HAL
open science

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

Mathis Petrovich, Michael J. Black, Gül Varol

To cite this version:

Mathis Petrovich, Michael J. Black, Gül Varol. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. International Conference on Computer Vision, Oct 2023, Paris, France. hal-04258206

HAL Id: hal-04258206

https://hal.science/hal-04258206v1

Submitted on 25 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2305.00976v2 [cs.CV] 25 Aug 2023

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

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

Abstract

In this paper, we present TMR, a simple yet effective approach for text to 3D human motion retrieval. While previous work has only treated retrieval as a proxy evaluation metric, we tackle it as a standalone task. Our method extends the state-of-the-art text-to-motion synthesis model TEMOS, and incorporates a contrastive loss to better structure the cross-modal latent space. We show that maintaining the motion generation loss, along with the contrastive training, is crucial to obtain good performance. We introduce a benchmark for evaluation and provide an in-depth analysis by reporting results on several protocols. Our extensive experiments on the KIT-ML and HumanML3D datasets show that TMR outperforms the prior work by a significant margin, for example reducing the median rank from 54 to 19. Finally, we showcase the potential of our approach on moment retrieval. Our code and models are publicly available at https://mathis.petrovich.fr/tmr.

1. Introduction

The language of movement cannot be translated into words.
Barbara Mettler

We ask the question whether a cross-modal space exists between 3D human motions and language. Our goal is to retrieve the most relevant 3D human motion from a gallery, given a natural language query that describes the desired motion (as illustrated in Figure 1). While text-to-image retrieval is a well-established problem within the broader vision & language field [39], there has been less focus on the related task of *text-to-motion* retrieval. Searching an existing motion capture dataset based on text input can often serve as a viable alternative to text-to-human-motion synthesis in many applications, while also providing the added benefit that the retrieved motion is guaranteed to be realistic. Additionally, once a cross-modal embedding is built to map text and motions into a joint representation space, both of the symmetrical text-to-motion and *motion-to-text* tasks

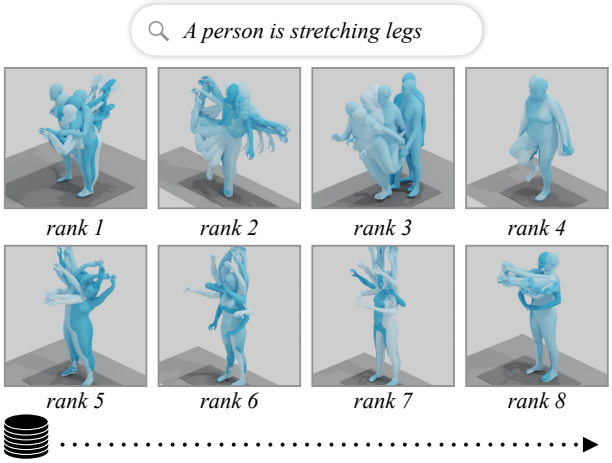


Figure 1. **Text-to-motion retrieval:** We illustrate the task of text-based motion retrieval where the goal is to rank a gallery of motions according to their similarity to the given query in the form of a natural language description.

can be performed. Such a retrieval-based solution has a range of applications, including automatically indexing large motion capture collections, and helping to initialize the cumbersome text labeling process, by assigning the nearest text to each motion.

Let us first differentiate text-to-motion *retrieval* from text-to-motion *synthesis*. Motion synthesis [5, 9, 36, 51] involves generating *new* data samples that go beyond the existing training set, while motion retrieval searches through existing motion capture collections. For certain applications, reusing motions from a collection may be sufficient, provided the collection is large enough to contain what the user is searching for. Unlike generative models for motion synthesis, which struggle to produce physically plausible, realistic sequences [5, 35, 36], a retrieval model has the advantage that it always returns a realistic motion. With this motivation, we pose the problem as a nearest neighbor search through a cross-modal text-motion space.

Early work performs search through motion databases to build motion graphs [4, 23] by finding paths between existing motions and synthesizing new motions by stitching motions together with generated transitions. If the motion database is

labeled with actions, the user can specify a series of actions to combine [4]. In contrast, our search database is *not* labeled with text. *Motion matching* [8], on the other hand, seeks to find the animation that best fits the current motion by searching a database of animations, doing motion-to-motion retrieval [45]. Our framework fundamentally differs from these lines of work in that our task is multi-modal, i.e., user query is text, which is compared against motions. The most similar work to ours is the very recent model from Guo et al. [15], which trains for a joint embedding space between the two modalities. This model is only used to provide a performance measure for motion synthesis tasks, by querying a generated motion within a gallery of 32 descriptions (i.e., motion-to-text retrieval), and counting how many times the correct text is retrieved[1]. While this can be considered as the first text-motion retrieval model in the literature, its main limitation is the low performance, in particular when the gallery contains fine-grained descriptions. We substantially improve over [15], by incorporating a joint synthesis & retrieval framework, as well as a more powerful contrastive training [34].

We get inspiration from image-text models such as BLIP [26] and CoCa [56], which formulate a multi-task objective. Besides the standard dual-encoder matching (such as CLIP [39] with two unimodal encoders for image and text), [26, 56] also employ a text synthesis branch, performing image captioning. Such a generative capability potentially helps the model go beyond 'bag-of-words' understanding of vision-language concepts, observed for the naive contrastive models [11, 57]. In our case, we depart from TEMOS [36] which already has a synthesis branch to generate motions from text. We incorporate a cross-modal contrastive loss (i.e., InfoNCE [34]) in this framework to jointly train text-to-motion synthesis and text-to-motion retrieval tasks. We empirically demonstrate significant improvements with this approach when ablating the importance of each task.

Text-motion data differs from its text-image counterparts particularly due to the nature of motion descriptions. In fact, for an off-the-shelf large language model, sentences describing different motions tend to be similar, since they fall within the same topic of human motions. For example, the text-text cosine similarities [48] after encoding motion descriptions from the KIT training set [37] are on average 0.71 on a scale between [0, 1], while this value is 0.56 (almost orthogonal) on a random subset of LAION [44] image descriptions with the same size. This poses several challenges. Typical motion datasets [15, 37, 38] contain similar motions with different accompanying texts, e.g., 'person walks', 'human walks', as well as similar texts with different motions, e.g., 'walk backwards', 'walk forwards'. With naive contrastive training [34], one would make all samples within a batch as negatives, except the corresponding label for a given anchor. In this work, we take into account the fact that there are potentially significant similarities between pairs within a batch. Specifically, we discard pairs that have a text-text similarity in their labels that is above a certain threshold. Such careful negative sampling leads to performance improvements.

In this paper, we illustrate an additional use case for our retrieval model – zero-shot temporal localization – and highlight this task as a potential future avenue for research. Similar to temporal localization in videos with natural language queries [12, 13, 19, 25, 41], also referred to as "moment retrieval", we showcase the grounding capability of our model by directly applying it on long motion sequences to retrieve corresponding moments. We illustrate results on the BABEL dataset [38], which typically contains a series of text annotations for each motion sequence. Note that the task is zero-shot, because the model has not been trained for localization, and at the same time has not seen BABEL labels, which come from a different domain (e.g., typically action-like descriptions instead of full sentences).

Our contributions are the following: (i) We address the little-studied problem of text-to-motion retrieval, and introduce a series of evaluation benchmarks with varying difficulty. (ii) We propose a joint synthesis and retrieval framework, as well as negative filtering, and obtain state-of-the-art performance on text-motion retrieval. (iii) We provide extensive experiments to analyze the effects of each component in controlled settings. Our code and models are publicly available[2].

2. Related work

We present an overview of the most closely related work on text-to-motion synthesis and retrieval, as well as a brief discussion on cross-modal retrieval works.

Text and human motion. The research on human motion modeling has recently witnessed an increasing interest in bridging the gap between semantics and 3D human body motions, in particular for text-conditioned motion synthesis. Unlike unconstrained 3D human motion synthesis [55, 60, 61], action-conditioned [16, 35], or text-conditioned [5, 9, 14, 22, 24, 36, 51, 58] models add semantic controls to the generation process. The goal of these works is to generate either deterministically [2, 14, 27] or probabilistically [36], motion sequences that are faithful to the textual description inputs. Note that this is different than gesture synthesis from speech [17], in that the text describes the motion content. However, despite remarkable progress in text-to-image synthesis [40, 42], text-to-motion synthesis remains at a very nascent stage. The realism of the synthesized motions is limited, e.g., foot sliding artifacts [36]. We turn to text-to-motion retrieval as another alternative, and perhaps complementary approach to obtain motions for a given textual description. Our focus is therefore different from the work on synthesis. However, we do make use of a motion synthesis branch to aid the retrieval task.

[1] While the paper [15] describes a motion-to-text retrieval metric, we notice that the provided code performs text-to-motion retrieval.

[2] https://mathis.petrovich.fr/tmr

Motion retrieval is relatively less explored. As briefly mentioned in Section 1, motion-to-motion retrieval (e.g., motion matching [8, 20, 45]) methods exist. However, the *text*-to-motion retrieval task is more challenging due to being cross-modal, i.e., nearest neighbor search across text and motion modalities. Within this category, the very recent work of Guo et al. [15] trains a retrieval model purely for evaluation purposes, and applies a margin-based contrastive loss [18], using Euclidean distance between all pairs within a batch.

Two works are particularly relevant to ours. Firstly, we build on the TEMOS [36] text-to-motion synthesis model, which also has a cross-modal embedding space. However, text and motion embeddings are encouraged to be similar only across positive pairs. We therefore add a contrastive training strategy to incorporate negatives, consequently improving its retrieval capability from a large gallery of fine-grained motions. Secondly, we compare to the aforementioned method of [15], whose motion-to-text retrieval model is adopted for measuring text-to-motion generation performance automatically by other works [10, 51, 59].

Cross-modal retrieval. Among widely adopted vision & language retrieval models, some successful examples include CLIP [39], BLIP [26], and CoCa [56] for images, and MIL-NCE [33], Frozen [7], and CLIP4Clip [30] for videos. They all use variants of cross-modal contrastive learning techniques, such as InfoNCE [34], which we also employ in this work. As discussed in Section 1, we draw inspiration from BLIP [26] and CoCa [56], which add synthesis branches to standard retrieval frameworks. Our approach is similar in spirit to these works in that we perform a cross-modal vision & language retrieval task, but differ in focusing on 3D human motion retrieval, which, to the best of our knowledge, has not been benchmarked.

3. Text-to-motion retrieval

In this section, we introduce the task and the terminology associated with text-to-motion retrieval (Section 3.1). Next, we present our model, named TMR, and its training protocol (Section 3.2). We then explain our simple approach for identifying and filtering incorrect negatives (Section 3.3), followed by a discussion of the implementation details (Section 3.4).

3.1. Definitions

Given a natural language query T, such as '*A person walks and then makes a right turn.*', the goal is to rank the motions from a database (i.e., the gallery) according to their semantic correspondence with the text query, and to retrieve the motion that matches best to the textual description. In other words, the task involves sorting the database so that the top ranked motions are the most relevant matches, i.e., creating a search engine to index motions. Additionally, we define the symmetric (and complementary) task, namely motion-to-text retrieval, where the aim is to retrieve the most suitable caption that matches a given motion from a database of texts.

3D human motion refers to a sequence of human poses. The task does not impose any limitations on the type of representation used, such as joint positions, rotations, or parametric models such as SMPL [28]. As detailed in Section 4.1, we choose to use the representation employed by Guo et al. [15] to facilitate comparisons with previous work. While our experiments are based on the SMPL skeleton, our method is applicable to any skeleton topology. One could alternatively use retargetting [1, 53, 54] to adapt to other body representations.

Text description refers to a sequence of words describing the action performed by a human in natural language. We do not restrict the format of the motion description. The text can be simply an action name (e.g., 'walk') or a full sentence (e.g., 'a human is walking'). The sentences can be fine-grained (e.g., 'a human is walking in a circle slowly'), and may contain one or several actions, simultaneously (e.g., 'walking while waving') or sequentially (e.g., 'walking then sitting').

3.2. Joint training of retrieval and synthesis

We introduce TMR, which extends the Transformer-based text-to-motion synthesis model TEMOS [36] by incorporating additional losses to make it suitable for the retrieval task. The architecture consists of two independent encoders for inputting motion and text, as well as a decoder that outputs motion (see Figure 2 for an overview). In the following, we review TEMOS [36] components and our added contrastive training.

Dual encoders. One approach to solving cross-modal retrieval tasks involves defining a similarity function between the two modalities. In our case, the two modalities are text and motion. The similarity function can be applied to compare a given query with each element in the database, and the maximum value would indicate the best match. In this paper, we follow the approach taken by previous work on metric learning, such as CLIP [39], by defining one encoder for each modality and then computing the cosine similarity between their respective embeddings. Such dual embedding has the advantage of fast inference time since the gallery embeddings can be computed and stored beforehand [32].

Our model is built upon the components of TEMOS [36], which already provides a motion encoder and a text encoder, mapping them to a joint space (building on the idea from Language2Pose [2]); this serves as a strong baseline for our work. Both motion and text encoders are Transformer encoders [52] with additional learnable distribution parameters, as in the VAE-based ACTOR [35]. They are probabilistic in nature, outputting parameters of a Gaussian distribution (μ and Σ) from which a latent vector $z \in \mathbb{R}^d$ can be sampled. While the text encoder takes text features from a pre-trained and frozen DistilBERT [43] network as input, the motion sequence is fed directly in the motion encoder. Note that when performing retrieval, we directly use the output embedding that corresponds to the mean token (μ^M for motion, μ^T for text).

Motion decoder. TEMOS is trained for the task of motion

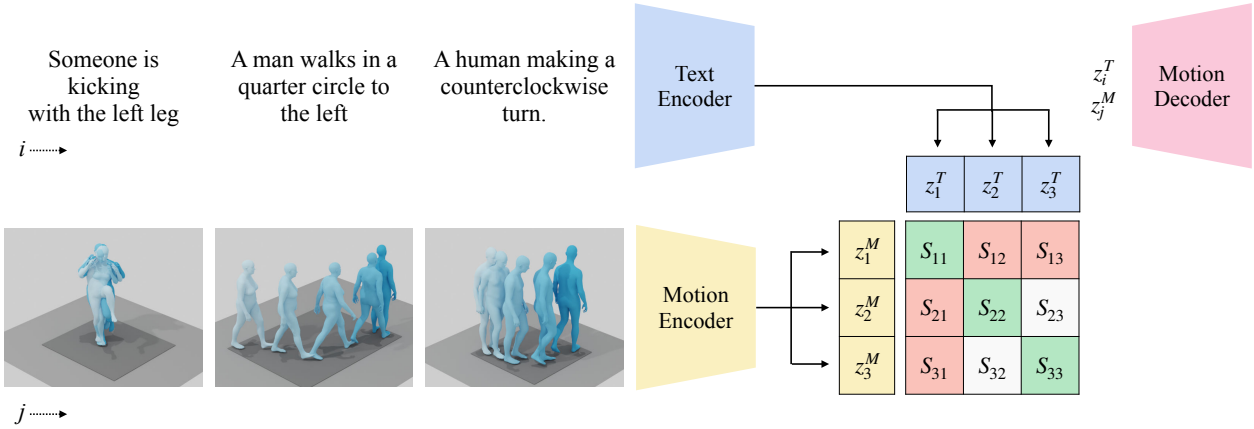


Figure 2. **Joint motion retrieval and synthesis:** A simplified view of our TMR framework is presented, where we focus on the similarity matrix defined between text-motion pairs within a batch. Here, we show a batch of 3 samples for illustration purposes. The goal of the contrastive objective is to maximize the diagonal denoting positive pairs (green), and to minimize the off-diagonal negative pair items that have text similarity below a threshold (red). In this example, remaining similarities S_{23} and S_{32} are discarded from the loss computation because there is high text similarity between T_2 and T_3. The rest of the model remains similar to TEMOS [36], which decodes a motion from both text z^{T_i} and motion z^{M_j} latent vectors. See text for further details.

synthesis and comes equipped with a motion decoder branch. This decoder is identical to the one used in ACTOR [35], which supports variable-duration generation. More specifically, it takes a latent vector $z \in \mathbb{R}^d$ and a sinusoidal positional encoding as input, and generates a motion non-autoregressively through a single forward pass. We show in Section 4.2 that keeping this branch helps improve the results.

TEMOS losses. We keep the same base set of losses of [36], defined as the weighted sum $\mathcal{L}_{\text{TEMOS}} = \mathcal{L}_{\text{R}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{E}}\mathcal{L}_{\text{E}}$. In summary, a reconstruction loss term $\mathcal{L}_{\text{R}}$ measures the motion reconstruction given text or motion input (via a smooth L1 loss). A Kullback-Leibler (KL) divergence loss term $\mathcal{L}_{\text{KL}}$ is composed of four losses: two of them to regularize each encoded distributions – $\mathcal{N}(\mu^M, \Sigma^M)$ for motion and $\mathcal{N}(\mu^T, \Sigma^T)$ for text – to come from a normal distribution $\mathcal{N}(0, I)$. The other two enforce distribution similarity between the two modalities. A cross-modal embedding similarity loss $\mathcal{L}_{\text{E}}$ enforces both text z^T and motion z^M latent codes to be similar to each other (with a smooth L1 loss). We set λ_{KL} and λ_{E} to 10^{-5} in our experiments as in [36].

Contrastive training. While TEMOS has a cross-modal embedding space, its major drawback to be usable as an effective retrieval model is that it is never trained with negatives, but only positive motion-text pairs. To overcome this limitation, we incorporate a contrastive training with the usage of negative samples to better structure the latent space. Given a batch of N (positive) pairs of latent codes $(z_1^T, z_1^M), \cdots, (z_N^T, z_N^M)$, we define any pair (z_i^T, z_j^M) with $i \neq j$ as negative. The similarity matrix S computes the pairwise cosine similarities for all pairs in the batch $S_{ij} = \cos(z_i^T, z_j^M)$. In contrast to Guo et al. [15], who consider one random negative per batch (and use a margin loss), we adopt the more recent formulation of InfoNCE [34], which was proven effective in other work [7, 26, 39]. This loss term can be defined as follows:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2N}\sum_i \left(\log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ij}/\tau} + \log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ji}/\tau} \right), \quad (1)$$

where τ is the temperature hyperparameter.

Training loss. The total loss we use to train TMR is the weighted sum $\mathcal{L}_{\text{TEMOS}} + \lambda_{\text{NCE}}\mathcal{L}_{\text{NCE}}$ where λ_{NCE} controls the importance of the contrastive loss.

3.3. Filtering negatives

As mentioned in Section 1, the descriptions accompanying motion capture collections can be repetitive or similar across the training motions. We wish to prevent defining negatives between text-motion pairs that contain similar descriptions.

Consider for example the two text descriptions, "*A human making a counterclockwise turn*" and "*A person walks quarter a circle to the left*". In the KIT-ML benchmark [37], these two descriptions appear as two different annotations for the same motion. Due to the flexibility and ambiguity of natural language, different words may describe the same concepts (e.g., 'counterclockwise', 'circle to the left').

During training, the random selection of batches can adversely affect the results because the model may have to push away two latent vectors that correspond to similar meanings. This can force the network to focus on unimportant details (e.g., 'someone' vs 'human'), ultimately resulting in decreased performance due to unstable behavior and reduced robustness to text variations.

To alleviate this issue, we leverage an external large language model to provide sentence similarity scores. In particular, we use MPNet [49] to encode sentences and compute similarities between two text descriptions. We then determine whether to filter a pair of text descriptions (t_1, t_2) if their similarity

is higher than a certain threshold, referring them as 'wrong negatives'. During training, we filter wrong negative pairs from the loss computation. Note that we refrain from defining them as positives, as the language model may also incorrectly say two descriptions are similar when they are not.

3.4. Implementation details

We use the AdamW optimizer [29] with a learning rate of 10^{-4} and a batch size of 32. Since the batch size can be an important hyperparameter for the InfoNCE loss, due to determining the number of negatives, we report experimental results with different values. The latent dimensionality of the embeddings is $d=256$. We set the temperature τ to 0.1, and the weight of the contrastive loss term λ_{NCE} to 0.1. The threshold to filter negatives is set to 0.8. We provide experimental analyses to measure the sensitivity to these added hyperparameters.

4. Experiments

We start by describing the datasets and evaluation protocol used in the experiments (Section 4.1). We then report the performance of our model on our new retrieval benchmark along with comparison to prior work (Section 4.2). Next, we present our ablation study measuring the effects of the additional contrastive loss, the negative filtering, and the hyperparameters (Section 4.3). Finally, we provide qualitative results for retrieval (Section 4.4), and our use case of moment retrieval (Section 4.5).

4.1. Datasets and evaluation

HumanML3D dataset (H3D) [15] provides natural language labels to describe the motions in AMASS [31] and HumanAct12 [16] motion capture collections. We follow the motion pre-processing procedure of [15], and apply the SMPL layer [28] to extract joint positions, canonicalize the skeletons to share the same topology (i.e., same bone lengths), then compute motion features (extracting local positions, velocities and foot contacts similar to Holden et al. [21]). The data is then augmented by mirroring left and right (both in motions and their corresponding texts). After this procedure, and following the official split, we obtain 23384, 1460, 4380 motions for the training, validation, and test sets, respectively. On average, each motion is annotated 3.0 times with different text. During training we randomly select one as the matching text, for testing we use the first text.

KIT Motion-Language dataset (KIT) [37] also come from motion capture data, with an emphasis on locomotion motions. It originally consists of 3911 motion sequences and 6278 text sentences. We pre-process the motions with the same procedure as in H3D. The data is split into 4888, 300, 830 motions for training, validation, and test sets, respectively. In this dataset, each motion is annotated 2.1 times on average.

Evaluation protocol. We report standard retrieval performance measures: recall at several ranks, R@1, R@2, etc. for both text-to-motion and motion-to-text tasks (identical to the R Precision metrics adopted by Guo et al. [15]). Recall at rank k measures the percentage of times the correct label is among the top k results; therefore higher is better. We additionally report median rank (MedR), where lower is better. Note that the *retrieval* performance is evaluated on a gallery of unseen real motions (i.e., test set). A study of the *synthesis* performance of TMR can be found in Appendix B.

We define several evaluation protocols, mainly changing the gallery set. (a) **All** the test set is used as a first protocol, without any modification. This set is partially problematic because there are repetitive texts across motions, or just minor differences (e.g., person vs human, walk vs walking). (b) **All with threshold** means we search over all the test set, but, in this case, we accept a retrieved motion as correct if its text label is similar to the query text above a threshold. For example, retrieving the motion corresponding to *"A human walks forward"* should be correct when the input query is *"Someone is walking forward"*. We set a high threshold of 0.95 (scaled between [0, 1]) to remove very similar texts without removing too many fine-grained details. Appendix A provides statistics on how often similar text descriptions appear in the datasets. (c) **Dissimilar subset** refers to sampling 100 motion-text pairs whose texts are maximally far from each other (using an approximation of the quadratic knapsack problem [3]). This evaluation can be considered as an easy, but clean, subset of the previous ones. (d) **Small batches** is included to mimic the protocol described by Guo et al. [15], who randomly pick batches of 32 motion-text pairs and report the average performance. An ideal evaluation metric should not have randomness, and a gallery size of 32 is relatively easy compared to the previous protocols.

4.2. A new benchmark & comparison to prior work

We present the performance of our model on this new retrieval benchmark, on H3D (Table 1) and KIT (Table 2) datasets, across all evaluation protocols. We also compare against prior work TEMOS [36] and Guo et al. [15]. For TEMOS, we retrain their model on both datasets to have a comparable benchmark since the original model differs in motion representation and lacks left/right data augmentation (and they only provide a KIT-pretrained model, not H3D). For [15], we take their publicly available models trained on these two datasets.

TEMOS in particular is not designed to perform well on retrieval, since its cross-modal embedding space is only trained with positive pairs. However, Guo et al. train contrastively with negatives as well, using a margin loss [18]. For all 4 evaluation sets with varying difficulties, TMR outperforms the prior work, suggesting our model better captures the finegrained nature of motion descriptions. The model of [15] is adopted as part of motion synthesis evaluation in several works. TMR may therefore provide a better alternative. Our significant improvements over the state of the art can be credited to (i) jointly training for synthesis and retrieval, (ii) adopting the more recent contrastive objective, InfoNCE [34], while (iii) carefully

Protocol	Methods	Text-motion retrieval						Motion-text retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	**TEMOS** [36]	2.12	4.09	5.87	8.26	13.52	173.0	3.86	4.54	6.94	9.38	14.00	183.25
	Guo et al. [15]	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	**5.68**	**10.59**	**14.04**	**20.34**	**30.94**	**28.00**	**9.95**	**12.44**	**17.95**	**23.56**	**32.69**	**28.50**
(b) All with threshold	**TEMOS** [36]	5.21	8.22	11.14	15.09	22.12	79.00	5.48	6.19	9.00	12.01	17.10	129.0
	Guo et al. [15]	5.30	7.83	10.75	14.59	22.51	54.00	4.95	5.68	8.93	11.64	16.94	69.50
	TMR	**11.60**	**15.39**	**20.50**	**27.72**	**38.52**	**19.00**	**13.20**	**15.73**	**22.03**	**27.65**	**37.63**	**21.50**
(c) Dissimilar subset	**TEMOS** [36]	33.00	42.00	49.00	57.00	66.00	4.00	35.00	44.00	50.00	56.00	70.00	3.50
	Guo et al. [15]	34.00	48.00	57.00	72.00	84.00	3.00	34.00	47.00	59.00	72.00	83.00	3.00
	TMR	**47.00**	**61.00**	**71.00**	**80.00**	**86.00**	**2.00**	**48.00**	**63.00**	**69.00**	**80.00**	**84.00**	**2.00**
(d) Small batches [15]	**TEMOS** [36]	40.49	53.52	61.14	70.96	84.15	2.33	39.96	53.49	61.79	72.40	85.89	2.33
	Guo et al. [15]	52.48	71.05	80.65	89.66	**96.58**	1.39	52.00	71.21	81.11	89.87	**96.78**	1.38
	TMR	**67.16**	**81.32**	**86.81**	**91.43**	95.36	**1.04**	**67.97**	**81.20**	**86.35**	**91.70**	95.27	**1.03**

Table 1. **Text-to-motion retrieval benchmark on HumanML3D:** We establish four evaluation protocols as described in Section 4.1, with decreasing difficulty from (a) to (d). Our model TMR substantially outperforms the prior work of Guo et al. [15] and TEMOS [36], on the challenging H3D dataset.

Protocol	Methods	Text-motion retrieval						Motion-text retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	**TEMOS** [36]	7.11	13.25	17.59	24.10	35.66	24.00	**11.69**	**15.30**	**20.12**	26.63	36.39	26.50
	Guo et al. [15]	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	**7.23**	**13.98**	**20.36**	**28.31**	**40.12**	**17.00**	11.20	13.86	**20.12**	**28.07**	**38.55**	**18.00**
(b) All with threshold	**TEMOS** [36]	18.55	24.34	30.84	42.29	56.39	7.00	17.71	22.41	28.80	35.42	47.11	13.25
	Guo et al. [15]	13.25	22.65	29.76	39.04	49.52	11.00	10.48	13.98	20.48	27.95	38.55	17.25
	TMR	**24.58**	**30.24**	**41.93**	**50.48**	**60.36**	**5.00**	**19.64**	**23.73**	**32.53**	**41.20**	**53.01**	**9.50**
(c) Dissimilar subset	**TEMOS** [36]	24.00	40.00	46.00	54.00	70.00	5.00	33.00	39.00	45.00	49.00	64.00	6.50
	Guo et al. [15]	16.00	29.00	36.00	48.00	66.00	6.00	24.00	29.00	36.00	46.00	66.00	7.00
	TMR	**26.00**	**46.00**	**60.00**	**70.00**	**83.00**	**3.00**	**34.00**	**45.00**	**60.00**	**69.00**	**82.00**	**3.50**
(d) Small batches [15]	**TEMOS** [36]	43.88	58.25	67.00	74.00	84.75	2.06	41.88	55.88	65.62	75.25	85.75	2.25
	Guo et al. [15]	42.25	62.62	75.12	87.50	**96.12**	1.88	39.75	62.75	73.62	86.88	**95.88**	1.95
	TMR	**49.25**	**69.75**	**78.25**	**87.88**	95.00	**1.50**	**50.12**	**67.12**	**76.88**	**88.88**	94.75	**1.53**

Table 2. **Text-to-motion retrieval benchmark on KIT-ML:** As in Table 1, we report the four evaluation protocols, this time on the KIT dataset. Again, TMR significantly improves over Guo et al. [15] and TEMOS [36] across all protocols and metrics.

Motion Recons.	InfoNCE	Margin	Text-motion retrieval				Motion-text retrieval			
			R@1↑	R@2↑	R@3↑	MedR↓	R@1↑	R@2↑	R@3↑	MedR↓
✗	✗	✓	15.06	22.17	25.78	12.00	8.19	11.57	16.39	19.50
✗	✓	✗	19.76	25.30	36.87	6.00	17.47	19.76	30.60	**9.50**
✓	✗	✗	18.55	24.34	30.84	7.00	17.71	22.41	28.80	13.25
✓	✗	✓	19.88	24.46	34.46	7.00	14.70	19.76	28.19	12.50
✓	✓	✗	**24.58**	**30.24**	**41.93**	**5.00**	**19.64**	**23.73**	**32.53**	**9.50**

Table 3. **Losses:** We experiment with various loss definitions (i) with/without the motion reconstruction, and (ii) the choice of the contrastive loss between InfoNCE and margin-based. We see that InfoNCE [34] is a better alternative to the contrastive loss with Euclidean margin [18] (employed by Guo et al. [15]). The reconstruction loss through the motion decoder branch further boosts the results.

eliminating wrong negatives. In the following, we ablate these components in controlled experiments.

4.3. Ablation study

The rest of the quantitative evaluation uses the '(b) All with threshold' evaluation protocol, on the KIT dataset.

Which losses matter? Table 3 compares several variants of TMR where we check (a) whether the jointly trained motion synthesis branch helps retrieval, and (b) how important the form of the contrastive loss is. When removing the synthesis branch and only using the the contrastive loss, we perform a deterministic encoding (i.e., with a single token instead of two tokens μ, σ). First, we see that the motion synthesis branch certainly improves results compared with only training using a contrastive loss (e.g., 41.93 vs 36.87 R@3). This possibly forces the latent vector to capture the full content of the input text (i.e., instead of picking up on a subset of words, or bag-of-words [11, 57] upon finding a shortcut that satisfies the contrastive loss). Second, in the presence of a contrastive loss, the InfoNCE formulation is significantly better than the margin loss employed by previous work [15] (41.93 vs 34.46 R@1). Note that for this experiment, we keep the same negative filtering for both the margin loss and InfoNCE (we provide additional experiments without the negative filtering in Appendix B).

Threshold	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
0.55	19.40	23.25	30.48	9.00	17.83	21.69	29.52	14.00
0.60	17.95	26.87	36.87	6.00	20.60	24.70	31.81	11.25
0.65	23.01	28.67	36.39	7.00	19.04	21.69	29.76	11.50
0.70	22.29	29.64	38.80	6.00	18.19	22.77	32.05	**9.00**
0.75	20.00	27.11	37.83	6.00	20.24	24.46	**34.22**	9.50
0.80	**24.58**	**30.24**	**41.93**	**5.00**	19.64	23.73	32.53	9.50
0.85	21.45	25.78	38.43	6.50	**20.84**	24.10	33.37	9.50
0.90	23.25	30.12	40.48	6.00	20.00	**25.18**	33.13	9.50
0.95	20.48	26.99	38.43	6.00	19.28	23.37	31.93	10.25
✗	22.17	27.83	36.02	7.00	16.75	21.33	32.17	11.50

Table 4. **Filtering negatives:** We compare several threshold values for filtering negatives from the loss comparison due to having similar texts. We observe that removing negatives based on text similarity above 0.8 (from a scale between [0,1]) performs well overall.

The effect of filtering negatives. As explained in Section 3.3, during training we filter out pairs whose texts are closer than a threshold in an embedding space, and do not count them in the contrastive loss computation. Note that we still keep each item in the batch for the motion synthesis objective. In Table 4, we perform experiments with a range of different values for this threshold selection. On a scale between [0, 1], a threshold of 0.8 has the best results, balancing keeping a sufficient number of negatives and removing the wrong ones. Without filtering at all, the performance remains at 36.02 R@3 (compared to 41.93). We provide statistics on the percentage of filtered pairs in Appendix A.

Hyperparameters of the contrastive training. We show the sensitivity of our model to several hyperparameters added when extending TEMOS: (i) temperature τ of the cross entropy of InfoNCE [34] in Eq.1, (ii) the λ_{NCE} weighting parameter, and (iii) the batch size, which determines the number of negatives. We see in Table 5 that the model is indeed sensitive to the temperature, which is a common observation in other settings. The weight parameter and the batch size are relatively less important while also influencing the results to a certain extent. An experiment with the latent dimensionality hyperparameter can be found in Appendix B.

4.4. Qualitative results

In Figure 3, we provide sample qualitative results for text-to-motion retrieval on the full test set of H3D. For each query text displayed on the left, top-5 retrieved motions are shown on the right along with their similarity scores. Note that the ground-truth text labels (at the bottom of each motion) for the retrieved motions are not used, and the gallery motions are unseen at training. For the first two examples with 'playing violin' and 'handstand', we retrieve the ground-truth motion at rank 1. We observe that the next ranked motions depict visually similar motions as well (e.g., 'cartwheel' involves standing on the hands). For the free-from prompt example 'Someone is swimming' (i.e., the exact text does not appear in the gallery), the three first motions resemble or involve the swimming action, whereas motions at ranks 4 and 5 are incorrect. We notice

Temp. τ	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
0.001	9.52	21.81	27.23	12.00	7.47	9.76	16.51	15.50
0.01	21.45	29.04	38.80	6.00	**21.08**	**27.11**	**33.61**	**9.50**
0.1	**24.58**	**30.24**	**41.93**	**5.00**	19.64	23.73	32.53	**9.50**
1.0	1.08	1.93	3.61	306.5	1.81	1.93	2.41	372.0

(a)

Weight λ_{NCE}	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
0.001	18.55	23.25	36.75	7.00	18.19	**24.34**	31.45	11.50
0.01	20.84	26.99	37.23	7.00	18.92	23.13	32.17	10.25
0.1	**24.58**	**30.24**	**41.93**	**5.00**	**19.64**	23.73	32.53	**9.50**
1.0	19.52	24.46	34.46	7.00	19.04	**24.34**	**35.06**	**9.50**

(b)

Batch size	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
16	**25.42**	**31.57**	40.12	6.00	**20.36**	24.10	**33.73**	**8.00**
32	24.58	30.24	**41.93**	**5.00**	19.64	23.73	32.53	9.50
64	20.24	26.51	38.19	6.00	19.52	**24.22**	32.05	9.50
128	18.55	28.80	36.75	7.00	14.94	18.43	26.14	11.50

(c)

Table 5. **Hyperparameters of the contrastive training:** We measure the sensitivity to the parameters τ (temperature), λ_c the weight of the contrastive loss, and the batch size. Note that the learning rate is proportionally altered when changing the batch size. We display a wide range of values to show the full trends.

that the incorrect motions have a low similarity (< 0.6), and the human bodies are rotated similarly as in swimming. More qualitative results can be seen in Appendix C, as well as our supplementary video on the project page.

4.5. Use case: Moment retrieval

While our focus is retrieval, once our model is trained, it can be used for a different use case. Here, we test the limits of our approach, by qualitatively evaluating the capability of TMR on the task of temporally localizing a natural language query in a long 3D motion sequence. This is similar in spirit to moment retrieval in videos [12, 13, 19, 25, 41]. It is also related to categorical action localization in 3D motions [50]; however, our input is free-form text instead of symbolic action classes.

In Figure 4, we show four examples, where we apply a model pre-trained on H3D on BABEL sequences. In each example, the queried text is displayed on the top. The x-axis denotes the frame number, the green rectangle represents the ground-truth location for the given action, and the dashed red line marks the localization with the maximum similarity. We simply compute the motion features in a sliding window manner. The similarity between the text label and a 20-frame window centered at each frame is shown in the y-axis as a 1D plot over time. Despite our model not being trained for temporal localization, we observe its grounding potential. Moreover, there exists a domain gap between BABEL label used at test time and H3D used for training. Quantitative evaluation and more qualitative examples are included in Appendix B.

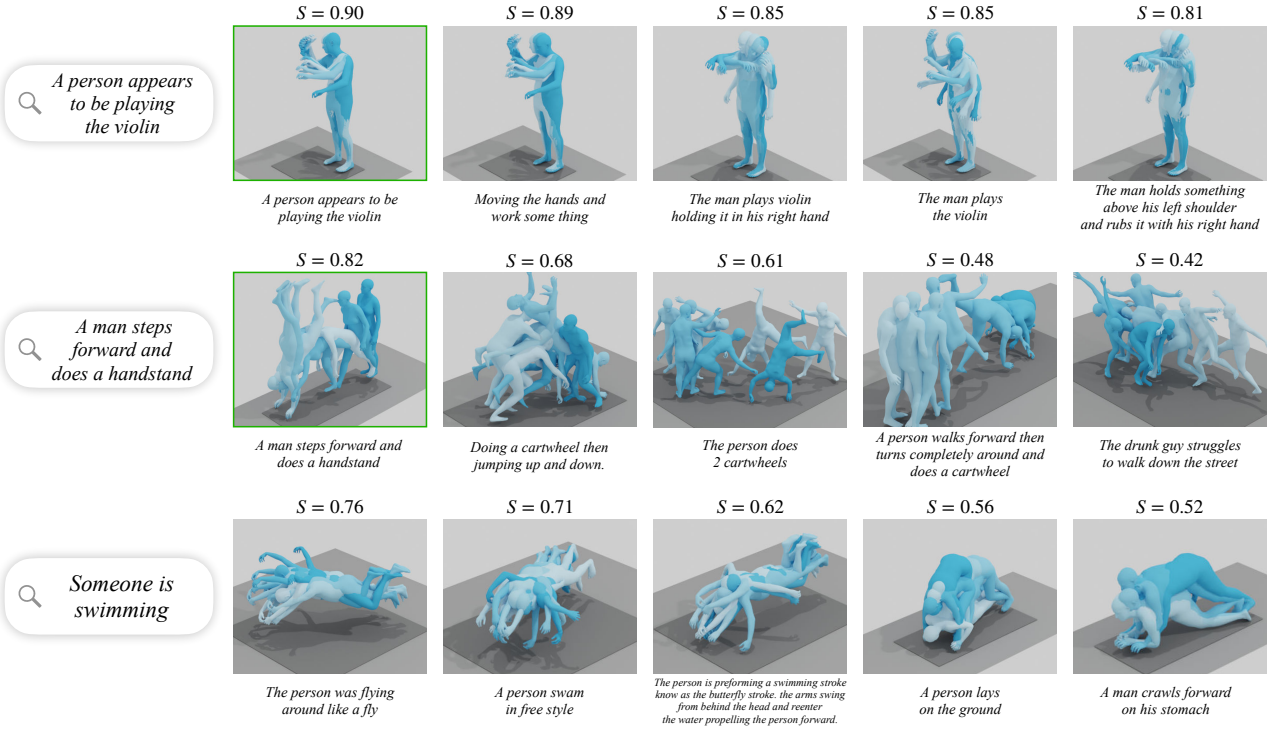


Figure 3. **Qualitative retrieval results:** We demonstrate example queries on the left, and corresponding retrieved motions on the right, ranked by text-motion similarity. The similarity values are displayed on the top. For each retrieved motion, we also show their accompanying ground-truth text label; note that we do not use these descriptions, but only provide them for analysis purposes. The motions from the gallery are all from the test set (unseen during training). In the first row, all top-5 retrieved motions correspond visually to 'playing violin' and the similarity scores are high >0.80. In the second row, we correctly retrieve the 'handstand' motion at top-1, but the other motions mainly perform 'cartwheel' (which involves shortly standing on hands), but with a lower similarity score <0.70. For the last example, we query a free-form text 'Someone is swimming', which does not exist in the gallery (but the word 'swim' does). The model successfully finds swimming motions among the top-3, and the other two motions involve the body parallel to the ground.

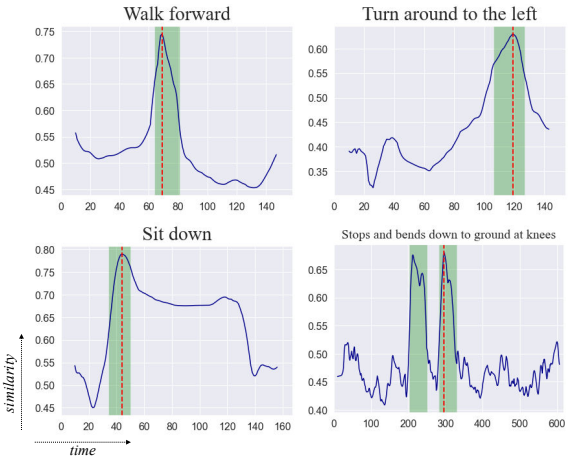


Figure 4. **Moment retrieval:** We plot the similarity between the temporally annotated BABEL text labels and the motions in a sliding window manner, and obtain a 1D signal over time (blue). We observe that a localization ability emerges from our model, even though it was not trained for temporal localization, and was not with the domain of BABEL labels. The ground-truth temporal span is denoted in green and the maximum similarity is marked with a dashed red line. More examples are provided in Appendix Figure A.2.

4.6. Limitations

Our model comes with some limitations. Compared to the vast amount of data (e.g., 400M images [44]) in image-text collections used to train competitive foundation models, our motion-text training data can be considered extremely small (e.g., 23K motions in H3D). The generalization performance of motion retrieval models to in-the-wild motions is therefore limited. Data augmentations such as altering text can potentially help to a certain extent (cf. [6]); however, more motion capture is still needed. Another limitation concerns the case where one wishes to replace motion synthesis by retrieving a training motion. In this use case, the model requires the full encoded database (i.e., encoded training set) to be stored in memory, which can be memory-inefficient. Note that the encoded representation is considerably smaller than the original motion data. This could be dealt with using a hierarchical structuring of the motions and language descriptions (e.g. into upper-body motions, lower-body motions, sitting motions, standing motions, etc.).

5. Conclusion

In this paper, we focus on the relatively little-studied problem of motion retrieval with natural language queries. We

introduce TMR, a framework to jointly train text-to-motion retrieval and text-to-motion synthesis, with a special attention to the definition of negatives, taking into account the fine-grained nature of motion-language databases. We significantly improve over prior work, and provide a series of experiments highlighting the importance of each component.

Future work may consider incorporating a language synthesis branch, along with the motion synthesis branch, to build a symmetrical framework, which could bring further benefits.

Acknowledgements. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011012129R2 made by GENCI. GV acknowledges the ANR project CorVis ANR-21-CE23-0003-01. The authors would like to thank Lucas Ventura and Charles Raude.

MJB Disclosure: https://files.is.tue.mpg.de/black/CoI_ICCV_2023.txt

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 2020. 3

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2Pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 2, 3

[3] Méziane Aïder, Oussama Gacem, and Mhand Hifi. Branch and solve strategies-based algorithm for the quadratic multiple knapsack problem. *Journal of the Operational Research Society*, 2022. 5

[4] Okan Arikan, David A. Forsyth, and James F. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics (TOG)*, 2003. 1, 2

[5] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal action composition for 3D humans. In *International Conference on 3D Vision (3DV)*, 2022. 1, 2

[6] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *International Conference on Computer Vision (ICCV)*, 2023. 8

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[8] Michael Büttner and Simon Clavet. Motion matching - the road to next gen animation. In *Nucl.ai, youtu.be/watch?v=z_wpgHFSWss&t=658s*, 2015. 2, 3

[9] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[10] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. MoFusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[12] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv:1907.12763*, 2019. 2, 7

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 7

[14] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[15] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5, 6, 11, 12

[16] 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 *ACM International Conference on Multimedia (ACMMM)*, 2020. 2, 5

[17] Ikhsanul Habibie, Mohamed Elgharib, Kripashindu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *SIGGRAPH*, 2022. 2

[18] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer Vision and Pattern Recognition (CVPR)*, 2006. 3, 5, 6

[19] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 7

[20] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 2020. 3

[21] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 2017. 5

[22] Sai Shashank Kalakonda, Shubh Maheshwari, Sarvadevabhatla, and Ravi Kiran. Action-GPT: Leveraging large-scale language models for improved and generalized zero shot action generation. *arXiv:2211.15603*, 2022. 2

[23] Lucas Kovar, Michael Gleicher, and Frédéric H. Pighin. Motion graphs. *ACM Transactions on Graphics (TOG)*, 2002. 1

[24] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. MultiAct: Long-term 3D human motion generation from multiple action labels. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[25] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[26] 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. 2, 3, 4, 12

[27] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *ViGIL NeurIPS Workshop*, 2018. 2

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person

linear model. *ACM Transactions on Graphics (TOG)*, 2015. 3, 5

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[30] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. 3

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[32] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[33] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2, 3, 4, 5, 6, 7

[35] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4

[36] 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, 3, 4, 5, 6, 11

[37] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 2016. 2, 4, 5, 11

[38] 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. 2, 11

[39] 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, 2, 3, 4, 12

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. 2

[41] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. In *Transactions of the Association for Computational Linguistics (TACL)*, 2013. 2, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[43] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Energy Efficient Machine Learning and Cognitive Computing NeurIPS Workshop*, 2019. 3

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. In *Data Centric AI NeurIPS Workshop*, 2021. 2, 8

[45] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *European Conference on Computer Vision (ECCV)*, 2002. 2, 3

[46] 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. 12

[47] 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. 12

[48] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2020. 2

[49] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2020. 4

[50] Jiankai Sun, Bolei Zhou, Michael J. Black, and Arjun Chandrasekaran. LocATe: End-to-end localization of actions in 3D with transformers. *arXiv:2203.10719*, 2022. 7

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017. 3

[53] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[54] Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, and Yinda Zhang. Neural pose transfer by spatially adaptive instance normalization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[55] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[56] 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. 2, 3, 12

[57] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 2, 6

[58] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2

[59] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv:2208.15001*, 2022. 3

[60] Yan Zhang, Michael J. Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv:2007.13886*, 2020. 2

[61] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *CVPR*, 2020. 2

APPENDIX

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 [37] and H3D [15]. Moreover, we illustrate the use case of moment retrieval on BABEL [38].

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.

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. [15] 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 $d = 256$ as in TEMOS [36]. In Table A.3, we experiment with this architectural design choice, and observe that $d = 128$ 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

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).

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. [15]	36.88	47.00	48.38	45.80	37.73	55.38
TEMOS [36]	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. [15] (Guo Syn.), TEMOS [36], 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. [15] (Guo Ret.), TEMOS, and our TMR retrieval branch. We use the protocol (d), i.e., 32 gallery size protocol from [15] . We make several observations: (i) TMR, when used for motion synthesis, performs better than or similar to Guo Syn. [15] 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.	Text-motion retrieval				Motion-text retrieval			
d	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.

already made by the text-to-image multi-modal models such as BLIP [26] and CoCa [56] which improve performance over contrastive-only CLIP [39] 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 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 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 [46, 47], 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

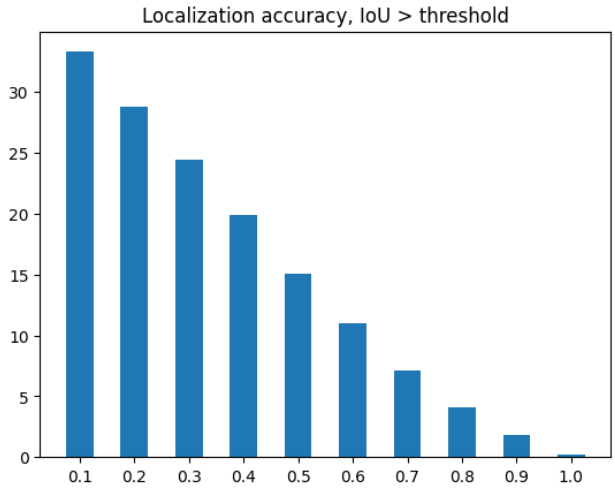


Figure A.1. **Moment retrieval (quantitative):** We plot the localization accuracy (y-axis) with various IoU thresholds (x-axis).

(c) considers the most dissimilar text subset of 100 motions. Protocol (d) is reported for completeness; it follows [15], 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.

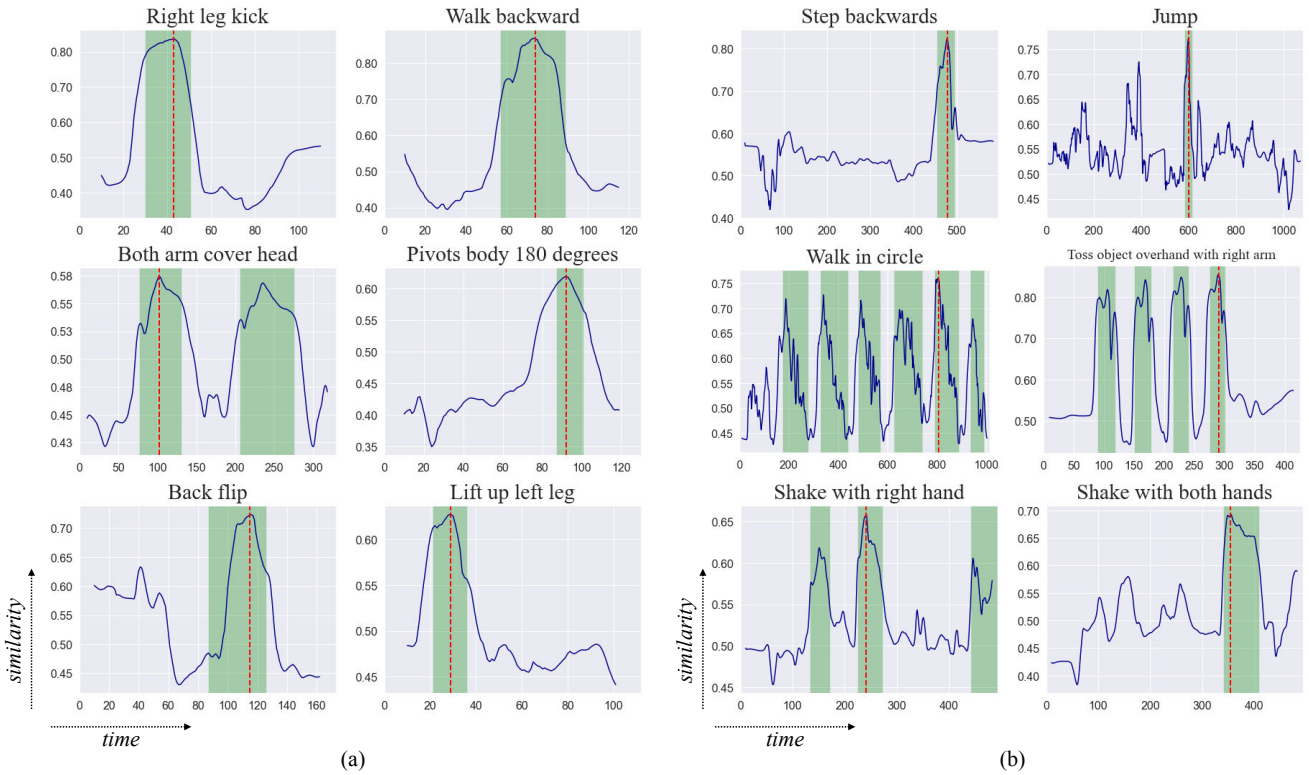
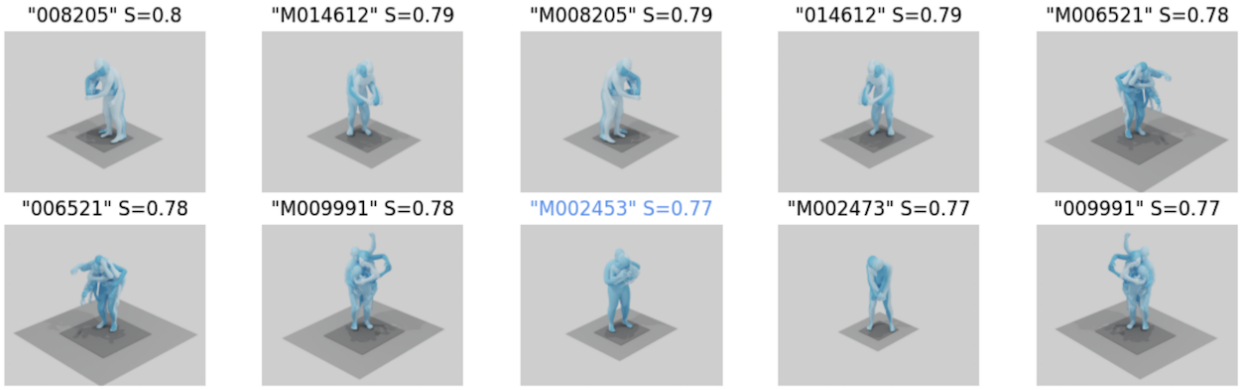


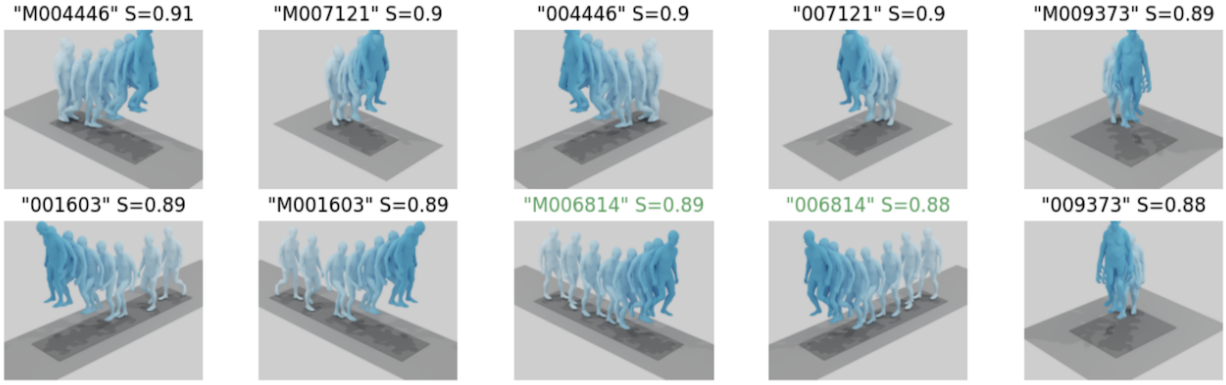
Figure A.2. **Moment retrieval (qualitative):** To complement Figure 4 of the main paper, (a) we provide six additional temporal localization results for various text queries on the BABEL dataset. (b) We further visualize six challenging examples when querying on very long motion sequences, i.e., more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

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

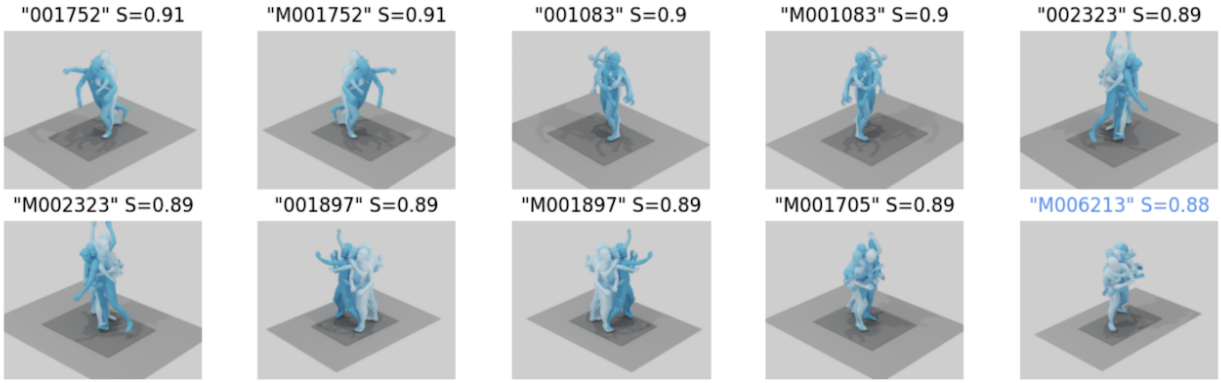
Query text: "a person begins to walk forward up the stairs"
Query keyid: "009903", rank: 31, rank with threshold: 8



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

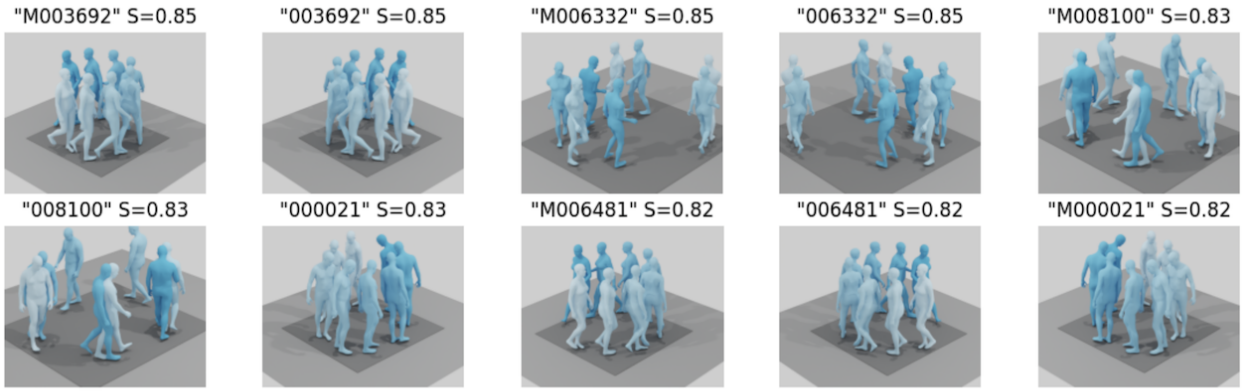
Figure A.3. **Protocols (a) and (b) using all 4,380 motions in H3D:** For each text query, we show the top 10 ranks for the text-to-motion retrieval. Our model generalizes to the concept of "rocking a baby" in the first example, even though this exact same text was not seen in the training set. In the 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



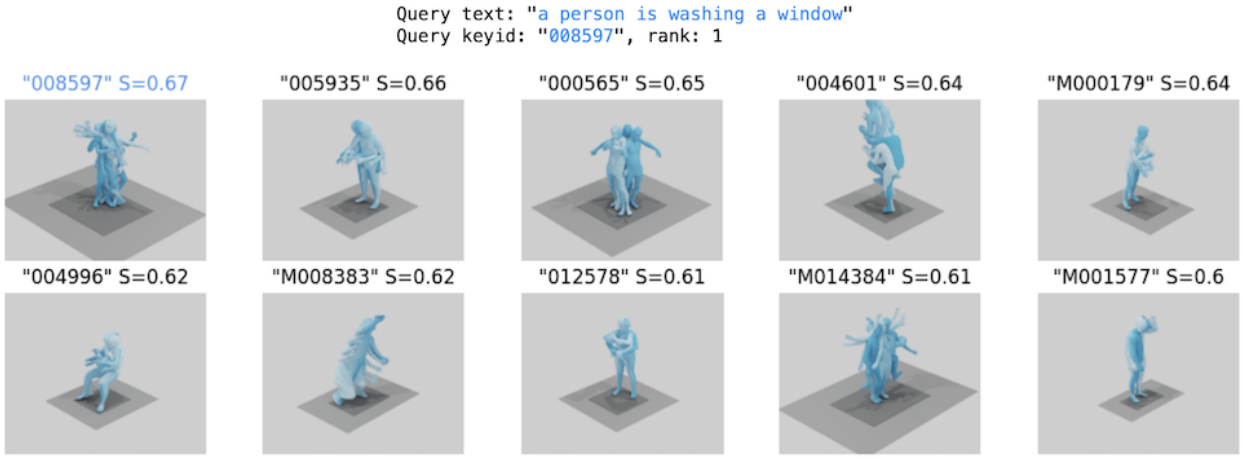
R1 001752: "a person stands still then they throw a football", TS=0.74
R2 M001752: "a person stands still then they throw a football", TS=0.74
R3 001083: "a person lifts object with two hands and throws with right hand.", TS=0.79
R4 M001083: "a person lifts object with two hands and throws with left 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

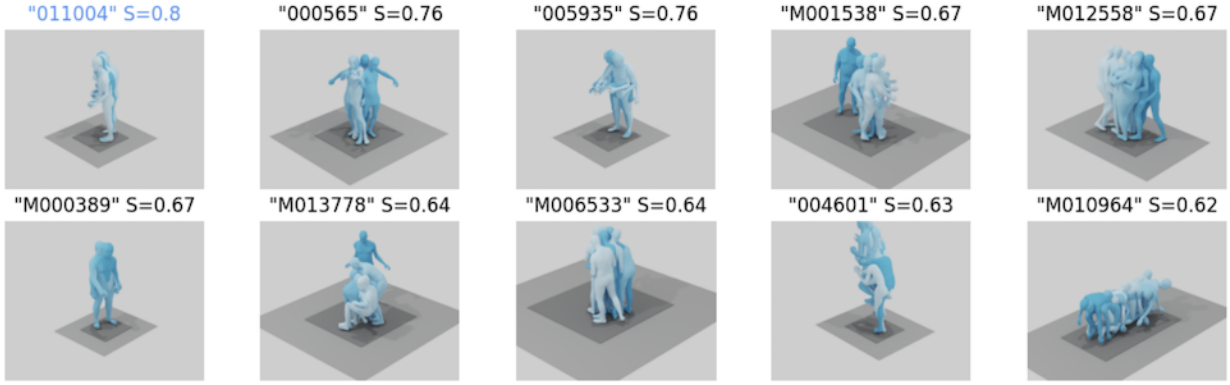


R1 M003692: "a person walks in a clock wise circle and stops were he began.", TS=0.82
R2 003692: "a person walks in a clock wise circle and stops were he began.", TS=0.82
R3 M006332: "a man walks in a counterclockwise circle.", TS=0.82
R4 006332: "a man walks in a clockwise circle.", TS=0.81
R5 M008100: "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.



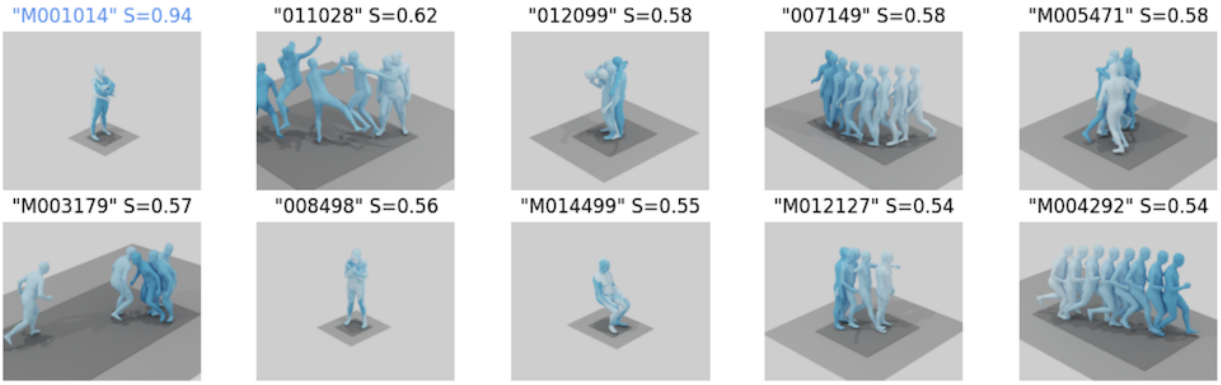
R1 008597: "a person is washing a window"
R2 005935: "place items in a line up"
R3 000565: "an off balance intoxicated man gestures at another person to the left. seemingly in an argument."
R4 004601: "someone is climbing a ladder, they walk up 3 steps and then back down."
R5 M000179: "a person holds 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"



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."

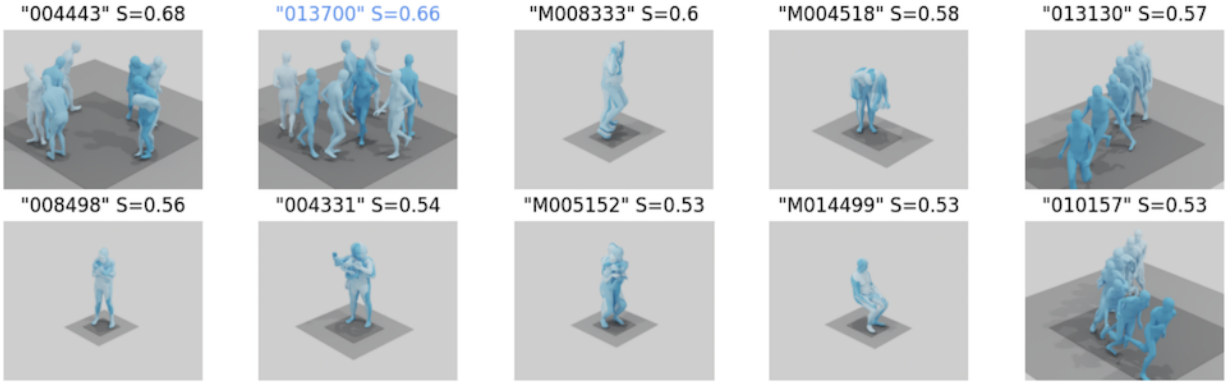
Figure A.5. **Protocol (c) using the most dissimilar 100 texts on H3D:** As there are fewer motions than in protocols (a)(b), and they are more likely to be different, we naturally observe a better performance.

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.