Towards Unified Human Motion-Language Understanding via Sparse Interpretable Characterization

Guangtao Lyu[1] Chenghao Xu[1] Jiexi Yan[1] Muli Yang[2] Cheng Deng[1]
[1] Xidian University [2] I2R, A*STAR
{guangtaolyu ,chx}@stu.xidian.edu.cn, {jxyan1995,muliyang.xd,chdeng.xd}@gmail.com

Introduction & Motivation

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.

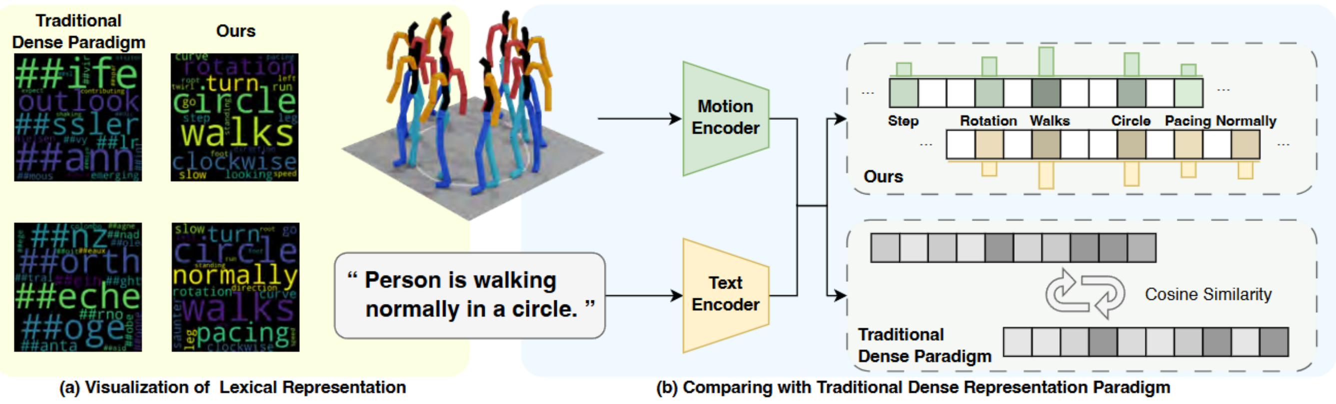

Figure 1: (a) Visualization of lexical representations of our framework and the traditional dense paradigm, and (b) conceptual comparison of our framework and the traditional dense paradigm. The color intensity reflects the higher values along the dimension.

Methods

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 tempordal ynamics 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.

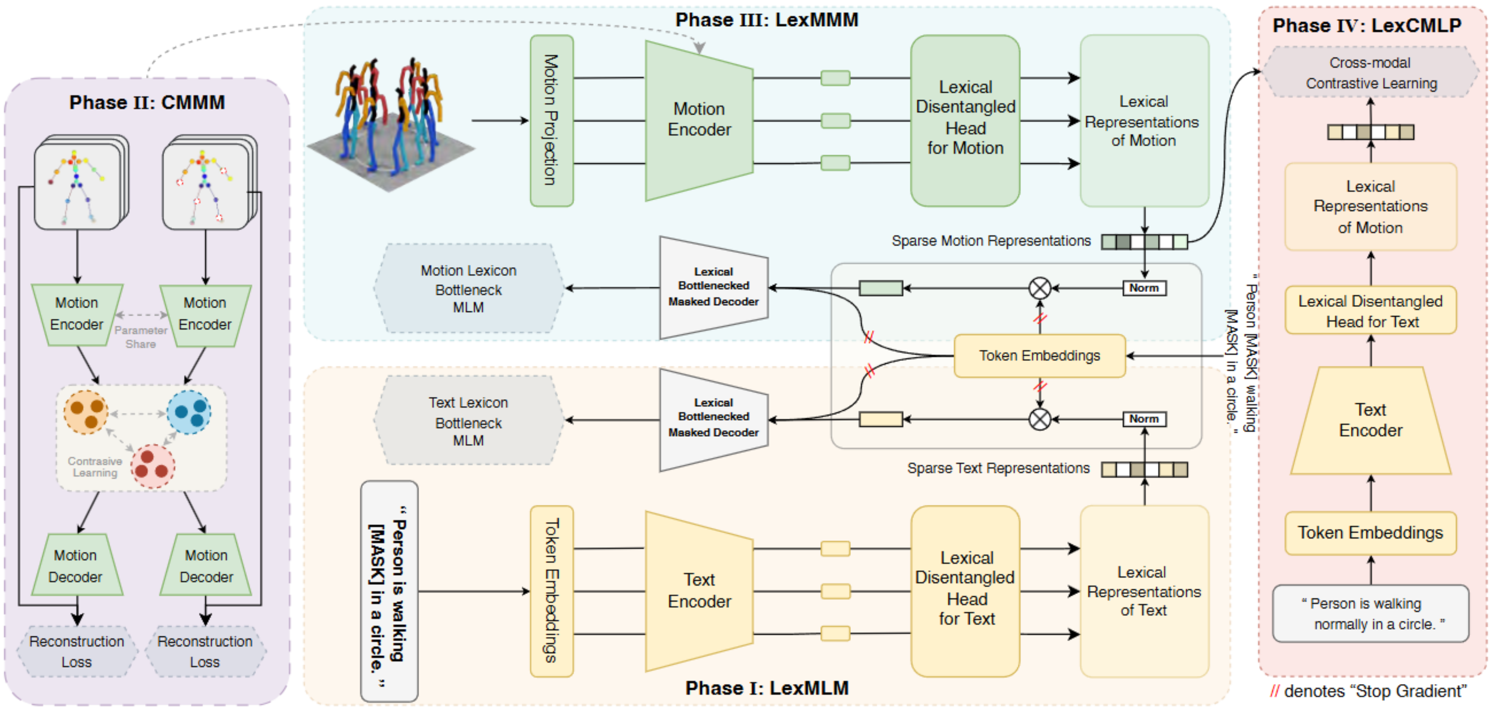

Figure 2: The framework of our method, including i) LexMLM, which enhances the language model's focus on high-entropy motion-related words; ii) CMMM, which captures spatial and temporal dynamics for robust motion representation; iii) LexMMM, which enables the motion model to identify semantic features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

Quantitative Results

Setting	Methods	Text to motion retrieval R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	Motion to text retrieval 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 R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	Motion to text retrieval 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.

Methods	HumanML3D Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	Bert Score↑	KIT-ML 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.

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 4: 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	Parameters	Text to motion retrieval R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	Motion to text retrieval 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 5: Ablation Study of the Motivation on the Motion-Text Retrieval Benchmark.

Interpretability & Visualization

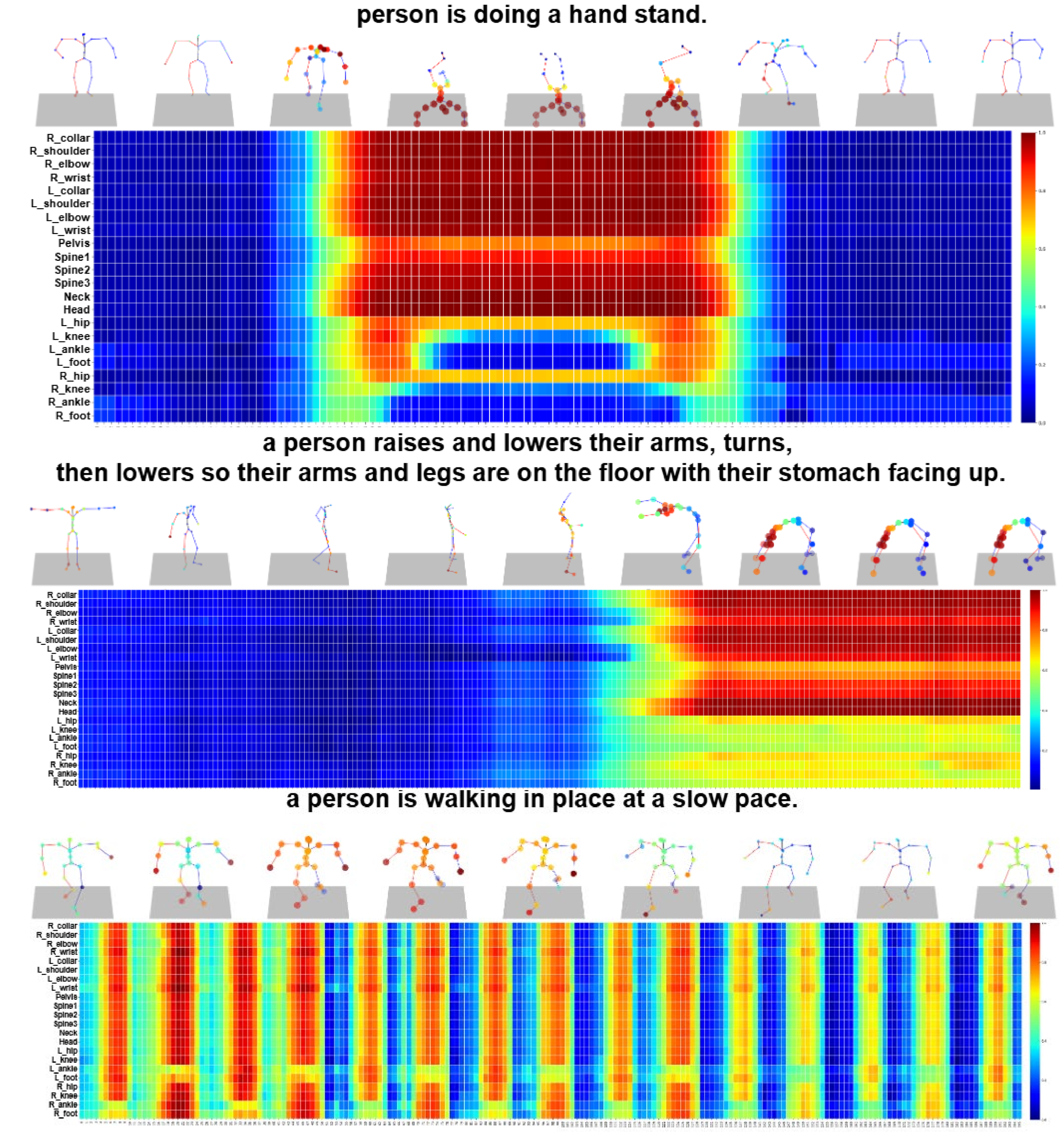

Figure 3: The PCA visualization of the spatiotemporal features extracted by our motion encoder.

Figure 4: Visualization of lexical representations.

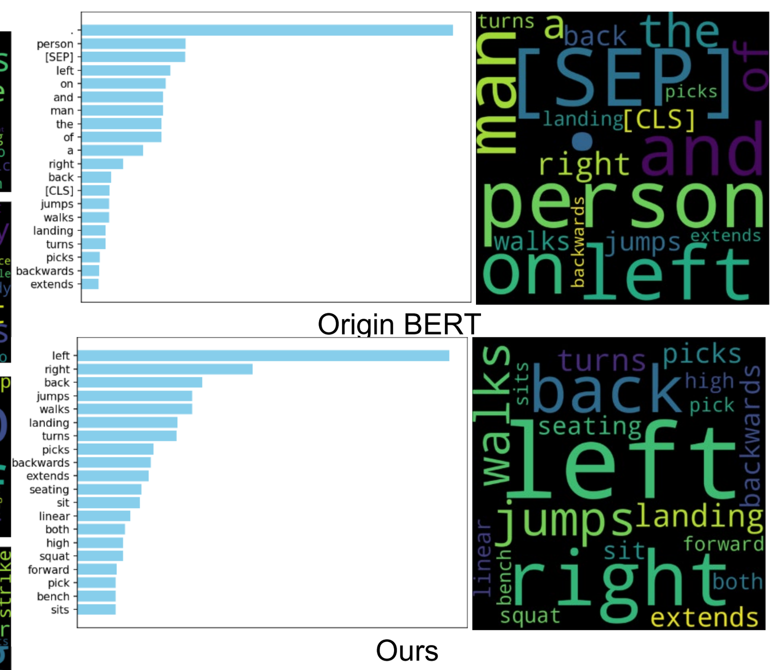

Figure 5: The distribution of high-importance words extracted by the original BERT and ours.

Qualitative Results

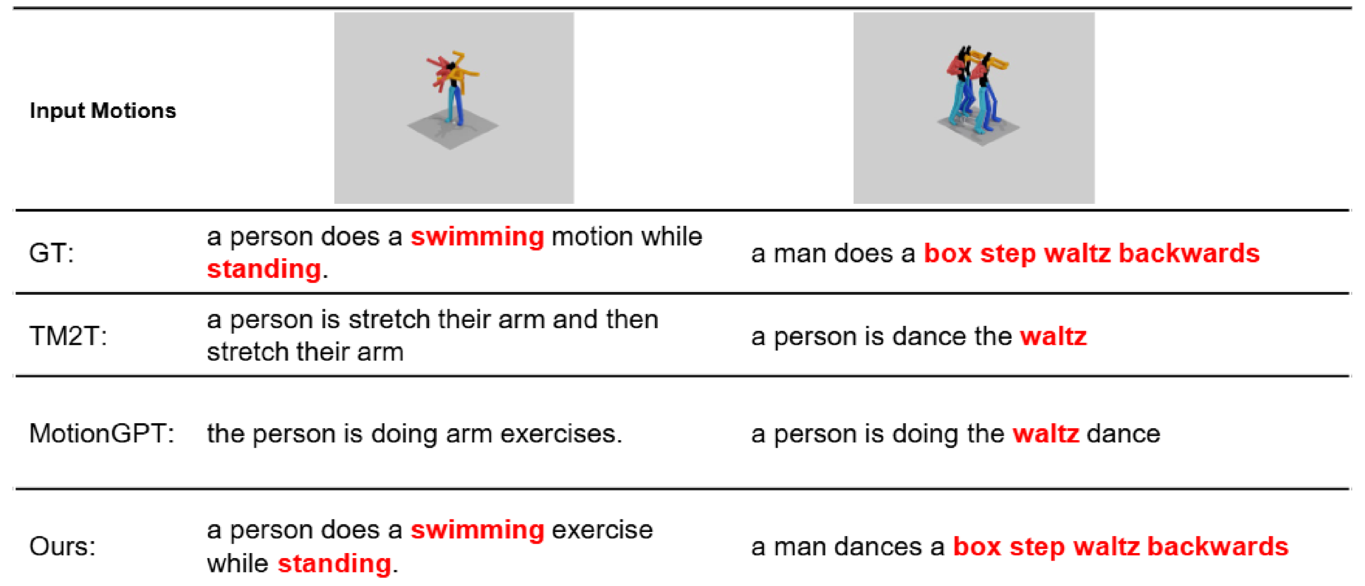

Figure 6: The motion captioning results. The red words highlight the keywords.

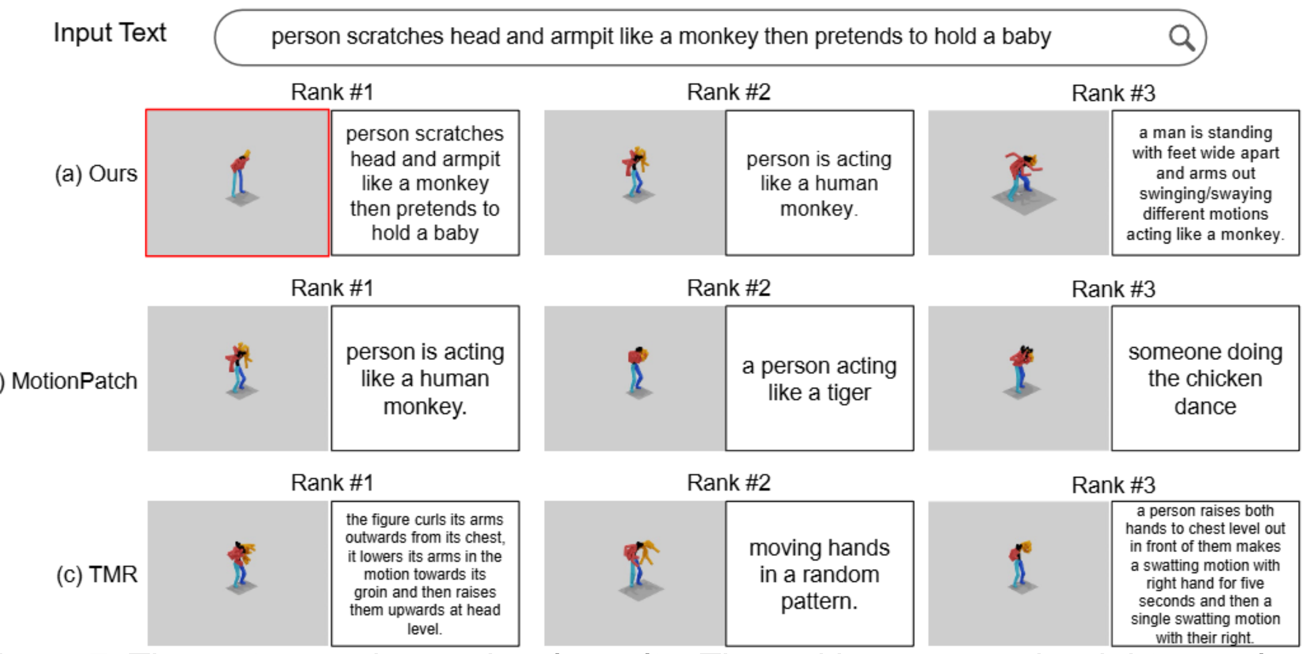

Figure 7: The text to motion retrieval results. The red box means the right sample.

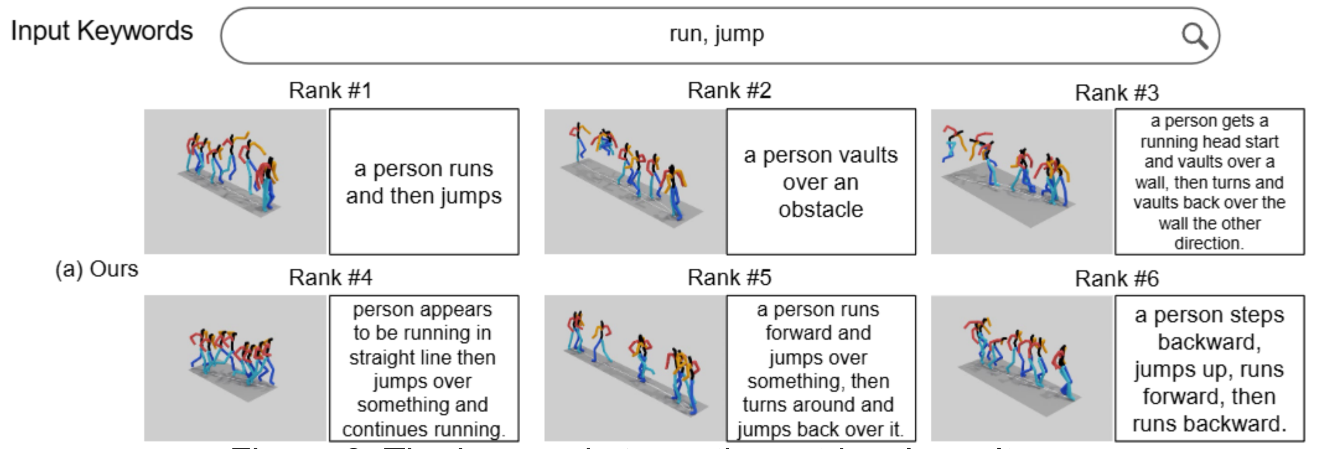

Figure 8: The keywords to motion retrieval results.

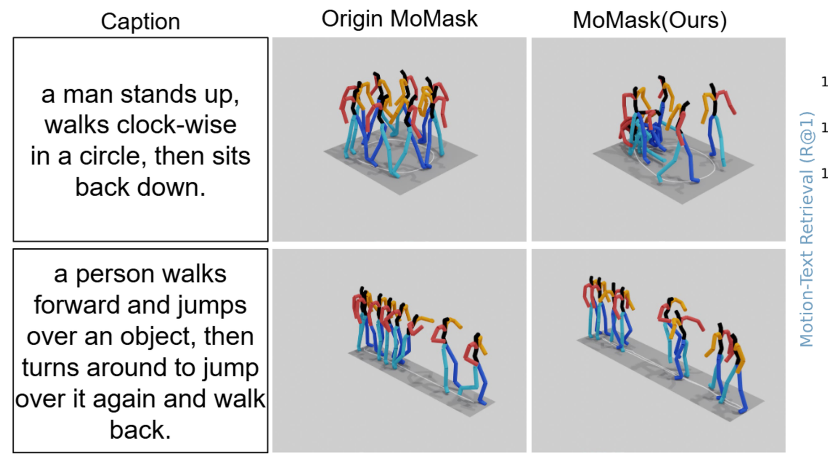

Figure 9: The Text to Motion Generation Results.

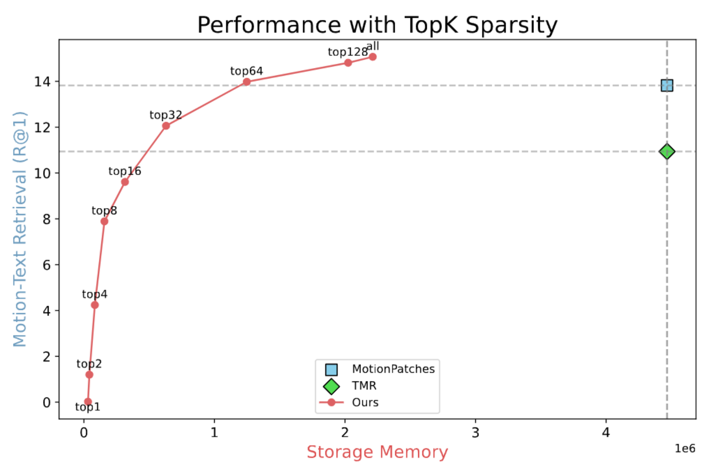

Figure 10: Sparsity for efficient retrieval.