Tri-Modal Motion Retrieval by Learning a Joint Embedding Space

Supplementary Material

1. More Experimental Details

Implementation Details. All experiments are conducted using a single NVIDIA A40 GPU with PyTorch [4]. Our video encoder is based on ViT-B/32 with 12 layers, complemented by a temporal transformer with 6 layers to consider time series information. The motion encoder comprises a 6-layer transformer encoder, while the text encoder is based on DistilBERT [6], supplemented by a temporal transformer to consider the word embedding positions, we finetune DistilBERT during the training process. We follow [7] and initialize DistilBERT and CLIP image encoder trained on the Kinetics-400 dataset. We sample 8 frames from a video sequence, configure the latent dimension C to 512, set ϵ to 0.8 and assign 0.1 to λ_{recon}. Training is conducted with a batch size B of 64 over 400 epochs. During the training process, we use AdamW optimizer [3] with a learning rate being 1e-4 and then linearly decaying to 1e-5 after the first 100 epochs. In the process of augmenting data, an image is first resized randomly, from which a crop measuring 256×256 pixels is extracted. Subsequently, this crop is subjected to a variety of transformations, including jittering of colors randomly, conversion to grayscale on a random basis, application of Gaussian Blur, flipping horizontally in a random manner following the implementation of RandAugment [1]. Our 2-modal version shares the same setting to 3-modal version, with the differences lying in the contrastive learning and modalities fusion between text and motion only.

Evaluation Metrics. Our evaluation of retrieval performance utilizes standard metrics, including recall at various ranks (*e.g.*, R@1, R@2) for both text-motion and video-motion tasks. A higher R-precision value indicates a more accurate retrieval. Additionally, we assess the median rank (MedR) of our results. MedR represents the median ranking position of the ground-truth result, with lower values indicating more precise retrievals. Following TMR [5], the four used evaluation protocols are outlined below: (i) **All** uses the entire test dataset as the retrieval database. However, the precision may be compromised as texts categorized into negative pairs could still convey similar meanings. (ii) **All with threshold** addresses the problem mentioned above, we set the threshold to 0.8, same to the negative filtering threshold. If the similarity between the retrieved motion and the ground-truth motion exceeds this threshold, the result is deemed accurate. (iii) **Dissimilar subset** retrieves motion from a refined subset. The database comprises 100 sampled pairs, with each pair being distinctly dissimilar. Consequently, it's relatively easier to retrieve the correct motion with this protocol compared to the prior ones. (iv) **Small batches** involves randomly selecting batches of 32 motion-text pairs and assessing the average performance.

2. Rendering RGB Videos for KIT-ML Dataset and HumanML3D Dataset

AMASS [2] dataset only contains motion capture data without any RGB videos. In this regard, for each motion capture sequence, an avatar is randomly picked from 13 different avatars shown in Figure 1, animated and rendered to form its corresponding RGB videos of size 512×512, obtained using one of the 4 predefined lightning conditions displayed in Table 1. These 4 lightning conditions represent the positions of top center, left and right top, left and right bottom with different strength of illumination ranging from 0 to 1. Additionally, we adjust the trajectory (global translations of each frame) of each sequence properly to avoid the rendered avatar out of the camera scope. Combining these adjusted global translations with the original SMPL pose and shape annotations, we obtain the new annotations for each sequence in AMASS. Combined with annotated text and SMPL motion, we obtain the associated RGB videos for HumanML3D and KIT datasets.

Lightning	Position	Color
1	$[0, 0, -300]$	$[1.0, 1.0, 1.0]$
2	$[-300, -300, -300]$ $[300, -300, -300]$	$[0.8, 0.8, 0.8]$ $[0.8, 0.8, 0.8]$
3	$[0, 0, -300]$ $[-300, 0, -300]$	$[1, 1, 1]$ $[0.4, 0.4, 0.4]$
4	$[300, 0, -300]$ $[-300, 0, -300]$	$[0.4, 0.4, 0.4]$ $[1, 1, 1]$

Table 1. Four predefined lightning conditions used for rendering in EventAMASS dataset.

3. Additional Experiments

In this section, we show our experiments on hyperparameters search of the contrastive training and ablation study conducted on the KIT-ML dataset using 'All with threshold' evaluation protocol.

Hyperparameters of the Contrastive Training. We conduct ablation studies on four hyper-parameters: the reconstruction weight λ_{recon}, the negative filtering threshold ϵ, the batch size B and the latent dimension C. The purpose

	Modal		DistilBERT		Negative Filtering			Reconstruction		text-to-motion retrieval				motion-to-text retrieval			
	MT	MTV	Frozen	Unfrozen	TMR	Soft Labels	Ours	No	feature fusion	R1↑	R5↑	R10↑	MedR↓	R1↑	R5↑	R10↑	MedR↓
TMR	✓	×	✓	×	✓	×	×	×	×	24.58	50.48	60.36	5.00	19.64	41.20	53.01	9.50
Row2	×	✓	✓	×	✓	×	×	×	✓	27.38	55.45	71.58	4.50	21.98	43.37	57.62	8.00
Row3	×	✓	×	✓	✓	×	×	×	✓	27.93	56.54	71.94	4.50	23.25	44.07	59.55	8.00
Row4	×	✓	×	✓	×	✓	×	×	✓	28.75	58.14	72.98	4.00	23.61	43.86	60.32	8.00
Row5	×	✓	×	✓	×	×	✓	✓	×	13.49	35.97	52.86	11.50	11.48	30.64	48.79	13.25
Row6	×	✓	×	✓	×	×	✓	×	×	28.56	56.17	72.57	4.00	24.21	43.99	61.73	7.50
Ours	×	✓	×	✓	×	×	✓	×	✓	**30.86**	**59.96**	**74.22**	**4.00**	**25.98**	**45.70**	**63.09**	**6.50**

Table 2. Ablation study on our tri-modal framework, frozen DistilBERT, negative filtering technique and motion reconstruction branch.

of the reconstruction loss is to ensure that no information is lost when translating between modalities. In our design of a custom attention mechanism, which aims to enhance alignment between the three modalities, we consider its relative weight in the overall loss to be significant. Upon testing four different values, we found that a weight of 0.1 yields the best performance. The ablation study for the reconstruction weight is shown in table 3. For the negative filtering threshold ϵ, we aim to identify an optimal value that ensures the model effectively aggregates data pairs containing similar information. After testing five different values, we find that a threshold of 0.8 yields the optimal outcome. The ablation study for the negative filtering threshold is shown in table 4. For the batch size B and the latent dim C, we test with four commonly used values respectively, we find that the best value for the batch size and the latent dimension is 64 and 512. The ablation studies are shown in table 5 and table 6.

Ablation Study. We do ablation study on DistilBERT, negative filtering and motion reconstruction branch. Results are presented in table 2.We first compare the difference between frozen or unfrozen DistilBERT during training in Row2 and Row3, we find that unfrozen DistilBERT can improve our performance. The results in Row2 and TMR show that the performance can be improved by our tri-modal framework without extra technique. The results in Row3 and Ours show that our negative filtering technique is better than that of TMR. The results in Row4 and Ours show that our negative filtering technique is better than soft labels, which is a common technique used in contrastive learning community. Results in Row5 and Ours show that motion reconstruction branch is important for the model training. Results in Row6 and Ours show that our proposed feature fusion effectively enhance the retrieval performance.

Re Weight λ_{recon}	**Text-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓	**Video-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓
0.001	18.11	25.16	34.30	13.00	22.41	31.49	46.53	10.00
0.01	27.26	37.78	**48.91**	6.00	34.19	47.31	58.88	4.00
0.1	**30.86**	**41.80**	48.63	**4.00**	**36.91**	**49.80**	**60.94**	**3.00**
1.0	24.15	32.28	41.12	9.00	31.76	43.58	54.29	7.00

Table 3. **Ablation study for the reconstruction weight.** We test four values for the reconstruction weight, the research outcome shows that 0.1 is the most appropriate value.

Threshold ϵ	**Text-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓	**Video-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓
0.70	25.90	37.46	45.79	5.00	31.21	45.37	58.44	5.00
0.75	27.18	40.23	48.21	**4.00**	34.15	47.29	59.71	4.00
0.80	**30.86**	**41.80**	**48.63**	**4.00**	**36.91**	**49.80**	**60.94**	**3.00**
0.85	26.43	38.22	47.81	5.00	33.97	46.38	57.47	5.00
0.90	22.19	32.33	41.48	7.00	29.29	44.62	58.12	6.00

Table 4. **ablation study for the negative filtering threshold.** We test five values for the negative filtering threshold, showing that 0.8 is the most appropriate value.

Batch size B	**Text-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓	**Video-to-motion retrieval** R@1↑	R@2↑	R@3↑	MedR↓
16	25.12	37.69	46.18	6.00	31.35	45.77	55.36	6.00
32	29.17	41.55	**48.98**	**4.00**	36.77	**49.96**	60.74	**3.00**
64	**30.86**	**41.80**	48.63	**4.00**	**36.91**	49.80	**60.94**	**3.00**
128	27.35	39.29	47.38	5.00	34.98	48.58	59.58	4.00

Table 5. **ablation study for the batch size.** We test four values for the batch size, showing that 64 is the most appropriate value.

4. Attention Score

In our approach, we leverage motion as the query to extract relevant information from both text and video modalities. This strategic extraction is quantified by computing the respective weights of text, video, and motion in the final representation, found to be 0.1147, 0.2170, and 0.6684, respectively. This weighting underscores the substantial contribution of each modality to our model. Notably, as videos are rendered and animated from motion sequences, and given that text and motion often exhibit a considerable spatial distance, the weight of the video modality is higher than that of text.

5. Visualization of the Embedding Space

We utilize t-SNE to transform the embeddings of 30 samples and visualize them in figure 2, where different markers represent different modalities. Results show that the three modalities of each sample locates close to each others.

Figure 1. **Front and Back Views of 13 Avatars Used in Video Synthesis.**

Latent dim	Text-to-motion retrieval				Video-to-motion retrieval			
C	R@1↑	R@2↑	R@3↑	MedR↓	R@1↑	R@2↑	R@3↑	MedR↓
128	24.68	38.94	45.53	6.00	31.13	46.25	58.97	5.00
256	29.71	41.34	47.95	**4.00**	34.38	47.83	59.72	4.00
512	**30.86**	**41.80**	**48.63**	**4.00**	**36.91**	**49.80**	**60.94**	**3.00**
1024	29.18	40.94	48.21	4.00	33.96	48.33	59.49	4.00

Table 6. **ablation study for the latent dimension.** We test four values for the latent dimension, showing that 512 is the most appropriate value.

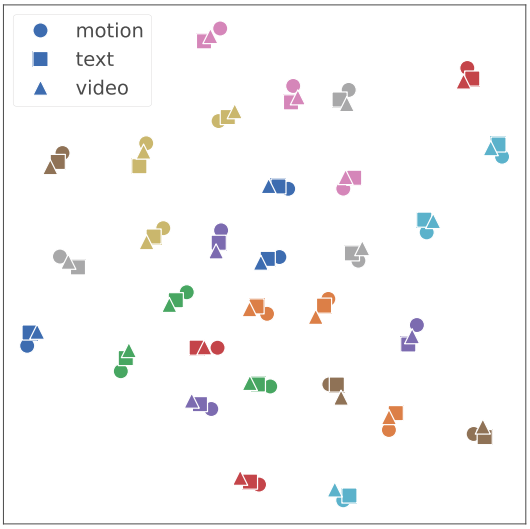

Figure 2. **Visualization of our Joint Embedding Space**

6. Additional Qualitative Results

In Figure 3, we present additional qualitative results. For the text-to-motion retrieval task, four supplementary outcomes are included. The first row features two randomly selected text descriptions from our test dataset. These results demonstrate our model's proficiency in accurately retrieving the corresponding ground-truth motion at the top rank. In the second row, we introduce two text descriptions not found in the existing database. However, these texts contain motions like "hop" and "swim", which can be found in the database. Intriguingly, our model displays its capability to generalize by precisely retrieving motion sequences that match the actions described in these novel texts. In the case of synthetic video-to-motion retrieval, we use four randomly chosen videos as query inputs, showcasing motions such as "picking up", "running around", "kicking", and "swinging". Impressively, for all four queries, our model identifies the correct ground-truth motion as the top result. Lastly, in the real-life video-to-motion retrieval scenario, we input four distinct videos featuring different individuals. Our model exhibits exceptional accuracy in successfully demonstrating the exact motions depicted in each video.

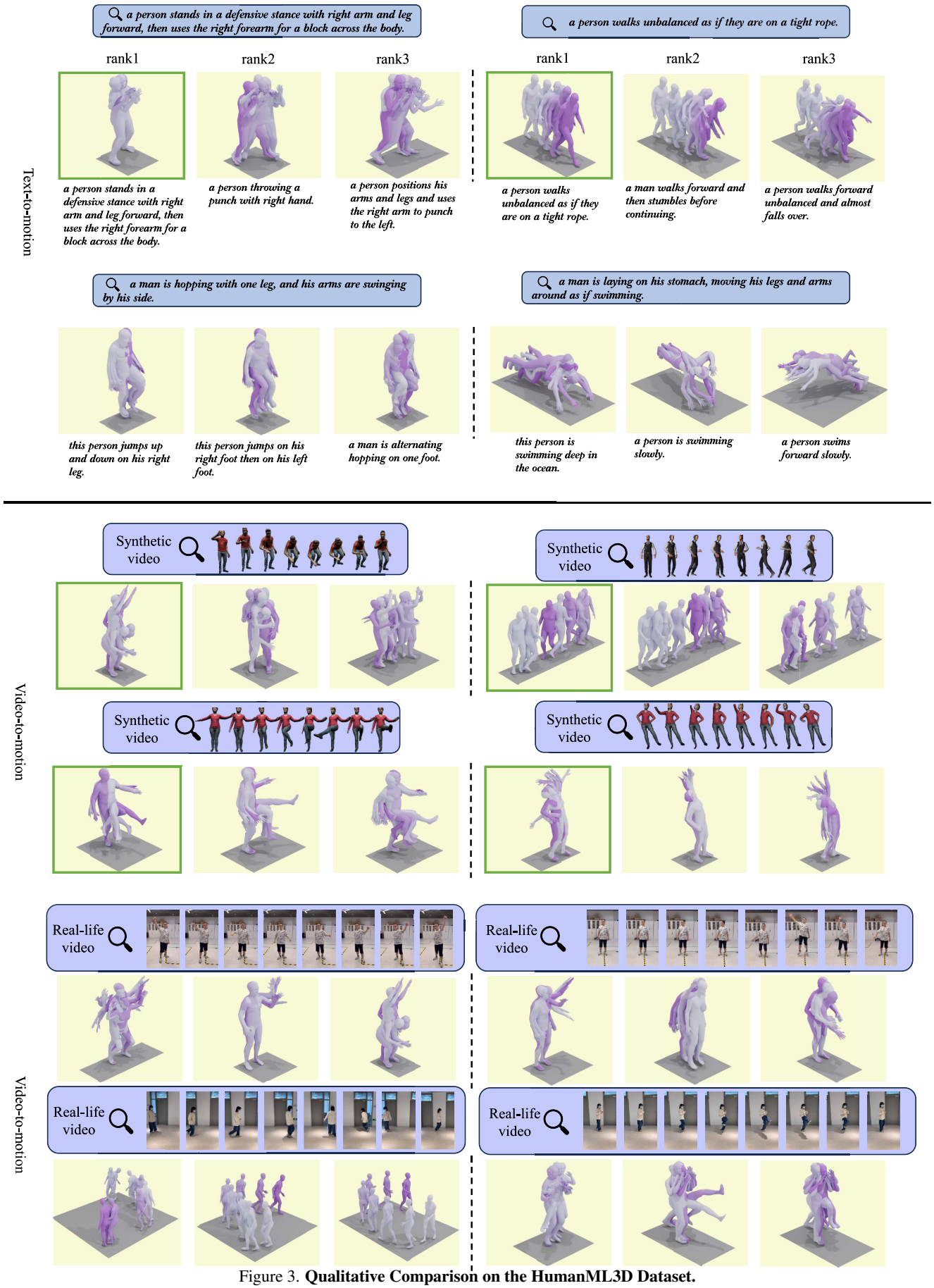

Figure 3. **Qualitative Comparison on the HumanML3D Dataset.**

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*, 2020. 1

[2] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[4] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[5] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 1

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, 2019. 1

[7] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2847–2855, 2023. 1