Text-to-Motion Retrieval: Towards Joint Understanding of Human Motion Data and Natural Language

Nicola Messina*
ISTI-CNR
Pisa, Italy
nicola.messina@isti.cnr.it

Jan Sedmidubsky*
Masaryk University
Brno, Czechia
xsedmid@fi.muni.cz

Fabrizio Falchi
ISTI-CNR
Pisa, Italy
fabrizio.falchi@isti.cnr.it

Tomáš Rebok
Masaryk University
Brno, Czechia
rebok@ics.muni.cz

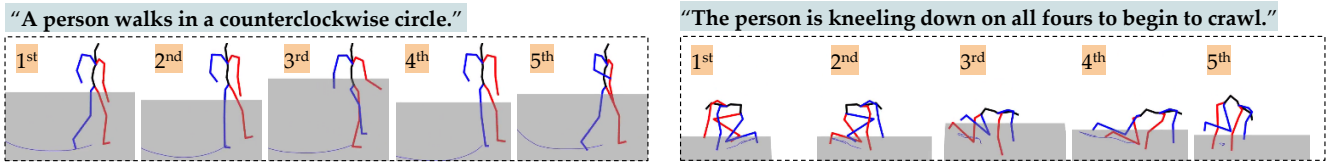


Figure 1: Five motions retrieved for two different queries specified by free text (CLIP as text encoder, MoT as motion encoder).

ABSTRACT

Due to recent advances in pose-estimation methods, human motion can be extracted from a common video in the form of 3D skeleton sequences. Despite wonderful application opportunities, effective and efficient content-based access to large volumes of such spatio-temporal skeleton data still remains a challenging problem. In this paper, we propose a novel content-based text-to-motion retrieval task, which aims at retrieving relevant motions based on a specified natural-language textual description. To define baselines for this uncharted task, we employ the BERT and CLIP language representations to encode the text modality and successful spatio-temporal models to encode the motion modality. We additionally introduce our transformer-based approach, called Motion Transformer (MoT), which employs divided space-time attention to effectively aggregate the different skeleton joints in space and time. Inspired by the recent progress in text-to-image/video matching, we experiment with two widely-adopted metric-learning loss functions. Finally, we set up a common evaluation protocol by defining qualitative metrics for assessing the quality of the retrieved motions, targeting the two recently-introduced KIT Motion-Language and HumanML3D datasets. The code for reproducing our results is available here: https://github.com/mesnico/text-to-motion-retrieval.

CCS CONCEPTS

• **Information systems** → **Novelty in information retrieval**; *Language models*; *Evaluation of retrieval results*; *Data access methods.*

KEYWORDS

human motion data, skeleton sequences, CLIP, BERT, deep language models, ViViT, motion retrieval, cross-modal retrieval

*Both authors contributed equally to the paper.

This work is licensed under a Creative Commons Attribution International 4.0 License.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592069

ACM Reference Format:
Nicola Messina, Jan Sedmidubsky, Fabrizio Falchi, and Tomáš Rebok. 2023. Text-to-Motion Retrieval: Towards Joint Understanding of Human Motion Data and Natural Language. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539618.3592069

1 INTRODUCTION AND RELATED WORK

Pose-estimation methods [12] can detect 3D human-body keypoints in a single RGB video stream. The keypoints detected in individual frames constitute a simplified spatio-temporal representation of human motion in the form of a so-called *skeleton sequence*. As indicated in [40], the analysis of such representation opens unprecedented application potential in many domains, ranging from virtual reality, through robotics and security, to sports and medicine. The ever-increasing popularity of skeleton data calls for technologies able to effectively and efficiently access large volumes of such spatio-temporal data based on its content.

Research in skeleton-data processing mainly focuses on designing deep-learning architectures for classification of labeled actions [8, 24, 33] or detection of such actions in continuous streams [32, 42]. The proposed architectures are often learned in a supervised way based on transformers [1, 8, 9], convolutional [24], recurrent [42], or graph-convolutional [11, 33] networks. Recently, self-supervised methods are becoming increasingly popular as they can learn motion semantics without knowledge of labels using reconstruction-based [39, 47] or contrastive-based learning [21, 46].

The trained architectures can serve as *motion encoders* that express the motion semantics by a high-dimensional *feature* vector extracted from the last hidden network layer. This concept can be transferred to the *motion retrieval* task to support content-based access based on the *query-by-example* paradigm [5, 38, 40], which aims at identifying the database motions that are the most similar to a user-defined query motion. Besides balancing descriptiveness and indexability of the motion features, the most critical issue is to specify a convenient query motion example. The example can be selected from available skeleton sequences [39], drawn in a

visualization-driven graphical user interface [4], modeled by puppet interfaces [31], specified as a set of logical constraints [18], or artificially generated [10]. However, such a query example may not ever exist, or its construction requires professional modeling skills. This paper focuses on motion retrieval but simplifies query specification by enabling users to formulate a query by free text.

With the current advances in cross-modal learning, especially in the field of textual-visual processing, the trend is to learn common multi-modal spaces [28] so that similar images can be described and searched with textual descriptions [27]. A representative example is the CLIP model [36], which learns an effective common space for the visual and textual modalities. This allows the use of open vocabularies or complex textual queries for searching images.

Our work has many analogies with the text-to-video retrieval task [13, 22, 23, 41, 50], given that the moving skeleton also evolves in space and time. Despite the popularity of such powerful and versatile text-vision models, no effort has been made for the skeleton-data modality. Differently from video data, the skeleton is anonymized and avoids learning many common biases present in video datasets. To the best of our knowledge, there is only one approach [20] that relates to text-to-motion matching. However, it uses pre-training and tackles only the classification task. A few available datasets providing the training data for text-to-motion retrieval – e.g., the KIT Motion Language [35] and recently-released HumanML3D [15] datasets – are primarily used for motion generation from a textual description [16, 34, 44, 47, 48], where the idea is to align text and motion embeddings into a common space, but never explicitly handling the *text-to-motion* retrieval task.

Contributions of this Paper

We tackle the above-mentioned gap by introducing a novel *text-to-motion* retrieval task, which aims at searching databases of skeleton sequences and retrieving those that are the most relevant to a detailed textual query. For this task, we define evaluation metrics, establish new qualitative baselines, and propose the first text-to-motion retrieval approach. These initial contributions can be employed for future studies on this challenging yet unexplored task.

Specifically, one of the main paper contributions is the proposal of a fair baseline by adopting promising (1) *motion encoders* already employed as backbones in other motion-related tasks and (2) *text encoders* successfully applied in natural language processing (NLP) and text-to-image retrieval. The core of this baseline is a two-stream pipeline where the motion and text modalities are processed by separate encoders. The obtained representations are then projected into the same common space, for which a metric is learned in a similar way as in CLIP [36] or ALADIN [30] in the text-to-image scenario. The choice of a two-stream pipeline is strategic to make the approach scalable to large motion collections, as feature vectors extracted from both modalities can be easily stored in off-the-shelf indexes implementing efficient similarity search access.

Inspired by recent advances in video processing [3], we also propose a transformer-based motion encoder – the Motion Transformer (MoT) – that employs divided space-time attention on skeleton joints. We show that MoT reaches competitive results with respect to a state-of-the-art motion encoder, DG-STGCN [11], on both KIT Motion Language and HumanML3D datasets.

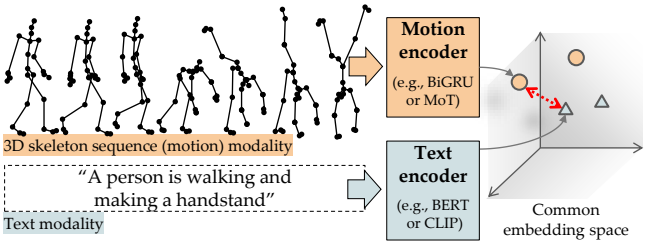


Figure 2: Schematic illustration of the learning process of the common space of both the text and motion modalities.

2 TEXT-TO-MOTION RETRIEVAL PIPELINE

The main idea of our approach is to rely on a two-stream pipeline, where motion and text features are first extracted through ad-hoc encoders and then projected into the same common space, as schematically illustrated in Figure 2. In this section, we sketch the whole pipeline which consists of the: (i) text encoder, (ii) motion encoder, and (iii) loss function used to optimize the common space.

2.1 Text Encoders

Inspired by recent works in NLP, we rely on two pre-trained textual models, namely BERT [19] and the textual encoder from CLIP [36].

BERT. We use the implementation from [14], which performed the task of motion synthesis conditioned on a natural language prompt. This model stacks together a BERT pre-trained module and an LSTM model composed of two layers for aggregating the BERT output tokens, producing the final text embedding. We take the final hidden state of the LSTM model as our final sentence representation. As in [14], the BERT model is fixed. At training time, we only update the LSTM weights.

CLIP. It is a recently-introduced vision-language model trained in a contrastive manner for projecting images and natural language descriptions in the same common space [36]. Here, we use the textual encoder of CLIP, which is composed of a transformer encoder [45] with modifications introduced in [37], and employs lower-cased byte pair encoding (BPE) representation of the text. We then stack an affine projection to the CLIP representation, which – similarly to the BERT+LSTM case – is the only layer to be trained.

2.2 Motion Encoders

Differently from the textual pipeline, which takes as input an unstructured natural language sentence, the input to motion encoder models is a vector $\mathbf{x} \in \mathbb{R}^{T \times J \times D}$, where T is the time length of the motion, J is the number of joints of the human-body skeleton, and D is the number of features used to encode each joint.

Bidirectional GRU. This architecture is widely adopted in time-series processing, and an early variant that used LSTM was applied to frame-level action detection in continuous motion data [6]. In particular, we first increase the dimensionality of the input – which is $D = 9$ in our case – by using a two-layer feed-forward network (FFN) before feeding it into the GRU: $\overrightarrow{\mathbf{z}}, \overleftarrow{\mathbf{z}} = \overleftrightarrow{\text{GRU}}(\text{FFN}(\mathbf{x}))$. Then, we compute the final motion embedding by concatenating the representations $\overrightarrow{z}$ and $\overleftarrow{z}$.

Upper-Lower GRU. To better learn semantics of different body parts, we adopt the model in [14] to independently process the upper and lower parts of the skeleton using two GRU layers.

DG-STGCN. This architecture [11] recently reached state-of-the-art results in motion classification. Their GCN module features a spatial module, built of affinity matrices to capture dynamic graphical structures, and a temporal module that performs temporal aggregation using group-wise temporal convolutions. We refer the reader to the original formulation [11] for further details.

MoT. Our proposed architecture that we built on top of the successful transformer-based video processing network ViViT [3]. In the original implementation, which processes a sequence of frames, the dimension J is the number of grid-arranged rectangular patches from each frame. In our case, instead, the spatial features come from the joints. Instead of using as J all individual skeleton joints, we first aggregate them obtaining features for five different body parts, similar to the pre-processing performed in Upper-Lower GRU. In this way, $J = 5$, which is far less than the total number of skeleton joints. This is beneficial from a computational point of view, and we found that this solution also reaches the best performance.

2.3 Optimization

We explore two widely-adopted metric learning loss functions, namely the symmetric triplet loss widely used in text-to-image [26] and the InfoNCE Loss, introduced for cross-modal matching in [49] and employed in CLIP [36] and recent cross-modal works [23]. We assume $(\mathbf{m}_i, \mathbf{c}_i)$ is the i-th motion and caption embedding pair, $S(\cdot,\cdot)$ is the cosine similarity, and B is the batch size.

The symmetric triplet loss is defined as:

$$\frac{1}{B}\sum_{i}^{B} \max_{j,j\neq i}[\alpha + S(\mathbf{m}_i,\mathbf{c}_j) - S(\mathbf{m}_i,\mathbf{c}_i)]_+ + \max_{j,j\neq i}[\alpha + S(\mathbf{m}_j,\mathbf{c}_i) - S(\mathbf{m}_i,\mathbf{c}_i)]_+$$

where $[x]_+ \equiv max(0, x)$ and α is a fixed margin. The index j identifies the hardest negative of the element with index i.

Info-NCE is basically a symmetric cross-entropy loss, defined as:

$$-\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(S(\mathbf{m}_i,\mathbf{c}_i)/\tau)}{\sum_j^B \exp(S(\mathbf{m}_i,\mathbf{c}_j)/\tau)} + \log \frac{\exp(S(\mathbf{m}_i,\mathbf{c}_i)/\tau)}{\sum_j^B \exp(S(\mathbf{m}_j,\mathbf{c}_i)/\tau)}$$

where τ is a temperature parameter learned during training.

3 EXPERIMENTAL EVALUATION

3.1 Metrics

Exact-search. Exact-search metrics leverage the intrinsic ground truth available in the employed datasets, where motions come with one (or more) textual descriptions. We can consider motions associated with the given textual query as the *exact* solutions, while all the other ones as irrelevant by default. In this context, the **recall@k** measures the percentage of queries that find the correct result within the first k elements in the results list, while the median and mean ranks represent the median and mean rank of the exact result computed among all the queries.

Relevance-based. There can exist motions relevant to a certain extent to the given textual query that are not paired in the dataset. In this context, the normalized Discounted Cumulative Gain (nDCG) metric is widely employed. The DCG takes into consideration the *relevance* a specific item has with the query, discounting it with a logarithmic factor that depends on the rank of that item: $\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{\mathrm{rel}_i}-1}{\log_2(i+1)}$. The nDCG normalizes DCG by its maximum theoretical value and thus returns values in the $[0, 1]$ range. We define the relevance similarly to previous works in image-to-text retrieval [7, 26, 29], that use a proxy relevance between textual descriptions, which is much easier to compute. In this work, we use two textual relevance functions: (i) the `SPICE` relevance [2] – a handcrafted relevance that exploits graphs associated with the syntactic parse trees of the sentences and has a certain degree of robustness against synonyms; and (ii) the `spaCy` relevance obtained from the *spaCy* Python tool, which implements a deep learning-powered similarity score for pairs of texts.

3.2 Datasets and Evaluation Protocol

We employ two recently introduced datasets, HumanML3D [15] and KIT Motion Language [35]. Both datasets carry one or more human-written descriptions for each motion. We employ the same pre-processing pipeline for both datasets – the one developed in the codebase of the HumanML3D dataset [15]. We employ $D = 9$ features to represent each joint: six features encoding continuous rotation representation plus three features encoding rotation-invariant forward kinematics joint positions.

KIT Motion-Language Dataset contains 3,911 recordings of full-body motion in the Master Motor Map form [43], along with textual descriptions for each motion. It has a total of 6,278 annotations in English, where each motion recording has one or more annotations that explain the action, like *"A human walks two steps forwards, pivots 180 degrees, and walks two steps back".*

HumanML3D is, in its essence, very similar to KIT Motion Language Dataset. However, it is a more recent dataset developed by adding textual annotations to already-existing and widely-used motion-capture datasets – AMASS [25] and HumanAct12 [17]. It contains 14,616 motions annotated by 44,970 textual descriptions.

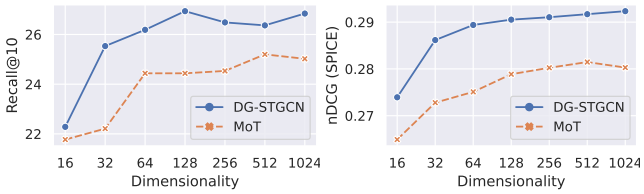
The results are reported on the test set of the respective datasets after removing possibly redundant queries. In particular, we use 938 and 8,401 textual queries to search among 734 and 4,198 motions for the KIT and HumanML3D datasets, respectively. For HumanML3D, these motions are obtained by splitting the originally provided ones using the available segment annotations associating a motion subsequence with the text that describes it. In this sense, HumanML3D enables a finer retrieval, as texts are more likely to describe the correct subsequence instead of the whole motion.

3.3 Results

We report text-to-motion retrieval results in Table 1, obtained with the InfoNCE loss (see Section 3.3.1 for a comparison of loss functions). The best results are competitively achieved by both DG-STGCN and our transformer-based MoT. The first remarkable insight is the superiority of CLIP over the BERT+LSTM on all the metrics in both datasets. With CLIP, the effectiveness of DG-STGCN and MoT over GRU-based methods is evident, especially on the KIT dataset, where the mean rank is almost 30 % lower. The nDCG metric, through the highly-semantic text-based relevance scores, confirms the trend of the recall@k values, suggesting that the CLIP model paired with GCNs and Transformers can both retrieve exact

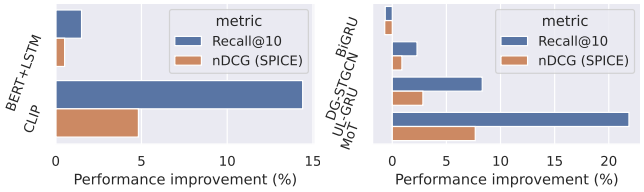
Table 1: Text-to-motion retrieval results on both the KIT Motion Language Dataset and HumanML3D Dataset. We report the best and the second-best results with bold and underlined font, respectively.

		KIT Motion Language Dataset							HumanML3D Dataset						
		Recall@k ↑			Rank ↓		nDCG ↑		Recall@k ↑			Rank ↓		nDCG ↑	
Text Model	**Motion Model**	r1	r5	r10	mean	med	SPICE	spaCy	r1	r5	r10	mean	med	SPICE	spaCy
BERT+LSTM	BiGRU	3.7	15.2	23.8	72.3	30	0.271	0.706	2.9	11.8	19.8	253.9	55	0.250	0.768
	UpperLowerGRU	3.2	15.7	25.3	90.2	34	0.263	0.697	2.4	10.5	17.7	285.7	68	0.242	0.763
	DG-STGCN	6.2	24.5	38.2	40.6	17	0.339	0.740	2.0	8.4	14.4	242.0	73	0.231	0.767
	MoT	5.3	21.3	32.0	51.1	20	0.318	0.723	2.5	11.2	19.4	234.5	51	0.247	0.768
CLIP	BiGRU	6.6	21.5	32.3	52.0	22	0.316	0.729	3.4	14.3	23.1	201.9	43	0.272	0.780
	UpperLowerGRU	6.4	22.0	32.2	52.3	22	0.321	0.732	3.1	12.6	20.8	200.4	47	0.269	0.779
	DG-STGCN	**7.2**	26.1	38.2	36.9	16	**0.355**	**0.751**	**4.1**	**16.0**	**26.5**	**159.6**	**33**	**0.291**	**0.789**
	MoT	6.5	**26.4**	**42.6**	**35.5**	**14**	0.352	0.748	3.5	14.8	24.5	166.2	38	0.280	0.785



(a) Recall@10 **(b) nDCG (SPICE)**

Figure 3: Performance varying space dimensionality.



(a) Text models dependence **(b) Motion models dependence**

Figure 4: Improvement of InfoNCE loss over Triplet loss.

and relevant results in earlier positions in the results list. Notably, from an absolute perspective, all the methods reach overall low performance on exact search, confirming the difficulty of the introduced text-to-motion retrieval task. This may be due to (i) some intrinsic limitations that are hard to eliminate – e.g., textual descriptions are written by annotators by possibly looking at the original video, which the network has no access to – or (ii) difficulties in capturing high-level semantics in motion or text data. In Figure 1, we report two qualitative examples of text-to-motion retrieval using CLIP + MoT, on HumanML3D. We can notice the potential of such natural-language-based approach to motion retrieval. Specifically, note how the approach is sensible to asymmetries – in the first case, where the *counterclockwise* adjective is specified in the query, only the correctly-oriented motions are returned in the first positions; in the second case, where no *right* or *left* is specified, both the original and mirrored motions are returned (e.g., the 1st and 2nd results).

3.3.1 Ablation Study on Loss Function and Space Dimensionality. In Figure 3, we report performance when varying the dimensionality of the common space, for the two motion models DG-STGCN and MoT employing the CLIP text model. We can notice how, on both metrics in Figure 3a/3b, the effectiveness remains quite high even for very small dimensions of the common space, with a negligible improvement after 256 dimensions. Specifically, with only 16 dimensions instead of 256, the performance drops by only about 6 % on nDCG with SPICE relevance and on average 15 % on Recall@10, considering both motion encoders. This suggests that the intrinsic dimensionality of the learned space is quite small, opening the way for further studies and feature visualization in future works.

In Figure 4, we also report the remarkable performance gain achieved by InfoNCE loss over the standard symmetric triplet loss. We can see how the InfoNCE loss induces the best results basically in all the configurations, confirming its power even in the under-explored text-motion joint domain. Breaking down the contributions of this variation on the text and motion models in Figures 4a and 4b respectively, we notice how the best gains are achieved by using the CLIP textual model and the MoT motion model.

4 CONCLUSIONS

In this paper, we introduced the task of *text-to-motion* retrieval as an alternative to the *query-by-example* search, and inherently different from the searching using a query label from a fixed pool of labels. We employed two state-of-the-art text-encoder networks, as well as widely adopted motion-encoder networks, for learning a common space and producing the first baselines for this novel task. We demonstrated that the CLIP text encoder works best also for encoding domain-specific natural sentences inherently different from image-descriptive ones, and that Transformers and GCNs obtain better motion representation than GRU-based encoders. In future works, we plan to train the models jointly on the two datasets and perform some cross-dataset evaluation to measure their generalization abilities and robustness. Other improvements include the use of video modality other than the motion and some unsupervised pre-training methods for boosting performance.

ACKNOWLEDGMENTS

This research was supported by ERDF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822), by AI4Media – A European Excellence Centre for Media, Society, and Democracy (EC, H2020 No. 951911), and by SUN – Social and hUman ceNtered XR (EC, Horizon Europe No. 101092612).

REFERENCES

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2020. A Spatio-temporal Transformer for 3D Human Motion Prediction. *arXiv* (2020). https://doi.org/10.48550/ARXIV.2004.08692

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, 382–398.

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 6836–6846.

[4] J. Bernard, N. Wilhelm, B. Krüger, T. May, T. Schreck, and J. Kohlhammer. 2013. MotionExplorer: Exploratory Search in Human Motion Capture Data Based on Hierarchical Aggregation. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2257–2266.

[5] Petra Budikova, Jan Sedmidubsky, and Pavel Zezula. 2021. Efficient Indexing of 3D Human Motions. In *International Conference on Multimedia Retrieval (ICMR)*. ACM, 10–18. https://dl.acm.org/doi/10.1145/3460426.3463646

[6] Fabio Carrara, Petr Elias, Jan Sedmidubsky, and Pavel Zezula. 2019. LSTM-based real-time action detection and prediction in human motion streams. *Multimedia Tools and Applications* 78, 19 (2019), 27309–27331. https://doi.org/10.1007/s11042-019-07827-3

[7] Fabio Carrara, Andrea Esuli, Tiziano Fagni, Fabrizio Falchi, and Alejandro Moreo Fernández. 2018. Picture it in your mind: Generating high level visual representations from textual descriptions. *Information Retrieval Journal* 21, 2 (2018), 208–229.

[8] Yi-Bin Cheng, Xipeng Chen, Junhong Chen, Pengxu Wei, Dongyu Zhang, and Liang Lin. 2021. Hierarchical Transformer: Unsupervised Representation Learning for Skeleton-Based Human Action Recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME51207.2021.9428459

[9] Yi-Bin Cheng, Xipeng Chen, Dongyu Zhang, and Liang Lin. 2021. Motion-Transformer: Self-Supervised Pre-Training for Skeleton-Based Action Recognition. In *2nd ACM International Conference on Multimedia in Asia (MMAsia)*. ACM, New York, NY, USA. https://doi.org/10.1145/3444685.3446289

[10] Z. Deng, Q. Gu, and Q. Li. 2009. Perceptually consistent example-based human motion retrieval. In *Symposium on Interactive 3D Graphics (SI3D)*. ACM, 191–198.

[11] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. 2022. DG-STGCN: Dynamic Spatial-Temporal Modeling for Skeleton-based Action Recognition. *arXiv* (2022). https://doi.org/10.48550/ARXIV.2210.05895

[12] Shradha Dubey and Manish Dixit. 2022. A comprehensive survey on human pose estimation approaches. *Multimedia Systems* (2022), 1–29. https://doi.org/10.1007/s00530-022-00980-0

[13] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).

[14] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1396–1406.

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

[16] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *European Conference on Computer Vision (ECCV)*. Springer Nature Switzerland, Cham, 580–597.

[17] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2021–2029.

[18] M. Kapadia, I-K. Chiang, T. Thomas, N.I. Badler, and J. T. Kider Jr. 2013. Efficient motion retrieval in large motion databases. In *Symposium on Interactive 3D Graphics and Games (I3D)*. ACM, 19–28.

[19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[20] Jihoon Kim, Youngjae Yu, Seungyoun Shin, Taehyun Byun, and Sungjoon Choi. 2022. Learning Joint Representation of Human Motion and Language. *arXiv preprint arXiv:2210.15187* (2022).

[21] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. 2020. MS2L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. In *28th ACM International Conference on Multimedia (MM)*. ACM, New York, NY, USA, 2490–2498. https://doi.org/10.1145/3394171.3413548

[22] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. 2022. Animating Images to Transfer CLIP for Video-Text Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1906–1911.

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[24] Na Lv, Ying Wang, Zhiquan Feng, and Jingliang Peng. 2021. Deep Hashing for Motion Capture Data Retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2215–2219. https://doi.org/10.1109/ICASSP39728.2021.9413505

[25] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.

[26] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–23.

[27] Nicola Messina, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–6.

[28] Nicola Messina, Davide Alessandro Coccomini, Andrea Esuli, and Fabrizio Falchi. 2022. Transformer-Based Multi-modal Proposal and Re-Rank for Wikipedia Image-Caption Matching. *arXiv preprint arXiv:2206.10436* (2022).

[29] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2021. Transformer reasoning network for image-text matching and retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 5222–5229.

[30] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*. 64–70.

[31] N. Numaguchi, A. Nakazawa, T. Shiratori, and J. K. Hodgins. 2011. A Puppet Interface for Retrieval of Motion Capture Data. In *Eurographics/ACM SIGGRAPH Symposium on Computer Animation (SCA)*. Eurographics Assoc., 157–166.

[32] Konstantinos Papadopoulos, Enjie Ghorbel, Renato Baptista, Djamila Aouada, and Björn E. Ottersten. 2019. Two-Stage RGB-Based Action Detection Using Augmented 3D Poses. In *18th International Conference on Computer Analysis of Images and Patterns (CAIP)*, Vol. 11678. Springer, 26–35. https://doi.org/10.1007/978-3-030-29888-3_3

[33] Wei Peng, Xiaopeng Hong, and Guoying Zhao. 2021. Tripool: Graph Triplet Pooling for 3D Skeleton-Based Action Recognition. *Pattern Recognition* 115 (2021), 107921. https://doi.org/10.1016/j.patcog.2021.107921

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis With Transformer VAE. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 10985–10995.

[35] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data* 4, 4 (2016), 236–252. https://doi.org/10.1089/big.2016.0028

[36] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* (2021). https://doi.org/10.48550/ARXIV.2103.00020

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[38] Jan Sedmidubsky, Petra Budikova, Vlastislav Dohnal, and Pavel Zezula. 2020. Motion Words: A Text-like Representation of 3D Skeleton Sequences. In *42nd European Conference on Information Retrieval (ECIR)*. Springer, 527–541.

[39] Jan Sedmidubsky, Fabio Carrara, and Giuseppe Amato. 2023. SegmentCodeList: Unsupervised Representation Learning for Human Skeleton Data Retrieval. In *45th European Conference on Information Retrieval (ECIR)*. Springer, Cham, 110–124. https://doi.org/10.1007/978-3-031-28238-6_8

[40] Jan Sedmidubsky, Petr Elias, Petra Budikova, and Pavel Zezula. 2021. Content-based Management of Human Motion Data: Survey and Challenges. *IEEE Access* 9 (2021), 64241–64255. https://doi.org/10.1109/ACCESS.2021.3075766

[41] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20020–20029.

[42] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2018. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Transactions on Image Processing* 27, 7 (2018), 3459–3471. https://doi.org/10.1109/TIP.2018.2818328

[43] Ömer Terlemez, Stefan Ulbrich, Christian Mandery, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. 2014. Master Motor Map (MMM)—Framework

and toolkit for capturing, representing, and reproducing human motion on humanoid robots. In *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 894–901.

[44] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. *arXiv* (2022), 1–12. https://doi.org/10.48550/ARXIV.2209.14916

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[46] Yang Yang, Guangjun Liu, and Xuehao Gao. 2022. Motion Guided Attention Learning for Self-Supervised 3D Human Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–13. https://doi.org/10.1109/TCSVT.2022.3194350

[47] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. *arXiv* (2023), 1–14. https://doi.org/10.48550/ARXIV.2301.06052

[48] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv* (2022), 1–16. https://doi.org/10.48550/ARXIV.2208.15001

[49] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747* (2020).

[50] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 970–981.