TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis Supplementary Material

Mathis Petrovich[1,2] Michael J. Black[2] Gül Varol[1]
[1] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
https://mathis.petrovich.fr/tmr

As mentioned in the main text, this appendix includes statistical analysis (Section A), additional experimental results (Section B), and further qualitative results (Section C).

Supplementary video. In addition to this appendix, we provide a video on our project page to allow viewing motions dynamically. In the video, we demonstrate qualitative results for text-to-motion retrieval on the two datasets KIT [4] and H3D [1]. Moreover, we illustrate the use case of moment retrieval on BABEL [5].

Code & Demo. We further provide the source code for training and evaluation, along with an interactive demo, which we make publicly available.

A. Statistics

Number of similar text descriptions in the test set. As mentioned in Section 4.1 of the main paper, the evaluation protocol (b) marks retrieved motions as correct if their corresponding text is similar to the queried text above a threshold of 0.95 (note that this threshold is different from the one used in training). Here, we report the total number of pairs that are above this threshold for each dataset. For KIT, on the 830 sequences of the test set, there are 344,035 unique pairs of texts ($830*829/2$) from which 2,467 of them are similar (about 0.7% of the data). For H3D, on the 4,380 sequences of the test set, there are 9,590,010 unique pairs of texts ($4380*4380/2$) from which 6,017 of them are similar (about 0.06% of the data).

Percentage of filtered negatives per batch during training. To complement Tables 4 and 5 of the main paper, in Table A.1, we compute the amount of negatives that are filtered on average per batch, depending on the threshold and the batch size. In our current setting, 17.29% of the negatives are discarded. We see that this rate remains similar across batch sizes.

Threshold	0.55	0.6	0.65	0.7	0.75	**0.8**	0.85	0.9	0.95
% filtered negatives	98.04	88.04	68.56	48.27	31.54	**17.29**	7.41	2.78	0.71

Batch size	16	**32**	64	128
% filtered negatives	17.02	**17.29**	16.96	17.28

Table A.1. **Percentage of filtered negatives per batch in KIT:** We compute the average percentage of negative pairs per batch that are discarded from the loss computation due to text similarity. The percentage decreases with higher thresholds as expected (top), but the batch size does not have a significant impact (bottom).

B. Additional experimental results

Motion synthesis results. As mentioned in Section 4.1 of the main paper, we evaluate the synthesis performance of TMR. In Table A.2, we compare the performance of TMR, TEMOS and Guo et al. [1] under various settings. See the caption for explanations and comments.

Latent dimensionality. As stated in Section 3.4 of the main paper, the dimensionality of the latent space is set to $\mathsf{d} = 256$ as in TEMOS [3]. In Table A.3, we experiment with this architectural design choice, and observe that $\mathsf{d} = 128$ brings overall better performance.

Contrastive-only baseline. As outlined in Section 4.3 of the main paper, we also experiment with the contrastive model without negative filtering, and present the results in Table A.4. The negative filtering overall improves the results both with the contrastive-only model and with the added synthesis branch (TMR). We note that the added synthesis branch empirically improves the results consistently. Similar conclusions were already made by the text-to-image multi-modal models such as BLIP [2] and CoCa [9] which improve performance over contrastive-only CLIP [6] by adding a text synthesis loss.

Moment retrieval. As presented in Section 4.5 of the main paper, we localize a textual query within a motion, by computing the similarity between the text and several temporal crops of the motion in a zero-shot manner (i.e., the model was not trained for this task, nor has it seen BABEL texts). Here, we provide additional qualitative results, and also report quantitative metrics.

In Figure A.2, we provide complementary qualitative results to Figure 4 of the main paper. At the right of Figure A.2 (b), we also show the localization potential on four very long sequences. As the search space gets larger, the similarity plot

Motions \ Eval	KIT-ML			H3D		
	Guo Ret.	TEMOS	TMR	Guo Ret.	TEMOS	TMR
Real motions	42.25	44.88	49.25	52.41	42.33	67.16
Guo Syn. [1]	36.88	47.00	48.38	45.80	37.73	55.38
TEMOS [3]	43.88	90.50	76.88	40.76	79.71	72.38
TMR	43.50	71.88	89.25	44.67	57.35	92.44

Table A.2. **Motion synthesis results:** We report R@1 text-to-motion retrieval performance of *generated* motions by the synthesis method of Guo et al. [1] (Guo Syn.), TEMOS [3], and our TMR synthesis branch, as well as the 'Real motions', on both KIT-ML (left) and H3D (right) benchmarks. Rows are different motion generation methods, columns are different retrieval evaluation models: retrieval method of Guo et al. [1] (Guo Ret.), TEMOS, and our TMR retrieval branch. We use the protocol (d), i.e., 32 gallery size protocol from [1] . We make several observations: (i) TMR, when used for motion synthesis, performs better than or similar to Guo Syn. [1] across all 3 retrieval evaluation models, showing we do not sacrifice synthesis performance. (ii) Evaluation with retrieval models that can also perform synthesis (TEMOS and TMR) favors motions generated by their own model. (iii) Certain numbers are better than Real motions, potentially because generations are sometimes more faithful to the input text, which may incompletely describe the real motion, or due to the bias mentioned in (ii).

Latent dim. d	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
64	18.80	28.67	38.43	6.00	18.07	21.81	31.45	9.50
128	**25.90**	**31.20**	40.72	6.00	**23.73**	**27.35**	**36.39**	**9.25**
256	24.58	30.24	**41.93**	**5.00**	19.64	23.73	32.53	9.50
512	23.13	28.43	35.42	7.00	20.36	26.39	33.61	10.50

Table A.3. **Latent dimensionality:** We experiment with the embedding space dimensionality, and observe that $d=128$ performs overall best. However, in all other experiments, we use $d=256$ as in TEMOS.

	NF	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
Contrastive-only	✗	19.16	**25.54**	33.13	8.00
Contrastive-only	✓	**19.76**	25.30	**36.87**	**6.00**
TMR	✗	22.17	27.83	36.02	7.00
TMR	✓	**24.58**	**30.24**	**41.93**	**5.00**

Table A.4. **Contrastive-only without negative filtering:** We report text-to-motion retrieval results on KIT-ML to analyze the impact of negative filtering (NF) on the contrastive-only baseline. First row is the supplemental result, the rest are from the main paper.

gets noisier; however, the maximum similarity still occurs at the ground-truth location (marked in green).

For the qualitative results, we display the similarity, centered for each frame, for a window size of 20 frames. Here, we also implement a temporal pyramid approach, where we use a sliding window, with window sizes varying between 10 and 60 frames, and a stride of 5 frames. For quantitative evaluation, we first obtain the predicted localization by selecting the window size and location that gives the best similarity with the text query. Then, we compute the temporal IoU (intersection over union) between the ground-truth segment and the predicted

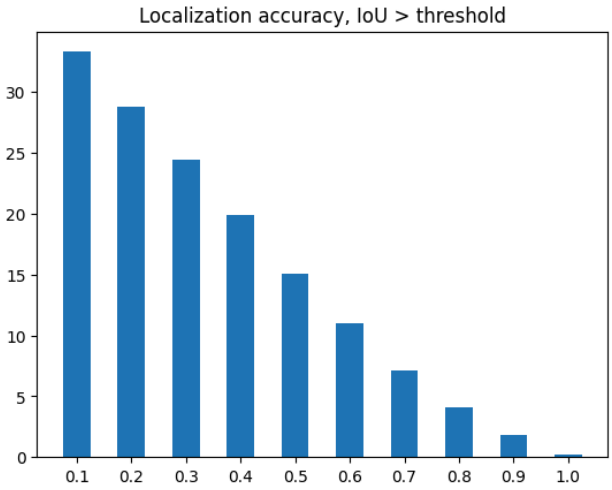


Figure A.1. **Moment retrieval (quantitative):** We plot the localization accuracy (y-axis) with various IoU thresholds (x-axis).

one. In Figure A.1, we report the localization accuracy, where a segment is counted as positive when it has an IoU more than a given threshold. We see that this simple approach can achieve reasonable results (20% of accuracy, with a threshold of 0.4). With a fixed window size of [20, 40, 60] frames, we obtain [17%, 19%, 14%], respectively. A dedicated localization method may consider moment proposal generation as in prior video localization work [7, 8], or a proposal-free approach that trains directly to regress temporal boundaries.

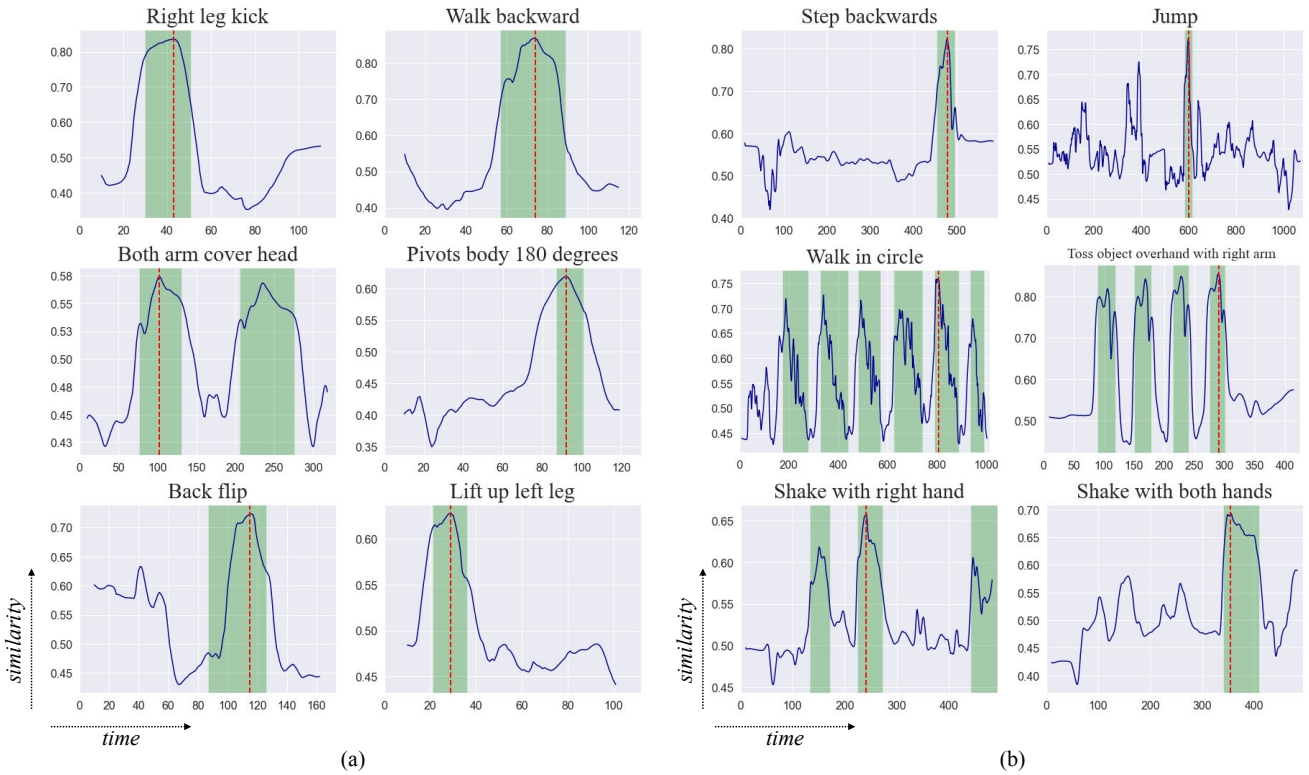
C. Additional qualitative results

In this section, we show qualitative results on the challenging H3D dataset for text-to-motion retrieval on the 4 proposed protocols described in Section 4.1 of the main paper. Protocols (a)(b) are used in Figures A.3 and A.4; (c) in Figure A.5; and (d) in Figure A.6. To reiterate, protocols (a) and (b) use all the test set (4380 motions) as gallery, but (b) marks a rank correct if the text similarity is above a threshold of 0.95. Protocol (c) considers the most dissimilar text subset of 100 motions. Protocol (d) is reported for completeness; it follows [1], and randomly samples batches of 32 motions. All examples are randomly chosen, (i.e., not cherry picked); therefore, are representative of the corresponding protocols.

Overall, we observe that our model is capable of retrieving motions that are semantically similar to the text descriptions. The performance naturally improves as we move from harder to easier protocols. Our detailed observations can be found in the respective figure captions.

References

[1] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human



(a) (b)

Figure A.2. **Moment retrieval (qualitative):** To complement Figure 4 of the main paper, (a) we provide six additional temporal localization results for various text queries on the BABEL dataset. (b) We further visualize six challenging examples when querying on very long motion sequences, i.e., more than 500 frames (25 seconds).

motions from text. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. 1

[3] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[4] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 2016. 1

[5] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

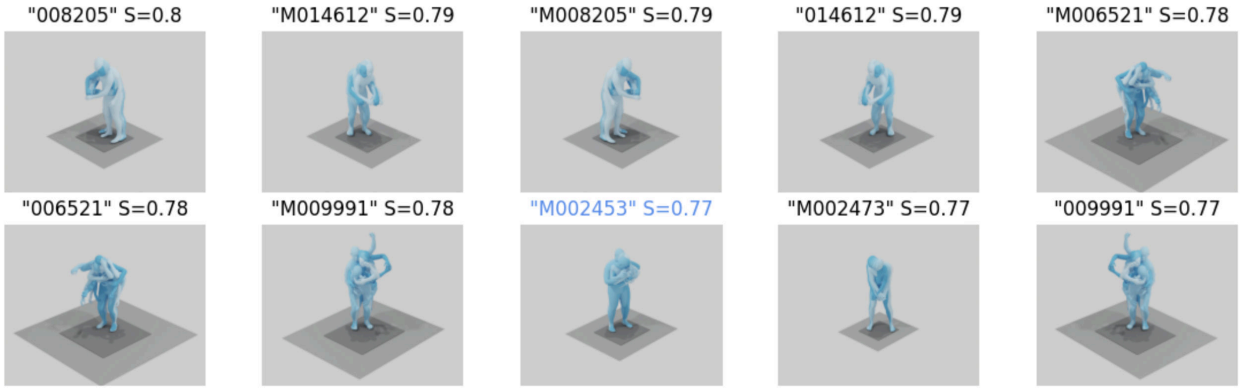
[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[7] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. VLG-Net: Video-language graph matching network for video grounding. In *International Conference on Computer Vision (ICCV)*, 2021. 2

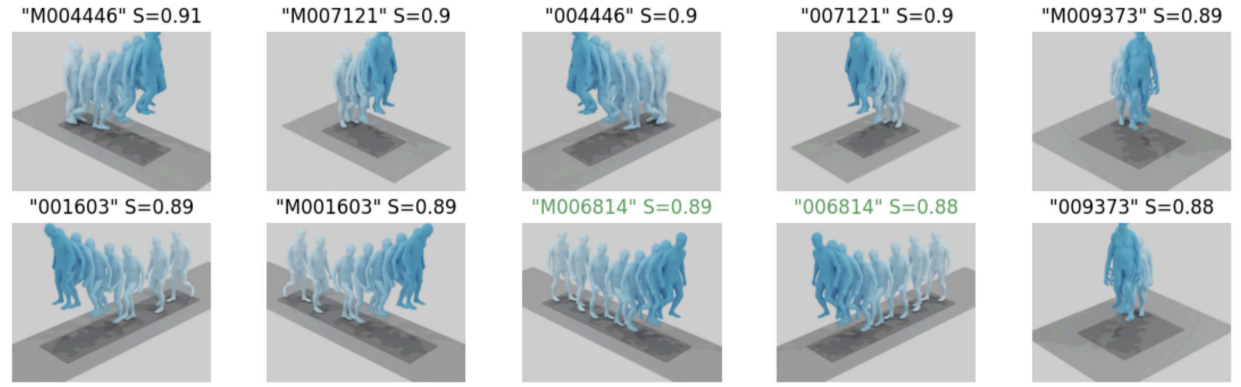
[9] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research (TMLR)*, 2022. 1

Query text: "a person is rocking a baby"
Query keyid: "M002453", rank: 8, rank with threshold: 8



R1 008205: "both the hand holding the right leg.", TS=0.61
R2 M014612: "a person rolls their arms and shoulders.", TS=0.63
R3 M008205: "both the hand holding the left leg.", TS=0.6
R4 014612: "a person rolls their arms and shoulders.", TS=0.63
R5 M006521: "moving hands in a random pattern.", TS=0.59
R6 006521: "moving hands in a random pattern.", TS=0.59
R7 M009991: "the man reaches his left hand into the air then shrugs and digs a hole and shrugs again.", TS=0.51
R8 M002453: "a person is rocking a baby", TS=1.0
R9 M002473: "a person is holding something in front of them and swings to the left.", TS=0.66
R10 009991: "the man reaches his right hand into the air then shrugs and digs a hole and shrugs again.", TS=0.51

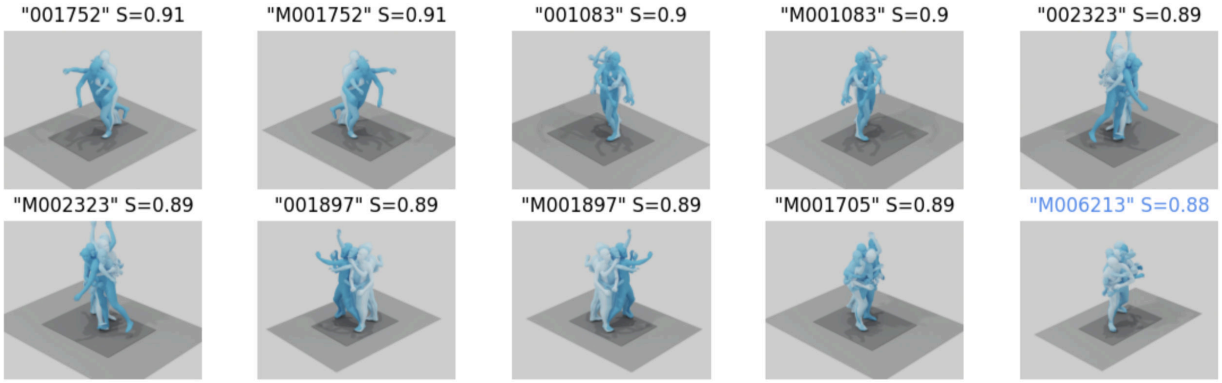
Query text: "a person begins to walk forward up the stairs"
Query keyid: "009903", rank: 31, rank with threshold: 8



R1 M004446: "the person is stepping on something.", TS=0.74
R2 M007121: "a person walks up four steps with their hands by their sides and their lean forward slightly as they go up the stairs and once they've stopped going up the stairs, they straighten up again", TS=0.8
R3 004446: "the person is stepping on something.", TS=0.74
R4 007121: "a person walks up four steps with their hands by their sides and their lean forward slightly as they go up the stairs and once they've stopped going up the stairs, they straighten up again", TS=0.8
R5 M009373: "a figure appears to climb stairs", TS=0.86
R6 001603: "a person walks forward then upwards.", TS=0.89
R7 M001603: "a person walks forward then upwards.", TS=0.89
R8 M006814: "a person walks forward and then up stairs", TS=0.97
R9 006814: "a person walks forward and then up stairs", TS=0.97
R10 009373: "a figure appears to climb stairs", TS=0.86

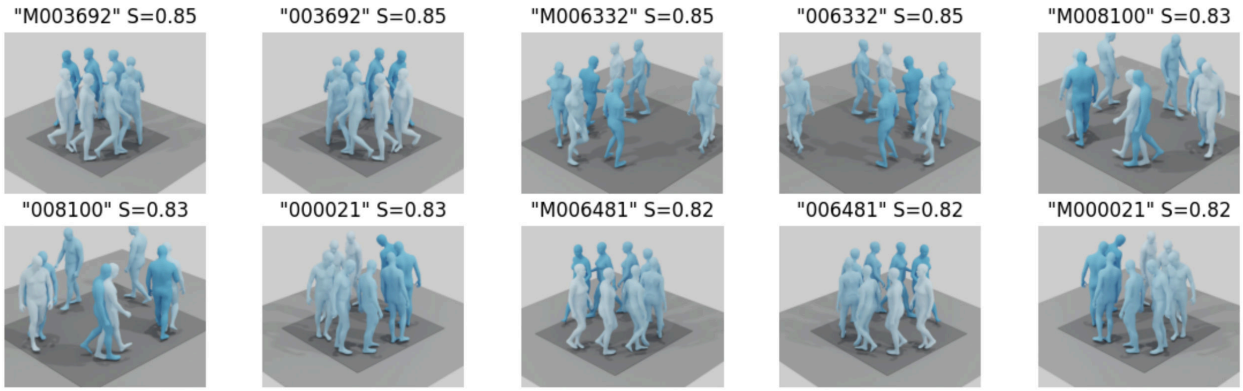
Figure A.3. **Protocols (a) and (b) using all 4,380 motions in H3D:** For each text query, we show the top 10 ranks for the text-to-motion retrieval. Our model generalizes to the concept of "rocking a baby" in the first example, even though this exact same text was not seen in the training set. In the second example, our model retrieves motions that are all coherent with the input query. However, according to evaluation protocol (a), the correct motion is ranked at 31. With the permissive protocol (b), we mark the rank 8 as correct, because their text similarity (TS) is higher than the threshold 0.95.

Query text: "a person winds up his arm and then pitches a ball."
Query keyid: "M006213", rank: 10, rank with threshold: 10



R1 001752: "a person stands still then they throw a football", TS=0.74
R2 M001752: "a person stands still then they throw a football", TS=0.74
R3 001083: "a person lifts object with two hands and throws with right hand.", TS=0.79
R4 M001083: "a person lifts object with two hands and throws with left hand.", TS=0.79
R5 002323: "a person standing up throws something forward from above their head, then throws something again forwa
rd from above their head with more force which makes them take one step forward with their right foot.", TS=0.74
R6 M002323: "a person standing up throws something forward from above their head, then throws something again forw
ard from above their head with more force which makes them take one step forward with their left foot.", TS=0.74
R7 001897: "person aims and throws a baseball", TS=0.85
R8 M001897: "person aims and throws a baseball", TS=0.85
R9 M001705: "a person is pitching a baseball.", TS=0.9
R10 M006213: "a person winds up his arm and then pitches a ball.", TS=1.0

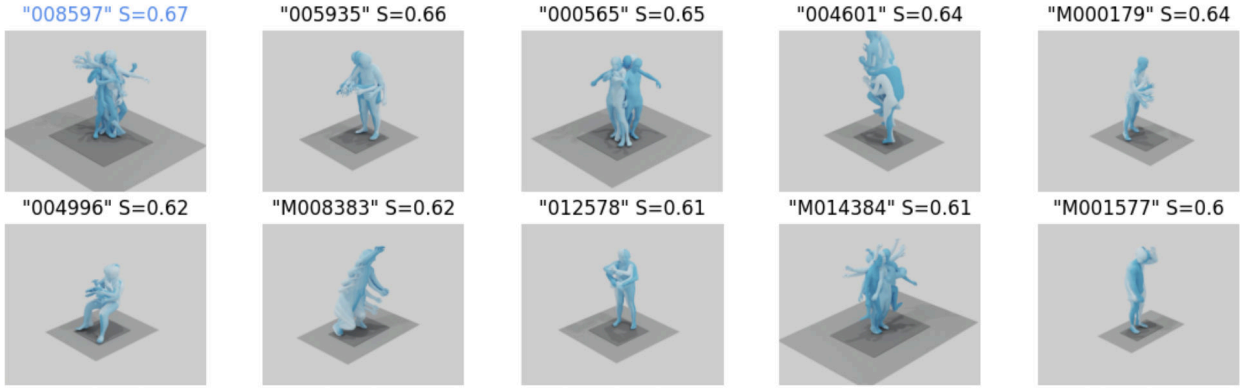
Query text: "walking in a circular pattern."
Query keyid: "013700", rank: 138, rank with threshold: 138



R1 M003692: "a person walks in a clock wise circle and stops were he began.", TS=0.82
R2 003692: "a person walks in a clock wise circle and stops were he began.", TS=0.82
R3 M006332: "a man walks in a counterclockwise circle.", TS=0.82
R4 006332: "a man walks in a clockwise circle.", TS=0.81
R5 M008100: "a person walks in a counter counterclockwise circle.", TS=0.85
R6 008100: "a person walks in a counter clockwise circle.", TS=0.84
R7 000021: "person is walking normally in a circle", TS=0.83
R8 M006481: "the person walks in a counterclockwise circle", TS=0.84
R9 006481: "the person walks in a clockwise circle", TS=0.81
R10 M000021: "person is walking normally in a circle", TS=0.83

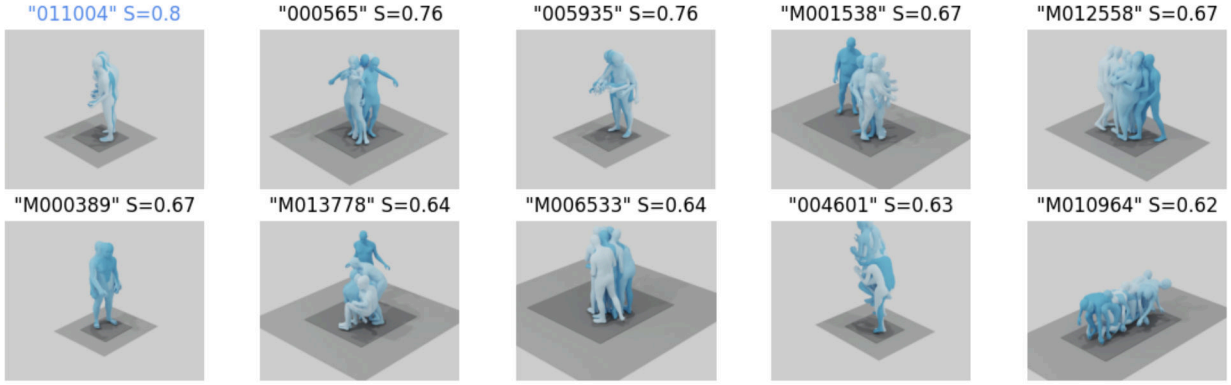
Figure A.4. **Protocols (a) and (b) using all 4,380 motions in H3D (continued):** On both examples, we see that our model retrieves reasonable motions, although the correct motions are ranked at 10 and 138.

Query text: "a person is washing a window"
Query keyid: "008597", rank: 1



R1 008597: "a person is washing a window"
R2 005935: "place items in a line up"
R3 000565: "an off balance intoxicated man gestures at another person to the left. seemingly in an argument."
R4 004601: "someone is climbing a ladder, they walk up 3 steps and then back down."
R5 M000179: "a person holds their left arm bent at the elbow and bends their right arm up and down"
R6 004996: "the person is sat down and their arms are shaking"
R7 M008383: "the person stands up while holding their right hand above their head."
R8 012578: "person person is planting vegetables."
R9 M014384: "a person balances on one foot while moving their other, and then switches."
R10 M001577: "a person scratching their head"

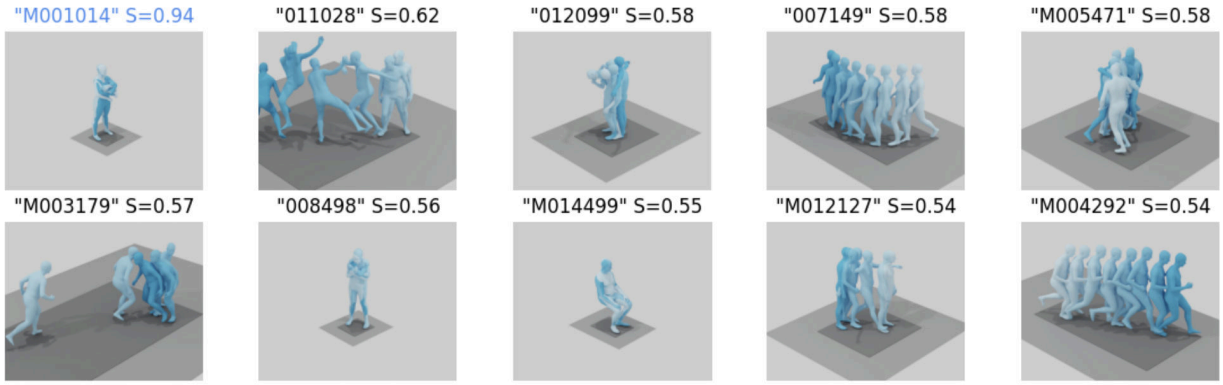
Query text: "a person picks something up in front of them moves it to the side then moves it back"
Query keyid: "011004", rank: 1



R1 011004: "a person picks something up in front of them moves it to the side then moves it back"
R2 000565: "an off balance intoxicated man gestures at another person to the left. seemingly in an argument."
R3 005935: "place items in a line up"
R4 M001538: "a person walks up and tosses something."
R5 M012558: "a person walks forward and then pulls something behind them."
R6 M000389: "a standing man loses a little bit of balance and his upper body leans and shakes toward his right."
R7 M013778: "a person sits down, turns to their left, then stands."
R8 M006533: "the person is walking backwards and then forwards."
R9 004601: "someone is climbing a ladder, they walk up 3 steps and then back down."
R10 M010964: "a person lowers and walks on all fours to the left."

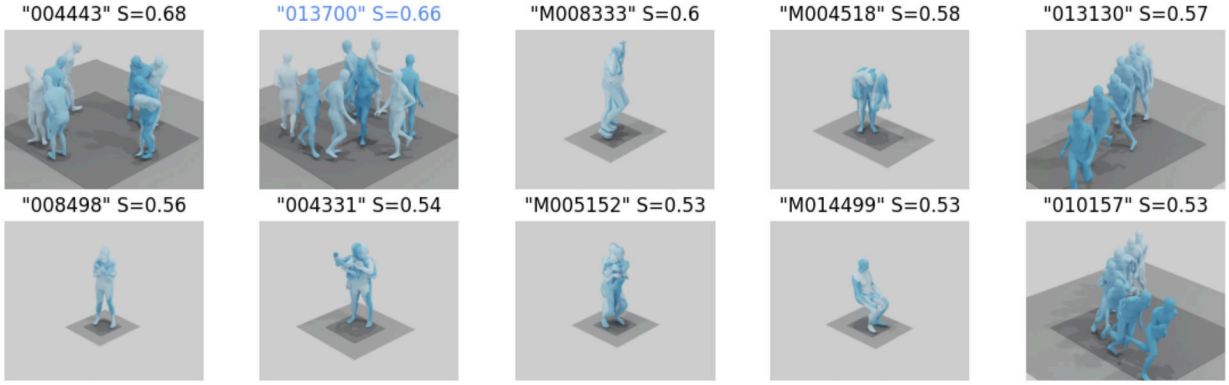
Figure A.5. **Protocol (c) using the most dissimilar 100 texts on H3D:** As there are fewer motions than in protocols (a)(b), and they are more likely to be different, we naturally observe a better performance.

Query text: "person has arms crossing."
Query keyid: "M001014", rank: 1



R1 M001014: "person has arms crossing."
R2 011028: "a man jumps then kicks the air whilst moving to the opposite end of the room."
R3 012099: "a person lifts something to their face and wobbles their body in circles."
R4 007149: "a person is walking at an angle to the right."
R5 M005471: "a person makes tiny steps in place with their hands over their head."
R6 M003179: "a person bends over to begin charging forward, turns around with arms raised, and charges back to original position."
R7 008498: "the person is shivering and then rubbing their hands together to stay warm."
R8 M014499: "a person brings his arms which were in the air along his body. his knees appear to be bent."
R9 M012127: "a man staggers backwards from a standing posture, swinging his arms, before ending in a standing posture."
R10 M004292: "someone running forward, moving forward"

Query text: "walking in a circular pattern."
Query keyid: "013700", rank: 2



R1 004443: "person walks to the right and bends down looking for something , takes a few steps and walks again and bends down again."
R2 013700: "walking in a circular pattern."
R3 M008333: "a man walks from side to side while holding his right forearm with right hand, and then walks back."
R4 M004518: "a person walks forward and rubs an object in front of them with their left hand."
R5 013130: "a person runs forward with one leg crossing in front of the other repetitively before coming to a stop."
R6 008498: "the person is shivering and then rubbing their hands together to stay warm."
R7 004331: "someone dusts a picture hanging on the wall with a cloth in their right hand, steadies the picture with their left hand, then finishes dusting it, and finally dusts all the way around the sides of the frame."
R8 M005152: "a person stands with their knees slightly bent and their hands pulled toward their chest, twists to one side then the other side, squats further, and stands back up."
R9 M014499: "a person brings his arms which were in the air along his body. his knees appear to be bent."
R10 010157: "the person was taking a left drive and then to the right."

Figure A.6. **Protocol (d) using random batches of size 32 on H3D:** As the gallery is very small, the correct motion tends to be at top ranks.