Supplementary Material: 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

1 Motion Data

As stated in the main manuscript, we mostly follow the setting in the original paper of TMR [9]. For pose representation, we converted each frame of J joints the in the motion sequence M into a 263 dimensional vector consisting of (r^{va}, r^{vx}, r^{vz}, r_h, $\mathbf{j}^p$, $\mathbf{j}^v$, $\mathbf{j}^r$, $\mathbf{f}$), where r^{va}, r^{vx}, r^{vz} and r_h are the angular velocity around y axis, velocity in x direction, velocity in z direction and height of the root on the XZ-plane, respectively. $\mathbf{j}^p \in \mathbb{R}^{3(J-1)}$, $\mathbf{j}^v \in \mathbb{R}^{3J}$ indicate the positions and velocities of the J joints. $\mathbf{j}^r \in \mathbb{R}^{6(J-1)}$ denotes the six-dimensional representation of rotation for each joint. $\mathbf{f} \in \mathbb{R}^4$ are the contact state of the feet.

For chronological event shuffling, we select sequences that contain multiple events. In HumanML3D dataset, multiple events are detected in 13,044 out of 23,384 sequences, 811 out of 1,460 sequences, 2,677 out of 4,380 sequences in the training, validation, and test sets, respectively.

2 Text-Motion Retrieval

2.1 TMR with Different Language Models

As the original method of TMR is equipped with DistilBERT language model to encode texts, we first observe how TMR behaves when combined with different language models. We trained each variation of TMR in a similar manner as the original model only using the original texts. Table 1 shows the results. As can be

Table 1: Retrieval results of TMR with different language models.

Models	Text-to-motion retrieval				Motion-to-text retrieval			
	R@1 ↑	R@3 ↑	R@5 ↑	MedR ↓	R@1 ↑	R@3 ↑	R@5↑	MedR ↓
DistilBERT	5.82	14.85	21.33	25.00	9.76	18.13	24.13	23.50
CLIP	4.22	10.58	15.92	39.00	6.52	13.37	17.91	37.50
t5-base	5.75	14.83	20.99	28.00	8.67	16.61	22.83	26.50
t5-large	5.25	13.71	19.48	29.50	9.06	16.67	21.10	28.00

Table 2: Comparison of retrieval results from training TMR with DistilBERT as language model using negative chronological samples under different batch size.

Batch Size	"orig → event"						CAR
	Text-to-motion retrieval			Motion-to-text retrieval			
	R@1 ↑	R@5 ↑	MedR ↓	R@1 ↑	R@5↑	MedR ↓	
16	6.27	22.29	24.00	10.10	24.43	23.50	99.70
32	5.75	20.69	25.00	10.45	24.29	25.50	99.59
64	5.73	22.01	25.00	9.15	22.99	24.00	99.07
128	6.04	21.67	25.00	9.49	24.68	23.50	98.73
256	6.07	22.60	25.00	9.88	24.91	24.00	98.36

seen, without training with the hard negative samples, all the variants perform similarly, showing no preference towards larger models. It is interesting to note that the CLIP encoder, which is a popular choice in motion analysis, performs the worst among all the models.

2.2 Training with Negative Descriptions: Configuration

We describe how our models are fine-tuned using the shuffled events as hard negative samples. As stated in the main manuscript, the model used in this paper follows that of TMR [9] for fairness of comparison. The model consists of two Transformer encoders, each corresponding to text and motion. Following ACTOR [8], each Transformer is equipped with additional learnable parameters corresponding to μ and σ that determine the Gaussian distribution from which the latent vectors are sampled. The dimension d of the latent vectors is set at 256. The motion decoder is the same as in ACTOR.

For the experiment in Fig 3 and Table 1 in the main manuscript, we followed the original method of TMR by using all the token embeddings from the textual descriptions. That is, we preprocess the text by feeding the texts into the language models, DistilBERT [12], CLIP [10], t5-base, and t5-large [11]. We then attach the two learnable parameters to the preprocessed token embeddings from each model and conduct the training.

For training, the same loss from TMR [9] is used. In other words, the loss is $\mathcal{L}_{\text{TMR}} = \lambda_R\mathcal{L}_R + \lambda_{KL}\mathcal{L}_{KL} + \lambda_E\mathcal{L}_E + \lambda_C\mathcal{L}_C$, where $\mathcal{L}_R$, $\mathcal{L}_{KL}$, $\mathcal{L}_E$, and $\mathcal{L}_C$ are reconstruction loss, KL divergence loss, embedding similarity loss, and contrastive loss, respectively. TMR employs InfoNCE [7] in eq.(1) for the contrastive loss $\mathcal{L}_C$, whereas our training with chronologically negative samples uses eq.(3) for $\mathcal{L}_C$. The balancing weights λ_R, λ_{KL}, λ_E, λ_C are set to 1, 10^{-5}, 10^{-5}, and 0.1, respectively in the experiments.

We use the AdamW optimizer [6] with a learning rate of 10^{-4}, as in TMR. We set the batch size at 32, and are trained for 200 epochs for optimal performance. We show the effect of batch size for the "orig → event" scenario in Table 2. As can be seen, despite some fluctuations, there is little difference among the results. For a fair comparison between language models, we set the maximum

Table 3: Comparison of CAR and Motion-to-text retrieval with corrupted texts with different conventional models. The language model is fixed in this setting.

Models	CAR	Motion-to-text retrieval with corrupted text					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5↑	R@10 ↑	MedR ↓
Guo et al. [13]	29.32	0.41	0.62	0.75	1.25	2.24	559.25
TEMOS [30]	62.16	3.31	4.24	6.52	8.80	13.98	167.00
TMR [31]	62.23	7.07	9.65	12.96	18.09	26.60	37.50
Ours	**99.07**	**9.03**	**11.45**	**17.11**	**22.76**	**33.14**	**24.50**

number of tokens to 77, which is the limit of the CLIP language encoder. Other hyperparameters are set to be exactly the same as in TMR. The models are implemented in PyTorch, and are trained using a single Tesla A100 GPU.

2.3 Comparison to Prior Methods

Although TMR is the most advanced method available, we enlist the CAR metric and the corrupted motion-to-text retrieval results in "orig → event" setting, from methods with publicly available pretrained models, in Table 3. For fairness, we list the results from DistilBERT for "Ours". We note that the language model is kept fixed in this experiment.

2.4 Fine-tuning Language Models

When fine-tuning the language models to conduct experiments shown in Table 2 in the main manuscript, some modifications to TMR are required to enable the training.

Disabling VAE functionality requires text token embeddings to be compressed into a single vector. To conduct this, we simply use the first token embedding for outputs from DistilBERT, t5-base, and t5-large. For CLIP, we extract the feature corresponding to the EoT token embedding. Instead of using these feature vectors as input to the original Transformer-based encoder of ACTOR, we instead use this as input to a simple projection head, consisting of 2 linear layers with GELU [4] activation and layer normalization [2]. Motion encoder is kept the same but used without the VAE functionality. The weights for losses λ_R, λ_E, λ_C are kept the same. However, as the method no longer solves for the distributions, we omit the KL loss.

Disabling reconstruction loss removes the necessity of the motion decoder. We therefore remove the component. As the weight for the reconstruction loss λ_R is dominant in the other settings, we instead raise the weight of the contrastive loss λ_C to 1.0. If VAE is also disabled, we also remove the KL loss, as it is unnecessary. We set the learning rate 10^{-5} for optimizing the language models to enable refinement of text embeddings, and maintain 10^{-4} to optimize the rest of the model. To enable conducting experiments in the same setting we use a batch size of 32, as t5 models require larger memory space. To also observe the

capabilities of each language model, the maximum number of tokens is raised to 128 to enable better expression for DistilBERT, t5-base, and t5-large.

We note that to avoid providing additional hints regarding the chronology of events, we unified the articles "a, the" to "The" in the event descriptions, as these tend to appear at the start of a description.

2.5 Evaluation

We use the same evaluation metric as TMR [9]: Recall at various ranks (R@1, R@2, ...) and median rank (MedR). Recall at rank k indicates the percentage of cases where correct answers are returned within the top k results. We select the weights from the epoch that provided the best R@1 motion-to-text retrieval accuracy on the validation set. We follow the same evaluation protocols in TMR to evaluate the performance of the models:

- **(a) All**: The entire test set is used for the retrieval task without any modification. Slight changes in the text affect this metric (e.g., person/human, walk/walking). We rely on this metric in most of the tests.
- **(b) All w/ Threshold**: The entire test set is used for this retrieval task as well, except, if the text is similar to the query text above a threshold, the retrieval results are accepted as correct. The threshold is set to 0.95 in order to remove very similar expressions.
- **(c) Dissimilar subset**: From the entire test set, we select 100 motion-text pairs that are far apart from each other, determined by using an approximation of the quadratic knapsack problem [1].
- **(d) Small batches**: Proposed by Guo et al., [3], this scenario selects random batches of 32 motion-text pairs, and reports the average performance of the batches.

2.6 Full Results: Proposed Fine-tuning

Motion-to-Text Retrieval Including Corrupted Texts As the main paper only contained results from cases where language models were kept fixed in Table 1, we tested the fine-tuned language models trained through our proposal in the task of motion-to-text retrieval including the corrupted texts as retrieval candidates. We also conduct the CAR test using our fine-tuned models. Tables 4 and 5 show the results in the "orig $\rightarrow$ event" and "event $\rightarrow$ event" scenarios, respectively. For the last four rows in each table indicating models fine-tuned through our method, we select the variation that disables the VAE loss, which performs best for most models. We also enlist the numbers from TMR [9] trained with the same data (excluding the negative samples) for reference.

As can be seen from the tables, fine-tuning the language models achieves higher performance than the non-tuned results from Table 1 in the main manuscript. The negative samples reinforce the models to differentiate accurate descriptions from the corrupted ones, even in the difficult setting of "event $\rightarrow$ event", where all the descriptions consist of concatenated events. A similar tendency can be

Table 4: Motion-to-text retrieval results with both the original and corrupted texts, and CAR accuracy in "orig → event" scenario. Here, the model is asked to to distinguish the original text from the corrupted text obtained by shuffling events.

Model	Motion-to-text retrieval w/ corrupted events						CAR↑
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	
TMR [9]	7.76	10.01	14.51	20.14	28.54	33.50	65.45
DistilBERT	10.26	12.77	18.43	23.54	33.42	24.00	**99.59**
CLIP	8.07	9.97	14.30	19.80	28.44	33.50	99.07
t5-base	10.90	13.82	19.73	27.35	38.02	19.50	98.73
t5-large	**11.70**	**15.19**	**21.65**	**28.15**	**39.23**	**17.50**	98.36

Table 5: Motion-to-text retrieval results with both the original and corrupted texts, and CAR accuracy in "event → event" scenario. Here, the model is asked to to distinguish the event descriptions concatenated in the original order from the corrupted text obtained by shuffling events.

Model	Motion-to-text retrieval w/ corrupted events						CAR↑
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	
TMR [9]	6.61	8.92	12.73	17.88	26.53	35.00	64.29
DistilBERT	8.05	10.56	15.65	20.89	31.98	27.00	93.69
CLIP	6.98	8.58	12.84	18.04	26.30	44.25	92.53
t5-base	10.13	12.82	18.43	24.73	35.63	**20.50**	93.50
t5-large	**10.70**	**13.57**	**19.32**	**26.07**	**36.88**	**20.50**	**94.51**

seen in the CAR accuracy as well, slightly improving from the non-tuned results shown in Fig 3 in the main manuscript. As differentiating similar text with different chronological order requires richer textual representation, the larger language models perform better overall, with the best retrieval results attained with the tuned t5-large model.

Conventional Text-Motion Retrieval We then demonstrate the performance of the models fine-tuned by our proposal in the conventional retrieval task, where shuffled event descriptions do not exist in the test data.

In the main manuscript, we only listed the results from "(a) All" for the "orig → event" scenario due to the limitation of space. Table 9 shows the complete results of the fine-tuning experiment with different language models for the "orig → event" scenario, and Table 10 the results of the "event → event" scenario. As with the prior experiments, we unified the articles "a, the" to "The" in the event descriptions. This means that models in "orig → event" are tested with these rectified events, and "event → event" models are trained and tested with these events. As can be seen from the results, most of the cases perform better as the components are removed when fine-tuning the language models. Similar to the previous results, larger models such as t5-base and t5-large tend to perform better overall in comparison to other models. Although there are some

Table 6: Retrieval results from "event → event" limited to sequences with multiple events. between the original TMR model and its variations equipped with different language models, which are fine-tuned with chronological negative samples. All the models are trained including single-event sequences.

	Text-to-motion retrieval						Motion-to-text retrieval					
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5↑	R@10 ↑	MedR ↓
TMR [9]	8.33	16.77	22.86	30.74	42.32	15.00	13.90	17.93	24.54	31.57	43.29	14.50
DistilBERT	9.15	16.40	21.63	29.29	42.73	15.00	12.10	15.88	22.34	29.25	41.05	15.50
CLIP	6.43	11.92	15.50	21.59	32.01	27.00	9.64	11.88	16.88	22.08	30.59	38.50
t5-base	**10.16**	18.98	24.62	32.98	47.14	12.00	14.23	18.27	26.04	33.10	45.05	13.50
t5-large	9.94	**19.20**	**25.14**	**34.14**	**48.60**	**11.00**	**15.69**	**19.50**	**27.12**	**35.34**	**47.67**	**12.00**

Table 7: Motion-to-text retrieval results with both the original and corrupted texts in "event → event" scenario limited to sequences with multiple events. Here, the model is asked to to distinguish the event descriptions concatenated in the original order from the corrupted text obtained by shuffling events.

Model	Motion-to-text retrieval w/ corrupted events					
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR [9]	8.44	11.62	16.59	22.23	32.95	26.00
DistilBERT	10.65	14.42	20.32	27.19	38.40	20.50
CLIP	8.93	10.98	15.69	20.84	28.88	42.50
t5-base	13.11	16.77	23.98	31.12	43.00	15.00
t5-large	**14.49**	**18.38**	**25.18**	**33.10**	**44.94**	**14.00**

fluctuations, these two models perform best when the VAE loss is removed and the reconstruction loss is kept in the model. Despite performing slightly worse in most cases, as it is a more difficult scenario, the same trend can be observed from the results of "event → event" scenario.

Test Results Limited to Multiple Events We further investigate the effectiveness of our proposal by focusing on the retrieval results of the 2677 sequences with multiple events in the test set. Table 6 shows the retrieval results from TMR and the models fine-tuned with our strategy for "(a) All" evaluation protocol, and Table 7 demonstrates the results from the task of motion-to-text retrieval including the corrupted texts as retrieval candidates. The bottom 4 rows from our strategy are the variant that does not use the VAE loss for fine-tuning the language model. Similar to other experiments, the proposed method improves the larger language models in terms of retrieval accuracy. The margin of improvement of the proposal becomes larger when the task is more demanding in terms of chronological understanding, as can be seen in the results from Table 7, where chronological information is required to distinguish correct texts from corrupted ones.

Table 8: Comparison of baseline classifier accuracy and CAR of fine-tuned language models under different variations in "event → event" scenario. Baseline is trained to distinguish event descriptions concatenated in the original order from the shuffled ones, only from features of the corresponding encoded event descriptions.

Encoder	Method	No Edit	Article	Pronoun
DistilBERT	Baseline	87.22	76.91	72.77
	Ours	**93.54**	**93.69**	**85.51**
CLIP	Baseline	83.08	72.10	70.94
	Ours	**92.15**	**92.53**	**82.11**
t5-base	Baseline	85.84	73.70	67.46
	Ours	**93.50**	**93.50**	**84.58**
t5-large	Baseline	85.58	72.10	69.51
	Ours	**94.62**	**94.51**	**81.21**

2.7 Effect of Articles and Pronouns

In the main manuscript, we noted that the pronouns and articles may provide additional hints with regards to the chronology of events. To observe whether these elements are sufficient to determine if the order of events is correct or not, we design a simple test by training a binary classifier that learns to separate event descriptions concatenated in the original order and the shuffled descriptions. In other words, we examine whether the chronology can be determined solely from texts. We design the classifier to first encode the features from the language models in the same manner as TMR [9] without the VAE loss to achieve maximal descriptiveness. We then place a layer that receives the embedded text features and outputs the likelihood of two classes, whether they are accurate or corrupt. We train this baseline classifier under the same aforementioned protocol in Section 2.2 using the events concatenated in the correct order, and the shuffled ones as corrupt versions of the training set.

We prepare 3 cases to determine the effect of such articles and pronouns. "No Edit": This is the case where the decomposed events are used as they are. In other words, events concatenated in the correct order are used as the positive samples, and those shuffled as negatives. No other editing is applied to the texts. "Article": This indicates cases where indefinite articles are converted to definite one, as indefinite articles tend to appear at the initial event. After replacing "A" with "The", the events are treated the same as in the previous case. "Pronoun": This is the case where variations of pronouns addressing a person, such as "a figure, a person" is unified to "The person". The rectified events are then processed in the same manner as the prior cases. All models are trained with data processed under the same protocol, and tested with the similarly processed data. During testing, baseline models are randomly provided with texts of either events in the original order or shuffled events, and are asked to classify whether it is shuffled.

Table 8 shows the results from the analysis. The results from "Baseline-No Edit" demonstrate that there do definitely exist some leakage of chronological information solely from textual descriptions, as the baseline classifier is able to distinguish correct descriptions in approximately 85% of the cases, regardless of

the language model. However, our method is able to outperform the baseline in all the cases, demonstrating the ability of the proposal to extract chronological information from not only textual information, but from motion as well.

Our strategy to stop the leakage by replacing indefinite articles with definite ones has a remarkable impact on the baseline classifier, as the accuracy drops by more than 10%, despite being trained with the same rectification. However, models trained through our strategy tend to perform similarly to the unedited version, displaying the power of training with negative samples.

The most difficult case was observed when replacing pronouns indicating the actor in the sequence as "The person". The baseline model deteriorated further, while the model refined by our method also regressed. However, the proposed method still outperforms the baseline with all the language models.

We note that all these cases, including the baseline, still outperform the CAR results from TMR [9] in Fig. 3 in the main manuscript. This demonstrates that the prior methods are not able to take full advantage of the hints existing in the texts, indicating the necessity to delve further into the language-based aspects of motion-language models.

There are also possibilities of other leakages of chronological information. For example, objects that a person in the sequence is interacting with, can be mentioned with a pronoun such as "it" as well. These could be functioning as additional information. However, as covering all such criteria requires additional analysis from the linguistics perspective, as a recent research [5] has done in the image domain, we leave this as future work.

2.8 Additional Qualitative Results

We further introduce visualizations of examples from the experimental results. Fig. 1 presents examples from TMR and results from our hard negative training. As in the main manuscript, we select these examples from the motion-to-text retrieval experiment including the shuffled texts. The setting is therefore identical to Table 3 and Fig. 6 in the main manuscript.

In the first example, all three texts retrieved with our method describe a similar event of the person going backward and then turning to go backward again. However, TMR fails to capture the chronology as well as the direction of the movement, indicating forward movement in some captions. The second example contains an ambiguous motion of picking objects as if the person is cooking something. The proposed training scheme is able to capture the characteristics of the motion and retrieve the correct answer at the third rank. On the contrary, TMR is only capable of capturing the feature for the movement of picking something up, therefore retrieves various activities that include the motion of picking. The third case is where TMR has a better result than the proposed scheme. TMR is able to retrieve the correct text at rank 1, while the proposal is only able to pick the correct text as the third most likely text. However, all the texts in the results from the proposal resemble similar activity, whereas TMR returns a different activity that involves lifting something up as the third most similar text.

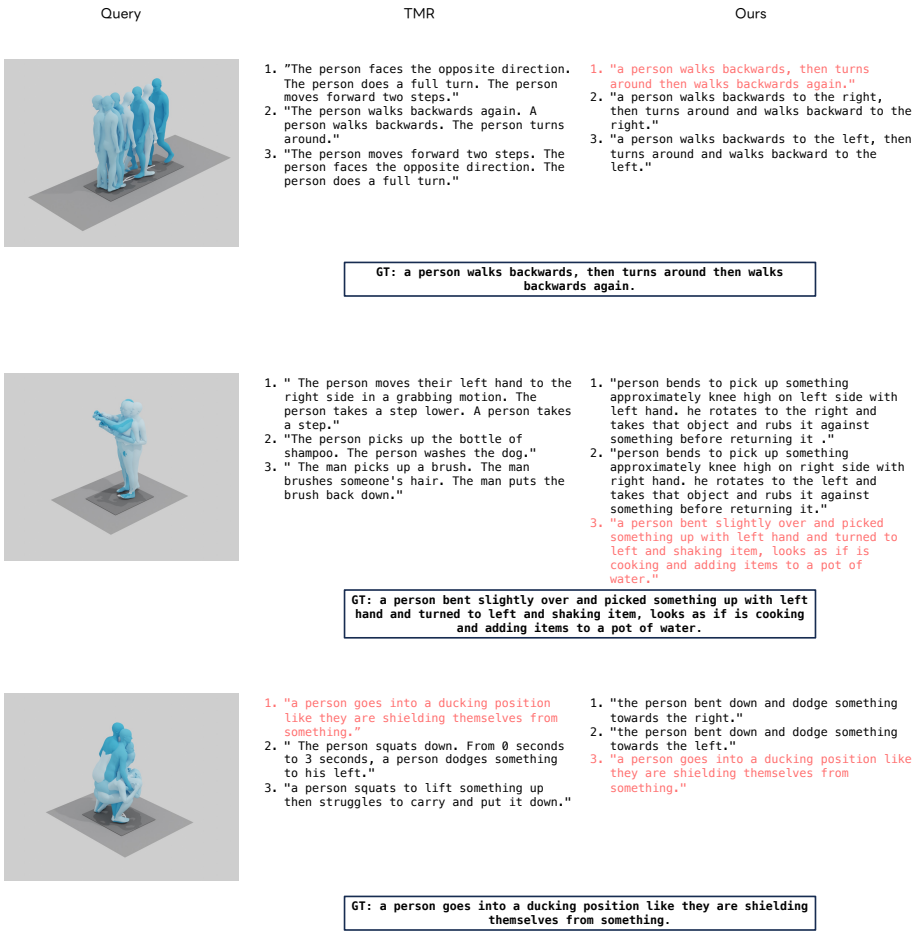

Fig. 1: Additional retrieval results with corrupted texts using TMR and the proposed training scheme. Pink texts indicate the successfully retrieved ground truth text.

3 Generation

3.1 Evaluation

Following the prior works on human motion generation, we evaluated the generation models using the following criteria:

– **(1) Top 1-3 retrieval precision (R-Precision)**: For each generated motion, its ground truth and 31 random descriptions are selected to form a pool These are ranked by the distance between the generated motion feature and the text features. The average accuracy at rank 1, 2, and 3 are measured.

- **(2) Frechet Inception Distance (FID)**: The distance between the feature distribution of the generated motions and that of the actual motions. The identical pretrained model as prior methods is used to extract the features.
- **(3) Multi-modal Distance (MM-Dist)**: Multi-modal distance measures the average distance between the generated motion features and their corresponding text features.
- **(4) Diversity**: Diversity is the variance of generated motion features across all the samples.
- **(5) Multi-modality (MModality)**: For each text, 10 pairs of motions are generated, and the average value of the mean distance of the pairs is recorded as Multi-modality.

4 Motion Visualization

As motion can be difficult to observe in still images, we attach a demonstration video including some of the results from the motion-to-text retrieval including the corrupted texts, and the results from motion generation on HumanML3D dataset. The examples show the event chronology is accurately reflected in the results obtained through models tuned by the proposed strategy.

5 Sample Code

The code will be released at https://github.com/kentfuji/ChronAccRet. We provide the training codes for building the motion-language model, and test codes to measure CAR, as well as the conventional evaluation metrics. We refer the readers to the README file in the code to run the experiments.

References

1. Aïder, M., Gacem, O., Hifi, M.: Branch and solve strategies-based algorithm for the quadratic multiple knapsack problem. Journal of the Operational Research Society **73**(3), 540–557 (2022)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. 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)
4. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
5. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In: NeurIPS (2023)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
7. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
8. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV (2021)

9. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023)
10. 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)
11. 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)
12. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS-W (2019)

Table 9: Comparison of "orig → event" retrieval results between the original TMR model and its variations with different language models, which are fine-tuned with chronological negative samples. "Tune" indicates whether the language model is fine-tuned, "VAE" whether VAE feature is used, and "Recon." whether a decoder reconstructs the poses.

Protocol	Method	Tune	VAE	Recon.	Text-to-motion retrieval						Motion-to-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	TMR [9]		✓	✓	5.82	11.02	14.85	21.33	32.76	25.00	9.76	12.34	18.13	24.13	33.23	23.50
	DistilBERT [12]	✓	✓	✓	5.25	9.92	13.46	19.96	30.98	29.00	8.69	11.09	16.51	22.45	31.89	28.00
		✓	✓		5.36	10.83	14.37	20.32	31.04	28.00	9.19	11.27	17.15	23.11	31.64	28.00
		✓		✓	6.11	11.25	15.74	22.08	32.78	24.00	10.26	12.77	18.43	23.54	33.42	24.00
		✓			6.55	12.25	16.04	22.99	34.60	22.00	11.18	13.57	19.37	25.52	36.38	21.50
	CLIP [10]	✓	✓	✓	4.68	8.78	11.79	16.72	25.91	41.00	8.05	9.97	14.46	18.66	27.24	39.50
		✓	✓		3.99	8.14	11.04	16.49	27.44	32.00	6.57	8.55	14.67	19.59	28.44	31.50
		✓		✓	4.58	8.67	12.00	17.91	27.62	35.00	8.07	9.97	14.30	19.80	28.44	33.50
		✓			5.57	10.61	13.87	20.00	30.06	28.00	8.69	10.77	16.42	22.06	31.14	27.50
	t5-base [11]	✓	✓	✓	6.39	12.73	16.56	22.81	34.63	23.00	10.56	13.41	19.64	26.21	36.09	22.50
		✓	✓		4.13	8.23	10.65	16.45	26.48	32.00	6.98	8.46	13.55	18.68	27.67	29.50
		✓		✓	6.98	12.82	18.02	24.98	36.75	19.00	10.90	13.82	19.73	27.35	38.02	19.50
		✓			5.50	10.52	14.60	21.62	33.42	22.00	9.88	12.18	17.86	24.77	35.36	21.50
	t5-large [11]	✓	✓	✓	6.57	12.20	16.15	22.97	33.92	24.00	9.74	12.57	19.34	25.62	36.59	22.00
		✓	✓		4.65	9.58	13.12	18.84	31.75	24.00	8.69	11.02	17.22	23.72	33.85	23.50
		✓		✓	**8.03**	**14.51**	**18.84**	**26.73**	**38.98**	**17.00**	**11.72**	**15.19**	**21.65**	**28.15**	**39.23**	**17.50**
		✓			6.52	11.98	16.56	24.16	36.50	20.00	10.56	13.25	19.73	25.91	36.86	19.50
(b) All w/ threshold	TMR [9]		✓	✓	12.32	16.79	22.99	30.18	42.34	15.00	13.64	15.99	22.47	28.95	38.44	18.00
	DistilBERT [12]	✓	✓	✓	10.99	14.32	20.07	27.01	39.07	19.00	12.07	14.12	20.96	27.60	37.34	21.50
		✓	✓		13.09	15.67	21.49	27.97	38.48	19.50	12.89	13.48	21.76	27.21	35.47	23.00
		✓		✓	13.00	16.83	23.79	30.00	41.42	16.00	14.14	16.20	23.08	28.42	38.57	19.50
		✓			13.21	16.77	23.22	31.68	42.95	15.00	15.01	17.22	24.06	30.34	41.29	17.50
	CLIP [10]	✓	✓	✓	10.29	13.48	18.77	23.95	33.42	28.00	11.18	13.18	18.82	23.29	32.53	31.50
		✓	✓		10.33	13.46	18.41	24.84	36.04	23.00	10.33	10.72	18.86	24.20	32.96	25.50
		✓		✓	9.79	13.39	19.30	25.41	35.17	22.00	11.52	13.34	19.07	25.36	34.24	25.50
		✓			12.23	16.20	21.15	28.22	39.23	18.50	12.45	14.03	20.80	26.53	35.79	23.00
	t5-base [11]	✓	✓	✓	13.89	17.72	23.59	30.13	43.16	15.00	14.30	16.29	24.02	30.70	40.69	17.50
		✓	✓		10.54	15.90	20.10	27.49	38.07	20.00	10.08	10.70	17.31	22.31	31.09	25.00
		✓		✓	**15.44**	**20.87**	27.40	34.99	47.92	**11.00**	14.99	17.93	24.95	32.62	42.84	15.00
		✓			11.86	15.17	22.51	30.43	43.07	14.00	13.48	14.92	21.88	28.97	39.28	17.50
	t5-large [11]	✓	✓	✓	13.21	17.68	24.16	31.23	42.52	15.00	13.32	15.81	23.56	30.29	41.65	17.00
		✓	✓		11.06	15.31	20.85	27.58	41.33	15.00	12.43	13.05	21.26	27.85	37.75	19.00
		✓		✓	15.19	20.55	**27.44**	**35.95**	**48.59**	**11.00**	**15.53**	**18.34**	**26.07**	**33.30**	**44.18**	**14.00**
		✓			13.53	18.91	27.05	35.01	47.58	12.00	14.94	16.74	24.13	30.45	41.20	15.50
(c) Dissimilar subset	TMR [9]		✓	✓	48.00	71.00	79.00	86.00	93.00	2.00	52.00	73.00	83.00	86.00	91.00	1.00
	DistilBERT [12]	✓	✓	✓	45.00	66.00	73.00	80.00	90.00	2.00	50.00	70.00	76.00	81.00	88.00	1.50
		✓	✓		52.00	63.00	69.00	76.00	82.00	**1.00**	50.00	64.00	74.00	80.00	83.00	1.50
		✓		✓	52.00	65.00	76.00	84.00	89.00	**1.00**	49.00	73.00	79.00	85.00	90.00	2.00
		✓			53.00	70.00	77.00	86.00	88.00	**1.00**	53.00	67.00	77.00	87.00	88.00	**1.00**
	CLIP [10]	✓	✓	✓	48.00	67.00	72.00	79.00	89.00	2.00	45.00	61.00	73.00	80.00	87.00	2.00
		✓	✓		50.00	68.00	70.00	81.00	87.00	1.50	52.00	66.00	72.00	80.00	89.00	**1.00**
		✓		✓	46.00	66.00	71.00	79.00	87.00	2.00	40.00	69.00	73.00	79.00	89.00	2.00
		✓			54.00	70.00	79.00	86.00	91.00	**1.00**	57.00	70.00	78.00	85.00	92.00	**1.00**
	t5-base [11]	✓	✓	✓	**56.00**	**76.00**	81.00	**87.00**	93.00	**1.00**	55.00	76.00	83.00	88.00	90.00	**1.00**
		✓	✓		46.00	66.00	76.00	85.00	**95.00**	2.00	54.00	69.00	76.00	88.00	**95.00**	**1.00**
		✓		✓	55.00	72.00	76.00	85.00	94.00	**1.00**	55.00	72.00	79.00	87.00	**95.00**	**1.00**
		✓			48.00	71.00	78.00	86.00	94.00	2.00	51.00	70.00	81.00	**89.00**	94.00	**1.00**
	t5-large [11]	✓	✓	✓	53.00	68.00	75.00	86.00	92.00	**1.00**	54.00	71.00	78.00	87.00	92.00	**1.00**
		✓	✓		54.00	72.00	75.00	82.00	88.00	**1.00**	55.00	70.00	75.00	82.00	89.00	**1.00**
		✓		✓	53.00	75.00	**82.00**	**87.00**	93.00	**1.00**	**59.00**	**79.00**	**87.00**	**89.00**	93.00	**1.00**
		✓			55.00	75.00	80.00	85.00	93.00	**1.00**	55.00	75.00	80.00	87.00	93.00	**1.00**
(d) Small batches	TMR [9]		✓	✓	70.19	83.74	88.82	93.18	96.83	1.00	70.64	84.65	89.46	93.32	96.67	1.00
	DistilBERT [12]	✓	✓	✓	66.77	80.52	85.95	90.49	94.64	1.02	67.93	81.55	86.47	90.90	94.57	1.03
		✓	✓		65.74	77.99	82.85	87.02	91.01	1.03	66.26	79.15	83.07	87.14	90.99	1.05
		✓		✓	68.20	82.34	87.29	91.10	94.73	1.02	69.69	82.94	87.39	90.92	94.39	1.03
		✓			70.10	82.69	87.07	91.13	94.32	1.01	70.19	81.98	86.66	90.69	94.23	1.01
	CLIP [10]	✓	✓	✓	60.36	75.36	81.57	87.52	93.50	1.12	60.83	75.41	81.73	87.64	93.57	1.07
		✓	✓		65.42	79.43	84.88	89.01	93.41	1.01	65.53	79.77	85.17	89.14	92.91	1.04
		✓		✓	63.16	77.83	83.90	89.01	94.64	1.06	64.12	78.97	84.60	89.83	95.23	1.07
		✓			67.97	81.34	86.61	91.15	95.53	1.02	67.66	81.93	87.00	91.10	95.44	1.01
	t5-base [11]	✓	✓	✓	69.84	82.96	88.57	92.86	96.51	1.01	71.08	84.26	89.21	92.88	96.49	**1.00**
		✓	✓		67.84	84.24	89.80	94.71	97.97	1.01	69.53	84.65	90.65	95.00	97.99	1.02
		✓		✓	**75.14**	87.86	**92.40**	**95.92**	**98.36**	**1.00**	74.11	87.96	**92.81**	**96.10**	**98.31**	**1.00**
		✓			72.19	86.22	90.94	94.50	97.45	**1.00**	72.54	86.86	91.20	94.59	97.58	**1.00**
	t5-large [11]	✓	✓	✓	70.69	83.90	88.69	93.07	96.08	1.02	71.56	84.79	89.21	93.04	96.08	1.01
		✓	✓		70.39	83.92	88.50	92.63	95.94	1.01	71.17	84.03	88.85	92.77	95.76	**1.00**
		✓		✓	74.95	87.52	91.42	94.71	97.22	**1.00**	**75.71**	87.66	91.70	94.98	97.42	**1.00**
		✓			74.45	**87.91**	92.36	95.35	97.79	**1.00**	75.00	**88.09**	92.38	95.85	98.13	**1.00**

Table 10: Comparison of "event → event" retrieval results between the original TMR model and its variations equipped with different language models, which are fine-tuned with chronological negative samples.

Protocol	Method	Tune	VAE	Recon.	Text-to-motion retrieval						Motion-to-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	TMR [9]		✓	✓	5.54	10.49	14.01	20.39	32.30	24.00	9.08	11.52	16.77	22.51	33.99	23.50
	DistilBERT [12]	✓	✓	✓	5.47	10.54	14.01	19.07	29.72	29.00	8.80	10.81	15.88	20.99	30.04	30.00
		✓	✓		4.24	8.96	12.04	17.88	28.4	29.00	7.07	9.10	13.80	19.37	28.90	29.50
		✓		✓	6.52	11.93	15.37	21.37	31.98	25.00	8.23	10.72	16.01	21.35	32.60	26.50
		✓			5.22	10.22	14.03	20.89	31.84	25.00	8.55	10.81	15.99	22.31	31.93	26.50
	CLIP [10]	✓	✓	✓	4.95	9.51	12.77	18.50	28.70	30.00	8.12	10.08	14.71	20.05	29.93	30.00
		✓	✓		3.79	7.53	10.56	15.90	26.07	35.00	6.71	8.78	13.66	18.93	26.57	36.50
		✓		✓	4.61	8.85	11.84	16.90	25.96	40.00	7.05	8.65	12.98	18.32	26.46	42.50
		✓			5.20	10.15	13.30	18.48	28.76	32.00	8.14	10.08	15.10	20.67	28.42	33.50
	t5-base [11]	✓	✓	✓	5.00	10.31	13.73	19.14	29.77	29.00	7.92	10.10	14.69	20.51	30.04	29.50
		✓	✓		4.81	8.90	12.43	18.39	30.20	27.00	7.12	9.60	14.48	20.64	30.22	27.50
		✓		✓	6.80	12.64	16.93	24.59	36.72	20.00	10.40	13.05	18.86	25.32	36.09	20.00
		✓			5.22	10.20	13.82	20.78	32.23	23.00	8.46	11.09	16.56	22.45	32.46	24.50
	t5-large [11]	✓	✓	✓	4.86	9.51	12.89	18.43	30.22	28.00	7.73	9.51	14.69	20.07	30.02	27.50
		✓	✓		4.70	8.87	12.20	18.64	30.20	26.00	7.37	9.60	14.83	20.80	31.09	26.00
		✓		✓	**7.21**	**14.26**	**18.39**	**25.34**	**38.55**	**19.00**	**10.93**	**13.89**	**19.80**	**26.44**	**37.34**	**20.00**
		✓			5.86	11.41	15.31	22.63	35.10	21.00	8.90	11.18	17.22	23.45	35.24	21.50
(b) All w/ threshold	TMR [9]		✓	✓	12.29	16.70	21.97	29.72	41.77	15.00	12.32	14.99	21.10	26.69	38.44	19.50
	DistilBERT [12]	✓	✓	✓	11.86	14.99	20.78	27.26	37.77	20.00	12.43	14.26	20.23	26.03	35.36	23.50
		✓	✓		9.95	14.30	19.59	26.53	37.16	20.00	10.63	11.27	17.79	23.24	33.12	23.50
		✓		✓	13.12	17.20	22.26	29.31	40.69	16.00	11.98	14.28	20.46	26.16	37.57	20.00
		✓			11.66	16.61	22.63	29.97	41.56	16.00	12.07	13.41	20.16	27.33	36.36	21.50
	CLIP [10]	✓	✓	✓	10.99	14.69	20.46	27.01	38.12	20.00	11.41	13.44	19.07	24.34	34.58	23.50
		✓	✓		8.92	11.61	17.04	23.13	34.01	24.50	10.36	11.20	17.45	23.18	31.16	29.00
		✓		✓	10.49	13.75	18.27	24.29	33.78	25.00	9.92	12.29	17.40	23.24	31.98	31.50
		✓			11.18	14.46	20.05	26.60	36.31	21.00	11.25	13.25	19.32	25.16	33.10	27.00
	t5-base [11]	✓	✓	✓	11.91	15.81	21.33	28.03	39.23	18.00	12.07	13.16	19.46	25.21	34.67	23.25
		✓	✓		11.84	16.24	22.17	29.11	42.13	18.00	10.45	11.36	17.15	23.70	33.60	23.50
		✓		✓	13.71	19.30	25.94	34.99	46.85	**12.00**	14.32	16.79	23.52	30.25	40.26	17.00
		✓			11.82	16.33	21.85	30.20	41.86	15.00	11.68	13.73	20.05	26.37	36.66	20.00
	t5-large [11]	✓	✓	✓	11.59	15.40	20.83	27.46	40.72	17.00	11.61	12.16	18.96	24.75	34.19	23.00
		✓	✓		11.77	15.49	20.73	28.74	40.33	16.00	10.86	12.02	18.75	25.14	35.15	22.00
		✓		✓	**14.23**	**20.37**	**26.30**	**34.33**	**47.90**	**12.00**	**14.71**	**17.29**	**24.68**	**31.36**	**41.67**	**16.00**
		✓			12.86	18.09	23.63	32.39	45.99	13.00	11.91	13.66	20.94	27.62	39.37	17.50
(c) Dissimilar subset	TMR [9]		✓	✓	52.00	66.00	74.00	81.00	88.00	1.00	46.00	68.00	75.00	81.00	86.00	2.00
	DistilBERT [12]	✓	✓	✓	50.00	73.00	78.00	85.00	90.00	1.50	52.00	69.00	75.00	82.00	89.00	**1.00**
		✓	✓		47.00	66.00	74.00	77.00	85.00	2.00	48.00	69.00	75.00	85.00	85.00	2.00
		✓		✓	53.00	67.00	76.00	85.00	93.00	**1.00**	52.00	68.00	**80.00**	85.00	91.00	**1.00**
		✓			48.00	68.00	78.00	85.00	90.00	2.00	47.00	68.00	79.00	86.00	90.00	2.00
	CLIP [10]	✓	✓	✓	46.00	64.00	72.00	85.00	91.00	2.00	44.00	69.00	74.00	84.00	90.00	2.00
		✓	✓		50.00	65.00	73.00	75.00	86.00	1.50	49.00	68.00	71.00	77.00	83.00	2.00
		✓		✓	44.00	61.00	69.00	76.00	80.00	2.00	45.00	58.00	66.00	75.00	80.00	2.00
		✓			47.00	62.00	73.00	79.00	84.00	2.00	46.00	62.00	72.00	76.00	86.00	2.00
	t5-base [11]	✓	✓	✓	49.00	67.00	74.00	82.00	89.00	2.00	44.00	62.00	73.00	82.00	89.00	2.00
		✓	✓		53.00	67.00	76.00	85.00	**95.00**	**1.00**	53.00	67.00	75.00	86.00	94.00	**1.00**
		✓		✓	49.00	71.00	77.00	**87.00**	94.00	2.00	51.00	69.00	79.00	**88.00**	**95.00**	**1.00**
		✓			51.00	67.00	74.00	83.00	94.00	**1.00**	50.00	66.00	75.00	83.00	92.00	1.50
	t5-large [11]	✓	✓	✓	45.00	67.00	76.00	86.00	92.00	2.00	50.00	**73.00**	**80.00**	86.00	92.00	1.50
		✓	✓		48.00	66.00	70.00	82.00	94.00	2.00	54.00	66.00	73.00	85.00	91.00	**1.00**
		✓		✓	**59.00**	**77.00**	**80.00**	**87.00**	94.00	**1.00**	53.00	**73.00**	79.00	86.00	93.00	**1.00**
		✓			54.00	72.00	77.00	**87.00**	**95.00**	**1.00**	**58.00**	69.00	76.00	86.00	94.00	**1.00**
(d) Small batches	TMR [9]		✓	✓	69.37	84.17	89.58	93.45	97.06	1.00	70.00	84.90	90.01	93.68	97.03	1.02
	DistilBERT [12]	✓	✓	✓	64.83	78.85	84.19	89.10	94.02	1.05	64.21	77.71	82.80	87.93	92.52	1.05
		✓	✓		66.70	80.59	85.99	89.90	93.34	1.03	66.42	79.74	85.06	89.03	92.79	1.01
		✓		✓	67.75	81.55	86.43	90.81	95.48	1.01	67.47	80.93	85.52	90.05	94.82	1.03
		✓			69.39	84.10	88.64	92.20	95.96	1.00	68.32	82.98	88.05	91.65	95.62	1.01
	CLIP [10]	✓	✓	✓	65.90	80.70	86.02	91.17	95.10	1.03	66.42	80.45	86.09	90.42	94.91	1.02
		✓	✓		62.73	75.96	81.66	86.11	91.24	1.08	62.00	75.16	80.27	84.99	89.37	1.08
		✓		✓	59.33	74.32	80.09	86.41	92.95	1.14	58.96	72.65	78.92	84.06	91.01	1.17
		✓			64.03	77.24	82.16	87.07	92.75	1.05	63.23	75.80	80.57	85.61	91.38	1.07
	t5-base [11]	✓	✓	✓	68.02	83.26	88.73	93.39	96.78	1.02	68.61	83.49	89.14	93.48	96.90	1.01
		✓	✓		69.02	83.65	89.67	93.29	96.74	1.01	69.30	83.62	89.55	93.59	96.76	**1.00**
		✓		✓	73.29	**86.93**	91.56	95.07	**97.97**	1.01	72.22	**86.54**	**91.35**	94.53	97.24	**1.00**
		✓			70.53	84.42	89.85	94.07	97.15	1.01	70.14	83.94	89.42	92.93	95.92	1.02
	t5-large [11]	✓	✓	✓	69.05	83.74	89.74	94.23	97.29	1.02	69.71	84.31	89.96	94.25	97.15	1.01
		✓	✓		70.39	83.92	88.50	92.63	95.94	1.01	71.17	84.03	88.85	92.77	95.76	**1.00**
		✓		✓	**74.00**	86.63	90.67	94.23	97.15	**1.00**	**73.31**	85.74	90.42	93.82	96.58	**1.00**
		✓			73.45	**86.93**	**91.77**	**95.23**	97.92	**1.00**	72.72	86.52	91.17	**95.07**	**97.56**	1.01