Chronologically Accurate Retrieval for Temporal Grounding of Motion-Language Models

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu

LY Corporation, Tokyo, Japan
{kent.fujiwara, mikihiro.tanaka, yu.qing}@lycorp.co.jp

Abstract. With the release of large-scale motion datasets with textual annotations, the task of establishing a robust latent space for language and 3D human motion has recently witnessed a surge of interest. Methods have been proposed to convert human motion and texts into features to achieve accurate correspondence between them. Despite these efforts to align language and motion representations, we claim that the temporal element is often overlooked, especially for compound actions, resulting in chronological inaccuracies. To shed light on the temporal alignment in motion-language latent spaces, we propose Chronologically Accurate Retrieval (CAR) to evaluate the chronological understanding of the models. We decompose textual descriptions into events, and prepare negative text samples by shuffling the order of events in compound action descriptions. We then design a simple task for motion-language models to retrieve the more likely text from the ground truth and its chronologically shuffled version. CAR reveals many cases where current motion-language models fail to distinguish the event chronology of human motion, despite their impressive performance in terms of conventional evaluation metrics. To achieve better temporal alignment between text and motion, we further propose to use these texts with shuffled sequence of events as negative samples during training to reinforce the motion-language models. We conduct experiments on text-motion retrieval and text-to-motion generation using the reinforced motion-language models, which demonstrate improved performance over conventional approaches, indicating the necessity to consider temporal elements in motion-language alignment.

Keywords: Human motion analysis · Motion-language model · Text-motion retrieval · Text-motion generation · Motion chronology

1 Introduction

Modeling human motion has become an important task in computer vision, especially with the emergence of applications for animating gaming characters and online avatars. To deal with the complex nature of human motion, natural language is gaining popularity as a medium to describe skeletal human motion, leading to 3D human skeletal motion datasets annotated with textual descriptions. This has led to proposals of novel tasks that attempt to establish a joint

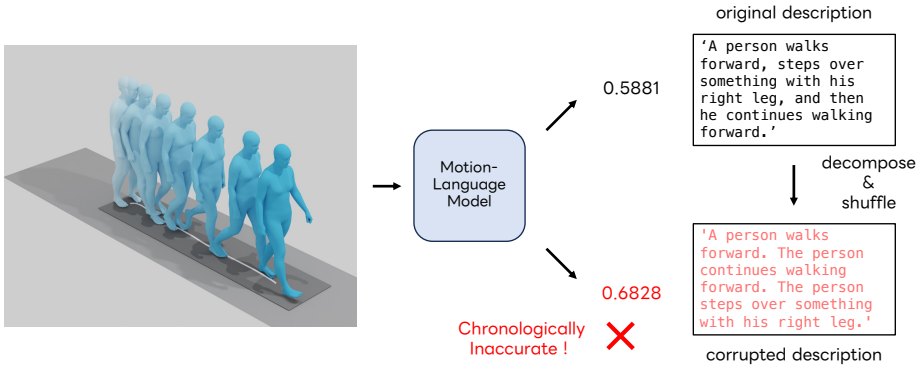


Fig. 1: Overview of Chronologically Accurate Retrieval test. Given a motion sequence, motion-language models trained on text-motion datasets are asked to retrieve the more relevant text from the ground truth and its shuffled version. Original texts are decomposed into events by off-the-shelf Large Language Models, which are randomly shuffled.

latent space for motion and language, including motion-text retrieval [27,38,43], and motion generation from texts [1,6,9,14,26,39,46,47].

Although various methods [27,38,43] demonstrate impressive performance in this field, we claim that the temporal element is not fully considered in most of the motion-language models. Establishing temporal alignment between motion and text becomes important when there are multiple events in a motion sequence, especially in motion recognition and generation, where the order of events needs to be accurate. However, there is currently no method that explicitly evaluates whether temporal alignment between motion and text is achieved or not.

To reveal the performance of the current state-of-the-art motion-language model in terms of understanding the chronology of events, we propose a simple test, which we call Chronologically Accurate Retrieval (CAR). This test assesses whether the motion-language model is capable of recognizing the order of events. We decompose the original motion descriptions into events, which are then shuffled to create chronologically inaccurate descriptions. We test the model to see whether it can retrieve the chronologically accurate description. The overview of CAR is shown in Fig. 1. CAR reveals that the current model falls short of understanding the temporal component of motion and descriptions. We also combine the state-of-the-art model with various language models to observe the tendencies among language models when used in motion-language models.

To address the issue of temporal alignment between motion and text, we further propose a simple strategy to refine the motion-language model. We enhance the current contrastive learning framework with the chronologically negative text samples. We employ the description derived from shuffled events as an incorrect description of the original motion, and append these shuffled descriptions as negative text samples. Using these negative samples, the motion-language model is trained to differentiate chronologically accurate descriptions from the incorrect ones, to achieve stronger temporal alignment between the two modalities.

We evaluate the refined motion-language model through text-motion retrieval and motion generation from texts. The results reveal that our proposal enhances the performance of the motion-language model in both of these tasks.

The contributions of our research are as follows:

- We propose to evaluate the temporal alignment between motion and language through Chronologically Accurate Retrieval (CAR), a novel test that measures how accurate models can differentiate original motion descriptions from chronologically incorrect ones given a motion.
- We reveal that current motion-language models fail to fully comprehend the temporal component of motion and language through CAR, even when larger language models are introduced to the motion-language model.
- We propose a simple solution to achieve better temporal correspondence between language and motion, where shuffled event descriptions are used as negative text samples to train and refine a motion-language model through contrastive learning. The resulting motion-language model achieves high performance in both text-motion retrieval and motion generation from text.

2 Related Work

2.1 Motion-Language Models

With the recent progress in cross-modal representation and the success of recent language models, language now plays a significant role in computer vision tasks. A vast amount of image data annotated with textual descriptions is employed in tasks such as image captioning [18, 19], segmentation [15], and recognition [30]. The development of diffusion models [7,32] has also led to an explosion of interest in attempting to generate images from textual descriptions [32].

The success of the cross-modal analysis in the image domain has recently garnered interest in establishing a cross-modal model connecting language and 3D human motion. Earlier research used 3D human motion datasets solely for computer vision tasks, such as action recognition [35], pose estimation [12], and motion prediction [21]. However, there has also been some interest in trying to annotate human motion with textual descriptions, as text is a convenient medium to describe the complex and varying nature of human motion. KIT-ML [29] is one of the first datasets to add textual description to 3D human motion data captured using motion capture devices. It is a collection of human motions consisting mainly of locomotion-related actions. Recently, HumanML3D [10] was released with textual annotations added to a large collection of motion data consisting of AMASS [23] and HumanAct12 [11] datasets. There are other small-scale motion datasets focusing on specific tasks, such as hand-object manipulation [37]. There are also attempts to enhance these datasets by capturing 3D human motion from videos and annotating the descriptions [20].

With the release of these larger-scale, motion-language datasets, there are various ongoing research efforts that attempt to capitalize on such datasets, as well as the capability of recent large language models (LLMs) [3, 40]. As skeletal

human motion data focuses solely on human motion without any background information, motion recognition [8, 36, 41] is one of the most popular research topics. There are tasks that attempt to classify what the motion is representing, which has recently evolved into retrieval tasks [27], where given a text or motion, the task is to find the most appropriate counterpart. Similar to the image domain, generating novel motion from semantic description [26,39,46,47] has also become a popular task. Beginning from action labels [25], the models have evolved to generate appropriate motions corresponding to given textual descriptions.

All of these tasks require establishing a robust motion-language latent space, where similar descriptions are located close to its corresponding motion representations. Despite the recent advances, there has been relatively little interest in exploring the temporal aspect of motions. Some methods [24, 42] attempt to localize the weakly assigned action labels to segments within the motion sequence to discover what actions are taking place where. Other attempts [28,34] have proposed to generate compound motions by connecting different actions smoothly. However, this does not lead to a common motion-language latent space, which is capable of representing the temporal component between language and motion. In this paper, we focus on the temporal alignment between text and motion, and demonstrate how the current motion-language models either misinterpret or overlook the temporal component of motion and language.

2.2 Alignment of Cross-Modality Latent Space

As previously stated, it is crucial for cross-modal models to establish accurate relationships between different representations. Achieving alignment between language and other modalities has become one of the major focal points of ongoing research. For example, in the image domain, some research has attempted to align text with images [5] so that accurate correspondence is achieved between the given caption and the generated image. These models are used for various downstream tasks such as segmentation, as the attention corresponding to the text input in the generation process can be interpreted as the region of interest [4]. Some studies also tried to discover misalignment between the modalities. One work of research [44] identified flaws in the popular CLIP model [30], where shuffling subject and object or other various elements in the texts did not lead to a substantial decline in the performance of the model.

There are also attempts to align text with sequential data. For example, in the field of video analysis, where certain events are localized within videos, some methods [13, 16, 45] attempt to identify the proper chronological correspondence between textual descriptions of events and videos. In the field of audio data analysis, there are also attempts to use language models to provide accurate descriptions to generate the corresponding sound sequences [2, 17]. In the field of human motion, as there is a lack of data that has frame-wise annotations indicating what actions are taking place at each moment, the temporal component of motions is currently being overlooked. Inspired by the recent research in these related fields, we uncover the deficiencies of the current motion-language models, and propose a simple solution to alleviating the temporal alignment issue.

3 Temporal Element in Motion-Language Latent Space

We first focus on analyzing the temporal element in motion-language latent space established by the state-of-the-art motion-language model. We choose the task of text-to-motion/motion-to-text retrieval to assess the alignment of the two modalities. We employ TMR [27], which is a recently proposed text-motion retrieval model that uses contrastive learning and a generative model to establish the correspondence between features of texts and human motion sequences.

3.1 Text-Motion Retrieval

Given a text description T, the goal of text-to-motion retrieval is to find a 3D skeletal motion sequence M that most closely resembles the provided text. Motion-to-text retrieval conducts the opposite task, where the closest textual description is retrieved when given a motion sequence.

TMR proposes to achieve this by following CLIP [30], training a model to learn a function $f(z^T, z^M)$ that provides the similarity between the outputs of motion and textual encoders. Here, z^T is a language feature obtained from a textual encoder, and z^M is the motion feature from a motion encoder. To train this model, given N samples of text-motion feature pairs $(z_1^T, z_1^M) \dots (z_N^T, z_N^M)$, the method calculates cosine similarities between all the combinations, $S_{ij} = (z_i^T, z_j^M)$, to establish a similarity matrix of features $\mathbf{S} \in \mathbb{R}^{N \times N}$. The similarity matrix is used to optimize the contrastive loss:

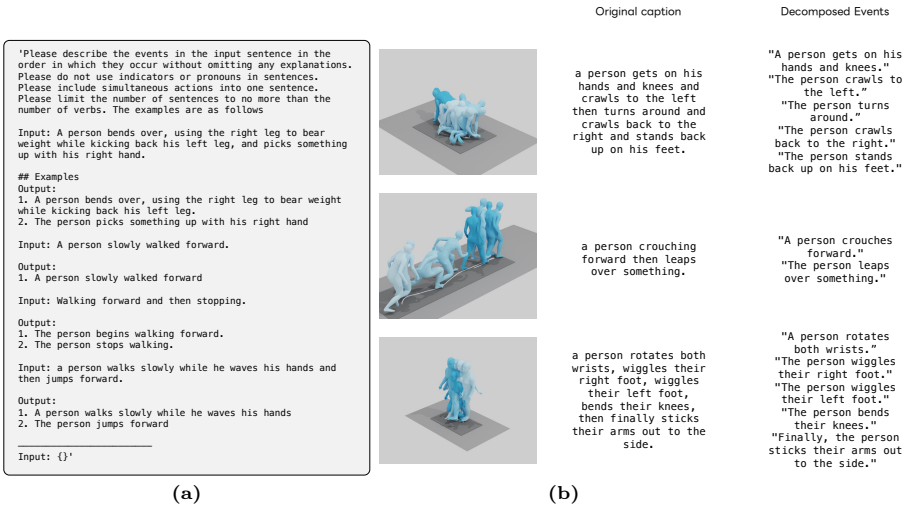
$$\mathcal{L} = -\frac{1}{2N} \sum_i \left(\log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ij}/\tau} + \log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ji}/\tau} \right), \tag{1}$$

where τ is a temperature parameter. With the support of additional constraints, the model learns to bring together motions and texts that correspond to each other, while returning low similarity values to the wrong pairs.

3.2 Chronologically Accurate Retrieval

TMR demonstrates robust performance in text-motion retrieval tasks. However, the analysis does not reveal whether the model is capable of understanding the temporal relationships between language and motion. This is due to the scale of the dataset, where the variety in the motion sequences does not necessarily consider chronological variations. Although motion datasets are usually augmented by mirroring the human body pose, motion augmentation by changing the order of actions is a challenging task, likely resulting in undesirable noise.

To analyze the chronological understanding of motion-language models, we focus on language and propose a simple task that tests whether the models understand the temporal element of the motion-language relationship. For this task, we propose to decompose the original textual descriptions provided for the motions into events contained in the description. We conduct the event decomposition using an off-the-shelf LLM, GPT3.5. The command prompt to



(a) **(b)**

Fig. 2: (a) Command prompt used to decompose motion captions into events. We provide the prompt as input to a Large Language Model, GPT3.5, and add the captions in the dataset to decompose descriptions into events. (b) Examples of original motion captions and decomposed events. Corresponding motions are shown on the left.

generate the decomposed events is shown in Fig. 2a. By presenting exemplar decompositions in the prompt, the LLM is able to decompose textual descriptions of motions into sequences of events, as presented in Fig. 2b.

We then form a pair of descriptions, consisting of the ground truth text and a chronologically incorrect description. The incorrect text is obtained by shuffling the events, if more than one, to generate a new description with a different chronological order. We then provide the motion-language model with a motion and calculate the similarity between the motion feature and the two descriptions, and choose the one with the higher likelihood to be the likely description for the motion. We call this task, Chronologically Accurate Retrieval (CAR) test.

3.3 Analysis

We conduct the Chronologically Accurate Retrieval test on the HumanML3D dataset [10]. The dataset consists of motions from AMASS [23] and HumanAct12 [11] motion datasets, and textual descriptions assigned to each of the motion sequences. Following prior works, we also augment the dataset by mirroring the human body pose and assigning them with the adjusted descriptions. The official dataset split consists of 23,384, 1,460, and 4,380 motion sequences for training, validation, and test sets, respectively. As each motion contains 3.0 descriptions on average, we select one random description during training and the first one for testing. For each description corresponding to each motion sequence, we apply the decomposition using LLM to obtain the event descriptions. Out of 4,380 test set motions, multiple events are found in 2,677 sequences. In

this section, we only use this test subset of multiple event sequences to conduct CAR. For this test, we prepare the chronologically inaccurate sample for each sequence in the test subset by shuffling the order of the events, and concatenating them as a single text input. Given the original texts $\{T_1, \dots T_K\}$ and the chronologically inaccurate samples $\{C_1, \dots C_K\}$, CAR is calculated as

$$CAR = \frac{1}{K} \sum_{i}^{K} g\left(f(z_i^T, z_i^M), f(z_i^C, z_i^M)\right) , \tag{2}$$

where z^C is the feature of the chronologically inaccurate text, and the function g returns 1 if $f(z_i^T, z_i^M)$ is larger, and 0 otherwise.

As decomposed events could have different structural characteristics from the original texts, which may lead to an undesirable domain gap, we prepare two scenarios. The first, "orig $\rightarrow$ event", is the case where original HumanML3D captions are used as is for input. During the test, the model is tasked with distinguishing the original texts in the test set from the shuffled event descriptions obtained from decomposing and shuffling the same text. The second, "event $\rightarrow$ event", uses decomposed event descriptions concatenated into one text in the correct order to train the model. For testing, the model is asked to compare the correctly concatenated events, and its shuffled version.

As language plays an important component in this test, we also enhance TMR by employing different textual encoders to see whether textual embedding has an effect on the resulting motion-language latent space. We compare DistilBERT [33], the default language model used in TMR, with CLIP [30], which is a popular choice in motion analysis. Additionally, we introduce two variants of T5 language models [31], t5-base and t5-large, to observe whether the size of models has an effect. The token embeddings from the language models are concatenated with the learnable parameters to be used in the VAE-style encoder.

We mostly follow the settings of the original TMR. For motion representation, we use the representation consisting of relative joint positions and accelerations employed by Guo et al. [11], as in the original TMR. We maintain the same network structure as the original method, using two VAE-based encoders to derive features for motion sequences and texts. A motion decoder is also used to decode the original motion from both the features of the corresponding motion and text. The weights balancing the losses to train TMR are also kept the same as the original. We set the batch size to 32 in order to achieve better results for all the patterns using different language models under the same environment, and employ the AdamW optimizer [22] with the learning rate set at 0.0001. The setting led to more accurate results for the original method as well. Further details can be found in the supplementary material.

We show the performance in terms of CAR in Table 1. The first 4 rows of the rightmost column of each scenario show CAR for TMR equipped with different language models. Despite its success in the retrieval of the HumanML3D dataset, the model struggles to differentiate the original text from decomposed and shuffled event descriptions, as most models achieve success around 60% of the time, only marginally above the chance level of 50%. CAR accuracy does not show

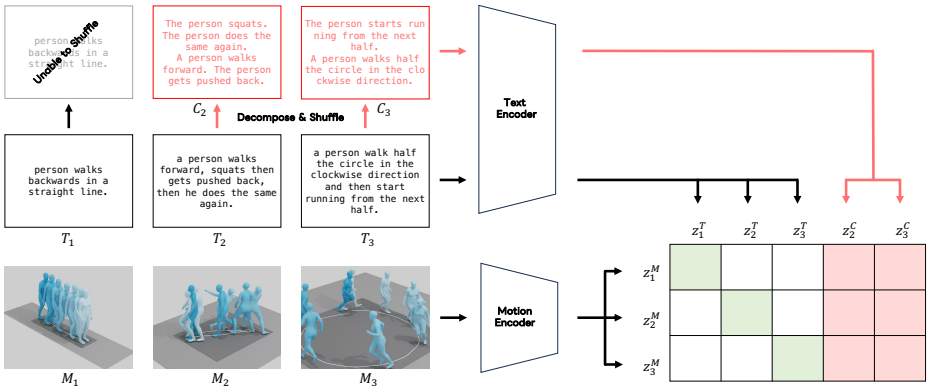


Fig. 3: Overview of the proposed contrastive learning scheme with chronological negative samples. We use the texts derived from shuffling the event order and employ them as negative text samples, corresponding to items indicated in pink.

much change in "event → event" scenario either. This indicates that the current motion-language models fail to capture the intricate temporal relationship between language and motion. This issue of temporal alignment must be addressed to further enhance the performance of future motion-language models.

4 Contrastive Learning with Shuffled Events

In this paper, we take the initiative and propose a simple solution to achieving better correspondence between textual descriptions and motions in terms of chronology. Inspired by the recent research that attempts to establish better correspondence between language and images in contrastive learning [44], we propose to reinforce the model by incorporating chronologically incorrect descriptions obtained by shuffling the decomposed events in the previous analysis.

4.1 Shuffled Events as Hard Negative Chronological Samples

Because the event-based shuffled descriptions are accurate in terms of the words used but not in terms of chronological order, they serve as crucial resource for motion-language models to recognize the inaccuracies in the chronology of events. We, therefore, propose to utilize these shuffled event descriptions as the hard negative samples corresponding to the original motion.

We consider a batch of N texts $\{T_1, \dots T_N\}$ and motions $\{M_1, \dots M_N\}$. Before encoding them into features and calculating the similarities between them, we select texts that contain more than one event and decompose them into event descriptions following the method described in Section 3. Here, we indicate the number of text-motion pairs that contain multiple events as K. The decomposed K texts are then shuffled randomly to obtain chronologically incorrect descriptions C_i, which corresponds to the original text T_i.

Table 1: Comparison of CAR accuracy and motion-to-text retrieval results with both the original and corrupted texts. We insert chronologically inaccurate texts as candidates for retrieval. Ours indicate models trained with the hard negative samples.

Method	Encoder	Train with Negatives	Motion-to-text retrieval									
			"orig → event"					"event → event"				
			R@1 ↑	R@5↑	R@10 ↑	MedR ↓	CAR	R@1 ↑	R@5↑	R@10 ↑	MedR ↓	CAR
TMR [27]	DistilBERT		7.89	19.82	28.40	33.75	64.81	6.93	17.08	26.23	36.00	65.04
	CLIP		6.18	16.54	24.48	43.00	63.17	6.73	17.63	26.89	40.25	62.87
	t5-base		6.71	17.63	28.15	35.00	66.42	6.14	15.58	24.02	36.00	63.91
	t5-large		7.76	20.12	29.36	34.00	66.72	5.86	16.40	25.84	35.00	66.83
Ours	DistilBERT	✓	9.38	**23.31**	**34.10**	**24.00**	99.33	**8.90**	**21.49**	**31.68**	**27.50**	92.90
	CLIP	✓	8.19	20.69	30.36	29.50	98.88	8.01	20.28	29.65	34.00	91.37
	t5-base	✓	8.55	23.08	33.10	25.00	**99.74**	7.6	20.14	30.13	30.00	91.78
	t5-large	✓	**9.65**	22.88	32.71	25.00	**99.74**	6.91	19.89	29.29	**27.50**	**93.09**

Once the shuffled texts $\{C_1, \dots C_K\}$ are obtained, we calculate the similarity matrix in the same manner as TMR and other CLIP-based contrastive learning approaches. However, instead of the similarity matrix being $\mathbf{S} \in \mathbb{R}^{N \times N}$, our modified similarity matrix is $\tilde{\mathbf{S}} \in \mathbb{R}^{N \times (N+K)}$, with additional columns corresponding to the shuffled text features z_i^C. We calculate the same row-wise and column-wise cross entropy loss, but exclude the column-wise loss for columns corresponding to the shuffled texts, as they do not have any corresponding motion. Therefore, the loss Eq. (1) becomes

$$\mathcal{L} = \mathcal{L}_{t2m} + \mathcal{L}_{m2t}\,, \tag{3}$$

where

$$\mathcal{L}_{t2m} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp \tilde{S}_{ii}/\tau}{\sum_j^N \exp \tilde{S}_{ij}/\tau}\,, \tag{4}$$

$$\mathcal{L}_{m2t} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp \tilde{S}_{ii}/\tau}{\sum_j^{(N+K)} \exp \tilde{S}_{ji}/\tau}, \tag{5}$$

and $\tilde{S}_{i,j}$ is the i,j-the element in the similarity matrix $\tilde{\mathbf{S}}$. This allows the model to use the shuffled events as negative samples, and train representations that differentiate chronologically corrupted descriptions from the true ones. An overview of the proposed scheme using shuffled texts as negative samples is shown in Fig. 3

4.2 Text-Motion Retrieval

To evaluate the effectiveness of our approach, we conduct the text-motion retrieval task by including the chronologically inaccurate texts as hard negative samples. We follow the same protocol as the previous experiment. We employ the same variations of TMR and train them using the additional negative samples.

CAR Accuracy. Firstly, we show whether the proposed training scheme achieves robustness in terms of chronological accuracy. The bottom half of the rightmost column of each scenario in Table 1 shows the CAR accuracy of the motion-language model trained by our proposed method. In both "orig → event" and "event → event", we employ the shuffled event descriptions as hard negative

samples to train the model in the aforementioned manner. Remarkably, all the models successfully learned to distinguish original texts from the chronologically shuffled versions in both of the scenarios at accuracy above 90%. The results demonstrate that our proposal to use chronologically shuffled events as negative samples works favorably towards understanding texts in terms of chronology.

Motion-to-text retrieval including shuffled texts. We additionally conduct a more challenging task of motion-to-text retrieval using all the texts, including the chronologically incorrect versions. Given a motion, the model retrieves the most similar text from all the descriptions including those generated by shuffling the event orders. The same models from the prior experiments are used in this comparison. We conduct the motion-to-text retrieval in both of the aforementioned scenarios. Text candidates for retrieval in "orig $\rightarrow$ event" includes the original text and the shuffled decomposed events, whereas in "event $\rightarrow$ event", the candidates are the event descriptions concatenated in the correct order, as well as the wrong order. Table 1 shows the results. The conventional motion-language model is not capable of distinguishing texts that resemble similar events with different chronological order in either of the scenarios. On the contrary, our training scheme allows all the methods to achieve high retrieval accuracy despite having more text candidates for retrieval, even when the texts are identical except for the order, as in "event $\rightarrow$ event" setting. This further demonstrates that the negative samples are effective at training the models to be better aware of the chronology of events in human motion sequences.

Fine-tuning language models. Experiments up to this point in this paper follow the settings and protocols proposed in TMR, where language embeddings are preprocessed. As the proposed training scheme involves enhancing the descriptive ability of the motion-language models, especially in terms of language, we also test whether fine-tuning the language model used to establish the motion-language latent space in our proposed scheme has any effect on the outcome.

Using the original model of TMR, we attach the aforementioned language models and enable fine-tuning of the language models. We then train the model by removing elements from the original model to observe the changes to the final retrieval results. We follow the same evaluation protocol in TMR to evaluate the performance of the models. Here, we rely on "All" criterion to assess the retrieval performance, where the entire test set is used for the retrieval task without any modification. We conduct the experiments in the "orig $\rightarrow$ event" setting. Results for other metrics, "All with Threshold", "Dissimilar Subset", "Small batches", and results from "event $\rightarrow$ event", are in the supplementary material.

Table 2 shows the results. Quite remarkably, by removing components from the original method, the proposed training scheme is able to better distinguish chronological differences in the description. Removing the VAE textual encoder especially played an important role. As chronologically inaccurate descriptions include expressions that are very similar in terms of vocabulary used, but differ significantly in terms of order, representing sentence features as a sample from a distribution likely has a detrimental effect at recognizing and comprehending chronology. We can also observe a trend that larger language models

Table 2: Comparison of retrieval results between the original TMR model and its variations equipped with different language encoders, which are further fine-tuned with our chronological negative samples. "Tune" indicates whether the language model is fine-tuned, "VAE" whether VAE feature is used, and "Rec." whether motion decoder is used to reconstruct the original poses.

Method	Encoder	Tune	VAE	Rec.	Text-to-motion retrieval				Motion-to-text retrieval			
					R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5↑	R@10 ↑	MedR ↓
TMR [27]	DistilBERT		✓	✓	5.82	21.33	32.76	25.00	9.76	24.13	33.23	23.50
Ours	DistilBERT	✓	✓	✓	5.25	19.96	30.98	29.00	8.69	22.45	31.89	28.00
		✓	✓		5.36	20.32	31.04	28.00	9.19	23.11	31.64	28.00
		✓		✓	6.11	22.08	32.78	24.00	10.26	23.54	33.42	24.00
		✓			6.55	22.99	34.60	22.00	11.18	25.52	36.38	21.50
	CLIP	✓	✓	✓	4.68	16.72	25.91	41.00	8.05	18.66	27.24	39.50
		✓	✓		3.99	16.49	27.44	32.00	6.57	19.59	28.44	31.50
		✓		✓	4.58	17.91	27.62	35.00	8.07	19.80	28.44	33.50
		✓			5.57	20.00	30.06	28.00	8.69	22.06	31.14	27.50
	t5-base	✓	✓	✓	6.39	22.81	34.63	23.00	10.56	26.21	36.09	22.50
		✓	✓		4.13	16.45	26.48	32.00	6.98	18.68	27.67	29.50
		✓		✓	6.98	24.98	36.75	19.00	10.90	27.35	38.02	19.50
		✓			5.50	21.62	33.42	22.00	9.88	24.77	35.36	21.50
	t5-large	✓	✓	✓	6.57	22.97	33.92	24.00	9.74	25.62	36.59	22.00
		✓	✓		4.65	18.84	31.75	24.00	8.69	23.72	33.85	23.50
		✓		✓	**8.03**	**26.73**	**38.98**	**17.00**	**11.72**	**28.15**	**39.23**	**17.50**
		✓			6.52	24.16	36.50	20.00	10.56	25.91	36.86	19.50

tend to perform better when performing fine-tuning. The additional textual information plays an important role for large language models in incorporating fine-grained details of chronological differences among event descriptions. These result demonstrate that by allowing both modalities to form better alignment with each other, the proposed training scheme allows the motion-language model to establish a more robust motion-language latent space.

Qualitative Analysis. To further analyze the resulting motion-language model, we visualize some results from the motion-to-text retrieval including the chronologically inaccurate descriptions to demonstrate the performance of the original TMR compared to the model trained using the negative samples. Fig. 4 shows the query motions along with the retrieval results and the ground truth captions. These concrete examples demonstrate that understanding the chronology of events can also lead to better alignment between text and motion.

4.3 Motion Generation from Textual Descriptions

To further evaluate the effectiveness of our proposal, we utilize the fine-tuned language model from our proposal in the task of motion generation from texts.

Quantitative Analysis. We first employ multiple human motion generation models to assess the descriptive capability of the language model fine-tuned through our proposal. In this analysis, we compare 3 recent models; Motiondiffuse [47], T2M-GPT [46], and ReMoDiffuse [48] as the base generative models. As all of these methods rely on CLIP text encoder to obtain textual features, we compare 3 variations. The first variant is the original version of each method, which uses the unaltered CLIP encoder. The second employs a CLIP encoder

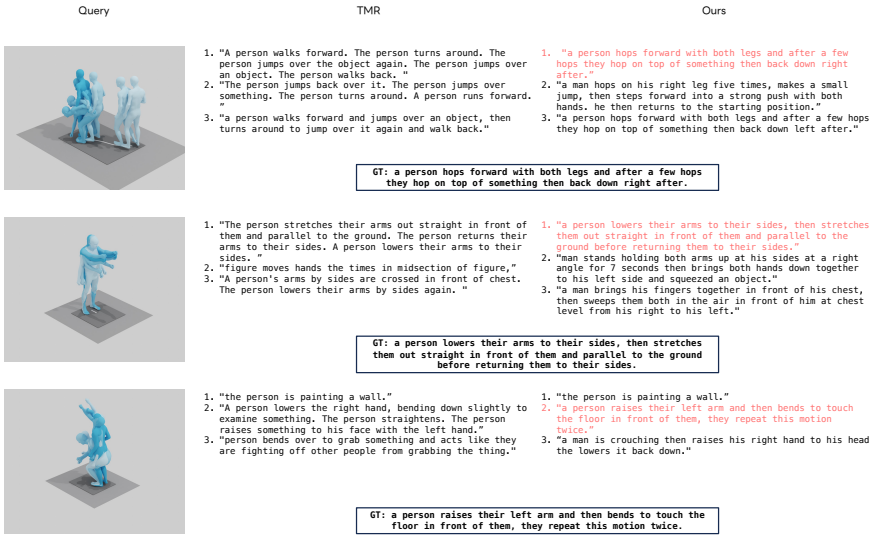


Fig. 4: Comparison of retrieval results with corrupted texts using TMR and the proposed training scheme. Pink texts indicate the successfully retrieved ground truth text.

that is fine-tuned through TMR by enabling backpropagation to the text encoders. We note that the second variant does not use the negative textual samples to tune the text model. The last utilizes the CLIP text encoder that is fine-tuned through our proposal, in other words, the encoder is fine-tuned by also incorporating the negative chronological sample texts. Following the prior works, we evaluate the models using Frechet Inception Distance (FID), Diversity, multi-modality (MModality), Top 1-3 retrieval precision (R-Precision), and Multi-modal Distance (MM-Dist). Detailed definitions of these metrics are introduced in the supplementary material. We conduct generation 10 times for each model, and the statistics of the results are shown in Table 3. As can be seen from the results, fine-tuning language models has a positive effect on the outcome of motion generation. Additionally, from the bottom rows of the table, we can observe that our proposal of using negative chronological text samples improves the metrics further in all of the generation models in this experiment.

We further analyze the effect of different language models fine-tuned through our proposal on the generated motions. We select T2M-GPT [46] as the base generation model, as the method uses a single token from the text encoder as the initial token for motion generation, which facilitates the customization. We employ the feature corresponding to the initial textual token to be used for the initial motion token of T2M-GPT. We compare the performance of different language models fine-tuned through our proposal by using negative chronological samples. As can be seen from the results in Table 4, all of the models demonstrate remarkable improvement in performance over baseline methods, including one that utilizes a fine-tuned language model without the negative samples.

Table 3: Comparison of performance of motion generation models and their variants. "Tune" indicates CLIP text encoders fine-tuned from text-motion retrieval tasks. "Neg" indicates the usage of negative chronological samples in contrastive learning.

Models	Tune	Neg	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity↑	MModality↑
			Top-1	Top-2	Top-3				
Real motion			$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Motiondiffuse			$0.486^{\pm.005}$	$0.681^{\pm.003}$	$0.783^{\pm.003}$	$0.651^{\pm.025}$	$3.079^{\pm.011}$	$9.466^{\pm.110}$	$2.204^{\pm.086}$
T2M-GPT			$0.489^{\pm.004}$	$0.677^{\pm.003}$	$0.774^{\pm.003}$	$0.116^{\pm.005}$	$3.287^{\pm.013}$	$9.771^{\pm.104}$	$1.710^{\pm.061}$
ReMoDiffuse			$0.489^{\pm.005}$	$0.676^{\pm.004}$	$0.774^{\pm.003}$	$0.140^{\pm.074}$	$3.287^{\pm.013}$	$9.247^{\pm.101}$	$2.633^{\pm.101}$
Motiondiffuse	✓		$0.506^{\pm.003}$	$0.697^{\pm.002}$	$0.797^{\pm.003}$	$0.414^{\pm.023}$	$2.981^{\pm.022}$	$9.606^{\pm.099}$	$2.582^{\pm.088}$
T2M-GPT	✓		$0.498^{\pm.003}$	$0.691^{\pm.005}$	$0.785^{\pm.003}$	$0.107^{\pm.008}$	$3.051^{\pm.019}$	$9.679^{\pm.076}$	$1.687^{\pm.085}$
ReMoDiffuse	✓		$0.525^{\pm.003}$	$0.719^{\pm.002}$	$0.814^{\pm.003}$	$0.146^{\pm.009}$	$2.839^{\pm.012}$	$9.301^{\pm.095}$	$2.362^{\pm.103}$
Motiondiffuse	✓	✓	$0.513^{\pm.004}$	$0.710^{\pm.005}$	$0.804^{\pm.003}$	$0.367^{\pm.018}$	$3.115^{\pm.016}$	$9.535^{\pm.080}$	$2.640^{\pm.091}$
T2M-GPT	✓	✓	$0.528^{\pm.004}$	$0.717^{\pm.003}$	$0.806^{\pm.002}$	$0.070^{\pm.006}$	$2.918^{\pm.017}$	$9.659^{\pm.145}$	$1.265^{\pm.145}$
ReMoDiffuse	✓	✓	$0.525^{\pm.005}$	$0.719^{\pm.006}$	$0.813^{\pm.004}$	$0.116^{\pm.010}$	$2.881^{\pm.011}$	$9.248^{\pm.121}$	$2.449^{\pm.125}$

Table 4: Comparison of performance of motion generation with different language models using T2M-GPT as the base model. "Tune" indicates fine-tuning language models through backpropagating TMR, and "Neg" indicates using negative chronological samples. Note that the original T2M-GPT relies on CLIP encoder for the initial token.

Encoder	Tune	Neg	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity↑	MModality↑
			Top-1	Top-2	Top-3				
Real motion			$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
CLIP			$0.489^{\pm.004}$	$0.677^{\pm.003}$	$0.774^{\pm.003}$	$0.116^{\pm.005}$	$3.287^{\pm.013}$	$9.771^{\pm.104}$	$1.710^{\pm.061}$
	✓		$0.498^{\pm.003}$	$0.691^{\pm.005}$	$0.785^{\pm.003}$	$0.107^{\pm.008}$	$3.051^{\pm.019}$	$9.679^{\pm.076}$	$1.687^{\pm.085}$
DistilBERT	✓	✓	$0.528^{\pm.007}$	$0.717^{\pm.008}$	$0.810^{\pm.004}$	$0.074^{\pm.012}$	$2.915^{\pm.020}$	$9.499^{\pm.197}$	$1.599^{\pm.075}$
CLIP	✓	✓	$0.528^{\pm.004}$	$0.717^{\pm.003}$	$0.806^{\pm.002}$	$0.070^{\pm.006}$	$2.918^{\pm.017}$	$9.659^{\pm.145}$	$1.265^{\pm.145}$
t5base	✓	✓	$0.530^{\pm.005}$	$0.719^{\pm.002}$	$0.810^{\pm.003}$	$0.087^{\pm.009}$	$2.896^{\pm.010}$	$9.716^{\pm.246}$	$1.536^{\pm.156}$
t5large	✓	✓	$0.529^{\pm.007}$	$0.718^{\pm.005}$	$0.812^{\pm.006}$	$0.074^{\pm.008}$	$2.904^{\pm.029}$	$9.681^{\pm.112}$	$1.670^{\pm.077}$

To evaluate the chronological accuracy of motion generation, we take advantage of the autoregressive characteristics of T2M-GPT and compare the cumulative likelihood of returning the ground truth motion tokens when using true texts and shuffled ones as input of T2M-GPT. In this experiment, we used the CLIP model for the text encoder, with ours fine-tuned by disabling the VAE loss using the negative events. The original model returned a better likelihood for the true text inputs in only 61.9% of test set motions compared to the shuffled inputs. However, the figure was 89.9% for the model tuned by our method, indicating the effectiveness of considering chronology also during generation.

Qualitative Analysis To qualitatively assess the generated motions, we present some of the example motions generated by the models. Fig 5 shows the generated results. We base the comparison on the same model T2M-GPT [46]. The top row shows the motions generated by the original T2M-GPT, and the bottom row the motions generated using T2M-GPT with the fine-tuned text-encoder trained with our proposal. When the prompt demands a sequence of compound actions, the model fails to fully capture all the events, for example, the person is not picking anything in the first case, and is not descending in the fourth example. There are also cases where chronology is wrong, as in the second example. On the other hand, the fine-tuned language model trained using our proposal is able to capture each event in the prompt in many of the cases.

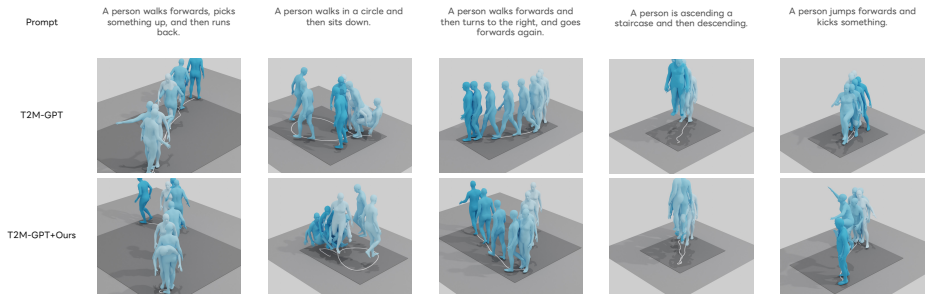


Fig. 5: Comparison of generated motions. Top: T2M-GPT. Bottom: T2M-GPT with our fine-tuned t5-large encoder. Texts at the top represent the input prompts.

5 Limitations

Although we achieved better correspondence between motion and language with our proposal in terms of chronology, we noticed that the existence of pronouns and articles in texts may provide implicit hints regarding the chronology of the events, as the original texts use definitive articles and avoid repetitive pronouns if a certain subject appears more than once in the text.

To observe the effect of these elements, we compared the CAR metric under different settings. The CAR metric for the best fine-tuned model t5-large in the original experiment was at 94.62%. First, we unified the articles 'a, an, the' at the start of sentences and events to 'The' in order to conceal the implicit information regarding when the event is taking place. This led to the CAR of 94.51% for the same fine-tuned model. There are also words that points to the person in the sequence, such as 'she' and 'he'. When additionally replacing them with 'The person' in the test set, the CAR metric for the same model fell to 81.21%. We discuss this issue in detail in the supplementary material.

Also, correspondence between individual words and motion is yet to be achieved, as sentences are compressed to a single feature. Exploring methods to localize the correspondence between texts and the frames of the motions is one of future works. As the proposal mainly focused on manipulating textual descriptions, we would also like to reinforce the current proposal by augmenting motion sequences so that motion event orders can also be manipulated.

6 Conclusion

We focused on the temporal element of motion and language, and uncovered that current approaches mostly overlook the factor of chronological alignment between text and motion, through a novel metric assessing event chronology. We proposed a simple solution to achieve better correspondence between language and motion, demonstrating the importance of chronology through various experiments. We hope to inspire other research to focus more on compound actions to establish a better representation for human motion, a complex medium of data.

References

1. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 3DV (2022)
2. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al.: Audiolm: a language modeling approach to audio generation. IEEE/ACM TASLP (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
4. Burgert, R., Ranasinghe, K., Li, X., Ryoo, M.S.: Peekaboo: Text to image diffusion models are zero-shot segmentors. arXiv preprint arXiv:2211.13224 (2022)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. TOG (2023)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
8. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: CVPR (2022)
9. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV (2021)
10. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
11. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: ACM MM (2020)
12. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI (2013)
13. Jung, M., Jang, Y., Choi, S., Kim, J., Kim, J.H., Zhang, B.T.: Overcoming weak visual-textual alignment for video moment retrieval. arXiv preprint arXiv:2306.02728 (2023)
14. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. arXiv preprint arXiv:2211.15603 (2022)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
16. Ko, D., Choi, J., Ko, J., Noh, S., On, K.W., Kim, E.S., Kim, H.J.: Video-text representation learning via differentiable weak temporal alignment. In: CVPR (2022)
17. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., Adi, Y.: Audiogen: Textually guided audio generation. In: ICLR (2023)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
19. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
20. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. In: NeurIPS (2023)
21. Liu, X., Yin, J., Liu, H., Yin, Y.: Pisep 2: pseudo-image sequence evolution-based 3d pose prediction. The Visual Computer (2022)

22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
23. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
24. Miki, D., Chen, S., Demachi, K.: Weakly supervised graph convolutional neural network for human action localization. In: WACV (2020)
25. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV (2021)
26. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: ECCV (2022)
27. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023)
28. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation (2024), arXiv preprint arXiv:2401.08559
29. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data (2016)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS-W (2019)
34. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: PriorMDM: Human motion diffusion as a generative prior. In: ICLR (2024)
35. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR (2016)
36. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)
37. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: ECCV (2020)
38. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV (2022)
39. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
40. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
41. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
42. Yu, Q., Fujiwara, K.: Frame-level label refinement for skeleton-based weakly-supervised action recognition. In: AAAI (2023)
43. Yu, Q., Tanaka, M., Fujiwara, K.: Exploring vision transformers for 3d human motion-language models with motion patches. In: CVPR (2024)
44. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bag-of-words models, and what to do about it? In: ICLR (2023)

45. Zhang, H., Liu, D., Lv, Z., Su, B., Tao, D.: Exploring temporal concurrency for video-language representation learning. In: ICCV (2023)
46. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
47. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
48. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023)