Text Motion Translator: A Bi-Directional Model for Enhanced 3D Human Motion Generation from Open-Vocabulary Descriptions

Yijun Qian[1], Jack Urbanek, Alexander Hauptmann[1], and Jungdam Won[2]

[1] Carnegie Melon University
[2] Seoul National University

Abstract. The field of 3D human motion generation from natural language descriptions, known as Text2Motion, has gained significant attention for its potential application in industries such as film, gaming, and AR/VR. To tackle a key challenge in Text2Motion, the deficiency of 3D human motions and their corresponding textual descriptions, we built a novel large-scale 3D human motion dataset, **LaViMo**, extracted from in-the-wild web videos and action recognition datasets. LaViMo is approximately 3.3 times larger and encompasses a much broader range of actions than the largest available 3D motion dataset. We then introduce a novel multi-task framework **TMT** (Text Motion Translator), aimed at generating faithful 3D human motions from natural language descriptions, especially focusing on complicated actions and those not existing in the training set. In contrast to prior works, TMT is uniquely regularized by multiple tasks, including Text2Motion, Motion2Text, Text2Text, and Motion2Motion. This multi-task regularization significantly bolsters the model's robustness and enhances its ability of motion modeling and semantic understanding. Additionally, we devised an augmentation method for the textual descriptions using Large Language Models. This augmentation significantly enhances the model's capability to interpret open-vocabulary descriptions while generating motions. The results demonstrate substantial improvements over existing state-of-the-art methods, particularly in handling diverse and novel motion descriptions, laying a strong foundation for future research in the field.

Keywords: Text2Motion, 3D Human Motion Generation, Generative Model, Multimodal

1 Introduction

The generation of 3D human motions conditioned on natural language descriptions, commonly known as *Text2Motion*, has become a significant area of research due to its potential applications in industries such as film, gaming, and AR/VR. Recent years have seen numerous successful attempts [2, 5, 9, 18, 29, 32, 37], with various machine learning models such as Variational Autoencoders (VAE) [19], Diffusion models [15], and Generative Adversarial Networks (GAN) [10] being

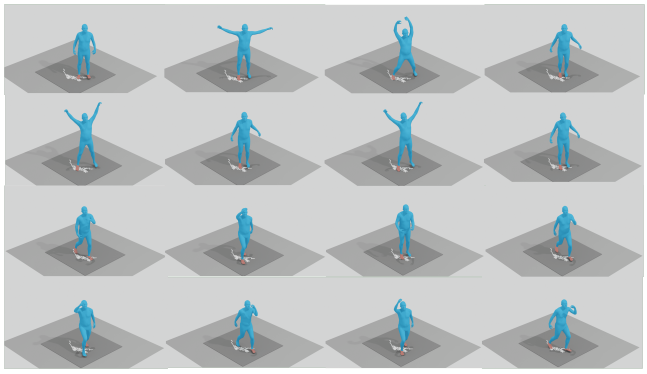

Fig. 1: This demonstrates the bi-directional generation capability of our model. Initially, our model generates the displayed motion using the black-colored textual description (Text2Motion). Subsequently, our model generates red-colored textual descriptions (Motion2Text) based on the displayed motion. The qualitative results can be best seen in the supplementary materials.

utilized. These efforts were enabled by publicly available 3D human motion datasets such as AMASS [26], and motion-text paired datasets like BABEL [31] and HumanML3D [11]. However, the quality of 3D human motions generated by existing models still fall short of the standards set by recent generative models in both the image [34] and language [7] domains.

A key factor contributing to this limitation is the lack of sufficient data. Models for text-conditioned image or video generation often benefit from initial weights pre-trained on large datasets. However, the largest publicly available 3D motion dataset only covers a small fraction of human behaviors observed in daily life. Furthermore, the largest publicly available textual description dataset only covers a portion of this. For instance, AMASS, the largest 3D motion dataset, contains approximately 11,000 motions, while a popular pre-training video dataset, IG-65M [8], consists of over 65 million videos. The disparity in size and diversity is even more pronounced considering that high-quality motion capture data collection often requires specialized equipment, actors, and manual data clean-up, making it challenging to match the scale of datasets available for other modalities. The constraint of limited action descriptions presents another hurdle in developing a generative model for Text2Motion. Current generative modeling techniques heavily depend on paired data. Consequently, models trained with a limited number of action descriptions are prone to overfitting, resulting in a lack of generalization. This leads to a significant drop in performance when the model encounters unseen descriptions. While previous research has attempted to improve robustness by breaking down complex actions into segments of atomic actions [32], this approach still has limitations as it requires known descriptions for atomic (i.e., single) actions.

In addition to dataset challenges, previous models are not particularly user-friendly, often requiring substantial prior knowledge of Text2Motion to generate plausible 3D human motions. For instance, many approaches [3,5,18,32] not only require natural language descriptions, but the duration of each atomic action as well. If the provided duration does not align with what was included in the training dataset, the model may fail to generate plausible motions. As the length of the input description increases, from a single verb to a complete sentence or from a sentence to a paragraph, the challenge of determining a reasonable duration grows exponentially.

In this paper, we present a novel framework designed to address the challenges above. Firstly, we introduce a large-scale motion dataset **LaViMo** (Large-scale Video Mocap), comprising more than 140,000 motion clips. This dataset is extracted from in-the-wild RGB web videos and publicly available action recognition datasets, leveraging recent advancements in 3D pose estimation. This approach allows us to circumvent the need for expensive motion capture equipment and extensive recording procedures. Additionally, we develop a method to augment textual descriptions of actions using Large Language Models (LLMs). This augmentation increases the semantic diversity of existing descriptions, enabling the model to handle open-vocabulary descriptions robustly when generating 3D human motions.

Additionally, we develop **TMT** (Text Motion Translator), a bi-directional text-conditioned motion generation model inspired by recent breakthroughs in language translation studies. The key idea is to formulate text-conditioned 3D human motion generation as a pure language translation problem. Both motions and textual descriptions are treated as discrete tokens, and their bi-directional translation is facilitated by a language model. As LLMs can determine when to stop generating sentences using a special token indicating the end-of-sentence, our model can also autonomously halt the motion generation process. To construct discrete tokens for motions, we introduce a motion VQ-VAE that encodes 3D motions into discrete embedding tokens and reconstructs 3D motions from these tokens. Unlike previous works, TMT is regularized by multiple tasks (Text2Motion, Text2Text, Motion2Motion, and Motion2Text). This multi-task regularization not only equips the model with additional capabilities, such as motion in-painting, but also enhances its robustness, motion modeling, and semantic understanding. We conduct a series of experiments, including 3D motion generation from lengthy and challenging action descriptions composed of open-vocabulary elements. Our qualitative and quantitative experiments demonstrate that our framework outperforms other state-of-the-art methods significantly in generating natural and semantically correct motions. Ablation studies further confirm the validity of our proposed method. Finally, we showcase experiments on the inverse of the original task, Motion2Text, made possible by our unique problem formulation. In a nutshell, our contribution are three-folded:

- The creation of a large 3D human motion dataset (more than 140k motions) extracted from in-the-wild RGB web videos and publicly available action recognition datasets.

- The introduction of a sample-free training strategy using Large Language Models (LLMs) to enhance semantic diversity and to enable the generation of natural 3D human motions with open-vocabulary descriptions.
- We introduce TMT, a bi-directional text-conditioned 3D human motion generation model capable of generating both 3D human motion and natural language descriptions when conditioned on the other modality. In addition, our TMT does not require the duration of each atomic action and is optimized by multi-task regularization.

2 Related Works

Most text-conditioned motion generation methods based on deep learning have pursued a similar idea, aiming to bridge semantic representations of text (*text-encoding*) and 3D motion features (*motion-encoding*) through a shared latent space [1, 9, 12, 16, 18, 20, 22–24, 29, 40]. Subsequently, these methods implement a decoder that translates latent embeddings into 3D human motions (*motion-decoding*). Then, text-conditioned motion generation can be performed by combining *text-encoding* and *motion-decoding* processes. Following the idea, various conditional generative models have been explored. For instance, TEMOS [29] introduces a VAE [19] while MDM [37] proposes a 3D human motion generation network based on diffusion [15]. These models were trained via supervision on motion-text paired datasets [11,30], which cover simple actions only and may include ambiguous descriptions. With the release of the BABEL dataset [31], which offers detailed action-level descriptions of complex motions existing in various motion datasets, several studies have emerged aiming to generate realistic 3D human motions from lengthy and semantically rich descriptions [3, 20, 32, 40]. TEACH [3] proposes a method to generate long motions by including previous action segment into conditional generation process. SINC [4] applies a motion augmentation algorithm which combines the body parts of different motions with the help of LLMs. The most recently published study, EMS [32], introduces a method that robustly generates motions from very long sentences, which may include tens of verbs and adverbs. This is achieved by factorizing the input sentence into a series of atomic actions, and the generated atomic actions (i.e., motion corresponding to each verb) are then combined sequentially in the final stage via optimization with loss function on naturalness.

Despite of the advancements achieved by above studies, their generation qualities are still far from what we typically expect from recent generative models in both image [34] and language [7] domains. This implies that there is still much room for improvement across various aspects, including data, model, and training algorithms. We argue that securing data is the biggest hurdle in text-conditioned motion generation.

3 Methods

Given natural language descriptions as input, text-conditioned motion generation (i.e., Text2Motion) is a task to generate full-body 3D motions that corre-

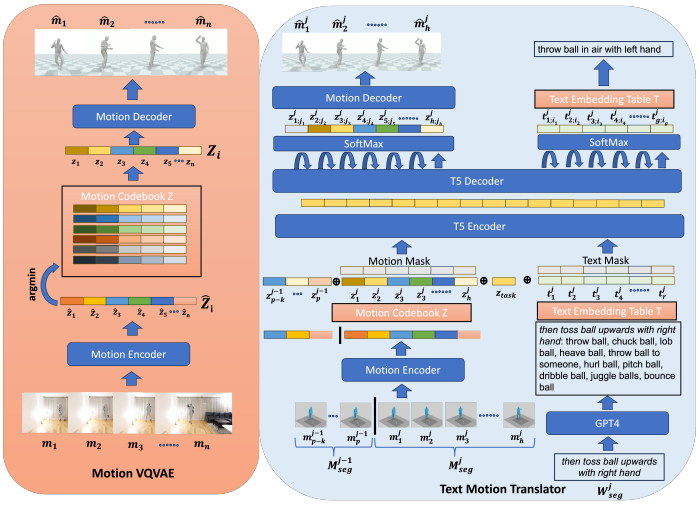

Fig. 2: Model Structure Overview: We pretrain our Motion VQ-VAE on the combination of our LaViMo and existing AMASS datasets, then train our generation module on motion-text paired dataset.

spond to the semantic meanings of the given descriptions. We are particularly interested in generating natural motions given open-vocabulary descriptions for unseen actions besides generating motions whose description has been seen in the training set.

As discussed in Section 1, two main challenges exist: the scarcity of data concerning both 3D human motions and their corresponding textual descriptions, and the model's requirement for detailed duration as input, affecting its usability. To solve the first challenge, we build a new 3D human motion dataset which is approximately 4 times larger than the existing dataset by recent advances in 3D pose estimation, and develop a method of augmenting existing textual descriptions for human actions (see Section 3.1 for the detail). To solve the second challenge, we develop a new model architecture inspired by language translation tasks. Our novel model produces higher quality motions than existing state-of-the-art models, more importantly, it has the capability to generate motions without the duration of actions as input. The model architecture and its training method will be detailed in Section 3.2 and Section 3.3, respectively.

3.1 Dataset Preparation

Building LaViMo Dataset We build a new 3D motion dataset from in-the-wild videos collected from the web as well as the videos existing in NTU120k [25] dataset. We extract approximately 140k motion clips corresponding to 130 hours of motion, which is approximately 3.3 times larger than the largest AMASS dataset. We have named our novel dataset **LaViMo**, a shortened version of large-scale video mocap dataset.

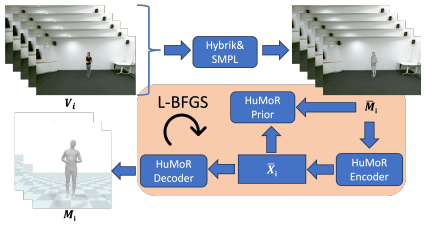

Fig. 3: 3D pose estimation pipeline overview: we balance the trade-off between performance and time spent by combining Hybrik Module and test-time optimization of HuMoR.

Figure 3 illustrates an overview of 3D pose estimation pipeline that we use. Given a RGB video $V_i = (f_1, f_2, \cdots, f_n)$, the pipeline is to extract a 3D human motion $M_i = (m_1, m_2, \cdots, m_n)$, where each $f_j \in \mathbb{R}^{H \times W \times 3}$ represents a single RGB frame, and each $m_j \in \mathbb{R}^D$ represents the 3D pose extracted from frame f_j. We basically adopt the pipeline of HuMoR [35], which runs a test-time optimization on top of the pretrained variational autoencoder (VAE) consisting of three modules, a prior $p_\theta(x_t|m_{t-1})$, an encoder $p_\theta(x_t|m_{t-1}, m_t)$, and a decoder $p_\theta(m_t|m_{t-1}, x_t)$, where x_t is a VAE latent embedding that models a transition between two poses. The process of test-time optimization is to search a series of optimal latent embeddings $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$ with the loss function including two primary terms, the data term and the naturalness term. The data term enforces that the decoded 3D motion $(\bar{m}_1, \bar{m}_2, \cdots, \bar{m}_n)$ matches the input RGB video $(f_1, f_2, \cdots, f_n)$, where the visual features extracted from f_j are compared with the features extracted from the decoded motion projected onto the 2D image plane. The naturalness term prevents the decoded motion from being out-of-distribution from the training dataset, where the degree of being out-of-distribution are measured by plugging the currently optimized latent embeddings and the previously decoded motion into the prior. Once the loss function is computed, the current latent embeddings are updated by L-BFGS [35] algorithm. We recommend readers to refer to the original paper [35] for the details.

Although HuMoR provides one of the state-of-the-art results for extracting 3D motions from videos as input, the optimization process is very time-consuming. For example, it approximately takes more than 5 minutes to extract a 2 seconds long motion from a video clip on single V100 GPU. This computational complexity prohibits us to extract motions from large scale videos. We observed that the optimization process can be improved significantly by providing a better initial guess of latent embeddings. Instead of following the initialization process in the original HuMoR, which is based on VPoser [27] and an extra shallow Gaussian mixture model, we adopt HybrIK [21] for the initialization. HybrIK is a per-frame light-weight 3D pose estimator which extracts 3D pose by combining analytical inverse-kinematics (IK) with estimation from neural networks. As a result, it is extremely faster than other methods based on heavy optimization processes such as HuMoR although there exist some compromise in the final motion quality if it is solely used. Given an input video, we first

perform HybrIK to extract a sequence of 3D poses, they are then converted into the latent motion embeddings by using the encoder of HuMoR [35]. Finally, the embeddings become a initial guess for the test-time optimization. By doing so, we found that a 3D motion given a 2 seconds long input video can be obtained approximately within 30 seconds, of which motion quality is comparable to what created by HuMoR only.

Augmenting Textual Description Given natural language descriptions of human actions from existing datasets, specifically the BABEL dataset [31], we augment these descriptions using a large language model [7]. This augmentation aims to enable our model to robustly handle input textual descriptions comprising open-vocabulary, potentially encompassing richer or more complex expressions.

For each description of an atomic action (i.e., a single human behavior), we ask the GPT4 [7] to generate augmented descriptions with the prompt: **"Given a human motion, provide a list of motions that would be visually similar. Some examples: 1) fall down: slip, trip, take a spill, faint, collapse. 2) jerk: tap door with hand, hit hand on door, bang fist on door. <>"**. Although GPT4 can give us ample descriptions which are visually similar to the action corresponding to the prompted input description, it may still generate inaccurate or not-suitable descriptions for our task. This is because GPT4 does not precisely take the attribute settings into consideration. For example, it considers "throwing a ball in the air with the left hand" similar to the motion of "tossing a ball upwards with the right hand," even after including negative examples in the input prompt. For a static (stationary) motion like "sway arms", it often regards "swing arms while walking" a similar description, which is a dynamic (mobile) motion. Additionally, for the foot-related motion "kick ball", it often generates the hand-related motion like "hit ball with racket". To address these issues, we employ several heuristics to filter out misleading information. The heuristics we develop are as follows:

- For each atomic action, we ask the GPT4 model to answer if the original and augmented description are using the same body part.
- For each atomic action, we ask the GPT4 model to answer if the original and augmented description are both static (stationary) motion or both dynamic (mobile) motion.
- We inspect the augmented descriptions to see if there is a mismatch in semantics. For example, we check "left/right", "forward/backward", "slowly/fast", "static/dynamic", and etc.

These heuristics not only enhance the robustness in generating from open-vocabulary descriptions but also alleviate some visible artifacts present in the generated motions, such as foot-sliding or floating-in-the-air issues. We will provide quantitative and qualitative comparisons in the ablation study section.

3.2 Text Motion Translator

Figure 2 depicts the novel model structure we developed, **TMT** (Text Motion Translator), inspired by recent breakthroughs in language translation studies. The key idea is to formulate text-condition 3D human motion generation as a language translation problem. In our model, both motions and textual descriptions are treated as discrete tokens, and the bi-directional translation between them is facilitated by a language backbone, specifically T5 [33] in our implementation.

To produce discrete tokens for motions, we pre-train a VQ-VAE [39] by using our LaViMo dataset as well as existing AMASS dataset, which is shown on the left of Figure 2. It is composed of a motion encoder $\mathcal{E}$ and a motion decoder $\mathcal{D}$, both of of them are consisted of stacked CNN blocks with activation and pooling layers. The encoder takes in a 3D human motion M_i as input then produces a latent embedding $\hat{Z}_i = (\hat{z}_1, \hat{z}_2, ..., \hat{z}_n)$, and the decoder generates a 3D human motion $\hat{M}_i$ from the encoded latent embedding. In this encoding-decoding process, VQ-VAE simultaneously learns a codebook $Z = \{z_1, z_2, ..., z_B\} \in \mathbb{R}^{B \times D}$ containing B discrete latent embeddings (node) where D is the dimension of each embedding, and the encoded latent embedding $\hat{Z}_i$ is replaced by the nearest embedding existing in the codebook before the decoding process starts. By doing so, a discrete latent space is constructed, we can produce discrete tokens of 3D human motions by plugging them into the encoder of VQ-VAE. To produce discrete tokens for the textual description, the input text is firstly augmented by a LLM [7], the text embedding table T that comes with the language backbone applies to the augmented input.

The generation (i.e., translation) module of our model is shown on the right of Figure 2. Given a series of motion-text pairs, $(M^1_{seg}, M^2_{seg}, \cdots)$ and $(W^1_{seg}, W^2_{seg}, \cdots)$, where $M^j_{seg} = (m^j_1, m^j_2, \cdots)$ and $W^j_{seg} = (w^j_1, w^j_2, \cdots)$ represent a motion segment of j-th atomic action and its corresponding textual description, respectively, motion and text tokens are first obtained by the tokenization methods explained above. Let's assume that $Z^j_{seg} = (z^j_1, z^j_2, \cdots, z^j_h)$ and $T^j_{seg} = (t^j_1, t^j_2, \cdots, t^j_g)$ are the motion and text tokens for j-th atomic action, which contains h motion tokens, and g text tokens. To make motion tokens context sensitive, we insert a special motion token $< SOM >$ representing *start-of-motion* before the first token if the current atomic action is the first atomic action, otherwise we insert last k motion tokens $(z^{j-1}_{p-k}, ..., z^{j-1}_p)$ from the previous atomic action M^{j-1}_{seg}, where p implies that the previous atomic action contains p tokens. We insert a special motion token $< EOM >$ representing *end-of-motion* at the end if the current atomic action is the last atomic action, and insert a special motion token $< EOAM >$ representing *end-of-atomic-action* at the end of every atomic action sequence. To enhance the semantics of text tokens, we use LLM to create augmentated tokens $t^j_{g+1}, \cdots, t^j_r$, where $r > g$. The motion tokens Z^j_{seg} and text tokens T^j_{seg} are concatenated, followed by the application of task-dependent adaptive masks. Further explanation of this process will be provided in detail in Section 3.3. The tokens are then forwarded to the transformer

encoder of T5, and the transformer decoder of T5 autoregressively predicts the masked portions of tokens. Once the entire tokens are predicted, 3D human motions are reconstructed by the VQ-VAE decoder while the textual descriptions are recovered by the text embedding table T.

3.3 Training TMT

We first train the motion VQ-VAE, then train the remaining parts existing in the generation module while the encoder and decoder of the VQ-VAE are frozen.

Training of Motion VQ-VAE We use three loss functions to optimize the motion VQ-VAE, where we keep the degree of motion tokens and language tokens the same to enable the integration of motion tokens directly into the language backbone.

$$\mathcal{L}^i_{vqvae} = \mathcal{L}^i_{rec} + \mathcal{L}^i_{emb} + \mathcal{L}^i_{com} \tag{1}$$

$$\mathcal{L}^i_{emb} = \|sg(Z_i) - \hat{Z}_i\|_2 \tag{2}$$

$$\mathcal{L}^i_{com} = \|Z_i - sg(\hat{Z})_i\|_2 \tag{3}$$

$$\mathcal{L}^i_{rec} = \mathcal{L}1(M_i - \hat{M}_i) + \alpha\mathcal{L}1(\mathcal{V}(M_i) - \mathcal{V}(\hat{M}_i)) \tag{4}$$

where $\hat{M}_i$ is the decoded motion sequence, sg represents the stop gradient operation, $\mathcal{L}1$ represents the smooth l1 loss function, and $\mathcal{V}(M_i)$ represents the velocity of motion sequence M_i.

Training of Generation Module We design four distinctive supervised or self-supervised tasks for training the generation module, which are *Text2Text*, *Motion2Motion*, *Text2Motion*, and *Motion2Text*. Depending on the task type, different mask settings are applied to the motion and text tokens. Additionally, an extra special token $< task >$ is inserted during concatenation of the two different tokens to guide the generation module in understanding the specific task it is being asked to perform.

Text2Text Similar to the random masking tasks in Bert [41], we mask out all motion tokens and 15% of text tokens. The TMT is then required to reconstruct the masked text tokens given unmasked text tokens $\{t_{q \notin masked}\}$.

$$L^j_{T2T} = -\sum_{p \in masked} \sum_{i=0}^{L-1} \log p(t^j_{p;i} | \{t^j_{q \notin masked}\}) \tag{5}$$

where L is the number of text tokens in text embedding table, and $t^j_{p;i}$ represents that the p-th text token of T^j_{seg} is predicted as i-th token of the text embedding table T.

Motion2Motion Similar to the self-supervised training task on Text2Text, we also use the Motion2Motion task for the self-supervision. The same loss function can be used because motions are also represented by tokens in our model. More specifically, we mask out all text tokens and 50% of motion tokens. The TMT is then required to reconstruct the masked motion tokens given unmasked motion tokens $\{z_{q \notin masked}\}$.

$$L^j_{M2M} = -\sum_{p \in masked} \sum_{i=0}^{B-1} \log p(z^j_{p;i} | \{z_{q \notin masked}\}) \tag{6}$$

where B is the number of motion tokens in the codebook, and $z^j_{p;i}$ represents that the p-th motion token of M^j_{seg} is predicted as i-th token of the motion codebook Z.

Text2Motion We masked out all motion tokens $(z^j_1, \cdots, z^j_h)$ extracted from the target atomic action M^j_{seg}. The TMT is then required to predict the masked motion tokens in an auto-regressive manner given the text tokens T^j_{seg}, unmasked motion tokens $\{z^{j-1}_{p-k:k}\}$, and previously predicted motion tokens $\{z^j_{r|r<s}\}$.

$$L^j_{T2M} = -\sum_{s=1}^{h} \sum_{i=0}^{B-1} \log p(z^j_{s;i} | \{z^{j-1}_{p-k:k}\}, \{z^j_{r|r<s}\}, T^j_{seg}) \tag{7}$$

where k is the length of motion tokens inherited from the previous atomic action and h is the duration of current atomic action.

Motion2Text Similar to Text2Motion task, we masked out all text tokens, the TMT is required to predict the text tokens based on motion tokens Z^j_{seg} and previously predicted text tokens $\{t^j_{r|r<s}\}$. We only use the ground truth text tokens $\{t^j_{1:g}\}$ as targets during the training process.

$$L^j_{M2T} = -\sum_{s=1}^{g} \sum_{i=0}^{L-1} \log p(t^j_{s;i} | \{t^j_{r|r<s}\}, Z^j_{seg}) \tag{8}$$

4 Experiments

We primarily evaluate our model against previous state-of-the-art models on two benchmarks (HumanML3D [11] and BABEL [31]) using two streams of metrics. It's important to note that BABEL and HumanML3D are built on a portion of the AMASS datasets, and **no samples from the validation set of these two benchmarks were utilized** during the pretraining stage. Additionally, LaViMo is a 3D human motion dataset without paired descriptions. **As such, we cannot pretrain the majority of previous works on LaViMo, which require paired data for conditional training.** For a fair comparison, we also report results pretrained on AMASS only (annotated as Ours wo LaViMo). Due to

Method	R-Precision↑	FID↓	Diversity →	MultiModal-Dist↓
Gound Truth	0.62	$4e^{-3}$	8.51	3.57
ACTOR [28]	0.33	1.43	7.41	8.39
MACVAE [20]	0.34	1.36	7.16	8.18
EMS [32]	0.42	0.96	8.22	7.23
TEACH [44]	0.46	1.12	8.28	7.14
PriorMDM [36]	0.48	0.79	8.16	6.97
Ours	**0.53**	**0.72**	**8.33**	**5.88**
Ours wo LaViMo	0.49	0.82	8.17	6.73

Table 1: Comparing ours against previous SOTAs on BABEL.

Method	R-Precision↑	FID↓	Diversity →	MultiModal-Dist↓
Gound Truth	0.511	$2e^{-3}$	9.503	2.974
TM2T [13]	0.424	1.501	8.589	3.467
T2M [12]	0.457	1.067	9.188	3.340
EMS [32]	0.462	0.617	8.782	3.792
MotionDiffuse [43]	0.491	0.630	9.410	3.113
MDM [37]	0.320	0.544	9.559	5.566
MLD [6]	0.481	0.473	9.724	3.196
T2M-GPT [42]	0.491	**0.116**	9.761	3.118
MotionGPT [17]	0.492	0.232	**9.528**	3.096
PriorMDM [36]	0.481	0.60	9.62	**2.96**
Ours	**0.528**	0.184	9.437	3.091
Ours wo LaViMo	0.464	0.310	9.191	3.652

Table 2: Comparing ours against previous SOTAs on HumanML3D.

Method	APE_r↓	APE_t↓	APE_l↓	APE_g↓	AVE_r↓	AVE_t↓	AVE_l↓	AVE_g↓
TEMOS [29]	0.766	0.731	0.172	0.825	0.269	0.262	0.016	0.274
TEACH [3]	0.674	0.654	0.159	0.717	0.222	0.220	0.014	0.234
EMS [32]	0.434	0.423	0.116	0.495	0.173	0.168	0.011	0.181
SINC [4]	0.502	0.477	0.249	0.616	0.174	0.174	0.010	0.180
PriorMDM [36]	0.388	0.372	0.116	0.464	0.156	0.152	0.010	0.174
Ours	**0.237**	**0.218**	**0.096**	**0.285**	0.144	0.122	**0.008**	0.153
Ours wo LaViMo	0.297	0.282	0.114	0.348	0.155	0.137	0.011	0.176
Ours wo LLM	0.288	0.263	0.125	0.337	**0.142**	**0.118**	**0.008**	**0.151**
Ours with LLM, wo PP	0.349	0.327	0.128	0.414	0.167	0.152	0.012	0.176

Table 3: Comparing ours against previous SOTAs on BABEL dataset under APE & AVE metrics. PP represents post-processing.

page limitations, details of the datasets, evaluation metrics, and implementation details are provided in the supplementary materials.

As shown in Tables 1 and 3, we compare our method to previous state-of-the-art (SOTA) methods [3,4,20,28,29,32,36,44] on the BABEL dataset. Our model outperforms all previous SOTAs, including PriorMDM, which uses a heavier diffusion generative backbone. Specifically, for APE&AVE metrics, our model is on average 33% better than the runner-up and 13% better on the remaining four metrics. Without LaViMo, our model still achieves superior performance on nine out of twelve metrics, with comparable results for the remaining three metrics. These experiments demonstrate the superiority of both the LaViMo dataset and the TMT model.

Although TMT was specifically designed to generate 3D human motions for long and complicated actions given elaborate descriptions (including atomic action level descriptions), we also report the performance on the HumanML3D dataset, another benchmark used by many previous works. Unlike BABEL, HumanML3D only contains simple single-sentence descriptions for each motion. As shown in Table 2, our model still achieves comparable results against recent SOTAs, such as PriorMDM and MLD, which use heavier diffusion-based generative backbones, and MotionGPT, which applies an extra stage of human-in-the-loop fine-tuning. We also observe that the use of LaViMo brings more improvements on HumanML3D than BABEL, possibly because LaViMo contains more atomic actions whose distribution is more similar to that of HumanML3D.

5 Ablations

Method	Split	$APE_r\downarrow$	$APE_t\downarrow$	$APE_l\downarrow$	$APE_g\downarrow$	$AVE_r\downarrow$	$AVE_t\downarrow$	$AVE_l\downarrow$	$AVE_g\downarrow$
A&La + L w P	S	0.186	0.174	0.072	0.232	0.117	0.109	**0.007**	0.126
	U	**0.327**	**0.296**	**0.138**	**0.379**	**0.192**	0.145	**0.010**	**0.201**
A&La + L w P + Sin	S	0.203	0.195	0.087	0.248	0.129	0.117	**0.007**	0.133
	U	0.374	0.361	0.162	0.424	0.223	0.191	0.012	0.247
EMS [32]	S	0.302	0.293	0.097	0.357	0.158	0.142	0.009	0.166
	U	0.668	0.654	0.149	0.740	0.199	0.214	0.014	0.207
A&La	S	0.163	0.157	0.066	0.219	0.112	0.105	0.007	**0.117**
	U	0.510	0.451	0.229	0.546	0.195	**0.141**	**0.010**	0.211
A&La + L wo P	S	0.309	0.293	0.106	0.362	0.147	0.138	0.010	0.161
	U	0.420	0.387	0.167	0.506	0.202	0.176	0.015	0.204
A + L w P	S	**0.153**	**0.142**	**0.064**	**0.201**	**0.108**	**0.099**	**0.007**	0.119
	U	0.553	0.531	0.203	0.609	0.238	0.204	0.018	0.277

Table 4: Analyze the improvements bought by each module on generating seen (S) and unseen (U) actions. A represents AMASS, La represents LaViMo, L represents using LLM for semantic augmentation, P represents post-processing, and Sin represents single task (text2motion) training.

5.1 Application of Large Language Models

This section delves into the enhancements achieved through the incorporation of Large Language Models (LLMs) for semantic augmentation. As depicted in Ta-

ble 3, employing LLMs leads to on average 17.3% better performance. However, it's important to note that utilizing all augmented descriptions from the LLM without any post-processing (filtering) results in a notable decrease in performance for both APE and Average Variance Error (AVE) metrics, which verifies the effectiveness of our filtering heuristics.

5.2 Seen and Unseen Actions

Our study extends beyond evaluating results on the full validation set of BABEL; we also conduct experiments to assess the specific improvements each implementation brings to the generation of seen and unseen actions. An action is considered 'seen' if its description appeared in the training set, and 'unseen' otherwise. Table 4 provides insights into these distinctions.

Through comparing the first and second line of Table 4, we find that the usage of multi-task training brings 13% improvements for unseen actions and 8% for seen actions.

Comparing the first and third lines shows that our model outperforms more on generating unseen actions than seen actions against EMS [32] (double the performance when generating unseen actions).

Further analysis, comparing the first, fourth, and fifth lines, reveals that while the integration of LLM without post-processing (filtering) significantly degrades the quality of generated seen actions, however, it still enhances the generation of unseen actions. More importantly, incorporating the post-processing substantially mitigates negative effects for both seen and unseen actions.

A comparison between the first and the last lines of Table 4 highlights that pretraining on AMASS is more beneficial for seen actions, whereas pretraining on LaViMo shows a greater impact on unseen actions. This aligns with our hypothesis that the remaining AMASS samples may include motions akin to those in BABEL's validation set, while LaViMo, with its broader range of complex and irregular motions, likely enhances the robustness of the pretrained motion VQ-VAE.

5.3 Qualitative Analysis

Quantitative metrics often fail to capture the essence of evaluation, which can only be achieved through human assessment. Figure 4(a) shows, within a red rectangle, a qualitative comparison to EMS [32], when an input description 'stand up from lying on the ground' is given. In Figure 4(b), we emphasize the model's capability, enhanced by LLM augmentation, to accurately produce natural motions for the unseen action 'hold arms as in waltz' series. Additionally, Figure 4(c) showcases the effects of our post-processing, leading to issues like foot sliding and incorrect global orientations when it is ablated. An example within a red bounding box shows the model generating an unnatural 'walk back' motion with improper global orientation. We speculate that this is due to LLM-generated augmentation candidates such as 'walk' or 'walk forward,' which might not align precisely with the intended motion "walk backward". The qualitative results can be best seen in the supplementary materials.

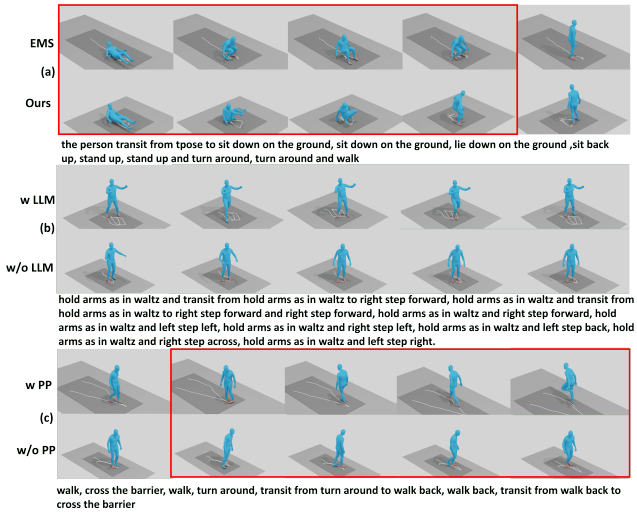

Fig. 4: Qualitative Results (a) Comparing between EMS and our Model. (b) Comparing between models trained with LLM augmentation (w LLM) and without LLM augmentation (w/o LLM). (c)Comparing between models trained with LLM augmentation with post-processing (w PP) and without post-processing (w/o PP).

6 Conclusion&Limitations

We introduce TMT, a novel text-conditioned 3D human motion generation model aimed at creating natural 3D human motions from open-vocabulary descriptions. It excels particularly with unseen actions by leveraging semantic augmentation through Large Language Models (LLMs). Additionally, we constructed LaViMO, a comprehensive 3D human motion dataset. Our approach demonstrates superior performance over existing state-of-the-art models from both quantitative and qualitative perspectives.

Despite our method significantly outperforming previous state-of-the-art models, particularly in generating unseen actions, it encounters challenges with action descriptions that fall outside the pretraining dataset's scope. This highlights a need for further improvement in our model's robustness and generalization capabilities, especially when compared to leading methods in text-conditioned image generation (e.g., DALLE2 [34]), text-conditioned video generation (e.g., Imagen [14]), and text-conditioned text generation (e.g., LLaMA2 [38] and GPT4 [7]). Our findings suggest that expanding the size and diversity of the pretraining dataset, particularly with coarse motions extracted from RGB videos, could still enhance the robustness of motion VQ-VAE.

7 Acknowledgments

Jungdam Won was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the

Korea government(MSIT) [NO.2021-0-01343-004, Artificial Intelligence Graduate School Program (Seoul National University)] and [IITP-2024-2020-0-01460, ITRC(Information Technology Research Center) support program]. He also got instrumental support from ICT(Institute of Computer Technology) at Seoul National University.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. 2018 IEEE International Conference on Robotics and Automation (ICRA) pp. 1–5 (2017)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. 2019 International Conference on 3D Vision (3DV) pp. 719–728 (2019)
3. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. 2022 International Conference on 3D Vision (3DV) pp. 414–423 (2022)
4. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Sinc: Spatial composition of 3d human motions for simultaneous action generation. ArXiv **abs/2304.10417** (2023), `https://api.semanticscholar.org/CorpusID:258236650`
5. Bie, X., Guo, W., Leglaive, S., Girin, L., Moreno-Noguer, F., Alameda-Pineda, X.: Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. arXiv preprint arXiv:2204.01565 (2022)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
7. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)
8. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12046–12055 (2019)
9. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1376–1386 (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
11. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
12. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
13. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A.A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. ArXiv **abs/2210.02303** (2022), `https://api.semanticscholar.org/CorpusID:252715883`

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. ArXiv **abs/2204.03458** (2022)
17. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36** (2024)
18. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349 (2022)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. arXiv preprint arXiv:2212.05897 (2022)
21. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)
22. Lin, A.S., Wu, L., Corona, R., Tai, K.W.H., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions (2018)
23. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.w.: Ohmg: Zero-shot open-vocabulary human motion generation. arXiv preprint arXiv:2210.15929 (2022)
24. Lin, X., Amer, M.R.: Human motion modeling using dvgans. ArXiv **abs/1804.10652** (2018)
25. Liu, J., Shahroudy, A., Perez, M., Wang, G., yu Duan, L., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 2684–2701 (2019), `https://api.semanticscholar.org/CorpusID:152282878`
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5441–5450 (2019)
27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
28. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10965–10975 (2021)
29. Petrovich, M., Black, M.J., Varol, G.: 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. pp. 480–497. Springer (2022)
30. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4 4**, 236–252 (2016)
31. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., for Intelligent Systems, M.J.B.M.P.I., Konstanz, U.: Babel: Bodies, action and behavior with english labels. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 722–731 (2021)
32. Qian, Y., Urbanek, J., Hauptmann, A.G., Won, J.: Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 2306–2316 (2023)
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. ArXiv **abs/2204.06125** (2022), https://api.semanticscholar.org/CorpusID:248097655
35. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11468–11479 (2021)
36. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. ArXiv **abs/2303.01418** (2023), https://api.semanticscholar.org/CorpusID:257279944
37. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023)
39. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
40. Wang, W., Zhe, X., Chen, H., Kang, D., Li, T., Chen, R., Bao, L.: Neural marionette: A transformer-based multi-action human motion synthesis system. arXiv preprint arXiv:2209.13204 (2022)
41. Xu, H., Durme, B.V., Murray, K.: Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. ArXiv **abs/2109.04588** (2021)
42. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023)
43. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
44. Zhu, M., Lin, X., Dang, R., Liu, C., Chen, Q.: Fine-grained spatiotemporal motion alignment for contrastive video representation learning. Proceedings of the 31st ACM International Conference on Multimedia (2023), https://api.semanticscholar.org/CorpusID:261493823