MotionGPT: Human Motion Synthesis with Improved Diversity and Realism via GPT-3 Prompting

Jose Ribeiro-Gomes*[1] Tianhui Cai*[2] Zoltán Á. Milacski[2] Chen Wu[2] Aayush Prakash[3]
Shingo Takagi[3] Amaury Aubel[3] Daeil Kim[3] Alexandre Bernardino[1] Fernando De La Torre[2]
[1]Instituto Superior Tecnico, Lisbon, Portugal [2]Carnegie Mellon University, USA [3]Meta, USA
{josepgomes@,alex@isr.}tecnico.ulisboa.pt {tcai2,zmilacsk,chenwu2}@andrew.cmu.edu
{aayushp,shingo.takagi,amauryaubel}@meta.com daeil.kim@gm.slc.edu ftorre@cs.cmu.edu

Abstract

*There are numerous applications for human motion synthesis, including animation, gaming, robotics, or sports science. In recent years, human motion generation from natural language has emerged as a promising alternative to costly and labor-intensive data collection methods relying on motion capture or wearable sensors (*e.g.*, suits). Despite this, generating human motion from textual descriptions remains a challenging and intricate task, primarily due to the scarcity of large-scale supervised datasets capable of capturing the full diversity of human activity.*

This study proposes a new approach, called MotionGPT, to address the limitations of previous text-based human motion generation methods by utilizing the extensive semantic information available in large language models (LLMs). We first pretrain a doubly text-conditional motion diffusion model on both coarse ("high-level") and detailed ("low-level") ground truth text data. Then during inference, we improve motion diversity and alignment with the training set, by zero-shot prompting GPT-3 for additional "low-level" details. Our method achieves new state-of-the-art quantitative results in terms of Fréchet Inception Distance (FID) and motion diversity metrics, and improves all considered metrics. Furthermore, it has strong qualitative performance, producing natural results. Code is available at https://github.com/humansensinglab/MotionGPT

1. Introduction

Human motion synthesis is an actively researched field that has vast potential applications in areas such as animation, gaming, robotics, medical research, and activity recognition. For instance, it can create more natural movements for animated characters [18], help develop prosthetic devices that mimic human movements [14] , and augment motion data to improve performance in downstream applications [50]. Traditional methods for synthesizing human motion rely on data collection with expensive and well-calibrated motion capture systems or wearable sensors [29, 31], as well as human labor to perform, clean and annotate the motion data. This makes it challenging to acquire large amounts of data for each activity.

Recent developments in generative modeling have led to various techniques that allow for the direct synthesis of motion sequences from natural language descriptions, with minimal human involvement and without the need for costly capture systems. These methods aim to generate plausible SMPL-compatible [26] parameters (*i.e.*, global translation, global orientation and joint rotation angles representing the body pose) that accurately reflect the given text. Popular approaches include CLIP-aligned autoencoders [29], and both Variational Autoencoders (VAEs) [4, 22, 35, 50] and Denoising Diffusion Probabilistic Models (DDPMs) [17, 46, 47, 51, 59] conditioned on text embeddings.

However, these methods have a major drawback of suffering from the moderate sizes and the ambiguous nature of the texts in their training datasets, which typically contain only coarse *"high-level"* descriptions. Consequently, they cannot generate detailed motions that correspond to fine-grained *"low-level"* texts, limiting their generalization, diversity, and controllability. For instance, as shown in Fig. 1, the "high-level" activity of *"greet a friend"* can involve a range of "low-level" variability, including a *"handshake"*, a *"wave"*, a *"hug"*, a *"kiss"*, or a *"bow"*, depending on culture, the situation, and personal relationships. It remains unclear how existing methods could capture such variability without larger training datasets with more detailed text annotations.

This paper introduces MotionGPT, a doubly text-conditional motion diffusion model that addresses the afore-

* denotes equal contribution

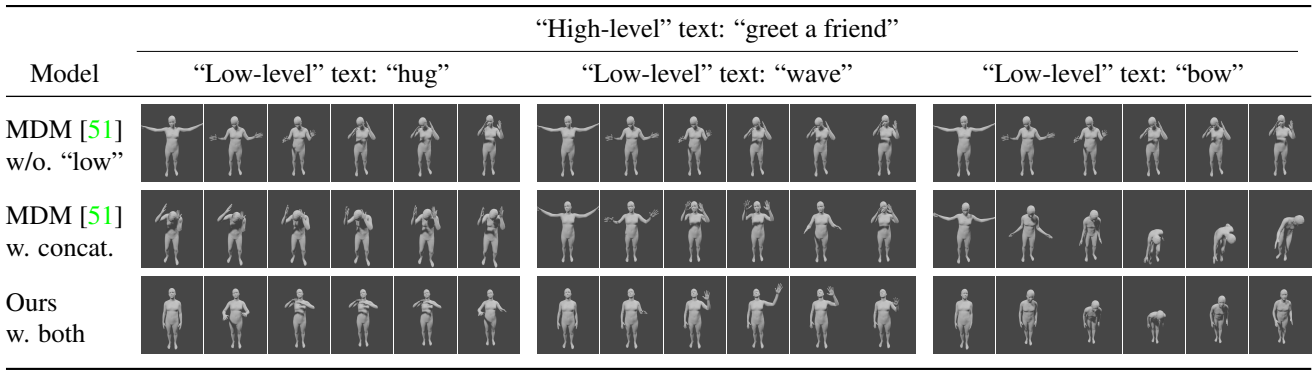


Figure 1. Example demonstrating the advantages of our proposed double text conditioning (best viewed zoomed in). For each scenario, we use the same coarse "high-level" text "greet a friend", but we vary the fine-grained "low-level" description: "hug", "wave", and "bow". MDM [51] uses only the "high-level" text and ignores the "low-level" completely. MDM with concatenated "high-" and "low-level" text generates more consistent but poor quality motions. Our model conditioned on both texts yields consistent and good quality motions.

mentioned limitation with zero-shot "common-sense" inference. We pretrain the motion diffusion model using both ground truth "high-level" HumanML3D [51] and "low-level" BABEL [36] annotations. During inference, the "low-level" texts are obtained by zero-shot prompting the GPT-3 [9] large language model (LLM) using the "high-" and "low-level" pairs of the training dataset along with a "high-level" query from the test dataset. This approach increases motion diversity by adding randomness and improves generalization by aligning test samples with the training dataset. Our proposed method achieves new state-of-the-art quantitative results in text-to-motion generation in terms of Fréchet Inception Distance (FID) and motion diversity, as well as strong qualitative performance.

Our contributions are summarized as follows:

- A motion diffusion model conditioned on both "high-" and "low-level" descriptions for fine-grained control over generated motions.
- Zero-shot prompting GPT-3 for "low-level" details of "high-level" texts, resulting in more motion diversity and better alignment with the training dataset.
- New state-of-the-art performance on the HumanML3D [51] benchmark in terms of FID and motion diversity.

2. Related Work

2.1. Human Motion Generation

Generative models for human motion synthesis can be categorized into two types: supervised and unsupervised methods.

Supervised methods aim to approximate input-to-target mappings from datasets of respective pairs. *E.g.*, one may estimate 3D human pose targets from monocular image [7, 33], video [23] or 2D pose [12, 28] inputs. However, in practice, target data may not always be readily available or sufficient for proper generalization.

Therefore, an alternative approach is to apply unsupervised learning by solely utilizing the input data. *E.g.*, predicting future motion frames can be achieved by masking out past frames from the input [3, 5, 10, 58]. Similarly, one may attempt to reconstruct the whole input sequence from a compressed representation via autoencoders or decoder-only models. Such schemes are either unconditional (purely unsupervised) or conditional (the latent code contains some supervisory signal). Unconditional methods have limited control over the generated motions, and are typically employed as motion priors in pose estimation [23, 33]. In contrast, conditional procedures offer more control. In particular, the emergence of labeled action recognition datasets such as NTU-RGB-D [44] has sparked a group of class-conditional solutions [11, 16, 34]. These models can generate a variety of motions from the same action class, but precise control over the outputs is still limited by the number of classes. As a solution, the emergence of motion datasets with natural language annotations, such as BABEL [36] and HumanML3D [15], has led to the development of text-conditional generative models that aim to take advantage of the compositionality of language. Notable examples include Conditional Variational Autoencoders (cVAEs) [4,35] and Conditional Generative Adversarial Networks (cGANs) [1]. More recently, Conditional Denoising Diffusion Probabilistic Models (cDDPMs) [21, 51, 59] have been proposed to address the many-to-many nature of text-to-motion generation. By replacing compression with iterative denoising, these models produce more diverse and detailed motions.

In this paper, we extend the Human Motion Diffusion Model (MDM) [51], a diffusion-based motion generative model that is conditioned on "high-level" text input. While

MDM can generate reliable motions when the text input is specific and similar to the training set, it struggles with ambiguous or difficult-to-generalize inputs, as shown in Fig. 4. To overcome this limitation, we propose a method that leverages the common-sense knowledge embedded in the GPT-3 language model to break down "high-level" test descriptions into more detailed and randomly-generated "low-level" variants that are similar to those of the training dataset. By incorporating this additional "low-level" text conditioning signal, we achieve better generalization due to the increased similarity to the training dataset, as well as more motion diversity thanks to the randomness.

2.2. Large Language Models and Zero-Shot Prompting

Large Language Models (LLMs) [9, 13, 37] are transformer models in Natural Language Processing (NLP) that are pretrained on vast amounts of text data. Pretraining is accomplished by solving unsupervised *pretext* tasks, such as masked word prediction, to learn generic feature representations. These features can then be used to solve downstream problems, such as question answering [20, 53], language translation [49, 62], sentiment analysis [24, 57], or guiding methods in other modalities by leveraging common-sense knowledge [8, 55].

One way to use the pretrained model is by fine-tuning it for the downstream task [19, 40, 41, 48]. However, this may require a considerable amount of training data.

Recently, it has been demonstrated that LLMs can also generalize to novel problems using zero-shot prompting [39, 42, 43]. The key idea behind this approach of in-context learning is to learn from analogy, wherein language models are able to learn tasks given only a few demonstration examples [9]. Though it has been used in other areas [30, 45], its use in human motion generation is novel. During inference, by providing a prompt comprising a small number of in-context input-target examples and an additional query input, the LLMs can identify the patterns in the examples and combine them with the pretrained semantic information to generate the query output. Prompt design is crucial for good performance [56, 60] and it has been shown that using more semantically similar examples to the query case yields better results [25].

In this work, we adopt zero-shot prompting for the GPT-3 LLM to augment "high-level" test motion descriptions into their random "low-level" counterparts with additional details on the action and increased similarity to the training dataset. We then condition a motion diffusion model on both texts.

3. Methods

In this section, we start by formulating the task (Section 3.1), followed by introducing our model, MotionGPT (Section 3.2), and our GPT-3 prompting (Section 3.3).

3.1. Problem Formulation

Our objective is to generate a sequence of 3D human motion parameters $\mathbf{x}^{1:N} = \left[\mathbf{x}^1, \dots, \mathbf{x}^N\right] \in \mathbb{R}^{N \times F}$ that corresponds to a given text conditioning $c \in \mathbb{R}^C$. Here, N denotes the variable temporal length of the motion sequence (*i.e.*, the number of frames) and F is the number of parameters per frame.

Text conditioning. The conditioning signal c is an embedding of a real-world text, such as a natural language sentence that describes an action or how it should be performed. Specifically, we utilize CLIP-embeddings of the descriptions contained within the HumanML3D [15] dataset. Throughout this study, we refer to these as the *"high-level"* texts.

Motion parameters. In line with related works in the field, we adopt a body motion representation that is compatible with SMPL [26], which includes 6D parent-relative joint rotation angles [61] for J body joints. Additionally, we incorporate a 3D global translation and a 4D global orientation in quaternion representation to facilitate spatial arrangement. To improve the training process, we also include extra redundant parameters, such as joint locations and angular velocities, as proposed by [15].

3.2. MotionGPT

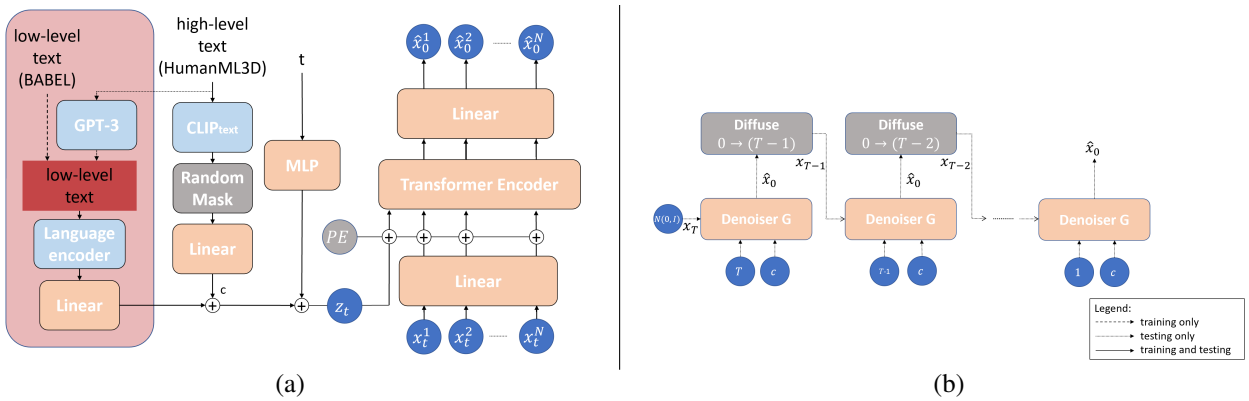
Our proposed MotionGPT architecture builds upon Human Motion Diffusion Model (MDM) [51] and introduces a second conditioning with a more detailed text, which we refer to as the *"low-level"* description, to enable fine-grained control over the embedding space and the generated motion. In addition to the "high-level" description of the motions, we train our model with ground-truth "low-level" texts, and use GPT-3 prompting (Section 3.3) during inference to achieve greater diversity and more similarity to the training data distribution. In this subsection, we provide an overview of our framework, which is illustrated in Fig. 2.

Diffusion model. We utilize a Denoising Diffusion Probabilistic Model (DDPM) [38, 51] for generating motions.

The forward Markov noising process initiates with the training sample $\mathbf{x}_0^{1:N} = \mathbf{x}^{1:N}$, and gradually adds noise to it, *i.e.*, for time steps $t = 1, \dots, T$:

$$q(\mathbf{x}_t^{1:N} | \mathbf{x}_{t-1}^{1:N}) = \mathcal{N}\left(\sqrt{\alpha_t} \cdot \mathbf{x}_{t-1}^{1:N}, (1 - \alpha_t) \cdot I\right). \quad (1)$$

The $\alpha_t \in (0, 1)$ values are constant hyperparameters. For small α_t, the forward process can be approximated as $\mathbf{x}_t^{1:N} \sim \mathcal{N}(0, I)$.



(a) (b)

Figure 2. Schematic diagram of our proposed MotionGPT architecture. Best viewed in color. "Low-level" conditioning contribution is highlighted in the pink box. (a) Denoising network G. The model receives a noisy motion sequence $\mathbf{x}_t^{1:N}$, a noising step t, and two text embeddings ("high-" and "low-level" encoded by CLIP [37] and a language model, respectively) as input, and predicts a clean motion $\hat{\mathbf{x}}_0^{1:N}$. We pretrain the network with ground truth text annotations, whereas during inference, we generate the "low-level" text using zero-shot prompting with GPT3. (b) Sampling. Given the text conditioning c and a Multivariate Standard Normal noise $\mathbf{x}_T$, we alternate between the denoising network G and the backward noising process with gradually decreasing number of steps.

Subsequently, the reverse denoising process trains a denoising neural network G to reconstruct the original motion input $\mathbf{x}_0^{1:N}$ by approximately inverting multiple steps of the forward process:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}_0^{1:N} \sim q(\mathbf{x}_0^{1:N}|c), t\sim[1,T]} \left[\|\mathbf{x}_0^{1:N} - G(\mathbf{x}_t^{1:N}, t, c)\|_2^2 \right]. \quad (2)$$

During inference, one may start with $\mathbf{x}_T^{1:N} \sim \mathcal{N}(0, I)$, then alternate between the denoising network G and the backward noising process with gradually decreasing number of steps $(T-1, \ldots, 1)$.

Architecture. Our denoising network architecture G is based on MDM [51] and employs a transformer [52] with four inputs: the training motion sample $\mathbf{x}^{1:N}$ of variable length, the encoding of the denoising step index t, the positional embedding of the temporal ordering of motion frames, and the conditioning vector c. These inputs are linearly projected into a common 512-dimension space and added together to form the input of the transformer encoder. We kindly refer the reader to MDM [51] for additional details.

Unlike MDM, which only uses the CLIP-embedding of the "high-level" text annotation of HumanML3D [15] as the conditioning vector c, we also incorporate the LLM-embedding of its "low-level" variant, and sum their linear projections. During training, we generate the "low-level" text by sorting and concatenating ground truth per-frame labels from the BABEL [36] dataset. During inference, we employ GPT-3 prompting (see Section 3.3). We experiment with various LLMs for the "low-level" embedding, such as DistilBERT [41], Sentence-T5 [32], and MiniLM [54]. Note that it is also possible to obtain the "low-level" text during inference through user input or test data from BABEL.

Training loss. Geometric losses [34, 35, 50, 51] are often utilized in motion generation to encourage natural and physically plausible motions. In this work, we incorporate regularization for joint positions (obtained via forward kinematics from the model-generated joint angles), joint angular velocity and foot contact (to prevent foot-skating). Specifically, we define the losses as follows:

$$\mathcal{L}_{pos} = \frac{1}{N} \sum_{i=1}^{N} \|FK(\mathbf{x}_0^i) - FK(\hat{\mathbf{x}}_0^i)\|_2^2, \quad (3)$$

$$\mathcal{L}_{vel} = \frac{1}{N-1} \sum_{i=1}^{N} \|(\mathbf{x}_0^{i+1} - \mathbf{x}_0^i) - (\hat{\mathbf{x}}_0^{i+1} - \hat{\mathbf{x}}_0^i)\|_2^2, \quad (4)$$

$$\mathcal{L}_{foot} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(FK(\hat{\mathbf{x}}_0^{i+1}) - FK(\hat{\mathbf{x}}_0^i)) \cdot f_i\|_2^2. \quad (5)$$

Here, the function $FK(\cdot)$ represents forward kinematics (converting from joint angles to joint positions), and $f_i \in \{0,1\}^J$ is the ground truth binary foot contact mask for each frame i.

The overall training loss is then defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{foot}\mathcal{L}_{foot}. \quad (6)$$

3.3. GPT-3 Prompting

During inference, we utilize zero-shot prompting (Section 2.2) for the GPT-3 LLM to generate "low-level" descriptions for each "high-level" test query. This approach has two benefits. Firstly, it increases the diversity of text prompts, resulting in a wider range of generated motions. Secondly, it allows for better generalization by achieving greater similarity to the training data distribution. This is done by breaking down hard "high-level" test set descriptions into similar "low-level" actions found in the training set.

Our prompting procedure is depicted in Fig. 3. We follow KATE [25] and retrieve the top-k most similar training examples for each "high-level" query $\hat{h}$. We accomplish this by precomputing a dictionary D of "high-level" Sentence-T5 [32] embeddings, *i.e.*, $D = z_{1:S}$, where S is the size of the training set. Then during inference, we obtain the embedding of $\hat{h}$, represented as $\hat{z}$, calculate its cosine similarity with the dictionary D, and retrieve the top-k most similar "high-level" training descriptions $H^p = h^p_{1:k}$. Finally, we form pairs of "high-level" texts in H^p with their corresponding "low-level" counterparts $l^p_{1:k}$ from the training dataset, and concatenate these pairs with $\hat{h}$ to create a prompt of the form $[h^p_1, l^p_1, h^p_2, l^p_2, ...h^p_k, l^p_k, \hat{h}]$. These pairs are used as examples for GPT-3, which generates the corresponding "low-level" description of $\hat{h}$. Given the stochastic nature of GPT-3, this generated description, which is then provided at inference time as the "low-level" conditioning for our model, is different every time, but still consistent with the provided query, thus enabling motion augmentation.

We bring to the reader's attention that this approach does not require fine-tuning GPT-3, nor manually engineering prompts.

4. Results

Details on the datasets and metrics are provided in Section 4.1. We present the results of the model for the case of text-to-motion task in Section 4.2. Given the work's emphasis on complex actions with dual conditioning, no action-to-motion tests or comparisons were performed. We include figures for qualitative inspection, comparing the current SOTA results [51] with ours. The prompt retrieval results using GPT-3 are shown in Section 4.3. Training was performed using a single NVIDIA RTX A4000 and took about 2-3 days. The model was trained with $T = 1000$ denoising steps and a cosine noise schedule. The weights on the language models were frozen during training.

4.1. Data and evaluation metrics

Data. During training, a batch of text-motion pairs is sampled for each iteration and is fed to their respective encoders. The AMASS [27] dataset was used for 3D motion capture data, containing high-quality motion capture data of multiple actions being performed. The HumanML3D [15] dataset was chosen for "high-level" text, featuring three to four descriptions of the actions contained in AMASS. For "low-level" text descriptions, during training, the BABEL [36] dataset was used, as it contains per-frame descriptions of the action, which were concatenated in a single sentence containing the sequence of actions present in the motion. When a corresponding BABEL label was not available, the HumanML3D [15] prompt was used for both high- and "low-level" conditioning.

For inference, only "high-level" descriptions must be provided. The "low-level" conditioning can be automatically generated by means of GPT-3, and, similar to the behavior of other proposed models, only a single text description of the motion is needed. This possibility allows for practical data augmentation, as multiple motions can be generated from a single text description due to the stochastic nature of GPT-3 prompts.

Another possibility is having both "high-" and "low-level" descriptions provided by the user. This is useful for cases where the user wishes to specify a particular variant of the motion being generated that includes a certain action, as shown in Fig. 1.

Evaluation metrics. We use the Fréchet Inception Distance (FID), Diversity, and MultiModality metrics, common for evaluating this task [15, 35, 50, 51]. We also include the metrics proposed and described in [15] and used in [15,51], namely R-precision and Multimodal Distance, which evaluate how much the motions fit the description.

4.2. Text-to-motion

Quantitative results. We compare our results with the SOTA text-conditioned motion generation method MDM [51], JL2P [2], Text2Gesture [6], and T2M [15]. We tested the proposed architecture using DistilBERT [41], Sentence-T5 [32] and MiniLM [54] as language encoders for the "low-level" text. We performed two sets of experiments, one using "low-level" descriptions from the BABEL [36] test set, and another using GPT-3 prompts as described in Section 3.3, using 10 examples when querying GPT-3. Both experiments were conducted using HumanML3D [15] and BABEL [36] (training) data, following the HumanML3D train-test split. For each set, we tested the three "low-level" encoders. When no "low-level"match was present in BABEL, we chose to repeat the "high-level" description for "low-level" as well. We summarize the results in Table 1.

Qualitative results. We further provide qualitative comparisons with the state-of-the-art approach MDM [51], in

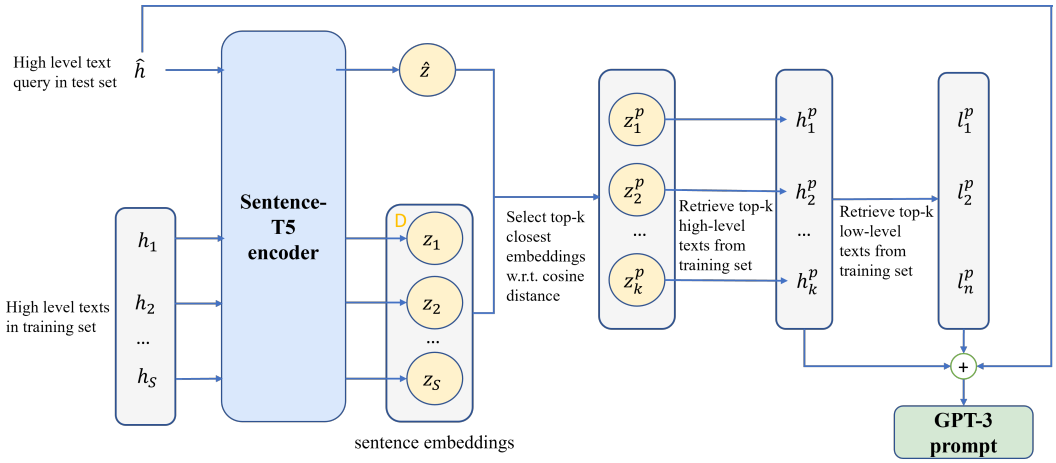


Figure 3. Overview of our GPT-3 Prompting Method. The Sentence-T5 encoder is used to map the "high-level" test query and the "high-level" texts of the training set to the embedding space. We retrieve the top-k "high-level" training texts that are closest to the test query in terms of cosine distance, as well as their respective "low-level" counterparts. The GPT-3 prompt is formed by concatenating the retrieved "high-" and "low-level" training texts and the test query.

Method	R-Precision ↑	FID↓	Multimodal Distance ↓	Diversity →	Multimodality ↑
Ground Truth [51]	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
JL2P [2]	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
Text2Gesture [6]	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
T2M [15]	$\mathbf{0.740^{\pm.003}}$	$1.067^{\pm.002}$	$\mathbf{3.340^{\pm.008}}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [51]	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799^{\pm.072}}$
Ours w. DistilBERT [BABEL]	$0.637^{\pm.007}$	$\mathbf{0.425^{\pm.051}}$	$5.319^{\pm.032}$	$\underline{9.511}^{\pm.078}$	$\underline{2.607}^{\pm.107}$
Ours w. Sentence-T5 [BABEL]	$0.631^{\pm.006}$	$0.783^{\pm.065}$	$5.420^{\pm.027}$	$9.227^{\pm.088}$	$2.464^{\pm.092}$
Ours w. MiniLM [BABEL]	$0.618^{\pm.006}$	$\underline{0.466}^{\pm.041}$	$5.523^{\pm.036}$	$9.800^{\pm.078}$	$2.523^{\pm.031}$
Ours w. DistilBERT [GPT-3]	$0.634^{\pm.008}$	$0.574^{\pm.077}$	$5.336^{\pm.035}$	$\mathbf{9.502^{\pm.079}}$	$2.457^{\pm.779}$
Ours w. Sentence-T5 [GPT-3]	$\underline{0.645}^{\pm.007}$	$0.571^{\pm.054}$	$\underline{5.267}^{\pm.021}$	$9.662^{\pm.062}$	$2.393^{\pm.043}$
Ours w. MiniLM [GPT-3]	$0.616^{\pm.007}$	$0.507^{\pm.046}$	$5.577^{\pm.032}$	$9.679^{\pm.082}$	$2.446^{\pm.032}$

Table 1. Quantitative experimental results on the HumanML3D test set. Results of baseline models are taken from [51]. ↑ represents higher scores being better, ↓ represents lower scores being better, and → indicates that closer to ground truth value is better. Each evaluation was performed 20 times, except for MultiModality, which was done 5 times. ± denotes the 95% confidence interval. Winning numbers are highlighted in **bold**. Second-best are underlined. Our doubly text-conditional model is trained with ground truth "high" and "low-level" text from the BABEL dataset. It performs better in terms of FID and diversity metrics, while maintaining competitive R-precision, Multimodal Distance, and Multimodality. We also ablated results for the "low-level" language encoder (DistilBERT, Sentence-T5, and MiniLM) and GPT-3 prompting.

Fig. 4. For additional examples, we kindly forward the reader to the supplementary material.

User study. 32 participants took part in a survey where they were shown motions from our model and MDM [51], and asked to choose which motion simultaneously fit the description better, and looked more natural. Each participant was shown 30 motions generated from descriptions randomly sampled from HumanML3D [15] training set. "Low-level" prompts were provided by the BABEL [36] dataset. The second language encoder for this task was MiniLM [54]. Our method was strictly preferred over MDM [51] for 61% of the motions presented, and there was no preference over either model for 30% of the motions.

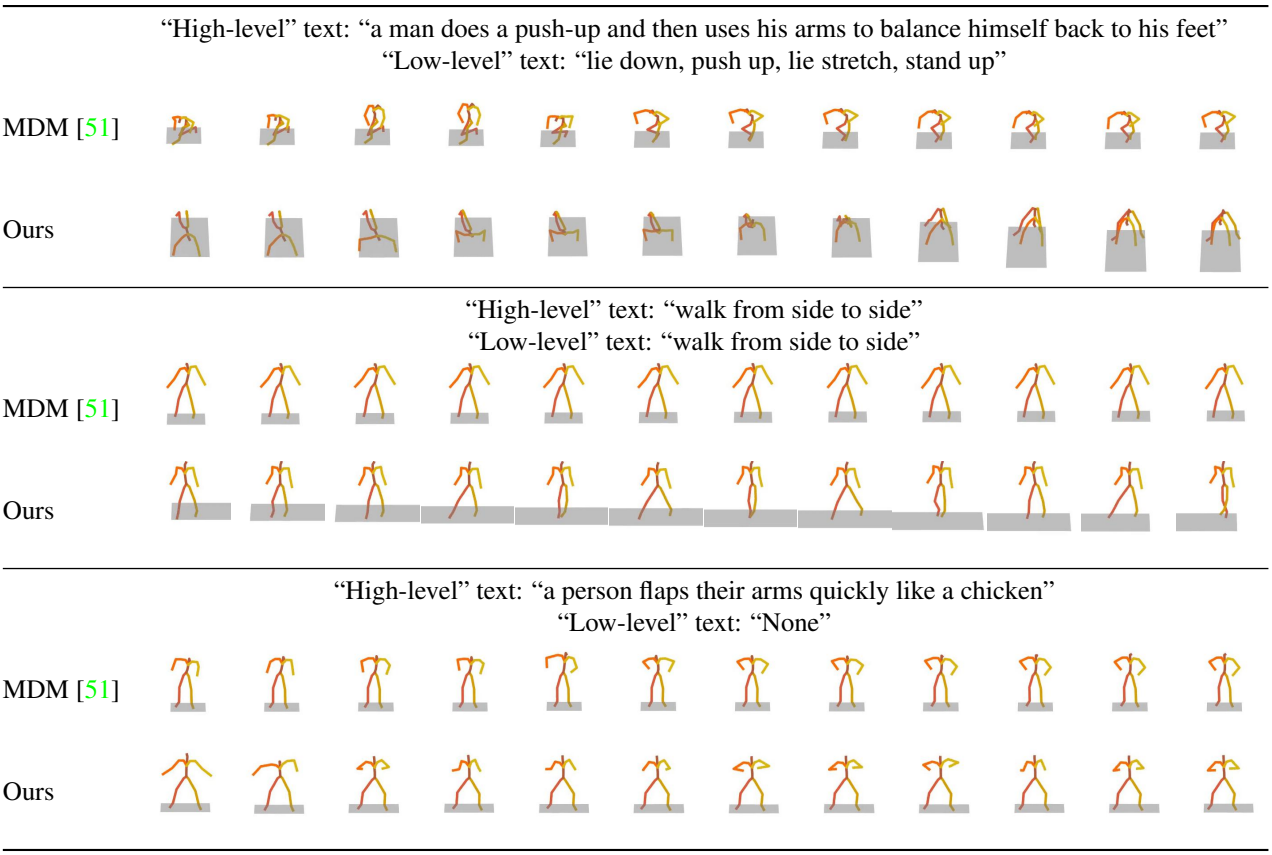


Figure 4. Qualitative experimental results on the HumanML3D test set. "Low-level" text is provided either by the BABEL dataset, by repeating the "high-level" variant, or by not specifying a "low-level" prompt and using 'None' as description. Our approach generates more detailed and realistic motions than MDM [51].

4.3. GPT-3 prompt retrieval

We conducted a comparison between the "low-level" text generated by GPT-3 using prompts retrieved through our method (Section 3.3), and the prompts generated by querying GPT-3 with random samples from the training set as examples. 500 testing samples were performed, and the average cosine similarity between the resulting "low-level" prompt and the ground truth description are used as the metric for evaluation. The results are as shown in Table 2, in which we also investigate the impact of the number of samples in the prompt on the similarity of the obtained "low-level" text. To assess the significance of the difference in average similarity, we performed a one-sided Kolmogorov-Smirnov test between the two methods with null hypothesis of our similarity being higher than random sampling. With the resulting p-value > 0.05, there is no evidence that the null hypothesis is not correct, thus should not be rejected.

We also show some examples of the generated prompts by GPT-3 when queried with descriptions from the test set of HumanML3D dataset, and which were used as "low-level" descriptions for the experiments in 4.2 in Table 3. The model was queried using the prompt "Convert scene to motion sequences according to the examples." and then given the examples and target "high-level" prompt for con-

	Cosine similarity		
Method	10 samples	20 samples	30 samples
Random	$0.8602^{\pm.057}$	$0.8611^{\pm.062}$	$0.8625^{\pm.061}$
Ours	$\mathbf{0.8688}^{\pm.063}$	$\mathbf{0.8735}^{\pm.064}$	$\mathbf{0.8744}^{\pm.065}$

Table 2. Average cosine similarity between the "low-level" texts generated by GPT-3 when using examples from our method vs. random samples from the training set. Similarity increases with higher number of examples for both cases, but our method remains consistently higher. A one-sided Kolmogorov-Smirnov test was conducted between the two methods, which shows no evidence in rejecting the hypothesis of our method's similarity being higher than random prompting (p-value<0.05).

high-level text	low-level text
a person, searching for something with their right hand, picks up the item with their left hand and places it in something by their head	stand, reach down with right hand, pick up object, move left arm to the back, pick up more object, move left arm to the back
a man using both hands to lift something off ground and places it back on ground in a slightly different position	stand, lift object, move object, set down object, stand
a man walks forward, then turns around and walks back before facing back and standing still	walk, turn, walk, stand still

Table 3. Table containing the generated "low-level" text when GPT-3 is prompted with examples from the HumanML3D test set. 10 examples pairs were used when querying GPT-3.

version. For these examples, 10 example sentences were provided, in order to mimic a more plausible scenario where the user wishes to minimize costs prompting GPT-3. For additional examples of generated prompts using GPT-3, we kindly forward the reader to the supplementary material.

5. Discussion

The proposed dual-conditioning approach is able to achieve metrics that either surpass or rival state-of-the-art results[2]. The better embedding of the prompt improves the R-precision and multimodal distance metrics, and achieves the best results for the FID and diversity metrics, as shown in Table 1.

This observation of improvement is not limited to the metrics evaluated using the BABEL dataset, which matches the data being used in training, but also GPT-3 generated "low-level" descriptions using the previously described method (Section 4.3 and Fig. 3).

Given the diversity of motions that may be described, the model is asked to interpret prompts not present in the training set. GPT-3 helps minimize this problem, by providing "low-level" prompts with actions that were in the training set. These generated prompts are in accordance with the training set distribution, as GPT-3 results can be better than BABEL results (Table 1), which use human annotations to describe the motions. Essentially, GPT-3 tells the model how to perform a new motion by breaking it down into actions present in the training examples.

Leveraging GPT-3, the potential for data augmentation is significant, as the user needs only specify the "high-level" description of the action, and multiple "low-level" prompts are provided. This results in a multitude of motions that are in accordance with the description, with slight variations (Fig. 1). Descriptions not present in training are realistically generated by means of this translation of new motions into previously known actions.

[2]It is worth noting that during the writing of this work, a similar work called T2M-GPT (https://mael-zys.github.io/T2M-GPT/) was published, setting a new state-of-the-art. While we do not directly compare with T2M-GPT, our claims regarding the efficacy of dual-conditioning in improving performance remain valid.

Another aspect of this dual-conditioning is that the user can be the one to provide the "low-level" description. This allows an additional level of control on the generated sequence, not present in current models. This ability may be interesting to explore the diversity in actions that depends on context and culture, with which current models may struggle (Fig. 1).

Failure cases for MotionGPT include: 1) prompts that require counting, e.g. "take 3 steps forwards" might walk indefinitely; 2) long sequences with many actions, where order and actions may sometimes be ignored, 3) motions very specific or ambiguous, such as "celebrate a goal like Ronaldo", and 4) interacting with 'hallucinated' objects, such as throw a ball, or sit on a chair.

The proposed approach does not require fine-tuning GPT-3 or manually crafting prompts. Nonetheless, delving into the potential ramifications of such techniques is a possible path for future investigation. While the GPT-3 model's outputs could align more closely with the dataset through fine-tuning, there exists a concern that such alignment might harm the model's capacity for effective generalization towards novel scenarios.

Overall, our proposed approach is able to outperform the current state of the art MDM [51], with the added benefit of allowing greater diversity of motions, as well as some degree of control over the generated motion.

6. Conclusion

In this work we propose MotionGPT, a novel approach to human motion synthesis that aims to leverage language models' common sense information, by allowing a dual-conditioning of the "high-level" description of a motion, and also a "low-level" description of the basic actions that compose it. This "low-level" description may be obtained by GPT-3, allowing an increased diversity of motions without additional input from user. It may be provided by the user instead, increasing their control over the action. We found that this approach resulted in a better embedding in latent space and increased motion diversity.

A limitation of this approach, due to the diffusion model architecture, is the long inference time. For this reason, the method does not scale to very long sequences. Another aspect worth pointing out is the dependence on datasets suited for this approach. The popularity of AMASS [27] motivated multiple annotations datasets, namely BABEL [36] and HumanML3D [15]. The availability of these datasets allowed our approach, and it would be interesting if other motion capture datasets could benefit from the same treatment.

Overall, the potential for data augmentation is significant and the information contained in LLMs is helpful for the task of human generation, by breaking down abstract and ambiguous motions, allowing for data augmentation and an increased diversity of generated motions.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920, 2018. 2

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 5, 6

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 2

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

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 2

[6] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021. 5, 6

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[8] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *6th Annual Conference on Robot Learning*, 2022. 3

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3

[10] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 2

[11] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 356–372. Springer, 2022. 2

[12] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 3

[14] Karthikeyan Duraisamy, Obiajulu Isebor, Alba Perez, Marco P Schoen, and D Subbaram Naidu. Kinematic synthesis for smart hand prosthesis. In *The First IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, 2006. BioRob 2006.*, pages 1135–1140. IEEE, 2006. 1

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 3, 4, 5, 6, 8

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[18] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 1

[19] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 3

[20] Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. Answer generation for retrieval-based question answering systems. *arXiv preprint arXiv:2106.00955*, 2021. 3

[21] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[24] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*, 2019. 3

[25] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-

context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. 3, 5

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 3

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5, 8

[28] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2

[29] Matteo Menolotto, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn, and Michael Walsh. Motion capture technology in industrial applications: A systematic review. *Sensors*, 20(19):5687, 2020. 1

[30] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022. 3

[31] Lars Mündermann, Stefano Corazza, and Thomas P Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of neuroengineering and rehabilitation*, 3(1):1–11, 2006. 1

[32] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 4, 5

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[34] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2, 4

[35] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 1, 2, 4, 5

[36] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021. 2, 4, 5, 6, 8

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[39] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021. 3

[40] Hayley Ross, Jonathon Cai, and Bonan Min. Exploring contextualized neural language models for temporal dependency parsing. *arXiv preprint arXiv:2004.14577*, 2020. 3

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3, 4, 5

[42] Teven Le Scao and Alexander M Rush. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493*, 2021. 3

[43] Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*, 2021. 3

[44] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[45] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States, July 2022. Association for Computational Linguistics. 3

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[47] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[48] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019. 3

[49] Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. *arXiv preprint arXiv:2004.14911*, 2020. 3

[50] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 1, 4, 5

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[53] Cunxiang Wang, Pai Liu, and Yue Zhang. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online, Aug. 2021. Association for Computational Linguistics. 3

[54] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 4, 5, 6

[55] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022. 3

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[57] Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*, 2021. 3

[58] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, pages 265–281, 2018. 2

[59] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2

[60] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023. 3

[61] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3

[62] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020. 3