ReMoGPT: Part-Level Retrieval-Augmented Motion-Language Models

Qing Yu, Mikihiro Tanaka, Kent Fujiwara

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

Abstract

Generation of 3D human motion holds significant importance in the creative industry. While recent notable advances have been made in generating common motions, existing methods struggle to generate diverse and rare motions due to the complexity of motions and limited training data. This work introduces ReMoGPT, a unified motion-language generative model that solves a wide range of motion-related tasks by incorporating a multi-modal retrieval mechanism into the generation process to address the limitations of existing models, namely diversity and generalizability. We propose to focus on body-part-level motion features to enable fine-grained text-motion retrieval and locate suitable references from the database to conduct generation. Then, the motion-language generative model is trained with prompt-based question-and-answer tasks designed for different motion-relevant problems. We incorporate the retrieved samples into the prompt, and then perform instruction tuning of the motion-language model, to learn from task feedback and produce promising results with the help of fine-grained multi-modal retrieval. Extensive experiments validate the efficacy of ReMoGPT, showcasing its superiority over existing state-of-the-art methods. The framework performs well on multiple motion tasks, including motion retrieval, generation, and captioning.

Introduction

In recent years, there has been notable advancement in the development of pre-trained large language models (LLMs), *e.g.*, GPT (Radford and Narasimhan 2018; Radford et al. 2019; Brown et al. 2020; Ouyang et al. 2022), BERT (Devlin et al. 2019), T5 (Raffel et al. 2020; Chung et al. 2022) and Llama (Touvron et al. 2023a,b). These innovations have enhanced the integration of language (Zhang et al. 2022; Touvron et al. 2023a), image (Radford et al. 2021; Wang et al. 2023; Li et al. 2022; Liu et al. 2023), 3D models (Youwang, Ji-Yeon, and Oh 2022; Mohammad Khalid et al. 2022; Cao et al. 2023), and multi-modal modeling including audio (Girdhar et al. 2023; Shukor et al. 2023), leading to impressive performance in various domains. Despite these improvements in LLMs, building a pre-trained model specifically for human motion and language is still in progress. Such a motion-language model would be able to solve various motion-related tasks through prompts, potentially benefiting diverse fields, including gaming, robotics, virtual assistants, and human behavior analysis.

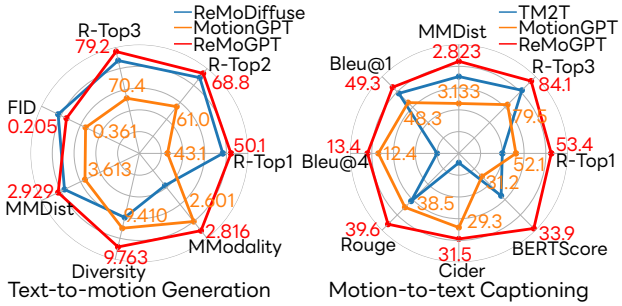

Figure 1: ReMoGPT achieves the state-of-the-art performance in text-to-motion generation and motion-to-text captioning.

Prior research on human motion has delved into diverse motion-related tasks, including motion generation (Guo et al. 2022a; Tevet et al. 2023), motion captioning (Goutsu and Inamura 2021; Guo et al. 2022b), and motion prediction (Yuan and Kitani 2020; Zhang, Black, and Tang 2021). As these works solely focus on each individual task that they were designed to solve, the resulting models cannot easily be exported to other motion-language tasks, despite the notable accomplishments. MotionGPT (Jiang et al. 2023) was proposed to solve all these individual tasks simultaneously, by converting motion clips into motion tokens and learning to generate the motion tokens and texts through fine-tuning of pre-trained language models (Raffel et al. 2020; Chung et al. 2022). However, the versatility comes at a cost, as the method shows limited performance when confronted with unconventional or infrequent conditions of text inputs.

In the context of motion generation, ReMoDiffuse (Zhang et al. 2023b) attempts to introduce a retrieval-augmentation pipeline, a common technique to enhance LLMs (Lewis et al. 2020), to address the limitation. However, motion diffusion models are sensitive to the scale in classifier-free guidance and cannot generate motion captions, limiting the range of applications. Additionally, ReMoDiffuse solely relies on the text-to-text similarity between captions using

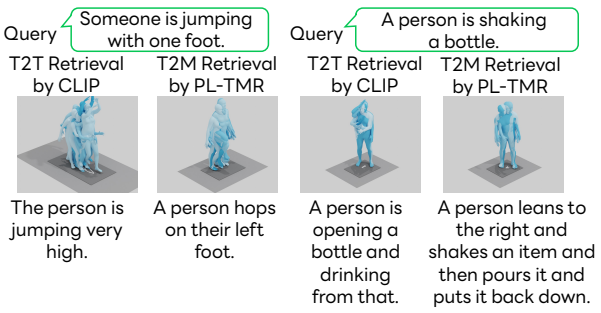

Figure 2: The comparison between samples obtained from text-to-text retrieval and text-to-motion retrieval.

CLIP (Radford et al. 2021) for the retrieval, making it difficult to retrieve the correct motions. This difficulty is caused by the diversity of motion data, where similar motions can be described by different captions (and vice versa), *e.g.*, "a person is walking forward briskly." and "the figure plants four steps leading with its left foot, with a fifth step not planted." both describe a sequence of walking. This limits the performance of retrieval-augmented generation.

In this paper, we introduce ReMoGPT, a generalized motion-language generative model enhanced with retrieval-augmentation, specifically designed to overcome the challenges outlined above. ReMoGPT is based on natural language models and uses motion tokens as the representation of motion clips. We propose a fine-grained multi-modal retrieval technique that considers the body-part-level motion features to select appropriate references from a database, and generate diverse and high-quality motion clips, achieving state-of-the-art performance in motion-related tasks as shown in Fig. 1. As Fig. 2 demonstrates, the proposed cross-modal part-level text-motion retrieval (PL-TMR) can achieve better results than text-to-text retrieval by CLIP in some cases. The retrieved samples are included in the prompts for effective guidance in motion generation and motion captioning without hindering the performance of each task nor limiting the range of applications.

We evaluate the efficacy of ReMoGPT on two standard motion generation benchmarks, namely HumanML3D (Guo et al. 2022a) and Motion-X (Lin et al. 2023). Extensive quantitative results demonstrate that ReMoGPT outperforms other existing motion-language models with the help of knowledge from the external database. With the use of multi-modal retrieval-augmented generation, ReMoGPT significantly enhances the generation quality for rare samples, demonstrating the generalizability of the proposed method.

We outline our contributions as follows:

- We propose ReMoGPT, a novel unified motion-language generative model that is augmented with multi-modal retrieval between motion and text, to solve various motion-related tasks.
- To enable efficient multi-modal retrieval-augmented generation, we propose a novel body-part-level text-motion retrieval (PL-TMR) model that captures fine-grained motion features.
- We evaluate ReMoGPT on several benchmarks and demonstrate the state-of-the-art performance of the proposed model across various tasks, including motion retrieval, generation, and captioning.

Related Works

Human Motion Generation. The task of generating motions deals with creating different and lifelike human movements using various inputs such as text (Ghosh et al. 2021; Guo et al. 2022a; Jiang et al. 2023), action (Petrovich, Black, and Varol 2021; Guo et al. 2020; Xin et al. 2023), and incomplete motion (Yuan and Kitani 2020; Zhang, Black, and Tang 2021; Ma et al. 2022; Tevet et al. 2023). Text-to-motion generation has recently garnered attention, as language is an intuitive interface for many users. MDM (Tevet et al. 2023) and MotionDiffuse (Zhang et al. 2024) generate motion using a diffusion-based generative model (Ho, Jain, and Abbeel 2020), which are also trained separately for various motion tasks. T2M-GPT (Zhang et al. 2023a) explores a generative framework using Vector Quantized Variational Autoencoders (VQ-VAE) (Oord, Vinyals et al. 2017) to quantize motion clips into discrete tokens, and train a Generative Pre-trained Transformer (GPT) for motion generation. MotionGPT (Jiang et al. 2023) considers human motion as a foreign language, by including the motion tokens transferred from motion clips in the prompt of language models to perform pre-training and instruction tuning of language models. The method is able to solve various motion-related tasks simultaneously, including motion generation and motion captioning. However, when the text inputs are sampled from unconventional or infrequent conditions, the performance of the generation model significantly degrades.

Human Motion Captioning. Describing human motion using natural language involves learning the correlation between motions and language, as demonstrated by (Takano and Nakamura 2015), which utilizes two statistical models. Additionally, recurrent networks, as explored in (Yamada, Matsunaga, and Ogata 2018; Plappert, Mandery, and Asfour 2018), have been employed for this purpose. More recently, TM2T (Guo et al. 2022b) introduces a novel motion representation method that condenses motions into a concise sequence of discrete variables. It then employs a neural machine translator (NMT) to establish mappings between the two modalities. However, the aforementioned MotionGPT (Jiang et al. 2023) is also able to solve this task, and performs better than TM2T (Guo et al. 2022b).

Text-Motion Retrieval. In recent years, vision-language foundation models have received substantial attention (Radford et al. 2021), propelled by the availability of large collections of image-text pairs gathered from the internet. These models have inspired works in motion analysis. TMR (Petrovich, Black, and Varol 2023) employs contrastive training during motion generation to align text features with motion features. MotionPatches (Yu, Tanaka, and Fujiwara 2024) proposes an image representation of motion sequences and uses pre-trained image models to extract motion features. Both methods map motion and language into the same feature space, enabling multi-modal retrieval.

Retrieval-Augmented Generation. Using retrieval to augment the generation process is a common technique for LLMs (Zhang et al. 2022; Touvron et al. 2023a) in natural language processing (Lewis et al. 2020). Recently, attempts to integrate retrieval-augmented generation methods into other domains have significantly increased. For instance, retrieval-augmented diffusion models (Blattmann et al. 2022) use user-assigned images instead of retrieval examples, enabling the effective transfer of artistic style from these images to the generated output.

In the motion domain, ReMoDiffuse (Zhang et al. 2023b) improves MotionDiffuse (Zhang et al. 2024) with a retrieval-augmentation pipeline by searching samples according to text-to-text similarity between captions. However, due to the diversity of motion and language, it is difficult to select the appropriate motion samples solely with text-to-text retrieval. Moreover, ReMoDiffuse is only applicable to motion generation. In this paper, we propose ReMoGPT, where fine-grained part-level multi-modal retrieval between motions and captions is incorporated in a unified motion-language generative model, to efficiently solve various motion-related tasks with one model.

Method

Preliminaries

Motion Tokenization. A motion with M frames, denoted as $m^{1:M} = \{m^i\}_{i=1}^{M}$ is first tokenized through a motion encoder $\mathcal{E}$ and a motion decoder $\mathcal{D}$. L motion tokens $z^{1:L} = \{z^i\}_{i=1}^{L}$, where $L = M/l$ and l is the downsampling rate, are obtained from $\mathcal{E}$, and these tokens can be decoded back into the motion as $\hat{m}^{1:M} = \mathcal{D}(z^{1:L}) = \mathcal{D}(\mathcal{E}(m^{1:M}))$ based on the VQ-VAE (Oord, Vinyals et al. 2017) architecture.

Motion-Language Models. Because a motion can also be tokenized, motion and language can be learned concurrently by merging the original text vocabulary $V_t = \{v_t^i\}_{i=1}^{K_t}$ with the motion vocabulary $V_m = \{v_m^i\}_{i=1}^{K_m}$, denoted as $V = \{V_t, V_m\}$, where V_m maintains the order of the motion codebook Z.

To solve the tasks of motion generation and captioning, (Jiang et al. 2023) proposed the use of a transformer-based model T5 (Raffel et al. 2020) as the motion-language model. The input of the model is a sequence of tokens $X_{in} = \{x_{in}{}^i\}_{i=1}^{L_{in}}$, where $x_{in} \in V$ and L_{in} denotes the input length. In the same manner, the output of the model is $X_{out} = \{x_{out}{}^i\}_{i=1}^{L_{out}}$, where $x_{out} \in V$, and L_{out} represents the output length.

The input tokens are processed by the transformer encoder, and the probability distribution of the potential next token at each step $p_\theta(x_{out} \mid x_{in}) = \prod_i p_\theta\left(x_{out}^i \mid x_{out}^{<i}, x_{in}\right)$ is predicted by the subsequent decoder in an autoregressive manner. Consequently, the training objective is as follows:

$$\mathcal{L}_{LM} = -\sum_{i=0}^{L_t-1} \log p_\theta\left(x_{out}^i \mid x_{out}^{<i}, x_{in}\right), \qquad (1)$$

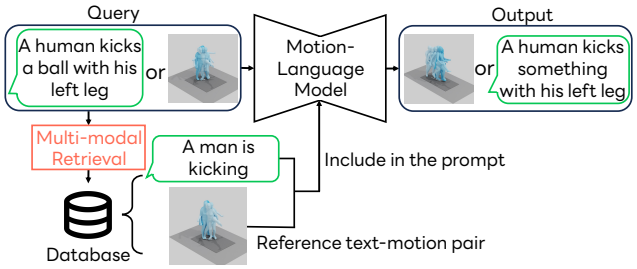

Figure 3: An illustration of the motion generation and captioning pipeline in ReMoGPT. Specifically, ReMoGPT trains a motion-language model to generate the output using the context of the retrieved motion-caption pairs.

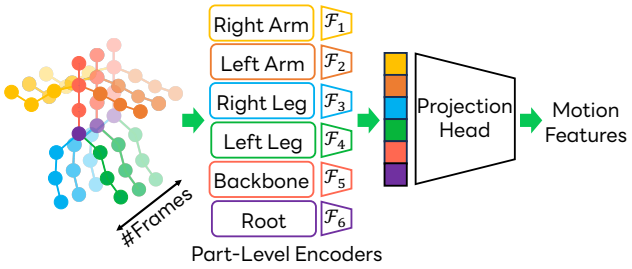

Figure 4: An overview of the proposed part-level motion encoder for text-motion retrieval.

which aims to maximize the log-likelihood of the data distribution, allowing the model to learn the relationships between motions and captions. During the inference phase, the next output token is recursively sampled from the predicted distribution $p_\theta\left(\hat{x_{out}}^i \mid \hat{x_{out}}^{<i}, x_{in}\right)$ until the end token is encountered.

ReMoGPT

We propose ReMoGPT, a unified motion-language generative model for solving various motion-related tasks in a universal format. Moreover, as shown in Fig. 3, ReMoGPT augments the existing models (Jiang et al. 2023) with the new capability of leveraging multi-modal "knowledge" from the external database using multi-modal retrieval, thus freeing the model from memorizing the appearance of rare entities.

Body Part-Level Text-Motion Retrieval. To perform multi-modal retrieval, we need to learn a function $s(m^{1:M}, t^{1:N})$ that computes the similarity between the motion $m^{1:M}$ and the caption $t^{1:N}$. The objective of $s(m^{1:M}, t^{1:N})$ is to yield a high similarity score for relevant motion-text pairs and a low score for irrelevant ones. To construct a robust multi-modal model that can incorporate fine-grained details, we propose a novel text-to-motion retrieval framework considering the body-part-level features of motion sequences. We separated the body into parts to capture fine-grained motion details that single whole-body embeddings might miss, as different parts often align with specific action semantics. Unlike existing methods TMR and Mo-

tionPatches that encode the full body into one single embedding with one motion encoder, implicitly learning inter-part relationships, our approach uses separate encodings with multiple lightweight encoders to explicitly model these relationships as shown in Fig. 4. This part-based method allows more precise retrieval of part-specific features, proving effective, especially with limited data.

3D human body models, such as SMPL (Loper et al. 2015), often utilize Kinematic Trees to represent the human skeleton through five chains as in (Jang, Park, and Lee 2022), corresponding to the limbs and backbone, for motion modeling. We adopt this approach and introduce an additional Root part to account for trajectories. Consequently, we divide the whole-body motion into six parts: Right Arm, Left Arm, Right Leg, Left Leg, Backbone, and Root. A motion sequence m is separated into part motions $m_p, p \in [1, \dots, P]$, where P is the number of parts.

Firstly, we use separated motion encoders $\mathcal{F}_p^M$, which are light-weight transformers (Vaswani et al. 2017), to obtain the embedding of p-th part motion, $\mathcal{F}_p^M(m_p)$. Then, we concatenate the embedding of each body part to build the embedding of the motion sequences as follows:

$$\mathcal{F}^M(m) = \text{Concat}[\mathcal{F}_1^M(m_1), \dots, \mathcal{F}_P^M(m_P)]. \quad (2)$$

Meanwhile, a language model $\mathcal{F}^T$ is also used to encode the caption t as $\mathcal{F}^T(t)$. Finally, projection heads are used to project the motion embedding and the text embedding into the same latent space, which are denoted as $\hat{\mathcal{F}}^M(m)$ and $\hat{\mathcal{F}}^T(t)$. Then, the similarity between the motion sequence and the caption is computed as follows:

$$s_{m\text{-}t} = \frac{\hat{\mathcal{F}}^M(m) \cdot \hat{\mathcal{F}}^T(t)}{\|\hat{\mathcal{F}}^M(m)\| \|\hat{\mathcal{F}}^T(t)\|}. \quad (3)$$

The similarity score $s_{m\text{-}t}$ increases when the motion is more closely aligned with the textual feature. Besides the motion-text similarity, text-motion similarity $s_{t\text{-}m}$, text-text similarity $s_{t\text{-}t}$ and motion-motion similarity $s_{m\text{-}m}$ can also be calculated in the same manner. To train the PL-TMR model, we employ contrastive learning on $s_{m\text{-}t}$ and $s_{t\text{-}m}$, optimizing the motion and text encoders $\mathcal{F}^M(t)$ and $\mathcal{F}^T(t)$, as well as the projection heads $\hat{\mathcal{F}}^M(t)$ and $\hat{\mathcal{F}}^T(t)$.

To build the retrieval database for ReMoGPT, we simply use all the training data as entities. In text-to-motion generation, where the prompt is the caption, we rank all elements according to the scores of $s_{t\text{-}m}$ and $s_{t\text{-}t}$, respectively. In the task of motion captioning, *i.e.*, the prompt is the motion, the scores of $s_{m\text{-}t}$ and $s_{m\text{-}m}$ are calculated instead. The top k similar samples in each score are then chosen as the retrieved samples (m_i, t_i). The retrieved original motions are then encoded as quantized tokens via VQ-VAE and included in the prompt for the following instruction tuning of motion-language models. Due to the efficiency of the retrieval model in fetching similar samples for top results, we find that using $k = 1$, *i.e.*, two retrieved samples in total, is sufficient to achieve good performance, as detailed in our experiment.

Instruction Tuning. To enhance the basic motion-language model (Jiang et al. 2023) with multi-modal retrieval, we simply perform instruction tuning by including the retrieved samples in the prompt. Following (Jiang et al. 2023), several instruction prompts are designed for motion generation and motion captioning as shown in Fig. 5. For instance, an instruction prompt for the motion generation task could be "Show me a motion that illustrates <Caption_Placeholder>" and for the motion captioning task, the instruction prompt could be "Describe the motion illustrated in <Motion_Placeholder>", where <Caption_Placeholder> denotes the caption, and <Motion_Placeholder> denotes a sequence of motion tokens generated by the motion tokenizer, respectively. Different from the existing method, we include the retrieval results as the context in the prompt to provide more informative features for the generation. Therefore, during the training phase, the goal is to maximize the log-likelihood of the data distribution as follows:

Text-to-Motion Generation	Motion-to-Text Captioning
### Input: Generate motion: A person is shaking a bottle. ### Context: Text: a person picks something up and shakes his hand up. Motion: <Motion_Placeholder_R1> Text: this person is shaking his hand with his right hand. Motion: <Motion_Placeholder_R2> ### Output: <Motion_Placeholder>	### Input: Generate text: <Motion_Placeholder> ### Context: Text: the person in sitting down eating dinner. Motion: <Motion_Placeholder_R1> Text: the person is chopping opinions. Motion: <Motion_Placeholder_R2> ### Output: A person is chopping vegetables.

Figure 5: Samples of prompt used for the instruction tuning in ReMoGPT. <Motion_Placeholder> denotes the motion tokens paired with the caption. <Motion_Placeholder_R1(R2)> denote the motion tokens of multi-modal retrieved motion-caption pairs.

$$\mathcal{L}_{LM} = -\sum_{i=0}^{L_t-1} \log p_\theta \left(x_{out}^i \mid x_{out}^{<i}, x_{in}, x_{rag} \right), \quad (4)$$

where x_{rag} is the samples selected by the multi-modal retrieval according to the source tokens. During the inference phase, the target tokens are also generated with the help of x_{rag} from the external database.

Experiments

Experimental Setup

Datasets. To show the effectiveness of the proposed method, we evaluate the proposed method under motion generation and motion captioning. We use two text-to-motion datasets: HumanML3D (Guo et al. 2022a) and Motion-X (Lin et al. 2023) in the experiments. HumanML3D is a dataset that includes 14,616 motion clips sourced from AMASS (Mahmood et al. 2019), along with 44,970 sequence-level textual descriptions. Motion-X is a more recent dataset with over 81,051 motion clips and captions, which is about 3 times as large as HumanML3D.

Text-to-motion retrieval

Methods	#Params	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR	82M	8.92	12.04	16.33	22.06	33.37	25.00
MotionPatches	152M	10.80	14.98	20.00	26.72	38.02	19.00
PL-TMR	118M	**11.00**	**17.02**	**22.18**	**29.48**	**43.43**	**14.00**

Motion-to-text retrieval

Methods	#Params	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
TMR	82M	9.44	11.84	16.90	22.92	32.21	26.00
MotionPatches	152M	11.25	13.86	19.98	26.86	37.40	20.50
PL-TMR	118M	**12.25**	**14.95**	**21.45**	**28.34**	**39.11**	**19.00**

Table 1: Results of text-to-motion and motion-to-text retrieval benchmark on HumanML3D.

HumanML3D is also a subset of Motion-X. Although KIT (Plappert, Mandery, and Asfour 2016) is another popular dataset, its scale is much smaller than HumanML3D, so we use Motion-X instead to illustrate the scalability of our method. For fair comparisons with previous works, we adopt the same motion representation as (Guo et al. 2022a), consisting of joint velocities, positions, and rotations. In our experiments, the proposed method is trained exclusively on either the HumanML3D or Motion-X datasets.

Evaluation Metrics. We first evaluate our text-motion retrieval model with Recall at various ranks (R@1, R@2, etc.) following (Petrovich, Black, and Varol 2023). Recall at rank k indicates the percentage of instances where the correct label appears within the top k results. Additionally, we calculate the median rank (MedR), where a lower value shows better performance. Then the performance of the proposed ReMoGPT is also evaluated following previous works (Guo et al. 2022a; Jiang et al. 2023) based on four aspects: (1) Motion quality with Frechet Inception Distance (FID), (2) Generation diversity with Diversity and multi-modality (MModality), (3) Text matching with the Top 1/2/3 precision of motion-retrieval (RPrecision) and Multi-modal Distance (MM Dist), (4) Linguistic quality in the task of motion captioning with BLEU (Papineni et al. 2002), Rouge (Lin 2004), Cider (Vedantam, Zitnick, and Parikh 2015), and BertScore (Zhang et al. 2020).

Implementation Details. For the text-motion retrieval model, we mainly follow the setting of MotionPatches (Yu, Tanaka, and Fujiwara 2024) and implement the motion encoder with a 4-layer Transformer (Vaswani et al. 2017) for each body part and the dimension of each part-level motion embedding is set to 512. For the motion-language model, we implement our method based on the code of MotionGPT (Jiang et al. 2023). The size of the codebook for the motion tokenizer is set as 512 and the temporal downsampling rate l is set as 4 in the motion encoder. As the base language model, T5 (Raffel et al. 2020) is used with 12 layers in both the transformer encoder and decoder. The AdamW optimizer is used in all the models for training. The motion tokenizers and the language model are first trained following (Jiang et al. 2023). Then, to perform the instruction tuning with retrieved samples, we further train the model with a learning rate of 10^{-4} and a mini-batch size of 16 for 200 epochs. All models are trained on 8 Tesla A100 GPUs.

Methods	RPrecision↑			FID↓	MMDist↓	Diversity↑	MModality↑
	Top1	Top2	Top3				
Real	0.511	0.703	0.797	0.002	2.974	9.503	-
TM2T	0.424	0.618	0.729	1.501	3.467	8.589	2.424
T2M	0.457	0.639	0.740	1.067	3.340	9.188	2.090
MotionDiffuse	0.491	<u>0.681</u>	<u>0.782</u>	0.630	3.113	9.410	1.553
MDM	0.320	0.498	0.611	0.544	5.566	9.559	<u>2.799</u>
MLD	0.481	0.673	0.772	0.473	3.196	<u>9.724</u>	2.413
T2M-GPT	0.491	0.680	0.775	**0.116**	3.118	9.761	1.856
ReMoDiffuse†	<u>0.492</u>	0.678	0.775	<u>0.137</u>	<u>3.091</u>	9.208	1.755
MotionGPT†	0.407	0.569	0.657	0.224	4.022	9.369	2.325
MotionGPT‡	0.431	0.610	0.704	0.361	3.613	9.410	2.601
ReMoGPT	**0.501**	**0.688**	**0.792**	0.205	**2.929**	**9.763**	**2.816**

Table 2: Results of text-to-motion generation on HumanML3D. MModality is empty for real motions because it is deterministic. The evaluation metrics are computed with the encoder used in (Guo et al. 2022a). The results of methods marked with † are re-evaluated with their official source code and released pre-trained models, and those marked with ‡ are re-trained with our codebase and tasks for fair comparison. **Bold** and <u>underline</u> indicate the best and the second best results.

Methods	RPrecision↑			FID↓	MMDist↓	Diversity↑	MModality↑
	Top1	Top2	Top3				
Real	0.248	0.371	0.453	0.001	4.895	6.850	-
MotionGPT	0.188	0.293	0.369	0.635	5.592	6.630	3.853
ReMoGPT	**0.235**	**0.360**	**0.435**	**0.352**	**5.083**	**6.902**	**4.153**

Table 3: Results of text-to-motion generation on Motion-X.

Results of Text-Motion Retrieval

In our evaluation of text-to-motion and motion-to-text retrieval benchmarks using the HumanML3D dataset (Table 1), we provide a comprehensive comparison against prior works, TMR (Petrovich, Black, and Varol 2023) and MotionPatches (Yu, Tanaka, and Fujiwara 2024). Remarkably, our model PL-TMR consistently outperforms these prior methods across all evaluation metrics. Although we introduced a separated motion encoder for each body part, the number of parameters of the proposed method is smaller than MotionPatches due to the usage of light-weight transformers. This indicates that our model effectively captures the nuanced nature of motion descriptions, providing more accurate and contextually relevant retrieval results with the body-part-level retrieval.

Results of Text-to-Motion Generation

The text-to-motion generation task involves generating human motion clips from a given text input. We compare our proposed ReMoGPT model with other state-of-the-art models (Guo et al. 2022b,a; Tevet et al. 2023; Xin et al. 2023; Zhang et al. 2023a), including ReMoDiffuse (Zhang et al. 2023b) and MotionGPT (Jiang et al. 2023).

In the experiments on HumanML3D, to ensure a fair comparison, we re-evaluate these two methods using their official source code and the released pre-trained models. Moreover, we re-trained MotionGPT with our codebase to fo-

Methods	RPrecision↑		MMDist↓	Length$_{avg}$↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
	Top1	Top3							
Real	0.523	0.828	2.901	12.75	-	-	-	-	-
TM2T	0.516	0.823	2.935	10.67	48.9	7.00	38.1	16.8	32.2
MotionGPT†	0.523	0.799	2.986	12.01	48.2	12.2	38.6	29.2	33.2
MotionGPT‡	0.521	0.795	3.133	12.08	48.3	12.4	38.5	29.3	31.2
ReMoGPT	**0.534**	**0.841**	**2.823**	**12.12**	**49.3**	**13.4**	**39.6**	**31.5**	**33.9**

Table 4: Results of motion-to-text captioning on HumanML3D.

Methods	RPrecision↑		MMDist↓	Length$_{avg}$↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
	Top1	Top3							
Real	0.208	0.412	4.997	10.02	-	-	-	-	-
MotionGPT	0.209	0.398	5.394	9.33	43.1	16.2	38.1	45.5	26.5
ReMoGPT	**0.216**	**0.414**	**5.121**	**9.71**	**45.7**	**17.2**	**38.7**	**50.2**	**27.9**

Table 5: Results of motion-to-text captioning on Motion-X.

cus on motion generation and captioning. Other results are sourced directly from their original papers or the scores reported in (Guo et al. 2022a). The results of text-to-motion generation on HumanML3D are shown in Table 2, and our method achieves the best results in most metrics. It is also noticeable that our method significantly improves the original MotionGPT with retrieval augmentation.

Moreover, the results of text-to-motion generation for Motion-X are summarized in Table 3. Because Motion-X is a newly released dataset, we only compare our method with the model trained with MotionGPT. The comparison shows that our method outperforms MotionGPT by a large margin, indicating the scalability of our proposal.

Results of Motion-to-Text Captioning

The motion-to-text captioning task involves generating a text description based on a given human motion sequence. We compare our method with the existing state-of-the-art methods TM2T (Guo et al. 2022b) and MotionGPT (Jiang et al. 2023). The performance is evaluated following the metrics used in (Guo et al. 2022b; Jiang et al. 2023). The results on HumanML3D in Table 4 illustrate that the proposed method outperforms MotionGPT in generating text descriptions of given motions. According to the comparisons between the proposed method and MotionGPT on Motion-X in Table 5, our model consistently outperforms the existing method even when the dataset becomes more challenging.

Analysis

Qualitative Results

In Fig. 6, we present the qualitative results for text-to-motion generation on the entire test set of HumanML3D. Each query text is displayed on the left, and on the right, we showcase the motions generated by ReMoDiffuse (Zhang et al. 2023b), MotionGPT (Jiang et al. 2023) and the proposed method. As a reference, we also calculate the text-to-text similarity of CLIP (Radford et al. 2021) between the query prompt and the captions in the training data and show the closest results under each query. As shown in Fig. 6, ReMoDiffuse and MotionGPT struggle to generate motions when the text description is novel to the model but similar

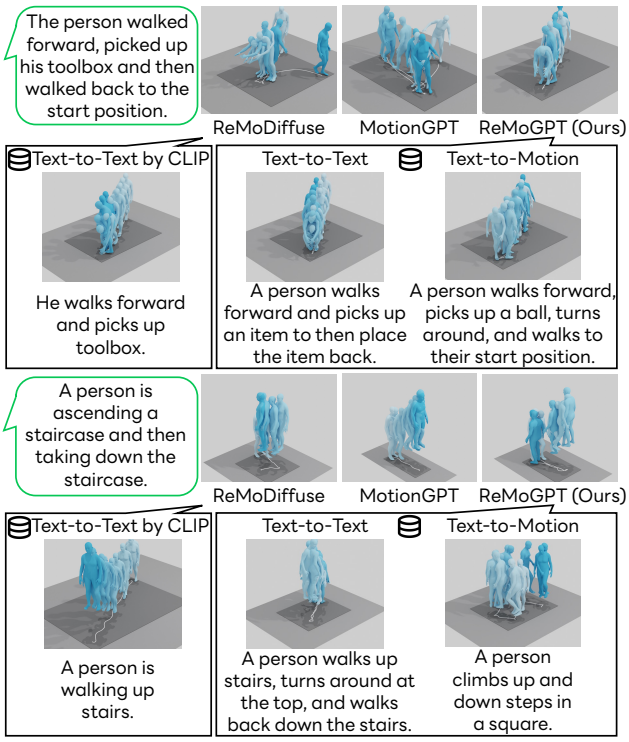

Figure 6: Qualitative results of text-to-motion generation. For each query, we show the rendered motions generated by each model. We also show the retrieved motions and captions as the reference.

motions are in the database. Specifically, because ReMoDiffuse only uses text-to-text similarity, the generated motions of ReMoDiffuse are closer to the retrieved captions rather than the query prompt. Meanwhile, our method can correctly find these motions and utilize them as the reference data for appropriate generation.

Ablation Studies

In this section, we explore various settings to better understand the factors influencing the performance of our model.

Retrieval Method. In the main experiments, we used text-to-motion and text-to-text retrieval for motion generation, and motion-to-text and motion-to-motion retrieval for motion captioning. We changed the modality of the retrieval and the number of retrieval samples on each task on HumanML3D dataset. The results are shown in Table 6. We observe that using the retrieval augmentation can improve the performance of the motion-language model. It is noticeable that the results of multi-modal retrieval are better than those of single-modal retrieval, which shows retrieval in both motion and language domains is important for the generation because different modalities can complement each other in some cases. We find that our method is not sensitive to the number of retrieved samples because similar samples are selected by the retrieval model in each modality. We also tried using four samples in total to augment the model, but the

Retrieval Modality	#Retrieved Samples	Text-to-Motion			Motion-to-Text		
		R-Top1↑	FID↓	MMDist↓	R-Top1↑	Bleu@4↑	Cider↑
-	-	0.431	0.361	3.613	0.521	12.421	29.349
T2T or M2M	1	0.472	0.232	3.393	0.522	12.721	30.135
T2T or M2M	2	0.469	0.235	3.381	0.524	12.887	30.239
T2M or M2T	1	0.479	0.231	3.282	0.528	12.925	31.054
T2M or M2T	2	0.481	0.224	3.180	0.529	12.915	31.421
Both	1+1	**0.501**	**0.205**	**2.929**	0.534	**13.412**	31.531
Both	2+2	0.498	0.207	3.086	**0.542**	13.391	**32.012**
Both (TMR)	1+1	0.493	0.218	3.012	0.532	13.312	31.231
Both (MotionPatches)	1+1	0.497	0.210	2.984	0.534	13.387	31.498

Table 6: Comparison of the retrieval methods on the HumanML3D dataset in motion generation and captioning.

Training Data	External Database	Text-to-Motion			Motion-to-Text		
		R TOP1↑	FID↓	MMDist↓	R TOP1↑	Bleu@4↑	Cider↑
HumanML3D	HumanML3D	**0.501**	0.205	2.929	0.534	13.412	31.531
	Motion-X	0.498	**0.189**	**2.921**	**0.541**	**14.145**	**33.021**
Motion-X	HumanML3D	0.223	0.608	5.318	0.209	16.731	49.821
	Motion-X	**0.235**	**0.352**	**5.083**	**0.216**	**17.216**	**50.244**

Table 7: Comparison of the external database for retrieval.

performance gain was limited. One possible reason for this may be that because the string of motion tokens is long, the T5 (Raffel et al. 2020) language model we used cannot handle these very long inputs effectively with the limited scale of motion data. We also compared the proposed method with existing multi-modal retrieval methods, and our part-level retrieval achieves the best performance.

External Database. We used the training set of each dataset as the external database for the retrieval augmentation. To illustrate the influence of the external database, we additionally used Motion-X as the database in the experiment of HumanML3D and vice versa. The results are shown in Table 7. Because HumanML3D is a subset of Motion-X, using Motion-X in the generation of HumanML3D can improve the performance of the model with retrieval augmentation. Although the performance of using HumanML3D as the database in the experiments of Motion-X degraded slightly, it still performed better than that of MotionGPT in the previous experiments.

Results of Rare Motion Generation

To evaluate the scalability of our proposed method in generating rare samples, and to ensure that a retrieval-augmented generation approach does not solely rely on memorization, we follow the metrics outlined in ReMoDiffuse (Zhang et al. 2023b) for text-to-motion generation of rare motions. The maximum cosine similarity between the given prompt and the captions of the training data is used in the calculation for rareness of each sample. If this similarity is larger, then rareness of the test data will be lower, and vice versa. We compute the MMDist and Diversity specifically for the top 5% of rare motions. Additionally, we calculate the average MMDist for each range of rareness, referred to as Balanced MMDist. The results are shown in Table 8, and our ReMoGPT outperforms other methods. We further show the retrieved and generated samples of rare motions, ranked in the top 5 in terms of rareness, in Fig. 7. Although the retrieved samples do not completely match the query, our method is still effective in generating proper motions from the hints these samples provide.

Method	MMDist ↓	Top 5% MMDist ↓	Balanced MMDist ↓	Diversity ↑	Top 5% Diversity ↑
MotionDiffuse	3.113	4.872	4.525	9.012	8.841
ReMoDiffuse	3.091	4.317	3.936	9.208	8.969
MotionGPT	3.613	4.421	4.161	9.354	8.889
ReMoGPT	**3.001**	**3.563**	**3.056**	**9.701**	**9.117**

Table 8: Results of rare motion generation on the HumanML3D dataset.

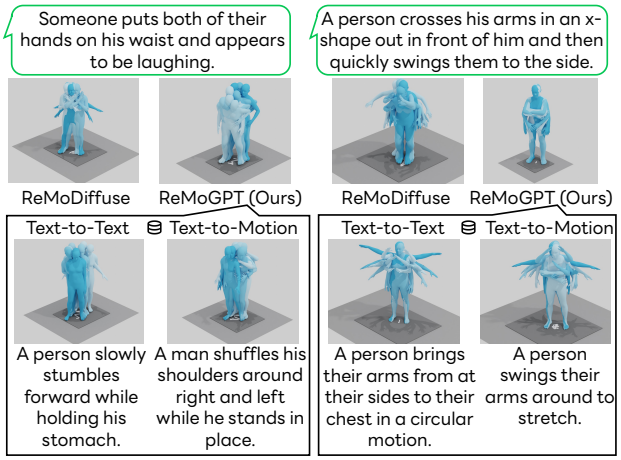

Figure 7: The retrieval and generation of rare motions.

Limitations

The proposed retrieval-augmented generation can improve generation performance with the external database. However, we find that the weaknesses of the proposed method include a limitation in handling completely new motions. If we want to generate completely new motions, *e.g.*, a yoga dataset (Tripathi et al. 2023) containing complex yoga actions, the model struggles to generate similar yoga motions even when using the dataset as the external database without training. One reason for this observation is the limited generalization ability of the motion tokenizer. If the motion is very different from the distribution of training data, the decoder of the motion tokenizer cannot reconstruct it correctly. Increasing the volume of training data to train the model is a potential solution to this problem in future work.

Conclusion

In this paper, we introduced a novel unified motion-language generative model, named ReMoGPT, to simultaneously solve text-to-motion generation and motion-to-text captioning. Our approach effectively addresses challenges in motion-related tasks with the augmentation of samples from the proposed PL-TMR. As a result, we have made significant advancements in motion generation and captioning even in cases where the samples can rarely be found in the data.

References

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-augmented diffusion models. In *NeurIPS*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *ICCV*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Goutsu, Y.; and Inamura, T. 2021. Linguistic descriptions of human motion with generative adversarial seq2seq learning. In *ICRA*.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *CVPR*.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Jang, D.-K.; Park, S.; and Lee, S.-H. 2022. Motion puzzle: Arbitrary motion style transfer by body part. *ACM TOG*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. In *NeurIPS*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. In *NeurIPS*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM TOG*.

Ma, H.; Li, J.; Hosseini, R.; Tomizuka, M.; and Choi, C. 2022. Multi-objective diverse human motion prediction with knowledge distillation. In *CVPR*.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*.

Mohammad Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia*.

Oord, A. V. D.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. *ICCV*.

Petrovich, M.; Black, M. J.; and Varol, G. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *ICCV*.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The kit motion-language dataset. *Big Data*.

Plappert, M.; Mandery, C.; and Asfour, T. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *CVPR*.

Radford, A.; and Narasimhan, K. 2018. Improving Language Understanding by Generative Pre-Training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

Shukor, M.; Dancette, C.; Rame, A.; and Cord, M. 2023. Unified model for image, video, audio and language tasks. *TMLR*.

Takano, W.; and Nakamura, Y. 2015. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Bermano, A. H.; and Cohen-Or, D. 2023. Human motion diffusion model. In *ICLR*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tripathi, S.; Müller, L.; Huang, C.-H. P.; Omid, T.; Black, M. J.; and Tzionas, D. 2023. 3D Human Pose Estimation via Intuitive Physics. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. In *CVPR*.

Xin, C.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; Yu, J.; and Yu, G. 2023. Executing your commands via motion diffusion in latent space. In *CVPR*.

Yamada, T.; Matsunaga, H.; and Ogata, T. 2018. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*.

Youwang, K.; Ji-Yeon, K.; and Oh, T.-H. 2022. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*.

Yu, Q.; Tanaka, M.; and Fujiwara, K. 2024. Exploring Vision Transformers for 3D Human Motion-Language Models with Motion Patches. In *CVPR*.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023a. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *T-PAMI*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023b. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. In *ICCV*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Zhang, Y.; Black, M. J.; and Tang, S. 2021. We are more than our joints: Predicting how 3d bodies move. In *CVPR*.