TOWARDS UNIFIED HUMAN MOTION-LANGUAGE UNDERSTANDING VIA SPARSE INTERPRETABLE CHARACTERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the comprehensive understanding of human motion has been a prominent area of research due to its critical importance in many fields. However, existing methods often prioritize specific downstream tasks and roughly align text and motion features within a CLIP-like framework. This results in a lack of rich semantic information which restricts a more profound comprehension of human motions, ultimately leading to unsatisfactory performance. Therefore, we propose a novel motion-language representation paradigm to enhance the interpretability of motion representations by constructing a universal motion-language space, where both motion and text features are concretely lexicalized, ensuring that each element of features carries specific semantic meaning. Specifically, we introduce a multi-phase strategy mainly comprising Lexical Bottlenecked Masked Language Modeling to enhance the language model's focus on high-entropy words crucial for motion semantics, Contrastive Masked Motion Modeling to strengthen motion feature extraction by capturing spatiotemporal dynamics directly from skeletal motion, Lexical Bottlenecked Masked Motion Modeling to enable the motion model to capture the underlying semantic features of motion for improved cross-modal understanding, and Lexical Contrastive Motion-Language Pretraining to align motion and text lexicon representations, thereby ensuring enhanced cross-modal coherence. Comprehensive analyses and extensive experiments across multiple public datasets demonstrate that our model achieves state-of-the-art performance across various tasks and scenarios.

1 INTRODUCTION

Understanding how we conceptualize and verbalize motion is a critical issue in the study of the human mind and communication. As a long-standing research hotspot, human motion understanding has led to the development of various tasks, including human motion-text retrieval (Petrovich et al., 2023; Yu et al., 2024), motion captioning (Guo et al., 2022b), and action recognition (Mahmood et al., 2019; Punnakkal et al., 2021), all of which have made significant progress in recent years.

Despite significant advancements in specific motion tasks, a considerable gap persists in achieving a nuanced comprehension of human motion, especially concerning fine-grained semantics and behavioral reasoning, since existing methods (Petrovich et al., 2023; Plappert et al., 2018; Zhou et al., 2024b) often lack interpretability in the context of motion understanding. For instance, as depicted in Fig. 1(a), when attempting to retrieve the motion associated with the phrase "Person is walking normally in a circle", the correct motion may be identified; however, the corresponding feature mapping within the unified cross-modal semantic space diverges significantly, failing to align with the semantic keywords essential for human comprehension. This semantic deficiency in the dense representations stems from their sole reliance on contrastive learning (Radford et al., 2021), which enforces cross-modal simply and rough alignment between dense text and motion embeddings primarily through cosine similarity, as illustrated in Fig. 1(b). Moreover, certain elements in text and motion are redundant and irrelevant. Such dense representations will incorporate this redundant information, which adversely affects the fine-grained semantic alignment of the representations. Consequently, it remains both crucial and challenging to further investigate interpretable representations of motion to bridge the understanding gap between language and human actions.

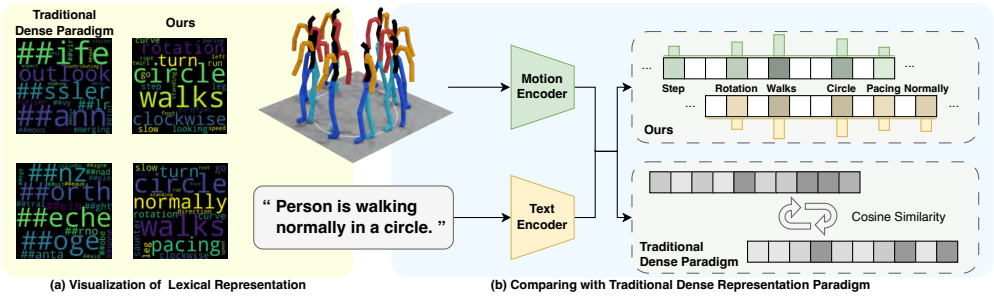


Figure 1: (a) Visualization of lexical representations of our framework and the traditional dense paradigm, and (b) conceptual comparison of our framework and the traditional dense paradigm. The color intensity reflects the higher values along the dimension.

Lexical representation (Formal et al., 2021; Shen et al.; Xiao et al., 2022) integrates interpretability and efficiency by employing vocabulary elements to represent data, with each dimension corresponding to a specific term. The magnitude of each dimension indicates the significance of the associated term, enabling an intuitive understanding of data that aligns with human cognition, while the activation of only relevant vocabulary ensures sparsity and efficiency. In contrast, dense representation encapsulates data in a whole vector without clear dimensional significance, which limits their interpretability and obscures the underlying semantics. Drawing from this concept, we propose leveraging lexical representation to embed both motion and text within a cohesive lexical space, thereby enhancing clarity and fostering a more nuanced and intuitive comprehension of motion.

In this paper, we present a novel human motion-language pre-training framework that incorporates lexical representation to extract aligned sparse representations, thereby improving the interpretability of motion representations for better human motion understanding. Our method employs a multi-phase training strategy consisting of four key phases: i) Lexical Bottlenecked Masked Language Modeling (LexMLM), which enhances the pretrained language model's focus on high-entropy motion-related words for capturing the motion semantics; ii) Contrastive Masked Motion Modeling (CMMM), which improves motion feature extraction by directly capturing spatial and temporal dynamics from skeletal motion; iii) Lexical Bottlenecked Masked Motion Modeling (LexMMM), which enables the motion model to identify the underlying semantic features of motion, facilitating improved cross-modal understanding; and iv) Lexical Contrastive Motion-Language Pretraining (LexCMLP), which aligns motion and text representations within a unified vocabulary space to enhance cross-modal coherence. Through these training phases, our motion and language models effectively prioritize critical motion information and lexicon, yielding aligned and semantically rich sparse lexicon representations. Additionally, we conduct comprehensive visualization analyses of both motion and text across various modalities and perspectives. Extensive experiments on publicly available datasets demonstrate that our proposed method achieves state-of-the-art performance across diverse tasks and scenarios.

Our contributions can be summarized as follows:

- We integrate the lexical representation paradigm into the motion-language representation framework, aligning both motion and text within a shared lexical vocabulary space. This integration significantly enhances interpretability and fosters a more intuitive and comprehensive understanding of human motion.
- We propose a novel pre-training framework that learns aligned and semantically rich sparse lexical features for both motion and text, allowing the respective models to efficiently concentrate on key motion information and core vocabulary.
- We utilize advanced visualization techniques, including feature visualization and word clouds, to conduct thorough analyses of our method across motion, language, and cross-modal dimensions, thereby showcasing its interpretability and effectiveness.
- Comprehensive evaluations on several public datasets reveal that our model achieves state-of-the-art performance, affirming the efficacy of the proposed approach, while also showcasing its robust versatility across a wide range of tasks and scenarios.

2 RELATED WORK

Human Motion Understanding. In the motion-language domain, human motion understanding encompasses key tasks such as motion-text retrieval (Fujiwara et al., 2024b; Athanasiou et al., 2024; Fujiwara et al., 2024a) and motion captioning. **Motion-Text Retrieval** focuses on retrieving relevant motion data based on textual queries. A notable approach in this area is TMR (Petrovich et al., 2023; Bensabath et al., 2024), which constructs a cross-modal embedding space (Yan et al., 2024; 2021; Xu et al., 2024) through CLIP-style contrastive learning and negative filtering. MotionCLIP (Tevet et al., 2022a) further enhances this task by leveraging the image-text joint space. MotionPatch (Yu et al., 2024) utilizes motion patches along with pre-trained ViT weights for effective motion-language modeling. **Motion Captioning** aims to generate descriptive captions for human motions. TM2T (Guo et al., 2022b) encodes motion sequences and employs a translation network to bridge the gap between motion and text tokens. MotionGPT (Jiang et al., 2024) treats human motions as a foreign language, enabling descriptions through an expanded vocabulary.

Lexical Representation. Lexical representation is celebrated for its interpretability and efficiency, making it a popular choice in information retrieval (Formal et al., 2021). This approach highlights the importance of vocabulary in text representation and employs sparsity techniques and inverted indexing systems to enhance retrieval speed and reduce latency. Unlike traditional methods such as TF-IDF and BM25 (Robertson et al., 1995), which lack learning capabilities, neural-model-based lexical weighting retrieval methods (Shen et al.) leverage language models for term-centric searches. These methods fall into two main categories: those using causal language models (CLM)(Radford et al., 2018; 2019; Brown et al., 2020; Nogueira et al., 2019; Touvron et al., 2023; Zhang et al., 2022) and those based on masked language models (MLM)(Devlin, 2018; Liu, 2019; Sanh, 2019). However, a notable limitation is their tendency to fine-tune pre-trained language models directly, often overlooking the mismatch between general language modeling and relevance-driven lexical weighting (Luo et al., 2023; Zhou et al., 2024a; Chen et al., 2023). Moreover, the inherent semantic gap between motion and text presents significant challenges in aligning them within a shared embedding space. To address these challenges, we propose a novel multi-phase training framework specifically designed for motion-language learning, supported by a comprehensive analysis of its effectiveness.

3 METHOD

The primary objective of our methodology is to achieve seamless alignment of motion and text within a unified semantic space that captures the fundamental semantic features of motion, facilitating a comprehensive and in-depth understanding of motion. An illustration of our approach is shown in Fig. 2. We will discuss the critical components of the model architecture in Section 3.1 and the multi-phase training method in Section 3.2.

3.1 MODEL ARCHITECTURE

Lexical Space. We utilize the word embedding space of pre-trained language models (PLMs) space as our interpretable lexical space to encapsulate semantically meaningful relationships between motion and language, aligning seamlessly with human cognition. In contrast to conventional techniques like TF-IDF and BM25 (Robertson et al., 1995), this approach leverages advanced natural language processing methodologies (Devlin, 2018; Radford et al., 2018), including pre-training strategies and models developed on extensive language datasets, thereby significantly enhancing its capacity to capture the nuanced semantic information inherent in real-world human motions.

Text Encoder. We utilize a pre-trained language model (Devlin, 2018) as our text encoder to effectively capture the semantic information from the input text. For an input motion caption of length L, denoted as $\mathbf{X} = [x_1, \ldots, x_L]$, the language model outputs text embedding with a shape of $\mathbb{R}^{L \times C_{\text{e}}}$, where C_{e} represents the dimensionality of the common embedding space.

Motion Encoder. In contrast to existing methods (Jiang et al., 2024; Petrovich et al., 2023) that utilize specifically designed input formats and encoders for motion synthesis tasks Zhang et al. (2023); Petrovich et al. (2022); Guo et al. (2024); Athanasiou et al. (2022) in motion captioning and

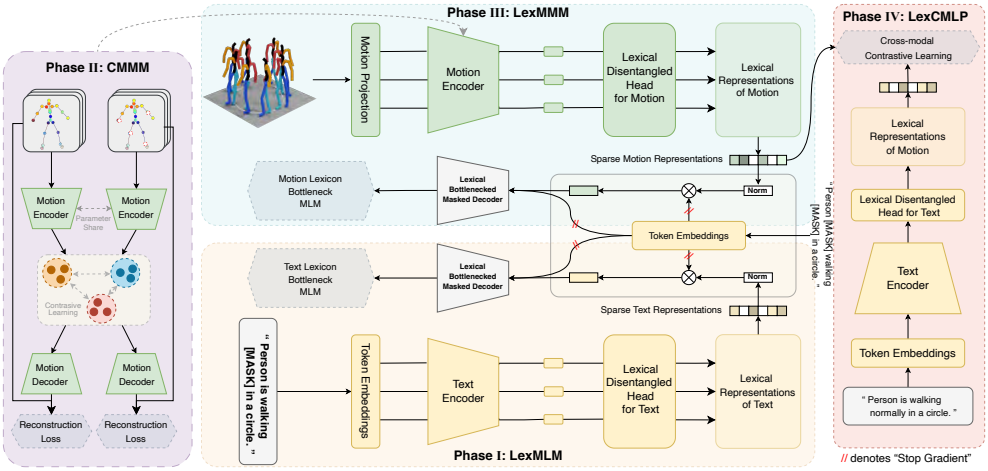


Figure 2: The framework of our method, including i) LexMLM, which enhances the language model's focus on high-entropy motion-related words; ii) CMMM, which captures spatial and temporal dynamics for robust motion representation; iii) LexMMM, which enables the motion model to identify semantic features and improve cross-modal understanding; and iv) LexCMLP, which aligns motion and text within a unified vocabulary space, ensuring cross-modal coherence.

motion-text retrieval, our approach directly employs sequences of skeleton keypoints as inputs and designs a transformer equipped with both spatial and temporal attention to enhance interpretability. Formally, the input 3D motion sequence is represented as $\mathbf{X} \in \mathbb{R}^{T\times J\times C_{\text{in}}}$, where T denotes the number of frames, J indicates the number of joints, and C_{in} specifies the input feature dimensions. We first project this sequence into a latent space $\mathbf{F}^0 \in \mathbb{R}^{T\times J\times C_{\text{f}}}$ and integrate learnable spatial joint position embeddings $\mathbf{P}^{\text{S}}_{\text{pos}} \in \mathbb{R}^{1\times J\times C_{\text{f}}}$ along with temporal position embeddings $\mathbf{P}^{\text{T}}_{\text{pos}} \in \mathbb{R}^{T\times 1\times C_{\text{f}}}$. The features are fed into the transformer to extract spatiotemporal information. Following feature extraction, a linear layer projects these features into the unified embedding space $\mathbf{E} \in \mathbb{R}^{T\times J\times C_{\text{e}}}$.

Lexical Disentanglement Head. The lexical disentanglement head is crafted to transform dense motion and text embeddings into sparse lexical representations. This process begins with the application of a language model head (LM-Head) (Devlin, 2018) space, which converts the motion and text from the embedding space into the lexicon space:

$$\boldsymbol{E}_{lex} = \text{LM-Head}(E_{embed}) \in \mathbb{R}^{|\mathcal{V}|}, \tag{1}$$

where $|\mathcal{V}|$ denotes the vocabulary size, E_{embed} denotes the motion or text embedding and $\boldsymbol{E}_{lex}$ encapsulates the lexical representations corresponding to the input embeddingsE_{embed}.

Subsequently, we adopt a strategy inspired by the SPLADE model (Formal et al., 2021) to reframe the data within a high-dimensional vocabulary framework:

$$p = \log(1 + \text{ReLU}(\text{MaxPool}(\boldsymbol{E}_{lex})) \in \mathbb{R}^{|\mathcal{V}|}, \tag{2}$$

where the ReLU activation ensures non-negativity of all values and provides sparsity, $\text{MaxPool}(\cdot)$ conducts a maximum pooling operation along the sequence dimension and saturation function $\log(1+\cdot)$ is utilized to prevent some terms from dominating the overall representation.

To further encourage sparsity, we follow (Paria et al., 2020) to introduce the FLOPs regularizer such that only a small number of token embeddings in V are non-zeros:

$$\mathcal{R}_{\text{FLOPs}} = \sum_{k\in V}\Big(\frac{1}{N}\sum_{i=1}^{N} p_k^{(i)}\Big)^2 \tag{3}$$

Consequently, the resulting vector p serves as the sparse lexicon representation.

Lexical Bottleneck Masked Decoder. A key challenge in learning the lexicon-importance distribution lies in the absence of direct supervision. Moreover, PLMs (Devlin, 2018; Radford et al.,

2018; 2019) are typically designed to reconstruct the entire text, which does not inherently prioritize motion-related key terms. To address this, we employ a Lexical Bottleneck Masked Decoder that reconstructs the entire text using dense representations derived from the sparse lexical representations. This approach enables the encoder to selectively emphasize motion-relevant tokens rather than treating all tokens with equal importance.

To facilitate this process, we propose a continuous bag-of-words (CBoW) (Shen et al.) derived from the lexicon-importance distribution p. Specifically, we define the bottleneck as follows:

$$E_{cbow} = p \cdot \text{sg}(W^{(te)}) \in \mathbb{R}^{|d|} \tag{4}$$

where $\mathbf{W}^{(\text{te})} = [\mathbf{e}^{(w_1)}, \mathbf{e}^{(w_2)}, \ldots] \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the learnable word embedding matrix from the language model, with d indicating the embedding dim C_e and $\mathbf{e}^{(w_i)} \in \mathbb{R}^d$ representing the word embedding for lexicon w_i, and $\text{sg}(\cdot)$ represents stop-gradient.

Finally, the bottleneck E_{cbow} is fed into a simple decoder to reconstruct the masked tokens.

3.2 Multi-Phase Lexical-Bottlenecked Pretraining

The primary objective of our model is to integrate motion and text within a unified semantic space, accurately reflecting the intricate relationships between these modalities. However, significant challenges arise from the inherent differences between motion and language. Motion encoders often struggle to translate motion into lexical representations, while pre-trained language models (Devlin, 2018; Radford et al., 2018; Achiam et al., 2023), designed for sentence reconstruction, do not prioritize vocabulary relevant to motion. Additionally, relying solely on contrastive learning for modality alignment does not guarantee semantic accuracy. To tackle these challenges and ensure meaningful semantic correspondences, we propose an approach that goes beyond traditional contrastive learning. Our multi-phase pre-training strategy progressively aligns motion and text, preserving the semantic structure and emphasizing the importance of motion-related vocabulary.

3.2.1 Lexical Bottlenecked Masked Language Modeling

Pre-trained language models, typically trained with Masked Language Modeling (Devlin, 2018) or Next Token Prediction (Radford et al., 2018; 2019), focus on recovering words from context. Consequently, they often assign higher scores to low-entropy words, such as articles and prepositions, while neglecting high-entropy words crucial for distinguishing motion content. This highlights the necessity for additional pre-training to adapt PLMs for more effective lexical representation.

To address this challenge, we propose Lexical Bottleneck Masked Language Modeling (LexMLM), which consists of three key components: a Text Encoder, a Lexical Disentanglement Head, and a Lexical Bottleneck Masked Decoder. Following standard practices, we pre-train the text encoder using the MLM objective. Given a masked input $\bar{x}$, where 30% of the tokens are replaced by the [MASK] token, the text encoder generates embeddings and reconstructs the masked tokens. These embeddings are then converted into a sparse, meaningful lexicon representation p by the Lexical Disentanglement Head, with sparsity enforced through a FLOPs regularizer. To further improve lexical representation and capture word importance without explicit labels, we propose the Bottleneck Masked Decoder to reconstruct the heavily masked input $\hat{x}$ using the learned lexicon representation p. We enhance this process with curriculum-masked modeling, gradually increasing the masking ratio from 50% to 100%. The whole LexMLM loss function is as follows:

$$\mathcal{L}_{\text{LexMLM}} = -\lambda_1 \sum_{\mathbb{D}} \sum_{j \in \mathbb{M}^{(\text{enc})}} \log P(\text{o}^j = x_j | \bar{x}) - \lambda_2 \sum_{\mathbb{D}} \sum_{j \in \mathbb{M}^{(\text{dec})}} \log P(\text{o}^j = x_j | \hat{x}; E_{cbow}) + \lambda_3 \mathcal{R}_{\text{FLOPs}}, \tag{5}$$

where $\mathbb{D}$ denotes the whole dataset, $\mathbb{M}^{(\text{enc})}$ and $\mathbb{M}^{(\text{dec})}$ denote the set of masked position in $\bar{x}$ and $\hat{x}$, o^j denotes the logit of x_j, and x_j refers to the original text token, the E_{cbow} is calculated as Eq.4.

3.2.2 Contrastive Masked Motion Modeling

In the motion-language domain, much of the existing research (Guo et al., 2022b; Petrovich et al., 2023; Jiang et al., 2024) has focused on generation using handcrafted features, leaving a gap in the development of motion encoders that can directly capture both the spatial and temporal dynamics

of skeletal motion for human motion understanding. To address this limitation, we propose a pre-training strategy for our motion encoder based on contrastive masked motion modeling (CMMM). In this approach, the motion sequence is processed through two parallel pathways: one with the full sequence X, and the other with a randomly masked portion of the input X_m. Both pathways share the same encoder weights, ensuring consistent feature extraction from both masked and unmasked inputs. To align the representations, we employ an InfoNCE loss (Oord et al., 2018) from contrastive learning, which encourages both pathways to produce similar embeddings, enhancing the encoder's ability to learn robust motion representations. Additionally, each pathway includes a simple decoder head that reconstructs the embeddings back into the motion sequence, $\hat{\mathbf{X}} \in \mathbb{R}^{T \times J \times C_{\text{in}}}$. This reconstruction process ensures that the learned features retain sufficient information to accurately recover the original sequence. By combining contrastive alignment with masked motion reconstruction, the encoder learns to capture robust motion features. The overall CMMM loss consists of reconstruction loss, velocity loss, and contrastive loss. The complete loss function is defined as:

$$\mathcal{L}_{\text{CMMM}} = \lambda_1 \sum_{t=1}^{T} \sum_{j=1}^{J} \parallel \hat{x}_{t,j} - x_{t,j} \parallel_2 + \lambda_2 \sum_{t=1}^{T-1} \sum_{j=1}^{J} \parallel \hat{v}_{t,j} - v_{t,j} \parallel_2 + \lambda_3 L_{info}(e_m, e_w), \tag{6}$$

where the $\hat{v}$ and v mean the velocity of $\hat{x}$ and x,and e_m and e_w are the embedding of x and x_m.

3.2.3 Lexical Bottlenecked Masked Motion Modeling

Despite the effectiveness of CMMM in extracting spatial-temporal motion embeddings, significant challenges persist in converting these embeddings into meaningful lexical representations. Unlike the extraction of lexical representations from text, this task requires a complex cross-modal conversion, where motion data must be translated into relevant textual representations. This process is inherently difficult due to the need to capture the intricate relationships between the dynamics of motion and the semantics of language, often resulting in the misalignment of representations.

To address this, we introduce Lexical Bottlenecked Masked Motion Modeling (LexMMM), composed of a Motion Encoder, a Lexical Disentanglement Head, and a Lexical Bottleneck Masked Decoder. The Motion Encoder extracts spatiotemporal motion embeddings, while the Lexical Disentanglement Head converts them into a sparse lexical representation, with a FLOPs regularizer applied to enforce sparsity. Similar to LexMLM, the Bottleneck Masked Decoder then reconstructs the masked text $\hat{X}$ input using the learned lexicon representation. We enhance the model with curriculum-masked modeling, gradually increasing the mask ratio from 50% to 100%. At full masking, the process can be interpreted as a motion captioning task. The whole LexMMM loss function is as follows:

$$\mathcal{L}_{\text{LexMMM}} = -\lambda_1 \sum_{\mathbb{D}} \sum_{j \in \mathbb{M}^{(\text{dec})}} \log P(\text{o}^j = x_j | \hat{x}; E_{cbow}) + \lambda_2 \mathcal{R}_{\text{FLOPs}}, \tag{7}$$

where $\mathbb{D}$ denotes the whole dataset, $\mathbb{M}^{(\text{dec})}$ denotes the set of masked position in $\hat{x}$, o^j denotes the logit of x_j, and x_j refers to the original text token, the E_{cbow} is calculated as Eq.4.

3.2.4 Lexical Contrastive Motion-Language Pretraining

While the previous training stages effectively established a solid basis for generating semantic lexicon representations from both motion and text encoders, the independent tuning of these components resulted in suboptimal parameter configurations, thereby constraining the in-depth alignment between motion and text representations for comprehensive human motion understanding.

To achieve this, we introduce a novel cross-modal alignment strategy called Lexical Contrastive Motion-Language Pretraining (LexCMLP). This method utilizes motion and text encoder to form contrastive pairs between sparse motion and caption lexicon representations. By applying a contrastive loss, we maximize the similarity between matching motion and text pairs while minimizing it for non-matching pairs. Additionally, to prevent semantic alignment in irrelevant spaces, the Lexical Bottleneck Masked Decoders from both LexMLM and LexMMM are retained as the lightly weighted auxiliary losses L^1_{aux} and L^2_{aux}. This fine-tuning approach strengthens the alignment of lexical representations across modalities, ensuring that the motion and text lexicons are effectively synchronized within the shared vocabulary space. The whole LexCMLP loss function is as follows:

$$\mathcal{L}_{\text{LexMMM}} = \lambda_1 L_{info}(p_m, p_t) + \lambda_2 \mathcal{R}_{\text{FLOPs}} + \lambda_3 L^1_{aux} + \lambda_4 L^2_{aux}, \tag{8}$$

where p_m and p_t represent the lexical representations of motion and text.

Setting	Methods	Text to motion retrieval						Motion to text retrieval					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
Dense	TEMOS	2.12	4.09	5.87	8.26	13.52	173.0	3.86	4.54	6.94	9.38	14.00	183.25
	T2M	1.80	3.42	4.79	7.12	12.47	81.00	2.92	3.74	6.00	8.36	12.95	81.50
	TMR	8.92	12.04	16.33	22.06	33.37	25.00	9.44	11.84	16.90	22.92	32.21	26.00
	MotionPatch	10.80	14.98	20.00	26.72	38.02	19.00	11.25	13.86	19.98	26.86	37.40	20.50
Lexicon	† TMR	7.83	10.42	15.04	20.93	31.94	26.50	8.68	10.32	15.68	21.37	30.91	27.50
	† MotionPatch	9.13	12.86	16.78	23.83	34.71	22.50	10.03	11.89	17.13	23.44	33.38	24.50
	Ours	**11.80**	**17.11**	**23.25**	**30.81**	**43.36**	**14.00**	**12.39**	**15.55**	**22.17**	**29.25**	**40.34**	**17.00**

Table 1: Results on the motion-text retrieval benchmark on HumanML3D. The symbol † indicates that the lexicon representation is used directly in place of the dense embedding.

Setting	Methods	Text to motion retrieval						Motion to text retrieval					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
Dense	TEMOS	7.11	13.25	17.59	24.10	35.66	24.00	11.69	15.30	20.12	26.63	36.39	26.50
	T2M	3.37	6.99	10.84	16.87	27.71	28.00	4.94	6.51	10.72	16.14	25.30	28.50
	TMR	10.05	13.87	20.74	30.03	44.66	14.00	11.83	13.74	22.14	29.39	38.55	16.00
	MotionPatch	14.02	21.08	28.91	34.10	50.00	10.50	13.61	17.26	27.54	33.33	44.77	13.00
Lexicon	† TMR	9.87	12.13	19.64	28.19	42.16	15.50	10.62	11.18	20.07	27.13	36.51	18.00
	† MotionPatch	10.82	18.48	26.38	31.02	46.51	12.50	11.53	15.11	24.92	30.18	40.52	15.00
	Ours	**15.13**	**23.74**	**31.61**	**36.81**	**54.12**	**8.00**	**15.01**	**19.47**	**30.06**	**35.63**	**47.53**	**10.50**

Table 2: Results on the motion-text retrieval benchmark on KIT-ML. The symbol † indicates that the lexicon representation is used directly in place of the dense embedding.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. To validate the effectiveness of our sparse lexical representations, we conduct experiments on two commonly used public datasets: the HumanML3D (Guo et al., 2022a) dataset and the KIT Motion-Language dataset (Plappert et al., 2016). The HumanML3D dataset extends the AMASS (Mahmood et al., 2019) and HumanAct12 (Guo et al., 2020) motion capture datasets by adding natural language annotations, comprising 23,384 motions for training, 1,460 for validation, and 4,380 for testing. The KIT-ML dataset is focused primarily on locomotion, derived from motion capture data, 4,888 motions for training, 300 for validation, and 830 for testing.

Evaluation Metrics. We validate the representation alignment using the motion-text retrieval task and assess the semantics through the motion captioning task. To evaluate the alignment performance, we follow previous methods (Petrovich et al., 2023) and use standard metrics, including Recall at various ranks (R@1, R@5, etc.), and the median rank (MedR) for both text-to-motion (t2m) and motion-to-text (m2t) retrieval tasks. For evaluating semantic capability, we employ the same linguistic metrics as previous works (Jiang et al., 2024) including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERTScore (Zhang et al., 2019).

Implementation Details. We utilize pretrained BERT (Devlin, 2018) as our text encoder and implement a transformer (Vaswani et al., 2017) with spatial and temporal attention mechanisms for the motion encoder. Our experiments employ the Adam optimizer (Kingma & Ba, 2014), with learning rates set to 10^{-5} for the text encoder, 10^{-4} for the motion encoder, and 10^{-3} for the Lexical Disentanglement Head and Lexical Bottleneck Masked Decoder. During the LexMLM phase, we train with a batch size of 128 for 50 epochs. In the CMMM phase, we use a batch size of 64 and train for 200 epochs. For the LexMMM phase, we freeze the lexical space and fine-tune the motion encoder to align with the language domain, using a batch size of 64 for 150 epochs. Finally, in the LexCMLP phase, we use a batch size of 64 and train for 20 epochs at a learning rate of 10^{-5}.

4.2 EXPERIMENTAL RESULTS

Motion-Text Retrieval Results. To evaluate the alignment between our text and motion lexical representations, we conduct motion-text retrieval benchmarks on the HumanML3D and KIT-ML

Methods	HumanML3D					KIT-ML				
	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	Bert Score↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	Bert Score↑
TM2T	48.9	7.00	38.1	16.8	32.2	35.1	6.2	28.7	28.9	30.4
MotionGPT	48.2	12.47	37.4	29.2	32.4	-	-	-	-	-
Ours	**49.7**	**13.62**	**39.2**	**53.1**	**33.1**	**43.4**	**8.9**	**35.2**	**65.3**	**31.2**

Table 3: Results on motion-to-text captioning benchmarks on HumanML3D and KIT-ML.

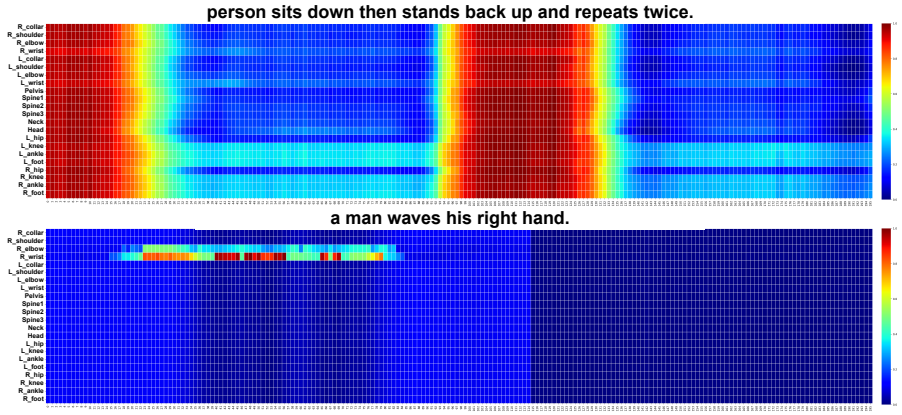


Figure 3: The PCA visualization of the spatiotemporal features extracted by our motion encoder.

datasets. We compare our proposed method with previous works, including TEMOS (Petrovich et al., 2022), T2M (Guo et al., 2022a), TMR (Petrovich et al., 2023), and MotionPatch (Yu et al., 2024). Additionally, we implement modified versions of TMR and MotionPatch by substituting their dense embeddings with lexicon representations. All evaluation metrics are consistent with those used in (Yu et al., 2024) to ensure a fair comparison. As illustrated in Tables 1 and 2, simply replacing dense embeddings with lexicon representations results in inferior performance, particularly for MotionPatch, indicating a significant loss of effectiveness from the pre-trained ViT (Dosovitskiy et al., 2020). In contrast, our method consistently outperforms existing approaches across all metrics in both datasets, demonstrating the effectiveness of our lexical representations and methods.

Motion Captioning Results. To evaluate the semantic capturing ability of our model, we conduct motion captioning benchmarks on the HumanML3D and KIT-ML datasets. We compare our approach with recent methods, including TM2T (Guo et al., 2022b) and MotionGPT (Jiang et al., 2024). As illustrated in Table 3, our method outperforms these contemporary techniques in generating text descriptions for specified motions. Notably, it excels in the CIDEr metric, which measures the model's effectiveness in capturing key information.

4.3 Ablation Studies

Ablation Studies on Multi-phase Training. We conducted ablation studies on the KIT-ML to evaluate the effectiveness of our multi-phase pertaining approach. We implemented a CLIP-style baseline similar to prior works (Petrovich et al., 2023), incorporating a motion encoder, a text encoder, and contrastive loss. Additionally, we replaced the motion decoder with language decoder for motion captioning. We progressively introduced different training phases to elucidate their contributions. The results reveal that the pre-training phases, CMMM and LexMLM, significantly enhance motion-text retrieval performance by improving the capabilities of both encoders. The LexMMM phase, which enriches the cross-modal understanding of the motion encoder, leads to notable gains in motion captioning metrics. The final alignment phase further elevates overall model performance. Detailed tuning for each phase is provided in the appendix.

Method	R@1	R@5	Cider
Baseline	7.32	24.73	0.2
+ LexMLM	8.82	26.17	0.8
+ CMMM	12.98	32.73	0.9
+ LexMMM	14.51	35.68	61.8
+ LexCMLP	15.07	36.22	65.3

Table 4: Ablation experiments on different pre-training phases.

Ablation Studies on Sparsity. Top-K Sparsifying (Shen et al.; Formal et al., 2021) adjusts the sparsity of lexicon-weighted representations, striking a balance between efficiency and effectiveness by retaining only the top-k weighted lexicons while setting others to zero. Applied exclusively during inference, this method introduces no additional training overhead. Fig. 4 illustrates the storage and retrieval performance across different sparsity levels on the KIT-ML dataset, where our model demonstrates superior storage efficiency and retrieval performance compared to previous approaches.

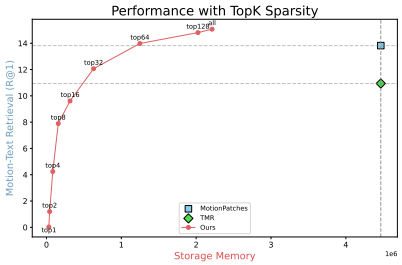
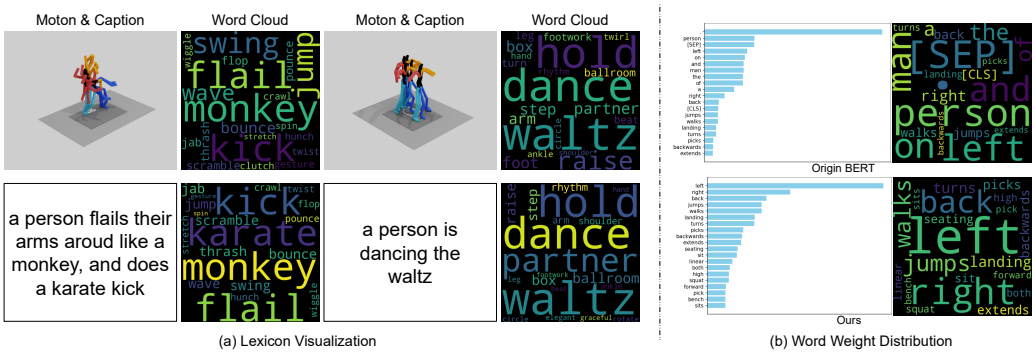


Figure 4: Results on KIT-ML dataset over different sparsity.



(a) Lexicon Visualization (b) Word Weight Distribution

Figure 5: (a)Visualization of lexical representations of input motion and caption, and (b) the distribution of high-importance words extracted by the original BERT and ours.

4.4 Interpretable Qualitative Analysis

Lexical Representation Visualization. We visualize the lexical representations of motion and text in Fig.5(a) using wordcloud(Oesper et al., 2011), where larger words represent higher lexical values. These visualizations show that the learned lexical representations align well with the motion semantics such as "monkey", "kick", "flail" and "dance waltz", highlighting the effectiveness of our method. Furthermore, the model accurately associates relevant words beyond those found in brief captions. For example, when inputting the "dance waltz" motion, our model learns classic movements like "hold" and related terms such as "partner", "ballroom" and "box step", reflecting its ability to efficiently explain the input motion by capturing comprehensive and contextually rich lexical information from both motion and text.

Motion Features PCA Visualization. We present the results of the principal component analysis (PCA) conducted on the motion spatial-temporal features extracted by our model. In this visualization, darker colors indicate stronger weights allocated by the model. As shown in Fig. 3, for the continuous local action "wave right hand", our model accurately maintains its focus on the right elbow and wrist. The subsequent frames are filled with blank spaces. In contrast, for the repetitive full-body motions "sit down" and "stand up", our model effectively captures this repetitive pattern and continuously attends to body parts such as the knee, ankle, and foot to a certain extent. This effective identification of relevant joints and dynamic emphasis aligns well with motion semantics, demonstrating that our model accurately extracts semantically relevant spatial-temporal motion features. Additional intuitive video demonstrations are provided in the supplementary materials.

Word Weight Distribution Visualization. Fig. 5(b) illustrates the frequency statistics of significant vocabulary before and after pre-training. Our text model demonstrates a stronger emphasis on motion-related vocabulary and relevant adjectives. In contrast, the original BERT (Devlin, 2018) tends to focus on low-entropy words such as "the", "of" and "a". The presence of terms like "person" and "man" highlights inconsistencies in motion caption styles within the dataset, including variations like "a person action", "the man action", "a figure action" and simply "action". As a result, the model also fixates on unnecessary details, such as punctuation variations in some sentences. Furthermore, the frequency distribution reveals imbalances, notably the significant disparity between "left" and "right". These underscore potential quality issues in the dataset and provide insights for future analysis and improvements of the dataset and methods.

5 Conclusion

In this paper, we introduce the lexical representation paradigm to the motion-language domain, mapping motions and texts into a shared vocabulary space. We propose a novel multi-phase pre-training framework that efficiently learns aligned, semantically correct sparse lexicon representation for both modalities, which significantly enhances interpretability and depth in human motion understanding. Comprehensive analyses and extensive experiments on multiple public datasets, demonstrate that our model achieves state-of-the-art performance across various tasks and scenarios.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pp. 414–423. IEEE, 2022.

Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing, 2024.

Léore Bensabath, Mathis Petrovich, and Gul Varol. A cross-dataset study for text-based 3d human motion retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1932–1940, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1855–1866, 2024.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander Toshev, Jonathon Shlens, Ruoming Pang, and Yinfei Yang. Stair: learning sparse text and image representation in grounded tokens. *arXiv preprint arXiv:2301.13081*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292, 2021.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. *arXiv preprint arXiv:2407.15408*, 2024a.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. *arXiv preprint arXiv:2407.15408*, 2024b.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pp. 580–597. Springer, 2022b.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pp. 1–21, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. In *Forty-first International Conference on Machine Learning*.

Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11206–11217, 2023.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.

Layla Oesper, Daniele Merico, Ruth Isserlin, and Gary D Bader. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, 6:1–4, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Biswajit Paria, Chih-Kuan Yeh, Ian EH Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. Minimizing flops to learn efficient sparse representations. *arXiv preprint arXiv:2004.05665*, 2020.

Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pp. 480–497. Springer, 2022.

Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9488–9497, 2023.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, and Daxin Jiang. Lexmae: Lexicon-bottlenecked pretraining for large-scale retrieval. In *The Eleventh International Conference on Learning Representations*.

Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. Lgtm: Local-to-global text-driven human motion diffusion model. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–9, 2024.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pp. 358–374. Springer, 2022a.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. *arXiv preprint arXiv:2205.12035*, 2022.

Chenghao Xu, Lyu Guangtao, Yan Jiexi, Yang Muli, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In *Proceedings of the 62st Annual Meeting of the Association for Computational Linguistics*, pp. 14734–14751, 2024.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023.

Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12465–12474, 2021.

Jiexi Yan, Cheng Deng, Heng Huang, and Wei Liu. Causality-invariant interactive mining for cross-modal similarity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3d human motion-language models with motion patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 937–946, 2024.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, and Lei Chen. Retrieval-based disentangled representation learning with natural language supervision. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366, 2024b.

Figure 6: Visualization of lexical representations of input motion and caption.

A More Results

A.0.1 Lexical Representation Visualization

We visualize the lexical representations of motion and text in Fig.6 using wordcloud, where larger words represent higher lexical values. These visualizations show that the learned lexical representations align well with the motion semantics such as “chicken”, “jump”, “swing” and “wave”, highlighting the effectiveness of our method. Furthermore, the model accurately associates relevant words beyond those found in brief captions, reflecting its ability to efficiently explain the input motion by capturing comprehensive and rich lexical information from both motion and text.

A.0.2 Motion Features PCA Visualization

We present the results of the principal component analysis (PCA) conducted on the motion spatial-temporal features extracted by our model in Fig.7. In this visualization, darker colors indicate stronger weights allocated by the model. Our model is able to focus on the correct body parts and track the motion changes along the temporal dimension. This effective identification of relevant joints and dynamic emphasis aligns well with the motion semantics, demonstrating that our model accurately captures semantically relevant spatial-temporal motion features.

A.0.3 The Results of Text to Motion Retrieval

We present the results of text-to-motion retrieval in Fig.9. Our method effectively identifies keywords such as “monkey” and “scratch” retrieving results that are more closely aligned with the target motion. Although TMR and MotionPatches retrieve similar actions, the motions still exhibit noticeable differences compared to the “monkey” actions. In Fig.10, we further illustrate the results of retrieval using multiple keywords, which may better reflect typical search practices. Our method ensures that the retrieved motions are simultaneously related to both “run” and “jump” while ex-

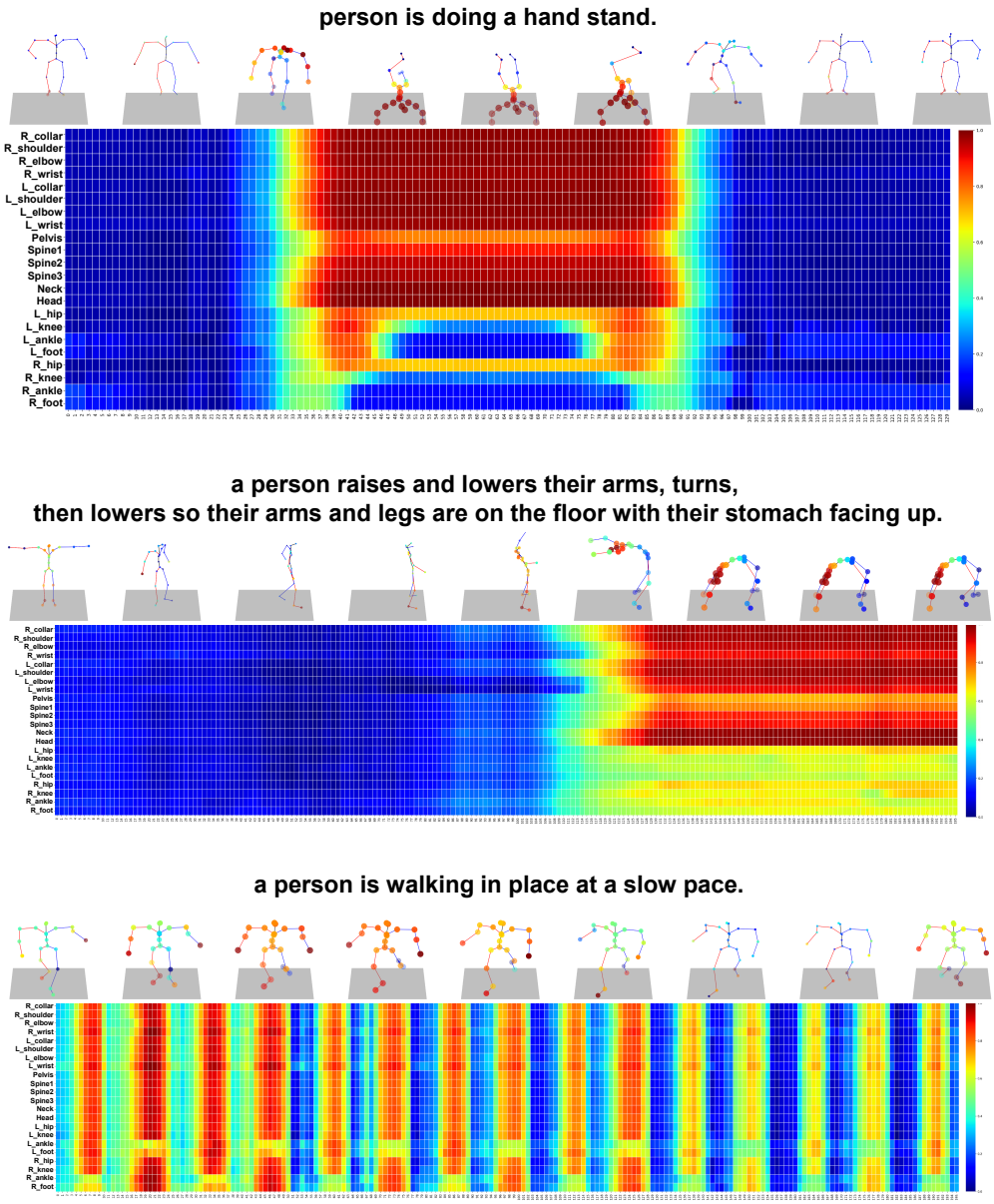


Figure 7: The PCA visualization of the spatiotemporal features extracted by our motion encoder.

isting methods tend to miss one of the keywords. These results demonstrate the effectiveness of our approach, especially in handling keyword-based retrieval. In addition, our model also supports non-parametric indexing, that is, without using a text encoder, the weight of the keyword is directly set to 1. It can also be combined with techniques such as sparse storage and inverted indexing to further improve the efficiency and effectiveness of retrieval.

A.0.4 The Results of Motion Captioning

We present the results of motion captioning in Fig.8. The captions generated by our method exhibit greater distinction, capturing more detailed and relevant keywords. For instance, in the dataset, swimming is a common action, but our approach effectively identifies specific details like “butterfly stroke” and “standing”. Similarly, for waltz, in addition to the general term, our method highlights unique movements such as the “box step”. This showcases the model’s ability to capture critical information, further demonstrating the effectiveness of our approach.

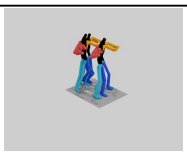
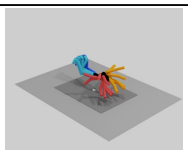
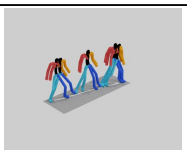
Input Motions		
GT:	a person does a **swimming** motion while **standing**.	a man does a **box step waltz backwards**
TM2T:	a person is stretch their arm and then stretch their arm	a person is dance the **waltz**
MotionGPT:	the person is doing arm exercises.	a person is doing the **waltz** dance
Ours:	a person does a **swimming** exercise while **standing**.	a man dances a **box step waltz backwards**
Input Motions		
GT:	the person is preforming a swimming stroke know as the **butterfly stroke**. the arms swing from behind the head and reenter the water propelling the person forward.	a man side-skips from left to right repeatedly
TM2T:	a person is swim in place	a person walk sideways to the left and then to the right
MotionGPT:	the man swims the breast stroke.	a man **sideskip** from **left to right**
Ours:	the person swims the **butterfly stroke**.	a man **side-skips** from **left to right repeatedly**

Figure 8: The motion captioning results. The red words highlight the keywords.

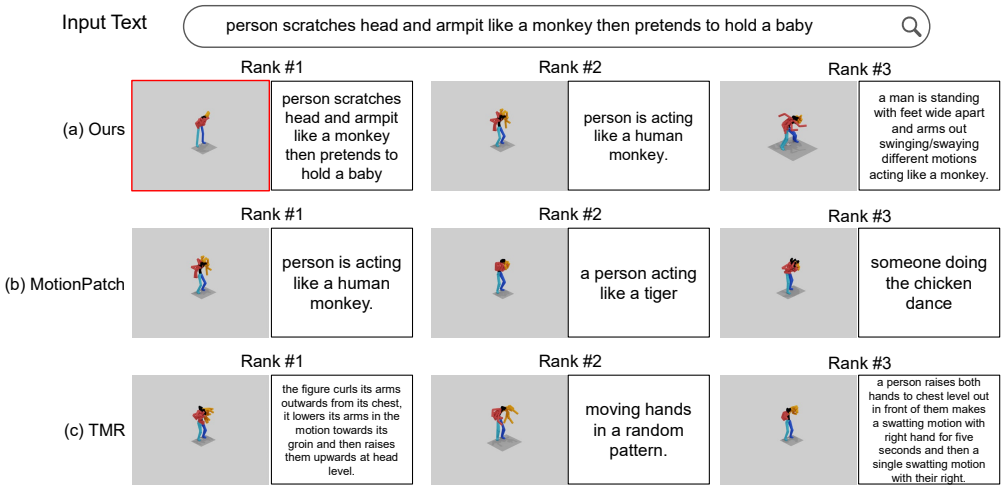


Figure 9: The text to motion retrieval results. The red box means the right sample.

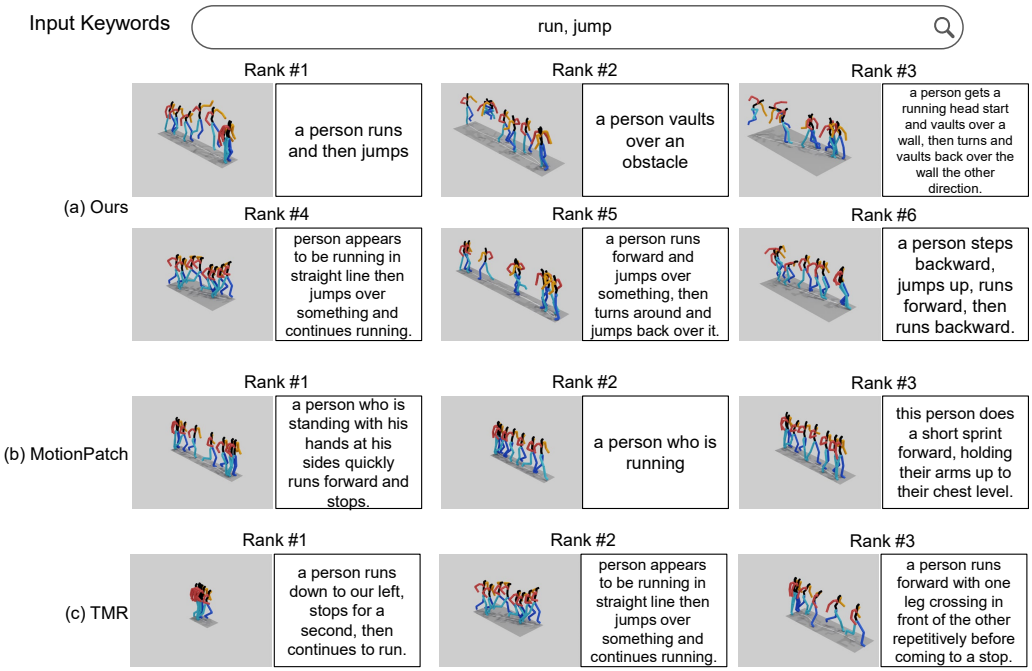


Figure 10: The keywords to motion retrieval results.

B Detailed Ablation Studies

B.0.1 Lexical Bottlenecked Masked Language Modeling

As shown in Table 5, we conducted ablation experiments on the hyperparameters of LexMLM loss using the KIT-ML dataset. The results highlight the importance of MLM pretraining for the text encoder and the significance of pretraining the Lexical Bottleneck Masked Decoder. Furthermore, assigning a larger weight to λ_2 improves performance, demonstrating that lexical bottlenecked masked modeling is more effective for extracting key terms. Overall, the ablation study illustrates the effectiveness of the different components in LexMLM.

λ_1	λ_2	λ_3	R@1	R@5	Cider
1	0	1e-4	7.91	24.87	0.2
0	1	1e-4	7.62	24.79	0.2
1	1	1e-4	8.13	25.31	0.4
0.1	1	1e-4	8.37	25.81	0.5
0.1	1	1e-5	8.31	25.66	0.5

Table 5: Ablation experiments of the hyperparameters of LexMLM on KIT-ML.

As shown in Table 6, we conducted ablation experiments on different mask ratios of the LexMLM loss using the KIT-ML dataset. Increasing the mask ratio in the encoder yields better results, while a higher mask ratio in the decoder further improves performance. This reflects the strength of both standard Masked Language Modeling and our Lexical Bottleneck Masked Language Modeling approach. However, directly setting the mask ratio to 100% harms the model's effectiveness. To address this, we adopted a curriculum learning strategy, gradually increasing the mask ratio from 50% to 100%. These results demonstrate the impact of varying mask ratios and the effectiveness of our method.

$\mathbb{M}^{(enc)}$	$\mathbb{M}^{(dec)}$	R@1	R@5	Cider
15%	15%	8.31	25.66	0.5
15%	30%	8.35	25.71	0.5
30%	30%	8.42	25.79	0.5
30%	50%	8.55	25.88	0.6
50%	50%	8.31	25.69	0.6
30%	70%	8.73	25.95	0.7
30%	100%	8.38	25.76	0.5
30%	50-100%	8.82	26.17	0.8

Table 6: Ablation experiments of the mask proportion of LexMLM on KIT-ML.

B.0.2 Contrastive Masked Motion Modeling

As shown in Table 7, we conducted ablation experiments on different mask ratios and the weights of the contrastive loss in CMMM using the KIT-ML dataset. For the reconstruction loss, we used both MSE loss and velocity loss, each with a weight of 1. Our primary focus is on demonstrating the impact of different mask ratios and contrastive loss weights. The results indicate that increasing the mask ratio appropriately improves performance. Moreover, incorporating the contrastive loss further enhances the retrieval capability of the model. These ablation studies highlight the effectiveness of both masked motion modeling and contrastive motion modeling, validating the overall performance of CMMM. of both masked motion modeling and contrastive motion modeling, validating the overall performance of CMMM.

mask	λ_3	R@1	R@5	Cider
0%	0	10.86	29.17	0.8
15%	0	11.48	29.93	0.8
30%	0	11.73	30.53	0.8
50%	0	11.97	31.02	0.8
60%	0	11.86	30.83	0.8
70%	0	11.81	30.81	0.8
50%	1	12.62	32.03	0.9
50%	0.1	12.98	32.73	0.9
50%	0.01	12.58	32.26	0.9

Table 7: Ablation experiments of CMMM on KIT-ML.

B.0.3 LEXICAL BOTTLENECKED MASKED MOTION MODELING

As shown in Table 8, we conducted ablation experiments on different mask ratios in LexMMM using the KIT-ML dataset. Increasing the mask ratio appropriately enhances the model's capability. We also applied curriculum learning, gradually raising the mask ratio from 40% to 100%. Through cross-modal pretraining with LexMMM, our method significantly improves the model's semantic understanding, helping it extract the correct keywords. This leads to improved retrieval performance. Overall, the experiments demonstrate the effectiveness of our LexMMM pretraining, particularly in enhancing semantic understanding.

$\mathbb{M}^{(\text{dec})}$	R@1	R@5	Cider
15%	13.85	34.71	49.5
30%	14.02	35.13	54.7
40%	14.36	35.39	57.3
50%	14.27	35.31	57.9
60%	14.12	35.15	57.1
70%	13.92	34.95	56.8
100%	13.57	34.12	51.9
40-100%	14.51	35.68	61.8

Table 8: Ablation experiments of the mask proportion of LexMMM on KIT-ML.

B.0.4 LEXICAL CONTRASTIVE MOTION-LANGUAGE PRETRAINING

As shown in Table 9, we conducted ablation experiments on different hyperparameters in LexCMLP using the KIT-ML dataset. A smaller weight was assigned to preserve the MLM loss obtained from the Lexical Bottleneck Masked Decoder, forcing the model to output accurate semantic information, which enhances its understanding and representational capacity. The experimental results demonstrate that our LexCMLP further improves the alignment of lexical representations and the model's ability to accurately grasp semantic information, proving the effectiveness of our method.

λ_1	λ_2	λ_3	λ_4	R@1	R@5	Cider
1	1e-4	1e-2	0	14.62	35.79	62.4
1	1e-4	0	1e-2	14.82	35.96	63.7
1	1e-4	1e-2	1e-2	15.07	36.22	65.3
1	1e-4	1e-3	1e-3	15.01	36.13	65.1

Table 9: Ablation experiments of the hyperparameters of LexCMLP on KIT-ML.

Text Encoder	Parameters	Text to motion retrieval						Motion to text retrieval					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
T5-Small	80M	4.40	7.14	10.02	14.19	21.25	60.00	5.66	6.66	9.88	13.85	19.79	68.00
T5-Base	250M	4.95	7.62	10.31	14.81	23.62	46.00	5.59	6.98	10.59	14.49	20.99	54.00
T5-Large	780M	5.82	8.72	11.66	16.93	26.27	37.00	6.69	8.33	12.16	16.70	23.53	45.00
T5-XL	3B	7.39	11.00	14.65	20.65	31.29	28.00	7.94	10.06	15.08	19.88	29.05	33.00
T5-XXL	11B	8.41	12.82	15.96	23.67	34.42	24.00	8.97	11.49	16.69	22.63	32.16	28.00
DistilBERT	66M	10.80	14.98	20.00	26.72	38.02	19.00	11.25	13.86	19.98	26.86	37.40	20.50

Table 10: Ablation Study of the Motivation on the Motion-Text Retrieval Benchmark Using the HumanML3D Dataset. While performance improves with the increasing size of T5 models, the largest T5-XXL still underperforms compared to the smaller DistilBERT.

C MORE TABLES AND FIGURES

C.0.1 MORE ABLATION STUDIES ABOUT THE MOTIVATION

To further clarify the motivation behind our method, we conducted an ablation study on the the existing state-of-the-art approach, MotionPatch Yu et al. (2024), using a series of text encoders of varying model sizes, ranging from T5-Small to T5-XXL Chung et al. (2024); Raffel et al. (2020). The experimental results are summarized in Table 10. It is evident that increasing the model size of T5 leads to improved performance. However, despite its significantly larger size, the performance of the T5 model remains substantially inferior to that of the original text encoder, DistilBERT Sanh (2019). This indicates that employing a more powerful text encoder does not necessarily enhance the alignment of semantic keywords critical for human understanding. These observations are supported by prior studies. For example, in the ablation study of MotionPatch Yu et al. (2024), a more advanced text encoder failed to deliver improved representational performance. Similarly, TMR Petrovich et al. (2023) advised the use of DistilBERT instead of the larger CLIP series text encoders. Additionally, TMR++ Bensabath et al. (2024) analyzed text annotations across various datasets (HumanML3D, KIT-ML, Babel) and demonstrated that while text augmentations can partially bridge the domain gap, significant differences persist between datasets. Hence, simply adopting a more powerful text encoder does not address the core issue identified in our study. A deeper and more nuanced understanding of the correlation between textual descriptions and motion is required. Consequently, the design of a lexical representation in our paper is both necessary and justified.

C.0.2 EXTENSION TO TEXT-TO-MOTION GENRATION TASK

Our method can be seamlessly integrated into text-to-motion (T2M) generation tasks. Currently, by simply replacing the text encoder in existing methods with our proposed text encoder to extract a more refined and sparse lexical representation, we observe consistent performance improvements across various generative architectures. Comparative experiments were conducted on multiple frameworks, including VAE-based methods (e.g., T2M Guo et al. (2022a)), diffusion-based approaches (e.g., MDM Tevet et al. (2022b)), autoregressive models (e.g., T2M-GPT Zhang et al. (2023)), and non-autoregressive models (e.g., MoMask Guo et al. (2024)). The results, summarized in Table 11, show notable gains, particularly in metrics evaluating motion-text alignment, such as R-precision (Top-1, Top-2, Top-3 accuracy) and Multimodal Distance (MM-Dist). These improvements are primarily attributed to two factors: (1) CLIP-based text encoders, pre-trained on image-text datasets, do not sufficiently focus on motion-related content, and (2) these encoders fail to capture temporal information during the text-image pretraining process. In contrast, our text encoder is specifically designed to identify motion-relevant keywords during pretraining and align them with a motion encoder that captures spatiotemporal motion information, leading to enhanced performance in T2M tasks. This aligns with recent observations in specialized T2M studies, such as HumanTomato Lu et al. and LGTM Sun et al. (2024), which highlight similar limitations of original CLIP text encoders. However, the majority of mainstream T2M research still emphasizes improving generation paradigms to achieve better results.

As shown in Fig. 11, when multiple actions are present in a sentence, the original MoMask often overlooks some actions, resulting in poor performance. This improvement arises because the pretrained CLIP-text encoder does not prioritize motion-related information, and the lack of temporal information in images makes it less sensitive to the sequential nature of actions. In contrast, ours

Paradigms	Methods	FID ↓	Top1 ↑	Top2 ↑	Top3 ↑	MM-Dist ↓
VAE	T2M	$1.087^{\pm.021}$	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$3.347^{\pm.008}$
	T2M§	$0.942^{\pm.009}$	$0.472^{\pm.004}$	$0.653^{\pm.002}$	$0.748^{\pm.003}$	$3.104^{\pm.006}$
Diffusion	MDM	$0.544^{\pm.044}$	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$5.566^{\pm.027}$
	MDM§	$0.524^{\pm.036}$	$0.357^{\pm.004}$	$0.536^{\pm.003}$	$0.643^{\pm.005}$	$5.212^{\pm.021}$
AR	T2M-GPT	$0.141^{\pm.005}$	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$3.121^{\pm.009}$
	T2M-GPT§	$0.133^{\pm.005}$	$0.506^{\pm.004}$	$0.684^{\pm.003}$	$0.781^{\pm.004}$	$3.002^{\pm.006}$
NAR	MoMask	$0.045^{\pm.002}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$2.958^{\pm.008}$
	MoMask§	$\mathbf{0.041}^{\pm.002}$	$\mathbf{0.532}^{\pm.002}$	$\mathbf{0.721}^{\pm.003}$	$\mathbf{0.814}^{\pm.002}$	$\mathbf{2.852}^{\pm.008}$
VAE	T2M	$3.022^{\pm.107}$	$0.361^{\pm.005}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.488^{\pm.028}$
	T2M§	$2.836^{\pm.062}$	$0.372^{\pm.004}$	$0.574^{\pm.004}$	$0.695^{\pm.005}$	$3.235^{\pm.016}$
Diffusion	MDM	$0.497^{\pm.021}$	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$9.191^{\pm.022}$
	MDM§	$0.482^{\pm.009}$	$0.214^{\pm.005}$	$0.319^{\pm.004}$	$0.418^{\pm.005}$	$8.682^{\pm.014}$
AR	T2M-GPT	$0.514^{\pm.029}$	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$3.007^{\pm.023}$
	T2M-GPT§	$0.502^{\pm.016}$	$0.423^{\pm.005}$	$0.641^{\pm.006}$	$0.752^{\pm.006}$	$2.927^{\pm.015}$
NAR	MoMask	$0.204^{\pm.011}$	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$2.779^{\pm.022}$
	MoMask§	$\mathbf{0.186}^{\pm.009}$	$\mathbf{0.441}^{\pm.006}$	$\mathbf{0.668}^{\pm.004}$	$\mathbf{0.792}^{\pm.005}$	$\mathbf{2.693}^{\pm.013}$

Table 11: Evaluation on the Text-to-Motion Generation Benchmarks: HumanML3D dataset (upper section) and KIT-ML dataset (lower section). The symbol § indicates that our text encoder is used to replace their original CLIP-text encoder.

text encoder focuses more on motion keywords, capturing multiple key movements and temporal information, yielding better results. Similarly, the importance of effective conditional encoders is being increasingly recognized in other domains, such as text-to-image generation and music-to-dance synthesis. In summary, simply replacing the original encoder with our method yields measurable performance improvements. This is an initial attempt. The motion-text field is rapidly evolving, with numerous related tasks emerging, such as Text to Human-Object Interaction(HOI) Motion Generation Xu et al. (2023), Text to Human-Scene Interaction (HSI) Motion Generation Cen et al. (2024), Text to Human-Human Interaction (HHI) Motion Generation Liang et al. (2024), and scenarios involving multiple people, controllable actions, long sequences, and variations in skeletal structures. These areas demonstrate significant potential. We aim to explore its application to a broader range of human motion-related tasks in the future.

C.0.3 Complexity Analysis of the Model

Computational Consumption of Model Training. We compare the training time of different models on the HumanML3D dataset using a single A6000 GPU and report the results in Table 12. During the training process, TM2T, MotionGPT, and our proposed method all require pretraining the motion encoder, which involves higher computational costs. Notably, our method ($\tilde{4}$0 hours) is significantly faster than TM2T (4 days) and MotionGPT (over 4 days). This demonstrates that while our multi-phase training strategy introduces additional steps compared to simpler methods, it remains computationally efficient relative to many existing motion-language approaches. In contrast, methods like TMR and MotionPatch leverage a pre-trained motion encoder and only train the alignment phase, resulting in reduced computational demands. To ensure a fair comparison, we replaced the dense heads in these methods with our Lexical Disentanglement Head, denoting them as †TMR and †MotionPatch, respectively. Our evaluation shows that the computational costs of †TMR and †MotionPatch remain nearly identical to the original TMR and MotionPatch implementations. While we attempted to utilize a pre-trained motion encoder in our method to reduce computational demands, this configuration resulted in significantly degraded performance. Therefore, future work will focus on fine-tuning the pre-trained motion encoder with our Lexical Disentanglement Head. This approach has the potential to effectively balance computational efficiency and performance.

Inference Efficiency. It is also important to note that in practical text-to-motion retrieval scenarios, motion features are pre-extracted and stored, meaning retrieval efficiency primarily depends on the text encoder. As shown in Fig. 4, our method supports sparsification techniques that reduce feature dimensions, improving computational efficiency during retrieval. Additionally, as demon-

strated in Fig. 10, our model supports keyword-based searches by directly activating specific feature dimensions with weights set to 1, thereby eliminating the need for a text encoder during retrieval. This capability significantly improves efficiency for targeted queries. When combined with rapid search techniques such as inverted indexing, our approach offers substantial potential for optimizing retrieval speed and scalability.

Methods	Training Epoch	Training Time
T2M	300 + 270 + 300 + 50	$\sim$ 3 days
TMR	500	$\sim$ 20 minutes
MotionPatch	50	$\sim$ 30 minutes
† TMR	500	$\sim$ 20 minutes
† MotionPatch	50	$\sim$ 30 minutes
TM2T	300 + 300	$\sim$ 4 days
MotionGPT	3000	$\gg$ 4 days
Ours	50 + 200 + 150 + 20	$\sim$ 40 hours

Table 12: Comparison of Training Epochs and Time on Motion-to-Text Captioning Benchmarks for HumanML3D. The symbol † indicates that the lexicon representation is directly used instead of a dense embedding. "$\gg$ 4 days" signifies that the training time for MotionGPT greatly exceeds 4 days.

C.0.4 Robustness and Effectiveness Analysis of the Learned Representation

We conducted two experiments on HumanML3D Guo et al. (2022a) and Motion-X Lin et al. (2024) to evaluate the robustness and effectiveness of the learned lexical representations. Since Motion-X includes data overlapping with HumanML3D, we removed the overlapping content to ensure a fair assessment of generalization.

Small Dataset Training, Large Dataset Testing. We trained our model on the smaller HumanML3D dataset and tested it on the larger Motion-X dataset. As shown in Table 14, our method outperforms existing approaches such as TMR and MotionPatch. Motion-X features more diverse scenarios, longer text descriptions, and richer semantics, which present challenges for methods like TMR and MotionPatch that rely on dense embeddings of global features. In contrast, our lexical representation captures keyword-level importance, allowing it to generalize better across diverse data.

Large Dataset Training, Small Dataset Testing. We trained our model on the larger Motion-X dataset and tested it on the smaller HumanML3D dataset. As shown in Table 13, all methods performed better than in the first experiment, underscoring the value of larger datasets. Importantly, our method continued to outperform others, further demonstrating its robustness and effectiveness.

However, unlike the image domain, where data is relatively easy to obtain, the human motion domain faces significant challenges in acquiring motion data. High-quality datasets primarily rely on motion capture (MoCap) systems. Datasets like KIT-ML and HumanML3D are among the largest and highest-quality MoCap datasets currently available. There are also datasets, such as Motion-X, that use algorithms to automatically extract human motion from videos. While this approach significantly expands dataset size, it often suffers from issues like jitter and anomalies. With future advancements in related technologies, it may become easier to extract more accurate human motion data from videos, enabling rapid dataset expansion. In addition, the quality of motion captions requires further improvement. As shown in Fig. 5(b), the distribution of keywords in datasets is imbalanced, and many descriptions are overly simplistic, failing to highlight meaningful differences. Phrases like "a person action" are common, resulting in significantly higher text similarity compared to other domains (e.g., as noted in TMR). While some datasets now include fine-grained annotations, many rely on AI-generated descriptions, such as those created by ChatGPT-like models, raising concerns about quality and consistency. Therefore, it is also necessary to deepen our understanding of human motion to improve dataset quality. Overall, constructing larger and higher-quality datasets remains a critical direction. We hope that our interpretable sparse lexical representations can enhance our understanding of motion, facilitating the creation of more accurate and detailed text annotations. This would improve the quality of datasets and advance human motion understanding and generation tasks.

Methods	Text to motion retrieval						Motion to text retrieval					
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR	0.06	0.11	0.24	0.39	0.51	829.00	0.08	0.16	0.21	0.37	0.56	861.50
MotionPatch	0.12	0.28	0.49	0.78	0.96	768.00	0.14	0.29	0.48	0.82	0.98	781.00
Ours	0.24	0.46	0.82	0.98	1.24	503.00	0.26	0.48	0.86	1.02	1.43	516.00

Table 13: Zero-Shot Text-to-Motion Retrieval Results. Methods are trained on HumanML3D and tested on Motion-X.

Methods	Text to motion retrieval						Motion to text retrieval					
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR	1.62	2.58	3.74	5.71	7.12	213.00	1.81	2.68	3.86	5.82	7.43	224.50
MotionPatch	1.86	2.78	3.92	5.87	7.96	201.00	1.92	2.84	4.12	6.03	7.64	208.50
Ours	2.76	3.82	5.13	7.42	8.76	182.00	2.83	3.96	5.28	7.89	8.98	186.00

Table 14: Zero-Shot Text-to-Motion Retrieval Results. Methods are trained on Motion-X and tested on HumanML3D.

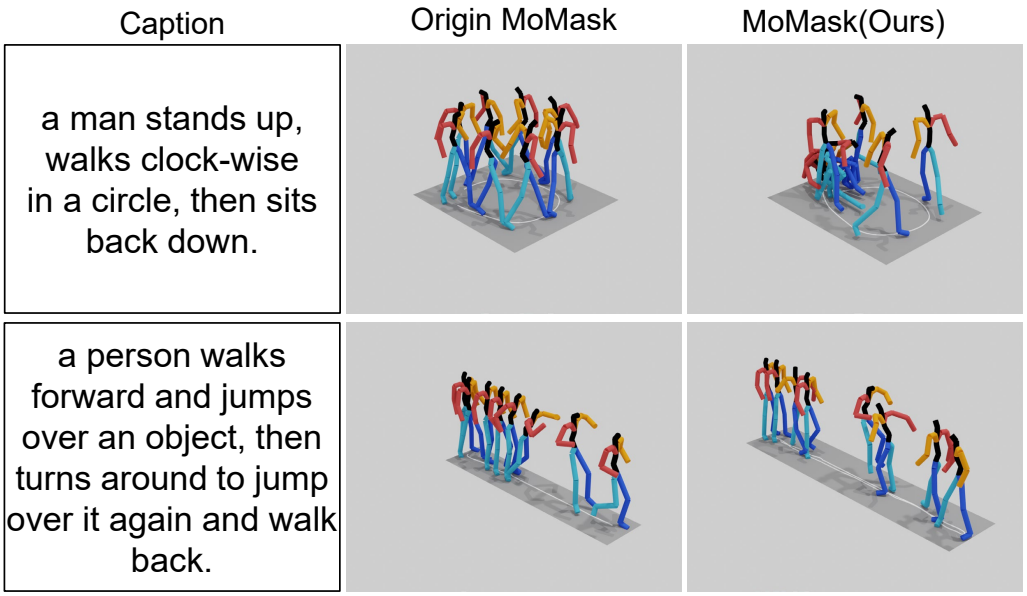


Figure 11: The Text to Motion Generation Results. Our text encoder replaces the original CLIP-text encoder in MoMask. When multiple actions are present in a sentence, the original MoMask often overlooks some actions, resulting in poor performance. This improvement arises because the pre-trained CLIP-text encoder does not prioritize motion-related information, and the lack of temporal information in images makes it less sensitive to the sequential nature of actions. In contrast, ours text encoder focuses more on motion keywords, capturing multiple key movements and temporal information, yielding better results.