y.

To mitigate the causal influence of emotion-irrelevant confounding factors C and further decouple the emotion information in the features, we adopt a front-door adjustment strategy to alleviate the issue where the model is confused by shallow motion semantics when identifying emotion information. Specifically, we use the do-operator (Pearl et al. 2018) to intervene on X = x to realize the causal effect of $\mathrm{X} \rightarrow \mathrm{Y}$ and reduce the influence of C on X, thereby blocking the back-door path $\mathrm{X} \leftarrow \mathrm{C} \rightarrow \mathrm{Y}$. Afterward, we use the front-door adjustment strategy to further compute the causal effect of $\mathrm{X} \rightarrow \mathrm{M} \rightarrow \mathrm{Y}$ ((e) to (f) in Figure 2), which can be mathematically written as $P(\mathrm{Y}|\mathrm{do}(\mathrm{X} = x)) =$:

$$\sum_{m} P(\mathrm{M} = m|\mathrm{X}) \sum_{x} P(\mathrm{X} = x)[P(\mathrm{Y}|\mathrm{X} = x, \mathrm{M} = m)] \quad (1)$$

Deconfounded Causal Attention for Emotion Decoupling. Next, we implement the front-door adjustment strategy via the attention mechanism to decouple and highlight emotion information in the input feature X. Specifically, we adopt the Normalized Weighted Geometric Mean (NWGM) (Srivastava et al. 2014; Xu et al. 2015) approximation to implement the front-door adjustment strategy in Equation 1:

$$P(\mathrm{Y}|\mathrm{do}(\mathrm{X} = x)) \approx \mathrm{Softmax}(\phi(\mathbf{s}_x, \mathbf{s}_m)) \quad (2)$$

where do is do-operator (Pearl et al. 2018). $\phi(\cdot)$ is the parameter network that simulates the predictive distribution $P(\mathrm{Y}|\mathrm{X} = x, \mathrm{M} = m)$. $\mathbf{s}_x = \sum_m P(\mathrm{M} = m|f(\mathrm{X}))\mathbf{x}$ and $\mathbf{s}_m = \sum_x P(\mathrm{X} = x|g(\mathrm{X}))\mathbf{m}$ represent the estimated results of self-sampling and cross-sampling, respectively. $f(\cdot)$ and $g(\cdot)$ are the networks that generate query embeddings for input features X. The generated query embeddings $\mathbf{x}$ and $\mathbf{m}$ correspond to the variables x and m, respectively. While self-sampling helps the model decouple emotion information, it remains susceptible to confounding factors. Cross-sampling is thus employed to learn true emotional commonalities from other samples.

We implement the operation of self-sampling and cross-sampling via the attention mechanism. Specifically, the self-sampling result $\mathbf{s}_x$ is computed by: $\mathbf{s}_x = \mathbf{V_m} \cdot \mathbf{Softmax}(\mathbf{Q_m}^\top \mathbf{K_m})$, where $\mathbf{Q_m}$ comes from $f(\mathrm{X})$. $\mathbf{V_m}$ and $\mathbf{K_m}$ are obtained by linear transformation of the current input features X. The cross-sampling result $\mathbf{s}_m$ is calculated by: $\mathbf{s}_m = \mathbf{V_g} \cdot \mathbf{Softmax}(\mathbf{Q_c}^\top \mathbf{K_g})$, where $\mathbf{Q_c}$ comes from $g(\mathrm{X})$. $\mathbf{V_g}$ and $\mathbf{K_g}$ are obtained from a global dictionary which is compressed from the training set. We perform K-means clustering (Wong 1979) on the training set samples to initialize this global dictionary. By learning emotion commonality representations from other samples, the cross-sampling result $\mathbf{s}_m$ can effectively decouple the emotion information in the input. Finally, we concatenate and integrate $\mathbf{s}_x$ and $\mathbf{s}_m$ using a Feed-Forward Network (FFN) with parameters $\mathbf{W}$ and b, and finally obtain the emotion-highlighted input $\mathbf{h} = \mathbf{W}(\mathrm{FFN}_1(\mathbf{s}_x) \oplus \mathrm{FFN}_2(\mathbf{s}_m)) + b$.

To better decouple the emotional commonality representation, we also design an emotional constraint loss. Specifically, we input the integrated features $\mathbf{h}$ into a simple emotion classification network. This network consists of several fully connected layers, each followed by a ReLU activation function to introduce non-linearity. Finally, the Softmax function outputs the probability distribution of emotional categories. We use the cross-entropy loss function to calculate the difference between the model's predicted emotional category probabilities and the true motion emotion label: $\mathcal{L}_{emo} = \mathrm{CrossEntropy}(y_e, \hat{y}_e)$, where y_e and $\hat{y}_e$ represent the true motion emotion label and the prediction result of the classification network, respectively.

Algorithm 1: ES-MoE for lifelong learning on scenario i

Input: Pre-trained mdoel parameters θ; Trained expert parameters $\{\Delta\theta_j \mid j = 1, \ldots, i-1\}$, i-th expert parameters $\Delta\theta_i$ to be trained
Output: Final updated model parameters θ'
Obtain the emotion-highlighted input $\mathbf{h}$
Construct the i-th expert parameters $\Delta\theta_i = \mathbf{A}_i\mathbf{B}_i$ for the current scenario based on the LoRA method
Calculate the expert weights W_i ▷ Eq.(4)
Aggregate the parameters of all experts to obtain the final updated model weights θ'

Scenario-Adapted Expert Constructing Block

Further, we design a scenario-adapted expert constructing block based on the MoE (Jacobs et al. 1991; Yu et al. 2024a) architecture to adapt to the gradually increasing motion scenarios and enable efficient knowledge transfer across different motion scenarios. Algorithm 1 introduces the process of constructing experts when encountering a new i-th motion scenario, and two main processes are detailed as follows:

LoRA Experts Construction. To enable the model to effectively learn and retain knowledge across different motion scenarios, we design multiple experts for different scenarios based on LoRA (Hu et al. 2022). Specifically, we train the model sequentially on n different motion scenarios $\{S_1, S_2, \ldots, S_n\}$. At each time step i, our training objective is:

$$\arg\min_{\theta'} \sum_{(t,\mathbf{mt}) \in D_i} \mathcal{L}_i(f_{\theta'}(t), \mathbf{mt}), \theta' = \theta, \Delta\theta_1, ..., \Delta\theta_i \quad (3)$$

where D_i represents the dataset of i-th scenario. $f_{\theta'}$ represents the model with weight θ'. $\Delta\theta_i$ represents the change in model parameters after training on the i-th scenario.

Given the simplicity and effectiveness of the LoRA method, we use the LoRA module with parameters represented as $\mathbf{A}_i \times \mathbf{B}_i$ to simulate the parameter updates after training on the i-th scenario. For each motion scenario S_i,

Unseen	#Daily Life	#Sports	#Dance	#Shows	#Game	#Animation	#Instrument Play	#Acrobatics	Total
	5407	2191	2068	1101	1618	1719	3903	1909	19916
Mixed	**#Scenario1**	**#Scenario2**	**#Scenario3**	**#Scenario4**	**#Scenario5**	**#Scenario6**	**#Scenario7**	**#Scenario8**	**Total**
	2489	2489	2489	2489	2489	2489	2489	2493	19916

Table 1: Statistics of our constructed L^2-EMG dataset. Unseen and Mixed denote two ways of splitting our dataset.

we assign a LoRA module as the motion generation expert for that scenario. Since motion scenarios are inherently interrelated, we adopt a Mixture of Experts (MoE) architecture to complete the motion generation task corresponding to the i-th scenario, which integrates the experts that have already been trained up to time step i, rather than solely relying on the experts trained at time step i. The model parameters after training at the i-th time step can be represented as: $\theta' = \theta + \sum_{j=1}^{i} W_j \Delta\theta_j = \theta + \sum_{j=1}^{i} W_j \mathbf{A}_j \mathbf{B}_j$, where θ represents pre-trained model parameters. $\Delta\theta_i = \mathbf{A}_i \mathbf{B}_i$ represents the parameters of the LoRA expert corresponding to the i-th scenario. W_i represents the weight of the i-th expert, measuring the expert's contribution.

Gating Network Construction. Next, we design a gating network $\mathrm{Gate}(\cdot)$ to generate weight W for each LoRA expert. Specifically, the implementation is divided into three steps: 1) First, we use orthogonal initialization (Saxe et al. 2014; Peng et al. 2024) to generate a set of mutually orthogonal expert keys, ensuring discrimination between different experts in the key feature space. 2) Then, we process $\mathbf{h}$, the emotion-highlighted input sequence in above Section, through a learnable down and up projection layer (Zhao et al. 2024) and map it to a query embedding aligned with the expert key dimension. 3) Finally, we compute the dot product similarity between the query vector and each expert key to generate the weight of each expert, which can be denoted as:

$$W_i = \mathrm{Gate}(\mathbf{h}, \mathbf{K}_i) = \frac{e^{\mathrm{QueryEmbedding}(\mathbf{h})^\top \cdot \mathbf{K}_i}}{\sum_{j=1}^{i} e^{\mathrm{QueryEmbedding}(\mathbf{h})^\top \cdot \mathbf{K}_j}} \quad (4)$$

where $\mathbf{h}$ represents the emotion-highlighted input. $\mathbf{K}_i$ represents the key vector of i-th expert.

At time step i, only the LoRA matrix and key vector of the i-th expert are trained, while the query embedding network remains trainable. To avoid over-reliance on previously trained experts, we randomly mask trained experts from participating in weight assignments depending on the training time point. Among the remaining experts, we select the top k most relevant ones based on their weight sizes and recalculate the weights using the Softmax function.

Optimization for ES-MoE

Our model training consists of two stages. First, we train a VQ-VAE–based motion tokenizer that encodes motion sequences into tokens and reconstructs the original motion. Second, we sequentially fine-tune the LLM using motion generation datasets from different scenarios.

Stage1. Motion Tokenizer Training. To train VQ-VAE as a motion tokenizer, we follow T2M-GPT to set the optimization goal $\mathcal{L}_{vq}$. $\mathcal{L}_{vq}$ consists of three main components: the motion reconstruction loss $\mathcal{L}_{re}$, the embedding loss $\mathcal{L}_{embed}$, and the commitment loss $\mathcal{L}_{commit}$, which can be denoted as: $\mathcal{L}_{vq} = \mathcal{L}_{re} + \mathcal{L}_{embed} + \mathcal{L}_{commit}$.

Stage2. Continual Scenario Tuning. In the second stage, we construct multiple L^2-EMG datasets (detailed in Section 4.1) to endow the model with the ability to continually generate emotional motions across diverse motion scenarios. Then, we sequentially fine-tune the model on these datasets from different scenarios using the instruction "Generate a motion sequence that aligns with the following emotional text description." The loss in the second stage can be expressed as: $\mathcal{L} = \mathcal{L}_{llm} + \lambda \mathcal{L}_{emo}$. Here, $\mathcal{L}_{llm}$ denotes the LLM next-token prediction loss, and $\mathcal{L}_{emo}$ denotes the emotion classification loss mentioned in Causal-Guided Emotion Decoupling Block, and λ controls the weight of $\mathcal{L}_{emo}$.

Experimental Settings

Dataset Construction. We build two datasets corresponding to ES-MoE's two training stages. **(1)** For the **motion tokenizer training** stage, we construct a text–motion pair dataset using EmotionalT2M (Yu et al. 2024b) and selected subsets of Motion-X (Lin et al. 2023). Samples contain an emotional motion description and its motion sequence. Since EmotionalT2M is small, we expand it with Motion-X, whose texts and emotion labels are merged following (Yu et al. 2024b) using ChatGLM (Zeng et al. 2023). The processed Motion-X data is then combined with EmotionalT2M. **(2)** For the **continual scenario training** stage, we encode motions into discrete motion tokens via the trained tokenizer and convert them into instruction data. We categorize all samples into eight scenarios based on annotated scenario labels, with two annotators and expert adjudication (Kappa = 0.85). **(3)** To evaluate L^2-EMG, we design two lifelong learning settings: the **Unseen L^2-EMG** dataset, which sequentially fine-tunes across eight scenario-specific subsets, and the **Mixed L^2-EMG** dataset, which randomly mixes scenarios to mimic real-world incremental data. Each subset has a primary scenario with fewer motions from others and is split into train/val/test (0.8/0.05/0.15). More details are provided in Table 1.

Evaluation Metrics. Following prior studies (Guo et al. 2022; Yu et al. 2024b), let $f_{i,j}, r_{i,j}, d_{i,j}, m_{i,j}, wf_{i,j}$ denote the widely-used motion generation metrics, i.e., FID, top-1 R-Precision, diversity, multimodality score, and weighted F1-score of the generated motions in scenario j after the model has been trained on scenario i, respectively. Based on these, the evaluation metrics for our ES-MoE are calculated as follows: **Average FID (AF)** of all generated motions with different scenarios after training on the final motion scenario N. It is computed by $AF_N = \frac{1}{N}\sum_{j=1}^{N} f_{N,j}$; **Average R-**

Backbone	Approach	Unseen L^2-EMG Dataset						Mixed L^2-EMG Dataset					
		AF↓	AR↑	AD↑	AMM↑	AWF↑	FR↓	AF↓	AR↑	AD↑	AMM↑	AWF↑	FR↓
LLaMA2 (Non-CL)	MTL	1.05	0.281	9.63	1.56	0.383	-	1.05	0.281	9.63	1.56	0.383	-
	SeqLoRA	3.58	0.157	8.83	1.10	0.198	6.70	2.61	0.195	9.32	1.07	0.239	5.17
LLaMA2 (CL)	LwF-LoRA	4.38	0.164	8.44	0.95	0.152	5.76	3.36	0.215	8.81	1.27	0.188	3.70
	EPI	2.21	0.180	**9.86**	1.47	0.256	4.89	1.99	0.207	9.13	1.33	0.283	1.51
	O-LoRA	2.35	0.214	9.29	1.18	0.282	2.75	2.20	0.236	9.82	1.29	0.296	-0.42
	Prog-Prompt	7.30	0.102	6.75	**1.84**	0.128	7.16	5.77	0.153	8.12	**1.87**	0.134	5.09
	SAPT	2.12	0.237	9.61	1.59	0.313	-0.54	1.65	0.245	9.47	1.82	0.327	-1.97
LLaMA2 (Ours)	**ES-MoE**	**1.89**	0.241	9.74	1.47	**0.340**	**-1.03**	**1.39**	**0.259**	**9.87**	1.65	**0.347**	**-3.03**
	w/o CGED	2.07	**0.248**	9.53	1.39	0.312	-0.75	1.70	0.247	9.44	1.77	0.305	-1.74
	w/o SAMoE	2.48	0.206	9.08	1.78	0.286	3.26	1.96	0.229	9.28	1.58	0.273	0.83
	w/o $\mathcal{L}_{emo}$	2.01	0.232	9.45	1.42	0.326	-0.86	1.52	0.237	9.75	1.52	0.336	-2.59

Table 2: Comparison of our ES-MoE approach with other approaches on the L^2-EMG dataset. '↑'('↓') indicates that the values are better if the metric is larger (smaller).

Precision (AR): The average text-motion match top-1 precision of all generated motions with different scenarios after training on the final motion scenario N. It is computed by $AR_N = \frac{1}{N}\sum_{j=1}^{N} r_{N,j}$; **Average Diversity (AD)** performance of all scenarios after training on the final motion scenario N. It is computed by $AD_N = \frac{1}{N}\sum_{j=1}^{N} d_{N,j}$; **Average MultiModality (AMM)** performance of all scenarios after training on the final motion scenario N. It is computed by $AMM_N = \frac{1}{N}\sum_{j=1}^{N} m_{N,j}$; **Average Weight F1-score (AWF)** to evaluate the emotion performance of generated motions proposed by (Yu et al. 2024b). We compute AWF after training on the final motion scenario N. It is computed by $AWF_N = \frac{1}{N}\sum_{j=1}^{N} wf_{N,j}$; **Forgetting Rate (FR)** of the model on the first $N-1$ motion scenarios after training on the final motion scenario N, measuring how much knowledge has been forgotten during the lifelong learning process. It is computed by $F_N = \frac{1}{N-1}\sum_{j=1}^{N-1}\left(\max_{k=j}^{N-1} r_{k,j} - r_{N,j}\right)$;

Implementation Details and Baselines. The baselines we choose include: Multi-Task Learning (MTL, which trains a model on multiple tasks simultaneously), SEQ-LoRA (sequentially fine-tuning the model using the LoRA method in a predefined order). LwF-LoRA (Li et al. 2018), EPI (Wang et al. 2023b), O-LoRA (Wang et al. 2023a), Prog-Prompt (Razdaibiedina et al. 2023), and SAPT (Zhao et al. 2024). In the continual scenario tuning stage, we use LLaMA 2 (7B) (llama.com/llama2) as the backbone of all baselines and ES-MoE, and fine-tune it with LoRA for a fair comparison. Following prior works (Wang et al. 2023a), we compute the results of all CL baselines three times with different scenario orders and take the average as the final score.

Results and Discussion

Main Experimental Results. Table 2 presents a comparative analysis of different approaches on the L^2-EMG dataset. From this table, we can see that: **1) Performance on Unseen L^2-EMG Dataset.** On the Unseen L^2-EMG Dataset, our ES-MoE approach outperforms other baselines on almost all metrics. For instance, compared to the best-performing baseline, SAPT, ES-MoE achieves better results on the AF, AR, and AD metrics, suggesting that it generates more natural, coherent, and text-aligned motions in diverse scenarios. On the AMM metric, ES-MoE achieves comparable results, indicating that the motions generated by ES-MoE are diverse and rich. Moreover, ES-MoE also achieves better performance in the AWF and FR metrics, further confirming its effectiveness in decoupling emotional information from motions and its flexibility in adapting to entirely new motion scenarios. **2) Performance on Mixed L^2-EMG Dataset.** On the Mixed L^2-EMG Dataset which better aligns with real-world applications, our ES-MoE approach also surpasses other baselines on most metrics. This indicates that ES-MoE is not only effective at decoupling emotional information and generating realistic and natural emotional motions but also adapts well to the real-world scenario.

Effectiveness Study for Emotion Decoupling. To validate ES-MoE's effectiveness in addressing the emotion decoupling challenge, we conduct ablation studies (Table 2). Specifically, **w/o CGED** and **w/o** $\mathcal{L}_{emo}$ denote settings where the Causal-Guided Emotion Decoupling Block (CGED) and the $\mathcal{L}_{emo}$ do not work, respectively. From this table, we can see that: 1) **w/o CGED** performs worse on both unseen and mixed datasets, especially in AWF score. This justifies the effectiveness of the CGED block in decoupling emotional commonality and enabling the generation of more emotionally expressive motions. 2) **w/o** $\mathcal{L}_{emo}$ also shows a performance drop in AWF and AR on both datasets. This further demonstrates the effectiveness of the CGED block in addressing the challenge of emotional decoupling.

Effectiveness Study for Scenario Adapting. To validate the capability of ES-MoE in addressing the scenario adaptation challenge, we conduct related experiments, as shown in Table 2. Specifically, **w/o SAMoE** refers to the setting where the Scenario-Adaptive Mixture of Experts(SAMoE) does not work. From this table, we can see that: **w/o SAMoE** shows a significant performance drop on both types of datasets compared to ES-MoE, with all metrics substantially decreased, especially AF and AR. This demonstrates that scenario-adapted experts can efficiently adapt to newly introduced motion scenarios, justifying the effectiveness of

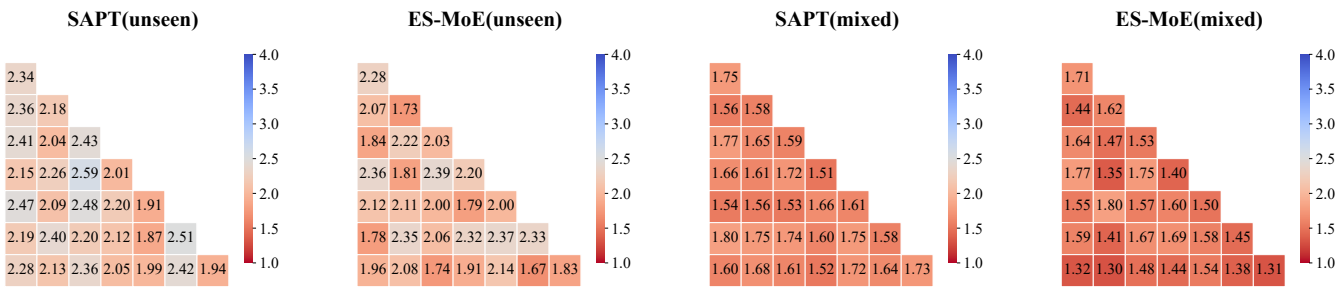

Figure 3: Detailed AF results of four approaches on L^2-EMG (Unseen), L^2-EMG (Mixed), where bluer represents severe forgetting of generation ability across different motion scenarios, indicating low performance; redder represents strong retention of generation ability across different motion scenarios, indicating high performance.

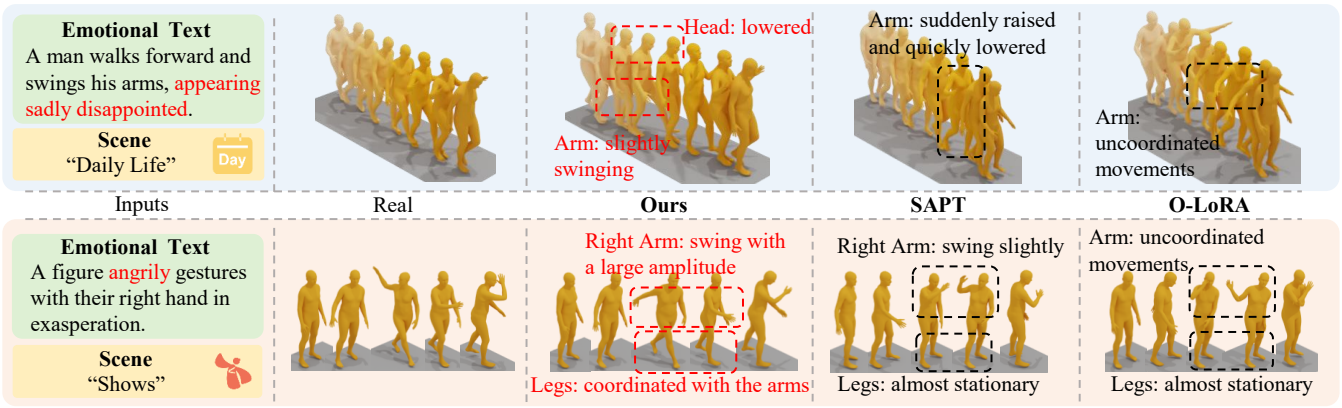

Figure 4: Visualization of motions generated by ES-MoE and other baselines. Red boxes indicate more precise emotional expressions achieved by ES-MoE, whereas black boxes indicate the limitations observed in motions generated by the baselines.

SAMoE in addressing the scenario adapting challenge.

Forgetting Analysis across Different Scenarios. We visualize the FID performance of SAPT (best baseline) and ES-MoE on Unseen and Mixed L^2-EMG datasets in Figure 3. Each matrix entry (i, j) denotes the FID score on the j-th scenario after training on the i-th scenario. Results show: 1) Final-scenario tests show no clear performance drop—sometimes improvement—indicating minimal forgetting in cross-scenario lifelong learning. 2) SAPT and ES-MoE achieve effective knowledge transfer, with ES-MoE consistently yielding lower FID, reflecting more accurate emotion decoupling and stronger scenario adaptability.

Qualitative Analysis via Visualization. Figure 4 shows visual comparisons of motions generated by ES-MoE and other methods. We observe: 1) All methods capture basic patterns such as walking and arm swinging, but O-LoRA often produces incoherent or uncoordinated motions—for example, in the first case, it swings both arms while walking, causing unnatural movement. 2) SAPT shows limited emotional understanding. In the second case, although it generates a coherent waving motion, the amplitude is too small to convey the intended angry emotion. In contrast, ES-MoE produces coordinated motions that accurately express the target emotion and match the scenario.

Conclusion

In this paper, we propose a new and challenging **L**LM-Centric **L**ifelong **E**mpathic **M**otion **G**eneration (L^2-EMG) task aimed at enhancing the lifelong learning ability of existing motion generation models in unseen scenarios. To address the L^2-EMG task, we propose an **E**motion-Transferable and **S**cenario-Adapted **M**ixture **o**f **E**xperts (ES-MoE) approach. The ES-MoE method consists of a causal-guided emotion decoupling block and a scenario-adapted expert constructing block, designed to tackle the challenges of the sustainable transfer of common emotional representations and the non-forgetting of scenario-specific motion characteristics, respectively. To comprehensively evaluate the ES-MoE approach, we construct multiple L^2-EMG datasets. Experimental results on the L^2-EMG dataset demonstrate the superior performance of ES-MoE compared to several state-of-the-art baselines. In our future work, we would like to transfer our ES-MoE approach to other tasks across diverse scenarios, e.g., human-scenario interaction motion generation (Jiang et al. 2024; Li et al. 2024), where scenario continual adaptation remains a key challenge in this task. Additionally, we also would like to use emotional motions as conditional guidance to assist humanoid robot control (Mao et al. 2024), in order to empower these humanoid robots with not only intelligence but also empathetic ability.

Acknowledgments

This work was supported by two NSFC grants, i.e., No. 62576234, No.62376178 and sponsored by CIPS-LMG Huawei Innovation Fund. This work was also supported by Collaborative Innovation Center of Novel Software Technology and Industrialization, and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Chen, C.; Zhang, J.; Lakshmikanth, S. K.; Fang, Y.; Shao, R.; Wetzstein, G.; Fei-Fei, L.; and Adeli, E. 2024. The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Motion. *CoRR*, abs/2412.10523.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of CVPR 2023*.

Feng, Y.; Lin, J.; Dwivedi, S. K.; Sun, Y.; Patel, P.; and Black, M. J. 2023. PoseGPT: Chatting about 3D Human Pose. *CoRR*, abs/2311.18836.

Gao, C.; Chen, K.; Rao, J.; Sun, B.; Liu, R.; Peng, D.; Zhang, Y.; Guo, X.; Yang, J.; and Subrahmanian, V. S. 2024. Higher Layers Need More LoRA Experts. *CoRR*, abs/2402.08562.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of Compositional Animations from Textual Descriptions. In *Proceedings of ICCV 2021*.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. *CoRR*, abs/2312.00063.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *Proceedings of CVPR 2022*.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. In *Proceedings of MM 2020*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of ICLR 2022*.

Isele, D.; and Cosgun, A. 2018. Selective Experience Replay for Lifelong Learning. In *Proceedings of AAAI 2018*.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Comput.*, 3(1): 79–87.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. In *Proceedings of NeurIPS 2023*.

Jiang, N.; He, Z.; Wang, Z.; Li, H.; Chen, Y.; Huang, S.; and Zhu, Y. 2024. Autonomous Character-Scene Interaction Synthesis from Text Instruction. In *Proceedings of SIGGRAPH 2024*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Li, B.; Zhao, Y.; Shi, Z.; and Sheng, L. 2022. DanceFormer: Music Conditioned 3D Dance Generation with Parametric Motion Transformer. In *Proceedings of AAAI 2022*.

Li, H.; Yu, H.; Li, J.; and Wu, J. 2024. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. *CoRR*, abs/2412.18600.

Li, Z.; sdaads; sdasdadas; sdadadad; and adsasdad. 2018. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12): 2935–2947.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. In *Proceedings of NeurIPS 2023*.

Liu, Y.; Xia, Z.; Zhao, M.; Wei, D.; Wang, Y.; Liu, S.; Ju, B.; Fang, G.; Liu, J.; and Song, L. 2023a. Learning Causality-inspired Representation Consistency for Video Anomaly Detection. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*.

Liu, Z.; Wu, S.; Jin, S.; Ji, S.; Liu, Q.; Lu, S.; and Cheng, L. 2023b. Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1): 681–697.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient Episodic Memory for Continual Learning. In *Proceedings of NeurIPS 2017*.

Mao, J.; Zhao, S.; Song, S.; Shi, T.; Ye, J.; Zhang, M.; Geng, H.; Malik, J.; Guizilini, V.; and Wang, Y. 2024. Learning from Massive Human Videos for Universal Humanoid Pose Control. *CoRR*, abs/2412.14172.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning Trajectory Dependencies for Human Motion Prediction. In *Proceedings of ICCV 2019*.

Oord. 2017. Neural Discrete Representation Learning. In *Proceedings of NeurIPS 2017*.

Pearl, J.; sdajnjad; asdad; sadadadad; and asdada. 2018. *The book of why: the new science of cause and effect*.

Peng, B.; Tian, Z.; Liu, S.; Yang, M.; and Jia, J. 2024. Scalable Language Model with Generalized Continual Learning. In *Proceedings of ICLR 2024*.

Razdaibiedina, A.; Mao, Y.; Hou, R.; Khabsa, M.; Lewis, M.; and Almahairi, A. 2023. Progressive Prompts: Continual Learning for Language Models. In *Proceedings of ICLR 2023*.

Saxe, A. M.; sddfs; sadsdsd; sdfsfs; and dsdfsfd. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of ICLR 2014*.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. In *Proceedings of ECCV 2022*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *Proceedings of ICLR 2023*.

Tseng, J.; dsaad; sdsadad; sadadassd; and dasda. 2023. EDGE: Editable Dance Generation From Music. In *Proceedings of CVPR 2023*.

Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X. 2023a. Orthogonal Subspace Learning for Language Model Continual Learning. In *Proceedings of EMNLP 2023*.

Wang, X.; Li, Q.; Yu, D.; Cui, P.; Wang, Z.; and Xu, G. 2022. Causal Disentanglement for Semantics-Aware Intent Learning in Recommendation.

Wang, Z.; Liu, Y.; Ji, T.; Wang, X.; Wu, Y.; Jiang, C.; Chao, Y.; Han, Z.; Wang, L.; Shao, X.; and Zeng, W. 2023b. Rehearsal-free Continual Language Learning via Efficient Parameter Isolation. In *Proceedings of ACL 2023*.

Wong, J. A. H. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of ICML 2015*.

Yang, H.; Su, K.; Zhang, Y.; Chen, J.; Qian, K.; Liu, G.; and Gan, C. 2024. UniMuMo: Unified Text, Music and Motion Generation. *arXiv preprint arXiv:2410.04534*.

Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; and He, Y. 2024a. Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters. In *Proceedings of CVPR 2024*.

Yu, T.; Wang, J.; Wang, J.; Luo, J.; and Zhou, G. 2024b. Towards Emotion-enriched Text-to-Motion Generation via LLM-guided Limb-level Emotion Manipulating. In *Proceedings of ACM MM 2024*.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; Tam, W. L.; Ma, Z.; Xue, Y.; Zhai, J.; Chen, W.; Liu, Z.; Zhang, P.; Dong, Y.; and Tang, J. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *Proceedings of ICLR 2023*.

Zhang, H.; Wang, J.; Luo, J.; Zhang, M.; and Zhou, G. 2025. Boosting LLM's Continual Sentiment Understanding for Low-Resource Languages. *IEEE Transactions on Audio, Speech and Language Processing*, 33.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023a. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. *CoRR*, abs/2301.06052.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024a. MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6): 4115–4128.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023b. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. In *Proceedings ofICCV 2023*.

Zhang, Y.; Huang, D.; Liu, B.; Tang, S.; Lu, Y.; Chen, L.; Bai, L.; Chu, Q.; Yu, N.; and Ouyang, W. 2024b. MotionGPT: Finetuned LLMs Are General-Purpose Motion Generators. In *Proceedings of AAAI 2024*.

Zhao, W.; Wang, S.; Hu, Y.; Zhao, Y.; Qin, B.; Zhang, X.; Yang, Q.; Xu, D.; and Che, W. 2024. SAPT: A Shared Attention Framework for Parameter-Efficient Continual Learning of Large Language Models. In *Proceedings of ACL 2024*.