Exploring Vision Transformers for 3D Human Motion-Language Models with Motion Patches

Supplementary Material

A. Motion Patches

In Fig. 3, we show the process of constructing motion patches for SMPL skeletons in the HumanML3D dataset. For the KIT-ML dataset, the skeleton structure is different but the process is the same as shown in Fig. 6. Because the motion patches use the kinematic chain of the skeleton to extract the spatial-temporal information in motion sequences, our model can be used in cross-skeleton recognition as detailed in Section 5.1 of the main paper.

B. Additional Experimental Results

B.1. Visualization of Attention Maps

In this paper, we find that pre-trained image ViT can help the learning of motion data with the proposed motion patches. As shown in Fig. 4, the motion patches can be regarded as a kind of spectrogram, where certain patterns related to motions can be observed. Pre-trained ViT helps detect these patterns, which makes transfer learning work. We additionally visualize the attention maps extracted from the ViT trained by our method in Fig. 7, where the important patterns are activated in the attentions. An analogous approach is audio recognition by rendering the spectrogram of audio as the input into pre-trained image models [12].

B.2. ViT Backbones

We evaluated our method with different ViT backbones. In the main paper, we used ViT-B/16 as the motion encoder. We additionally tried ViT with tiny, small, and large sizes provided in TIMM [2], and the results are shown in Table 8. We can find that ViT-Tiny and ViT-Small perform a little worse when compared to ViT-base in both datasets. The largest model, ViT-Large, performs well in the HumanML3D dataset, but not well in the KIT-ML Dataset, which may be due to the limited scale of the data. Overall, our proposed method works well on all the ViT backbones.

B.3. Motion and Text Encoders

In the paper, we employed the ViT pre-trained on ImageNet as the motion encoder and the pre-trained DistilBERT [44] as the text encoder. Additionally, we explored an alternative approach by utilizing the image encoder and text encoder of CLIP [41] as the motion and text encoders in our method for comparison. The results are shown in Table 9. We can

[2] https://github.com/rwightman/pytorch-image-models

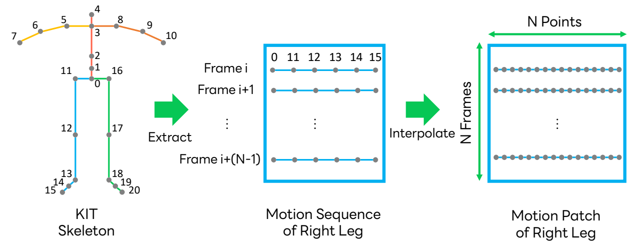

Figure 6. The process of building the motion patches for each motion sequence in KIT-ML. Different body parts are colored in different colors. We show the method to construct the motion patch of the right leg. The same process is applied to other body parts.

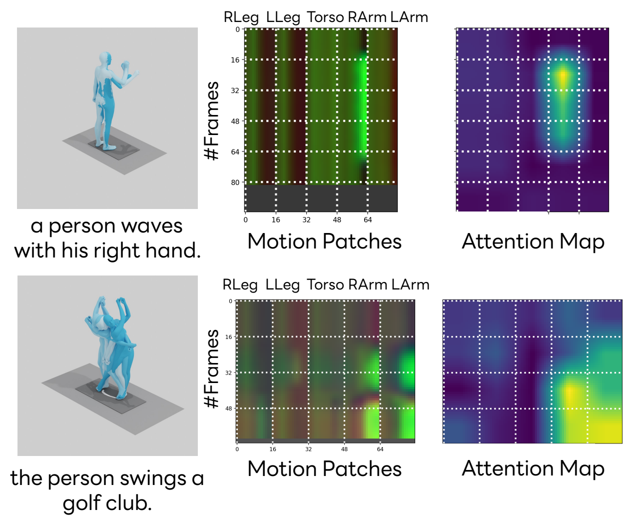

Figure 7. Visualization of attention maps extracted from ViT.

find that the pre-trained weights affect the performance of the model and the combination of ViT with ImageNet and DistilBERT achieved the best results. When the model of CLIP is used as the motion encoder or the text encoder, we find that the performance drops a little, which shows that CLIP is not effective for capturing motion representations. This might be because CLIP is pre-trained to focus on the semantic features of real-world images, while the motion patches resemble a type of spectrogram with color patterns.

B.4. Sizes of Motion Patches

Our investigation explores various motion patch sizes as detailed in Table 10. In addition to the 16×16 motion patches described in the paper, we have implemented our approach

Dataset: HumanML3D

ViT Size	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
Tiny	9.54	23.77	36.10	20.00	10.52	24.00	33.09	24.00
Small	9.63	24.64	36.85	19.00	10.30	24.04	33.77	23.00
Base	**10.80**	26.72	38.02	**18.00**	11.25	**26.86**	37.40	19.50
Large	10.47	**27.29**	**38.84**	19.00	**11.33**	26.82	**37.42**	**19.00**

Dataset: KIT-ML

ViT Size	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
Tiny	11.84	33.38	48.48	11.50	12.53	28.89	40.83	16.00
Small	12.06	33.45	49.00	11.00	13.94	28.80	39.51	17.00
Base	14.02	**34.10**	**50.00**	**10.50**	**13.61**	**33.33**	**44.77**	**13.00**
Large	**14.46**	32.53	42.77	15.00	13.49	28.80	38.31	18.00

Table 8. Results of retrieval with different ViT backbones.

Dataset: HumanML3D

Motion Encoder	Text Encoder	Text-motion retrieval				Motion-text retrieval			
		R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
ViT (ImageNet)	DistilBERT	**10.80**	**26.72**	**38.02**	**18.00**	**11.25**	**26.86**	**37.40**	**19.50**
ViT (ImageNet)	CLIP	9.66	24.12	35.47	21.00	10.37	24.50	34.35	24.00
ViT (CLIP)	DistilBERT	9.85	24.93	36.16	21.00	10.23	24.31	34.03	23.00
ViT (CLIP)	CLIP	6.84	18.57	29.45	32.00	7.82	19.12	27.41	35.00

Dataset: KIT-ML

Motion Encoder	Text Encoder	Text-motion retrieval				Motion-text retrieval			
		R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
ViT (ImageNet)	DistilBERT	**14.02**	**34.10**	**50.00**	**10.50**	**13.61**	**33.33**	**44.77**	**13.00**
ViT (ImageNet)	CLIP	13.01	33.29	49.76	11.00	12.66	31.45	41.45	16.00
ViT (CLIP)	DistilBERT	10.60	32.77	45.54	13.00	12.89	26.63	37.83	18.00
ViT (CLIP)	CLIP	10.48	26.51	36.75	24.00	11.69	23.61	30.48	36.00

Table 9. Results of retrieval with different motion and text encoders.

Dataset: HumanML3D

Patch Size	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
8×8	9.80	26.60	**38.15**	**18.00**	**11.74**	26.05	36.76	**19.00**
16×16	**10.80**	**26.72**	38.02	**18.00**	11.25	**26.86**	**37.40**	19.50
32×32	10.13	26.22	38.00	20.00	10.90	24.88	34.82	22.00

Dataset: KIT-ML

Patch Size	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
8×8	11.57	33.91	**50.84**	**10.00**	12.20	31.83	42.88	15.00
16×16	14.02	**34.10**	50.00	10.50	**13.61**	**33.33**	**44.77**	**13.00**
32×32	**14.34**	33.46	48.31	11.00	12.94	32.48	42.91	14.00

Table 10. Results of retrieval with different patch sizes.

using 8×8 and 32×32 motion patches. Interestingly, both 8×8 and 32×32 patches yielded favorable results. Nevertheless, it is s worth noting that the 16×16 patches consistently delivered the best overall performance.

B.5. Training Datasets

In Section 5.1, we demonstrated the effectiveness of our method in cross-skeleton recognition via zero-shot prediction and transfer learning. We further present the results of training our method using a combination of the HumanML3D and KIT-ML datasets in Table 11. These results indicate that our method can effectively learn from

Dataset: HumanML3D

Training Dataset	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
HumanML3D	**10.80**	26.72	38.02	**18.00**	11.25	**26.86**	**37.40**	**19.50**
Both	9.99	27.22	38.64	**18.00**	**11.37**	25.64	36.16	21.00
Both + FT	10.40	**27.70**	**38.91**	**18.00**	11.11	25.86	36.73	20.00

Dataset: KIT-ML

Training Dataset	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
KIT-ML	14.02	34.10	50.00	10.50	13.61	33.33	44.77	13.00
Both	12.53	35.30	50.96	10.00	13.13	32.28	43.71	14.00
Both + FT	**17.17**	**40.46**	**54.50**	**8.00**	**16.76**	**35.69**	**46.05**	**13.00**

Table 11. Results of retrieval with different training datasets. "Both" represent the combined datasets of the HumanML3D and KIT-ML datasets. "Both + FT" represents the model further fine-tuned on each dataset.

combined datasets and achieve competitive results on both datasets using a single model. This performance is comparable to the results obtained from separate models trained individually on each dataset. If we further fine-tune the model on each dataset, the proposed method can achieve state-of-the-art performance on the KIT-ML dataset.

C. Additional Qualitative Results

In this section, we present qualitative results of the text-to-motion retrieval and motion-to-text retrieval tasks with the comparisons between TMR [38] and the proposed method on the challenging HumanML3D dataset. The results of the text-to-motion retrieval are shown in Fig. 8. We can find that our method succeeded in finding the motion matching the text descriptions including the details, *e.g.*, "ducks" in the first sample and "with right arm up" in the second sample. Regarding the motion-to-text retrieval tasks shown in Fig. 9, each query motion is displayed on the left, and on the right, we showcase the top-5 retrieved text descriptions along with the ground-truth text labels of query motions. We successfully retrieved the ground-truth descriptions in the top-5 results, and the descriptions in the top-5 results seem to be reasonable to describe the motion sequences except for some mirror-augmented ones. When compared to the results of TMR [38], our method is better at catching the details of the motion such as "jumps twice" in the first sample and "moves backward then forwards" in the third sample.

D. Code

The code will be released at `https://github.com/YU1ut/MotionPatches`. We provide the training codes for building the proposed motion-language model and the test codes for text-to-motion retrieval and motion-to-text retrieval with the HumanML3D and KIT-ML datasets. Please refer to the README in the code repository for details.

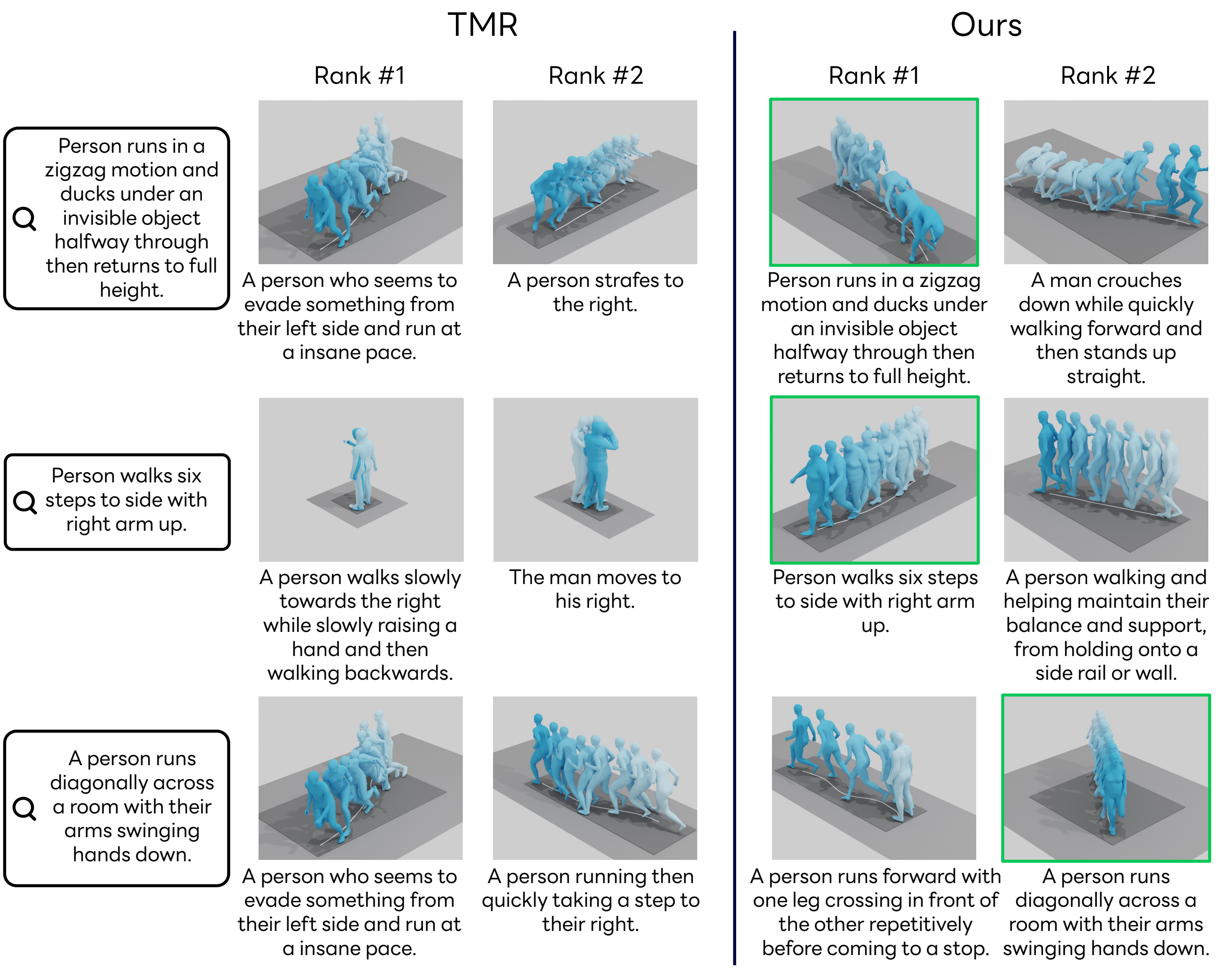

Figure 8. Comparisons of text-to-motion retrieval between TMR [38] and the proposed method. For each query, we show the retrieved motions ranked by text-motion similarity and their accompanying ground-truth text labels. Note that these descriptions are not used in the retrieval process. All motions in the gallery are from the test set and were unseen during training.

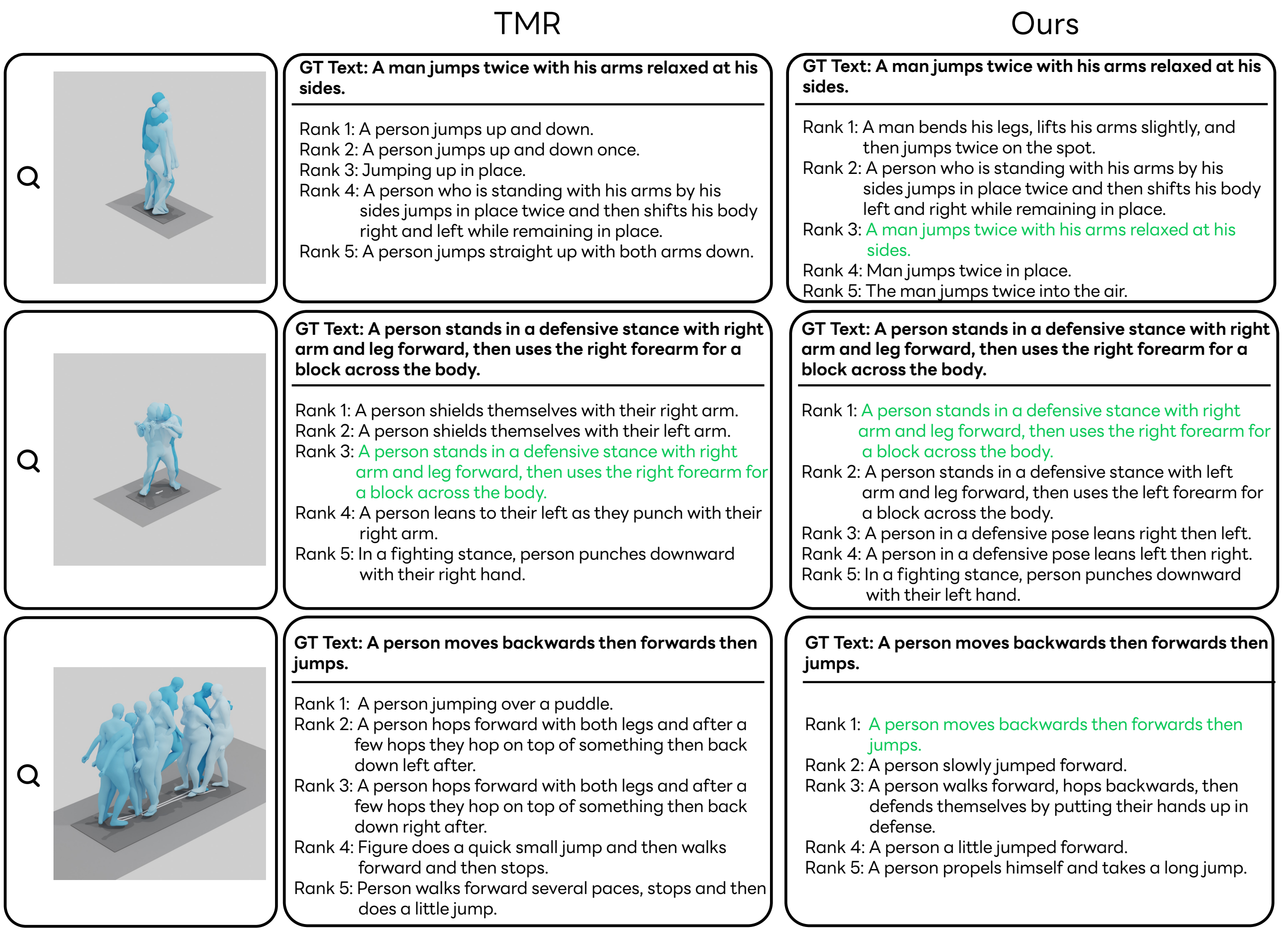

Figure 9. Comparisons of motion-to-text retrieval between TMR [38] and the proposed method. For each query motion, we show the retrieved descriptions ranked by motion-text similarity and their accompanying ground-truth text labels. Note that these ground-truth texts are not used in the retrieval process. All motions in the gallery are from the test set and were unseen during training. For all the samples, our proposed method retrieved reasonable descriptions.