🍅 HumanTOMATO: Text-aligned Whole-body Motion Generation

Shunlin Lu[*†♠◇] **Ling-Hao Chen**[*†♣◇]
Ailing Zeng[◇] **Jing Lin**[†♣◇] **Ruimao Zhang**[♠] **Lei Zhang**[◇] **Heung-Yeung Shum**[♣◇]

[*]Co-first author. Listing order is random. {shunlinlu0803, thu.lhchen}@gmail.com

[♣] Tsinghua University [◇] International Digital Economy Academy (IDEA) [♠] School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ)

[†] This work was done when S. Lu, L.H. Chen, and J. Lin were research interns at IDEA. Corresponding authors: R. Zhang and H.-Y. Shum, Project lead: A. Zeng.

Project page: https://lhchen.top/HumanTOMATO

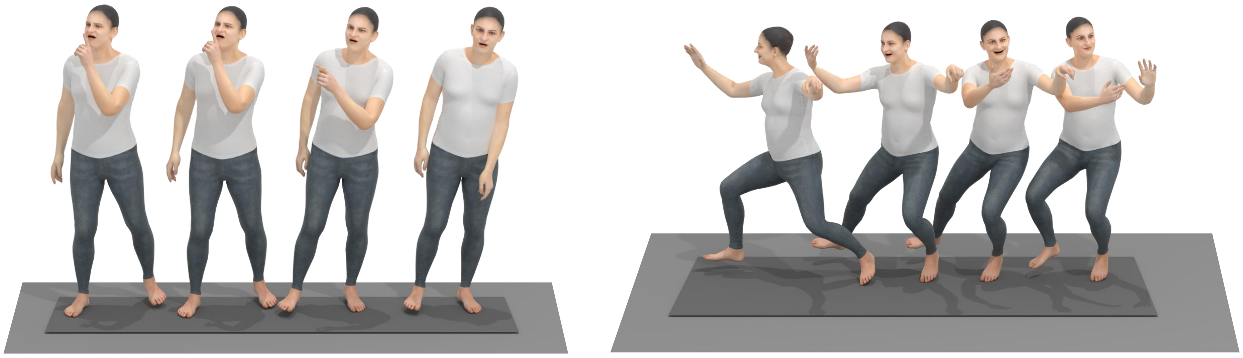

(a) stand and shush, angrily.
(b) Yang-style 40 form Tai Chi Competition routine step 34, happily.

Figure 1: The proposed HumanTOMATO can generate text-aligned whole-body motions with vivid and harmonious **face, hand, and body** motion. We show two generated motion keyframes based on the given texts.

Abstract

This work targets a novel text-driven **whole-body** motion generation task, which takes a given textual description as input and aims at generating high-quality and diverse facial expressions, hand gestures, and body motions simultaneously. Previous works on text-driven motion generation tasks mainly have two limitations: they ignore the key role of fine-grained hand and face controlling in vivid whole-body motion generation, and lack a good alignment between text and motion. To address such limitations, we propose a Text-aligned whOle-body Motion generATiOn framework, named HumanTOMATO, which is the first attempt to our knowledge towards applicable holistic motion generation in this research area. To tackle this challenging task, our solution includes two key designs: (1) a Holistic Hierarchical VQ-VAE (*aka* H^{2}VQ) and a Hierarchical-GPT for fine-grained body and hand motion reconstruction and generation with two structured codebooks; and (2) a pre-trained text-motion-alignment model to help generated motion align with the input textual description explicitly. Comprehensive experiments verify that our model has significant advantages in both the quality of generated motions and their alignment with text.

1. Introduction

Recent years have seen an explosion of huge demand for generating high-quality 3D human motions in many scenarios, such as games, films, animation, and robotics. To reduce laborious efforts in animation creation, recent studies (Tevet et al., 2023; Chen et al., 2023b; Zhang et al., 2022; 2023) attempt to generate human motions with textual description in a natural interactive way and have achieved rapid progress.

However, the generated motions from existing works are still unsatisfactory to meet real application needs. The problem is mainly due to two aspects. First, *existing text-driven motion generation models can only generate body-only mo-*

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

tions rather than whole-body motions, which are highly expressive yet much more challenging. On the one hand, the mentioned challenge comes from the limited availability of whole-body motion data. On the other hand, whole-body motion is much more complex, where fine-grained motions of body, hand, and face should be well generated. How to model whole-body human motions is still under-explored. Second, *the generated motions lack semantic alignment with the textual description.* Existing methods adopt CLIP (Radford et al., 2021) or Large Language Models (LLMs) (Raffel et al., 2020) to provide language guidance for motion generation (Zhang et al., 2022; Tevet et al., 2023; 2022; Jiang et al., 2023). However, their alignment supervision is provided at frame level and lacks sufficient understanding of a motion at its whole sequence level. As a result, they often fail to distinguish some scenarios, such as "`walking in a clockwise circle`" and "`walking in a counter-clockwise circle`", which requires understanding motions at sequence level rather than frame level. Such a drawback severely limits the ability to generate motions well-aligned with textual descriptions.

To tackle the above issues, we propose a novel Text-aligned whOle-body Motion generATiOn framework (HumanTOMATO), which includes two key designs. First, *a holistic hierarchical discrete modeling strategy for body and hand motions is proposed for reconstructing and generating whole-body motions vividly.* As whole-body motion is a kind of high-dimensional spatio-temporal signal, in the first stage, we propose a Holistic Hierarchical VQ-VAE (*aka* H^2VQ) to compress the motion into two-level discrete codes for body and hand, respectively. In contrast, a naïve solution that simply replaces body-only motions with whole-body motions or directly increases the size of the codebook is almost in vain. The key insight of our H^2VQ is learning informative and compact representations of fine-grained whole-body motions at very low bit rates. Moreover, the hand and body motions have different levels of amplitudes and details, which motivates us to model them separately. Based on the two-level discrete codes, in the second stage, we propose a Hierarchical-GPT to predict the hierarchical discrete codes of body and hand in an auto-regressive fashion. Extending the hierarchical modeling strategy of body-hand motions, we use an RVQ-based method for motion reconstruction and also generate facial motion with discrete codes auto-regressively. Second, *a pre-trained text-motion-alignment model is introduced to enhance the textual alignment of generated motions for the first time.* We pre-train a motion encoder and a text encoder, namely TMA (Text-Motion Alignment), in a contrastive learning manner (Radford et al., 2021) with pairwise text-motion data. Unlike previous work (Zhang et al., 2022; Tevet et al., 2023; 2022; Jiang et al., 2023) that relied on CLIP or LLMs embedding, our approach utilizes TMA text embedding as a language prior. In this way, the TMA provides a motion-aware language embedding for the Hirarchical-GPT to generate discrete motion codes more precisely. It is worth noting that, during training, merely supervising the prediction of discrete code tokens of body and hand is insufficient as it lacks supervision on the semantics of global motion sequence and leads to error accumulation in auto-regressive prediction. Thus, with the text-motion similarity measured by TMA, we additionally provide text-motion alignment supervision to supervise the alignment between generated motion sequences and texts explicitly.

With these key designs, compared with previous text-driven motion generation works, HumanTOMATO can generate whole-body motions semantically aligned with texts, as illustrated in Figure 1. To evaluate the alignment between generated motions and input texts, we further revisit the previous retriever used for evaluating text-motion alignment and find that its retrieval ability is worse than TMA. Hence, we introduce two new criteria (*TMA-R-Precision*$^{(256)}$ and *TMA-Matching-score*), which are more accurate and challenging to evaluate the text-motion alignment in this task.

We summarize our key contributions as follows:

- To the best of our knowledge, we propose the challenging Text-driven whOle-body Motion generATiOn task for the first time and design a model (**HumanTOMATO**) to generate vivid whole-body motions aligned with texts.
- To tackle the challenging whole-body motion generation problem, we introduce a H^2VQ for fine-grained body and hand motion reconstruction. Accordingly, we develop a Hierarchical-GPT combined with a facial motion generator to generate whole-body motions.
- To enhance the consistency and alignment between texts and motions, we pre-train text-motion-aligned encoders via a contrastive objective and introduce sequence-level semantic supervision to help motion-text alignment.
- We propose two new criteria (*TMA-R-Precision*$^{(256)}$ and *TMA-Matching-score*), which are more accurate and challenging for evaluating text-motion alignment.

We evaluate HumanTOMATO on both whole-body (Lin et al., 2023b) and body-only (Guo et al., 2022) motion generation benchmarks and answer four research questions based on our contributions. Comprehensive experiments affirm the vividness and alignment of our generated motions, outperforming competitors in motion reconstruction (32.7% $\downarrow$ MPJPE *v.s.* VQ) and motion generation metrics (9.2% $\uparrow$ TMA-R-Precision$^{(256)}$ Top3 *v.s.* the best baseline).

2. Related Work

Due to the page limitation, we leave more discussions on related work in Appendix A. There are three related research, including unconditional motion generation (Yan et al., 2019; Zhao et al., 2020; Zhang et al., 2020; Cai et al., 2021), text-driven motion generation (Petrovich et al., 2022; Zhang et al., 2022; Chen et al., 2023b; Guo et al., 2022), and co-speech motion generation (Yi et al., 2023; Zhi et al., 2023; Fan et al., 2022; Habibie et al., 2021). However, these works cannot generate whole-body motion from text. Besides, the effective quantization method and how to achieve higher text-motion alignment are not been carefully explored yet. Accordingly, we introduce our methodology as follows.

3. Methodology

3.1. Problem Formulation

We clarify notations and set up the novel research problem of text-driven **whole-body** motion generation. Given a text description $\mathbf{t}$ of a human motion, such as "*The man is playing the ukulele happily.*", the model should generate a vivid whole-body motion $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_L] \in \mathbb{R}^{L\times d}$ aligned with the text description, where L and d denote the number of frames and the dimension of the motion in each frame, respectively. As whole-body motion comes up with hand, body, and face motions, we can also decompose the $\mathbf{m}$ as $\{\mathbf{m}^H, \mathbf{m}^B, \mathbf{m}^F\}$ respectively, where $\mathbf{m}^H \in \mathbb{R}^{L\times d_h}, \mathbf{m}^B \in \mathbb{R}^{L\times d_b}, \mathbf{m}^F \in \mathbb{R}^{L\times d_f}, d = d_h + d_b + d_f$. Mathematically, we formulate the text-driven whole-body motion generation as follows:

$$\Theta^\star = \arg\max_{\Theta} P_\Theta(\mathbf{m} \mid \mathbf{t}), \tag{1}$$

where Θ denotes the model parameters and $P_\Theta(\cdot)$ denotes the motion distribution, respectively.

3.2. Learning Discrete Whole-body Representations

Vanilla Motion VQ-VAE. Motion VQ-VAE aims to learn discrete representations of human motions in an encoding-decoding fashion. Specifically, VQ-VAE recovers motions by using an auto-encoder and learns a codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^K$, where K denotes the codebook size and $\mathbf{e}(k)$ indicate the k-th embedded representation in the codebook. Given a vector $\mathbf{z}$ and the quantizer $\mathcal{Q}(\cdot;\mathcal{C})$, the quantized vector should be the element selected from the codebook $\mathcal{C}$ that can minimize the reconstruction error of $\mathbf{z}$ as,

$$\hat{\mathbf{z}} = \mathcal{Q}(\mathbf{z};\mathcal{C}) = \arg\min_{\mathbf{e}_k} \|\mathbf{z} - \mathbf{e}_k\|_2^2. \tag{2}$$

In a vanilla VQ-VAE, $\mathbf{z} = \mathtt{Enc}(\mathbf{m})$ indicates the latent code extracted from a motion encoder $\mathtt{Enc}(\cdot)$. Thus VQ-VAE can be optimized by,

$$\mathcal{L} = \|\mathbf{m} - \mathtt{Dec}(\mathcal{Q}(\mathbf{z};\mathcal{C}))\|_2^2 + \alpha\|\mathbf{z} - \mathtt{sg}(\hat{\mathbf{z}})\|_2^2, \tag{3}$$

where α is the hyper-parameter, $\mathtt{sg}(\cdot)$ is the stop-gradient operation and $\mathtt{Dec}(\cdot)$ indicate the motion decoder. Different from traditional methods, the codebook $\mathcal{C}$ in motion VQ-VAE is optimized by exponential moving average (EMA) and codebook reset (Code Reset) operations following Razavi et al. (2019); Van Den Oord et al. (2017); Zhang et al. (2023). While the discrete vector quantization of vanilla VQ-VAE is capable of compressing human motions, it falls short in minimizing quantization errors for detailed whole-body motion generation. In practice, an intuitive solution to address this would be to increase the size of the codebook. However, this scheme would evidently introduce additional computational cost and quickly encounter performance bottlenecks (see results in Section 4.4).

Holistic Hierarchical VQ-VAE. Recently, the Residual Vector Quantization technique, also known as RVQ (Barnes et al., 1996; Zeghidour et al., 2021; Yao et al., 2023), has significantly advanced the development of music generation task (Défossez et al., 2022; Copet et al., 2023). Technically, RVQ iteratively quantizes the quantization error at each level from the previous one, reducing quantization errors effectively while maintaining a low memory cost of the codebook (see Appendix C.2 for details). Motivated by this (Défossez et al., 2022), we propose a novel Holistic Hierarchical Vector Quantization scheme, shorted by H^{2}VQ, into the field of motion generation. Unlike RVQ, we incorporate the kinematic structure prior to the H^{2}VQ modeling, enabling it to learn compact representations of fine-grained whole-body motions at an extremely low bit rate. Given the distinct differences in amplitude and frequency between body and hand motions, we have further designed two separate encoders and codebooks to learn discrete representations for body and hand motions.

The architecture of our proposed H^{2}VQ is illustrated in Figure 2(a). In the encoding phase, we input hand and body motions, obtaining hand and body tokens through the hand encoder $\mathtt{Enc}^{\mathtt{H}}(\cdot)$ and body encoder $\mathtt{Enc}^{\mathtt{B}}(\cdot)$, respectively. The learned hand tokens are further quantized by the Hand Quantizer $\mathcal{Q}^H(\cdot\,;\mathcal{C}^H)$ as $\mathbf{z}^H$. Since the body motions are usually highly associated with some hand gestures (Ao et al., 2022), to train a more natural and coordinated body codebook, we fuse the body and hand tokens using the $\mathtt{Concat}(\cdot)$ and $\mathtt{Conv1d}(\cdot)$ operations. As shown in Figure 2, before this fusion operation, the quantized hand tokens undergo a transformation through a projection layer. After that, fused tokens are further quantized by Body Quantizer $\mathcal{Q}^B(\cdot\,;\mathcal{C}^B)$ as $\mathbf{z}^B$. Finally, the hand tokens $\mathbf{z}^H$ and body tokens $\mathbf{z}^B$ are concatenated together and fed into the Body-hand Decoder to reconstruct the body-hand motions precisely.

During the training phase, the primary goal is to reconstruct motions while concurrently updating the two codebooks through the EMA and Code Reset operations (Razavi et al., 2019; Van Den Oord et al., 2017; Zhang et al., 2023). In the inference phase, after obtaining quantized code indices, the Body-hand Decoder can generate body-hand motions

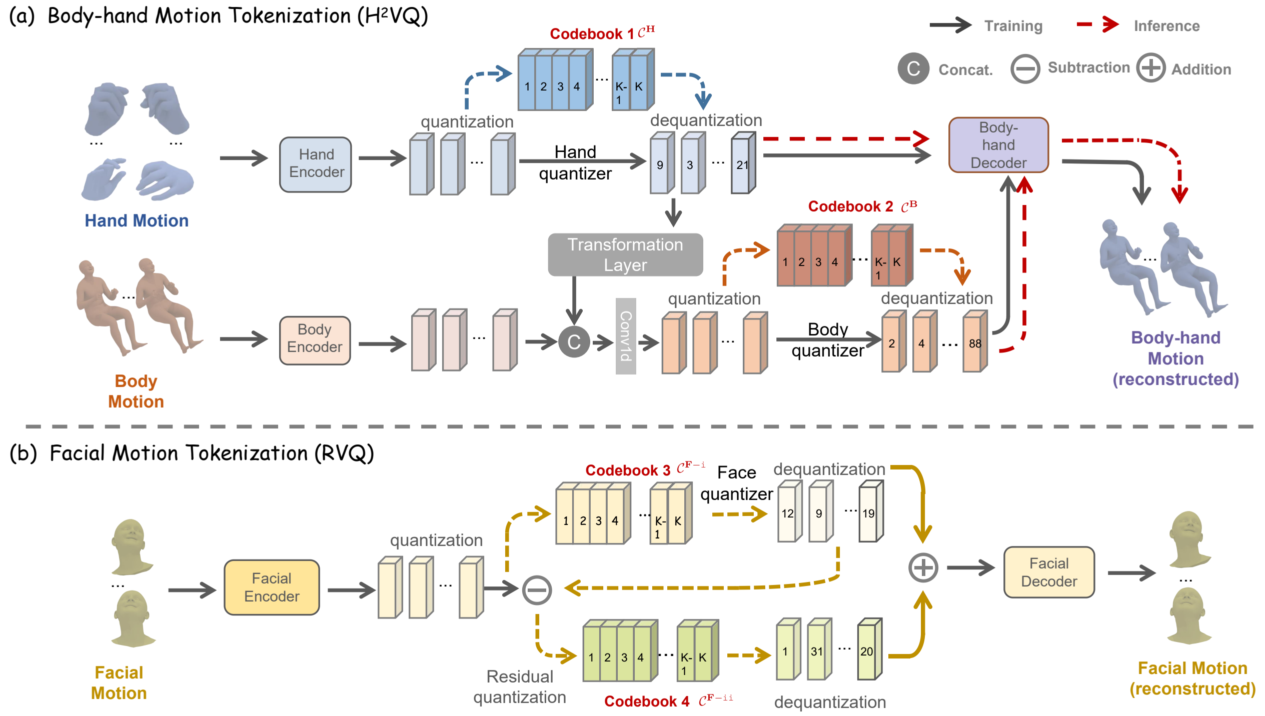


Figure 2: The framework overview of tokenization method for body-hand, and facial motions. (a) Holistic Hierarchical Vector Quantization (H^{2}VQ) to compress fine-grained body-hand motion into two discrete codebooks with hierarchical structure relations. (b) Residual Vector Quantization (RVQ) to compress facial motion into two discrete codebooks with hierarchical structure relations.

by querying the respective codebooks with obtained code indices. The detailed algorithmic flows for both training and inference phases can be found in Appendix C.

3.3. Hierarchical Whole-body Motion Generation

Given the two precise quantized codebooks of H^{2}VQ, the motion sequence should be generated by using the corresponding decoders and quantized codes. The previous popular approach is to predict code indices in GPT-like auto-regressive fashion (Zhang et al., 2023). Since the proposed H^{2}VQ requires the usage of two codebooks with structure relations, the aforementioned approach is not applicable. To better model the natural coherence of body-hand motions, we design a hierarchical discrete codes prediction module, named Hierarchical-GPT, which is illustrated in Figure 3(a), for generating body-hand motions.

Hierarchical-GPT. The Hierarchical-GPT is built upon a transformer-based architecture, where the first input token is a textual embedding. With the input body-hand motion $\mathbf{m}^B = [\mathbf{m}^B_1, \mathbf{m}^B_2, \cdots, \mathbf{m}^B_L]$ and $\mathbf{m}^H = [\mathbf{m}^H_1, \mathbf{m}^H_2, \cdots, \mathbf{m}^H_L]$, we have corresponding code indices, denoted as $\mathrm{I}^B = [\mathrm{I}^B_1, I^B_2, \cdots, \mathrm{I}^B_{L/r}, \texttt{End}]$ and $\mathrm{I}^H = [\mathrm{I}^H_1, \mathrm{I}^H_2, \cdots, \mathrm{I}^H_{L/r}, \texttt{End}]$, where 'End' indicates the end token and r denotes the down-sampling rate, which is used to convert the input motion sequence to discrete motion tokens. Therefore, as shown in Figure 3(a), the code indices prediction mechanism can be formulated as an auto-regressive prediction problem:

$$\begin{aligned} P(\mathrm{I}^{B,H}_{1,2,\cdots,L/r} \mid \mathbf{t}) &= \textstyle\prod_{s=1}^{L/r} P(\mathrm{I}^{B,H}_s \mid \mathrm{I}^{B,H}_{<s}, \mathbf{t}) \\ &= \textstyle\prod_{s=1}^{L/r} P(\mathrm{I}^B_s \mid \mathrm{I}^{B,H}_{<s}, \mathbf{t}) \cdot P(\mathrm{I}^H_s \mid \mathrm{I}^B_s, \mathrm{I}^{B,H}_{<s}, \mathbf{t}), \end{aligned} \quad (4)$$

where we first predict the body token index and then predict the hand token index at each down-sampled timestamp s. As shown in Figure 3(a), the first token is the textual embedding of the input text. Here we leverage a pre-trained text encoder to extract such an embedding. Please refer to Section 3.5 for more details. In practice, we train the prediction transformer with casual self-attention (Vaswani et al., 2017). As the Hierarchical-GPT aims to predict code indices, our model is optimized with the cross-entropy loss $\mathcal{L}_{CE}$. The training details are available in Appendix B.3.

3.4. Facial Motion Generator

Previous works (Richard et al., 2021; Fan et al., 2022; Habibie et al., 2021; Yi et al., 2023; Ng et al., 2024) hold the view that facial expression is partially independent of body and hand motions while highly related to the given facial descriptions and even speech. Moreover, the facial motion is represented in expression parameters, which is different from skeleton-based motions. Additionally, our experimental results in Section 4.3 (Table 6) also empirically verify the

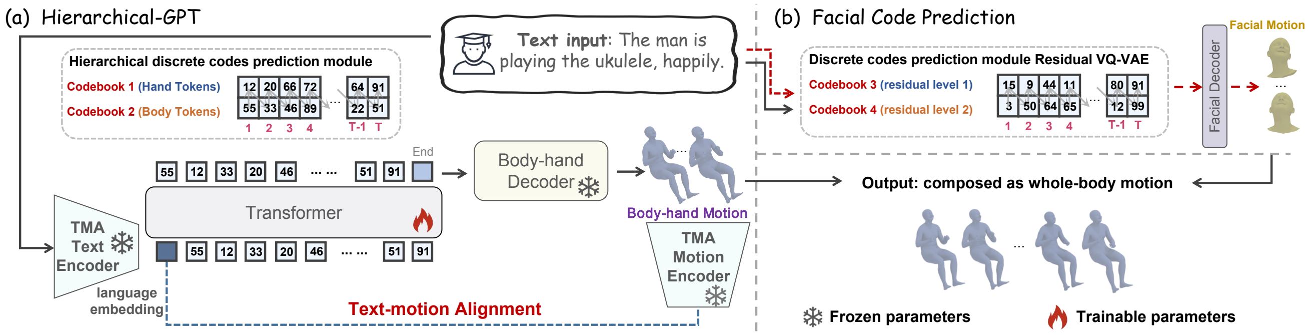


Figure 3: The code prediction mechanism of the (a) body-hand, and (b) facial motion generation. Both parts take textual description as input and predict tokens in an auto-regressive manner. The final whole-body motion is composed of both part motions decoded by the corresponding decoders.

rationality of modeling body-hand and facial motions separately. Based on these philosophical grounds, we generate the facial motion based on given expression texts separately. As shown in Figure 2(b), extending the hierarchical modeling strategy, we take the RVQ as the quantizers ($\mathcal{C}^{F-\texttt{i}}$ and $\mathcal{C}^{F-\texttt{ii}}$) to reconstruct facial motions. Due to residual quantization of the facial motion reconstruction, tokens in two codebooks enjoy a hierarchical structure: tokens from previous quantizers represent the rough facial motion and the consecutive ones represent details (Wang et al., 2023a; Barnes et al., 1996; Zeghidour et al., 2021). Therefore, similar to body-hand code prediction, we generate facial motion tokens in an auto-regressive fashion (Figure 3(b)):

$$
\begin{aligned}
&P(\mathrm{I}_{1,2,\cdots,L/r}^{F-\texttt{i},F-\texttt{ii}} \mid \mathbf{t}) = \textstyle\prod_{s=1}^{L/r} P(\mathrm{I}_s^{F-\texttt{i},F-\texttt{ii}} \mid \mathrm{I}_{<s}^{F-\texttt{i},F-\texttt{ii}}, \mathbf{t}) \\
&= \textstyle\prod_{s=1}^{L/r} P(\mathrm{I}_s^{\texttt{i}} \mid \mathrm{I}_{<s}^{F-\texttt{i},F-\texttt{ii}}, \mathbf{t}) \cdot P(\mathrm{I}_s^{F-\texttt{ii}} \mid \mathrm{I}_s^{F-\texttt{i}}, \mathrm{I}_{<s}^{F-\texttt{i},F-\texttt{ii}}, \mathbf{t}),
\end{aligned}
\tag{5}
$$

where `i` and `ii` are used to distinguish two codebooks.

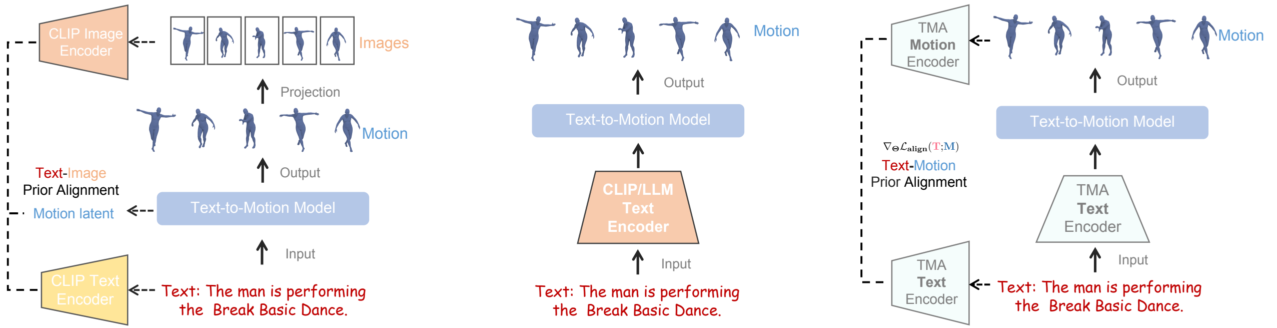
3.5. Pre-trained Text-motion Aligned Model as a Prior

In existing pre-trained models, there often exists a notable semantic gap between the representation of text and its corresponding motion due to the differences in the granularity of content representation between text and motion. For instance, text may involve a simple verb but its corresponding motion would require a sequence of actions. For the text-to-motion generation task, it is crucial to ensure that the textual embedding extracted from the text encoder is motion-aware. Therefore, we try to bridge the gap between text and motion representations, thereby obtaining a text embedding more conducive to driving motion generation.

As shown in Figure 4(a) and Figure 4(b), we can briefly divide previous attempts into two categories. The first is *supervision by an image-text aligned prior explicitly.* As there was no strong text-motion-aligned pre-trained model, MotionCLIP (Tevet et al., 2022) supervises the alignment between text embedding, image embedding, and motion embedding with the CLIP model. However, the image encoder of CLIP is a strong supervision of static image content understanding, which is quite different from dynamic motion. This supervision will cause the generated motion to be *over-smoothing*, even *stillness* (see Appendix E). Therefore, supervising generated motion via a text-image-aligned prior is inappropriate. The second is *learning with image-text aligned prior implicitly.* Existing attempts (Tevet et al., 2023; Zhang et al., 2023; 2022; Yuan et al., 2023) take the CLIP text embedding as the language prior to the text-motion model training. On the one hand, it learns the motion-text alignment implicitly with pairwise data, and there is no supervision to discriminate whether the generated motion is aligned with the text explicitly. On the other hand, the CLIP text embedding is aligned with visual content, lacking the understanding of dynamic motion clues, which cannot provide sufficient spatial-temporal information to generate text-aligned motion. Therefore, it is essential to introduce a text-motion-aligned pre-training method, ensuring that the trained text encoder can output textual embeddings more conducive to accomplishing text-to-motion generation tasks, instead of adapting from the image-text-aligned model.

Motivated by Petrovich et al. (2023), we pre-train a motion encoder and a text encoder via aligning Text and Motion in a contrastive way (Radford et al., 2021) through a Alignment target, named TMA. Different from previous work (Zhang et al., 2022; Tevet et al., 2023; 2022; Jiang et al., 2023), the text embedding of TMA plays the role of motion-aware language prior other than the embedding from CLIP or LLMs, which is beneficial for generating text-aligned motions. In this work, the TMA is re-trained by ourselves. We leave the training details in Appendix D.

Based on the pre-trained TMA, we further explore enhancing the alignment between the given text and generated motions from two aspects, which are shown in Figure 4(c). The first is *replacing the CLIP text encoder with the TMA text encoder.* Compared with the CLIP text encoder, the pre-trained TMA text encoder provides text embeddings aligned better with dynamic human motions. With the motion-aware language prior, our model can capture motion sequentiality, directions, and dynamics better than text-image-aligned language prior. The second is *introducing the motion-text*


CLIP Image Encoder
Images
Projection
Motion
Output
Text-Image Prior Alignment
Motion latent
Text-to-Motion Model
Input
CLIP Text Encoder
Text: The man is performing the Break Basic Dance.
Motion
Output
Text-to-Motion Model
CLIP/LLM Text Encoder
Input
Text: The man is performing the Break Basic Dance.
TMA Motion Encoder
Motion
Output
$\nabla_{\Theta}\mathcal{L}_{\text{align}}(\mathbf{T};\mathbf{M})$
Text-Motion Prior Alignment
Text-to-Motion Model
TMA Text Encoder
Input
TMA Text Encoder
Text: The man is performing the Break Basic Dance.

(a) Learning *image*-text aligned prior *explicitly*. (b) Learning *image*-text aligned prior *implicitly*. (c) Learning *motion*-text alignment *explicitly* (Ours).

Figure 4: Technical comparisons on introducing language priors of existing methods.

alignment supervision with TMA. When training, we feed the generated motion and the given text into the pre-trained TMA motion encoder and text encoder, respectively, to obtain both motion and text embeddings. Then, we calculate a contrastive loss $\mathcal{L}_{align}$ (Radford et al., 2021) for supervising the motion-text alignment. Accordingly, the weighted contrastive loss $\eta\mathcal{L}_{align}$ is added to the optimization objective, where η is the hyper-parameter. The proposed optimization objective provides explicit sequence-level supervision for the text-motion alignment.

3.6. Model Training and Inference

Model Training. In the first stage, similar to the vanilla VQ (Eqn. 3), H^{2}VQ is optimized by,

$$\begin{aligned}\mathcal{L} = \|\mathbf{m} - \texttt{Dec}\left(\mathcal{Q}^H(\mathbf{z}^H;\mathcal{C}^H), \mathcal{Q}^B(\mathbf{z}^B;\mathcal{C}^B)\right)\|_2^2 \\ +\alpha\left(\|\mathbf{z}^H - \texttt{sg}(\hat{\mathbf{z}}^H)\|_2^2 + \|\mathbf{z}^B - \texttt{sg}(\hat{\mathbf{z}}^B)\|_2^2\right).\end{aligned} \tag{6}$$

The RVQ of the facial motion auto-encoder is trained in a similar way. Besides, the codebooks are optimized by EMA and Code ReSet techniques. In the second stage, we train the Hierarchical-GPT with both the cross-entropy loss $\mathcal{L}_{CE}$ and the text-motion alignment loss $\mathcal{L}_{align}$, the overall loss as $\mathcal{L}_{CE} + \eta\mathcal{L}_{align}$.

Model Inference. In the inference phase, we first extract the text embedding from TMA. Then we feed the TMA textual embedding as the initial token into the Hierarchical-GPT, which then predicts discrete body and hand tokens in an auto-regressive fashion. The body and hand tokens are fed into the Body-hand Decoder to generate text-aligned human motion. Ultimately, incorporating the facial motions produced by the facial motion genrator, we output the comprehensive whole-body motions.

4. Experiments

In this section, we evaluate the proposed HumanTOMATO model on both whole-body and body-only motion generation benchmarks. Besides, we will also present ablations on each technical design of our method. We structure the experiments to answer the following research questions (RQs).

- **RQ1:** Does our proposed HumanTOMATO model outperform existing generation methods on the whole-body motion generation task?
- **RQ2:** How do hierarchical representations of body-hand motions help improve the quality of motion generation?
- **RQ3:** How does the pre-trained text-motion aligned model help the text-motion alignment?
- **RQ4:** Why are the proposed evaluation metrics on alignment between generated motions and given texts more accurate and challenging?

4.1. Datasets and Evaluation

4.1.1. Whole-body and Body-only Datasets

Motion-X (Lin et al., 2023b) is the largest 3D whole-body motion-text dataset, consisting of 95,642 high-quality human motions along with 95,642 text captions. In Motion-X, GRAB (Taheri et al., 2020) is a representative subset with vivid grab motions, which is used for our ablation study.

HumanML3D (Guo et al., 2022) is currently the largest 3D body-only motion-text dataset, which consists of 14,616 human motions along with 44,970 text captions.

We take the Motion-X dataset to evaluate the whole-body motion generation task and the HumanML3D dataset to perform ablations for verifying the generalizability of our proposed solution to the body-only motion generation setting. We follow Lin et al. (2023b); Guo et al. (2022) to split these datasets into training, validation, and test sets with proportions of 80%, 5%, and 15%.

4.1.2. Evaluation Details

We quantitatively evaluate the generated motions from three aspects. (1) **The quality of the generated motions.** *Frechet Inception Distance (FID)* is adopted to measure the gap between the distributions of the generated and real motions. (2) **Text-Motion Alignment.** *Matching-score* is used to measure the similarity between texts and the generated motions and *R-Precision*$^{(B)}$ is used to measure the motion-to-text retrieval accuracy in a B-size retrieval pairwise motion-text set. (3) **Generation diversity.** We use *Diversity* to evaluate the average extracted feature Euclidean distances among 300 randomly sampled motion pairs and use *MModality* to measure the generation diversity within the same given text.

To better evaluate the alignment between generated motions and texts, we additionally introduce new metrics to evaluate text-motion alignment from two aspects: (1) **More accurate evaluation.** Previous works used the feature extractor from Guo et al. (2022) to calculate the *R-Precision*$^{(B)}$ and *Matching-score*. However, its retrieval accuracy is not as accurate as the TMA described in Section 3.5 (comparison in Section 4.6). Therefore, we introduce *TMA-R-Precision*$^{(B)}$ and *TMA-Matching-score* to evaluate the text-motion alignment, where the feature extractor is replaced by TMA but not the retriever in Guo et al. (2022). (2) **More challenging evaluation metrics.** Retrieval of corresponding texts in a 32-size set is easier than in a larger size set. Therefore, we add a new retrieval setting as $B = 256$. The comparison between these two settings will be shown in Section 4.6.

We compare our HumanTOMATO with existing state-of-the-art baselines. (1) **TEMOS**: TEMOS (Petrovich et al., 2022) is a VAE-based text-to-motion generation framework. (2) **T2M-GPT**: The T2M-GPT (Zhang et al., 2023) method learns discrete representations for motions at first, and then introduces a GPT-like codes prediction mechanism in the second stage with CLIP prior. (3) **MotionDiffuse**: MotionDiffuse (Zhang et al., 2022)is a pioneering work that introduces diffusion models into the field of action generation, predicting noise in each iteration. (4) **MDM**: MDM (Tevet et al., 2023) is also early research using diffusion models to generate motion, predicting ground truth in each iteration. (5) **MLD**: Motivated by latent diffusion models (Rombach et al., 2022), MLD (Chen et al., 2023b) learns motion latent representations for motion VAE via a diffusion model. For facial motion generation, as this is the first attempt to generate whole-body motion, we take both cVAE-based (Petrovich et al., 2022) and diffusion-based (Tevet et al., 2023) methods as baselines. Both methods are extended from previous motion generation models. More details of facial motion generation baselines are in Appendix B.5.

4.2. Implementation Details

Motion Representation. For body-hand motion representations, inspired by the motion representation (H3D-Format) in Guo et al. (2022), we expand the body-only representation to holistic body-hand motion representation. Specifically, the i-th pose is defined by a tuple of root angular velocity $\dot{r}^a \in \mathbb{R}$ along the Y-axis, root linear velocities ($\dot{r}^x, \dot{r}^z \in \mathbb{R}$) on XZ-plane, root height $r^y \in \mathbb{R}$, local joints positions $\mathbf{j}^p \in \mathbb{R}^{3N-1}$, and velocities $\mathbf{j}^v \in \mathbb{R}^{3N}$, where N denotes the number of whole body joints, including both body joints and hand joints. For face motion representations, we follow the Flame Format (Kim et al., 2023) and use $\mathbf{f} \in \mathbb{R}^{50}$ to represent the face expression. Thus, we represent the whole-body motion as $m_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, \dot{r}^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{f}\}$. We conduct a set of ablation studies on HumanML3D based on VAE and VQ-VAE to justify the motion format. Please refer to Appendix B.1 for more details.

	FID↓	Top1↑	Top2↑	Top3↑	Diversity↑	Matching-score↓
GT	-	0.277	0.428	0.507	10.304	6.065
cVAE	1.530	0.084	0.114	0.165	6.316	9.973
Diffusion	3.342	0.064	0.109	0.155	**8.657**	13.410
Ours	**1.044**	**0.200**	**0.311**	**0.374**	7.175	**6.997**

Table 4: Comparison with baselines on facial motion generation.

	FID	Top1$^{(32)}$	Top3$^{(32)}$	TMA Top1$^{(256)}$	TMA Top3$^{(256)}$
Separate	2.209 (.047)	0.359 (.002)	0.666 (.002)	0.306 (.003)	0.552 (.002)
Ours	**1.174** (.015)	**0.416** (.009)	**0.703** (.007)	**0.399** (.000)	**0.638** (.004)

Table 5: Separate *v.s.* Holistic modeling strategy on the body-hand motion. The test Mean (±std.) values are reported.

Experiment Details. All our experiments are trained with the AdamW (Loshchilov & Hutter, 2019) optimizer using a fixed learning rate of 10^{-4} on 4× NVIDIA Tesla A100-80GB GPUs and are tested on 1× NVIDIA Tesla A100-80GB GPU. Training batch size is set to 256 for both H^{2}VQ and Hierarchical-GPT stages. Each experiment is trained for 6,000 epochs during H^{2}VQ stages and 2,000 epochs during Hierarchical-GPT stages. Two codebook sizes are both 512. Please refer to Appendix B for more details.

4.3. Main Results Analysis (RQ1)

We answer RQ1 from both quantitative and qualitative aspects. (1) *Quantitative results.* We quantitatively compare our method with baselines from body-hand motion generation quality, text-motion alignment, and diversity, which are shown in Table 1. The metrics show that our method enjoys good generation quality and text-motion alignment (9.2% ↑ TMA-R-Precision$^{(256)}$ Top3 *v.s.* best baseline). The mean values are reported in Table 1. The standard values are reported in Appendix F. (2) *Qualitative results.* We compare our method with MLD (Chen et al., 2023b) and T2M-GPT (Zhang et al., 2023). The comparison results shown in Figure 5 and Figure 6 demonstrate that our method has a stronger ability on the generation quality of different body parts (hand, body, and face). For the "`Flying Kick`" case, MLD and T2M-GPT fail to generate desirable motions, but our method achieves it. For the second case, MLD fails to generate "forward" motion, and motions generated by T2M-GPT walk backward first and finally walk forward. In contrast, HumanTOMATO generates a vivid motion corresponding to textual descriptions. For facial case analysis, our approach wins baselines on generation quality (0.486↓ on FID) and semantic alignment (20.9↑ on Top3). the results of cVAE tend to be over-smoothing,

	FID↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$			Matching Score ↓	TMA-Matching Score ↓	MModality↑	Diversity↑
		Top1↑	Top2↑	Top3↑	Top1↑	Top2↑	Top3↑				
GT	-	0.500	0.708	0.814	0.407	0.578	0.673	2.888	0.768	-	11.087
TEMOS	9.147	0.279	0.442	0.555	0.258	0.389	0.444	5.482	0.928	1.195	9.764
T2M-GPT	1.366	0.368	0.553	0.655	0.310	0.446	0.527	4.316	0.881	2.356	10.753
MDM	3.800	0.352	0.547	0.634	0.310	0.430	0.530	4.050	0.840	**2.530**	**11.400**
MLD	3.407	0.385	0.571	0.683	0.333	0.477	0.561	3.901	0.883	2.448	10.420
MotionDiffuse	**1.129**	0.391	0.587	0.695	0.368	0.493	0.584	3.950	0.829	1.654	10.580
HumanTOMATO	1.174	**0.416**	**0.603**	**0.703**	**0.399**	**0.555**	**0.638**	**3.894**	**0.809**	1.732	10.812

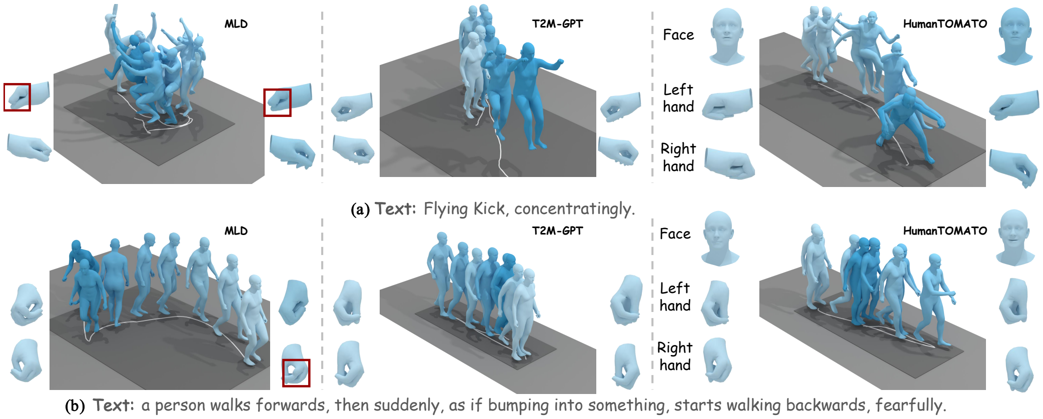
Table 1: Main results of motion generation on the Motion-X dataset.

Quantizer	v.-err.↓	FID↓	Top3↑
VQ	0.639	0.650	0.395
RVQ	**0.605**	**0.462**	**0.437**

Table 2: Ablation on different face tokenizers.

Codebook size	512	1024	4096
VQ	140.7	139.3	134.2
RVQ	110.9	111.2	116.8
H^2VQ	**93.0**	**83.9**	**84.9**

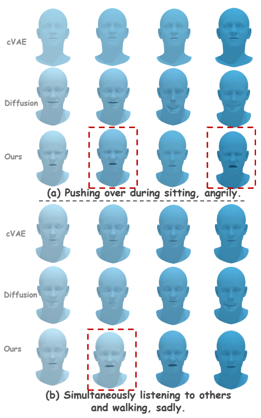
Table 3: Ablation on scaling the codebook size (MPJPE).



(a) Text: Flying Kick, concentratingly.

(b) Text: a person walks forwards, then suddenly, as if bumping into something, starts walking backwards, fearfully.

Figure 5: Qualitative comparisons with SOTA models trained on Motion-X. HumanTOMATO supports face motion generation and has better performance on natural hand motion generation and text-motion alignment.



(a) Pushing over during sitting, angrily.

(b) Simultaneously listening to others and walking, sadly.

Figure 6: Comparisons with baseline methods on face motion generation.

separate face modeling	body-hand				face	
	MPJPE↓	MPJPE-body ↓	MPJPE-hand↓	FID↓	v.-err. ↓	FID ↓
✗	108.31	72.02	42.50	0.45	1.368	1.406
✓	**92.97**	**62.34**	**37.20**	**0.20**	**0.605**	**0.462**

Table 6: Ablation on whether modeling with facial motion separately.

thereby diminishing the expressiveness of facial dynamics. Conversely, the diffusion-based method often results in inaccurate generations of jaw pose, subsequently distorting facial expressions. In contrast, our method demonstrates the capability to synthesize dynamic and vivid facial motions well aligned with the given facial texts. As shown in Table 2 (v.-err. means vertices error), our hierarchical design on face motions is better than vanilla VQ (0.188↓ FID). Furthermore, we explore the holistic modeling strategy of body and hand motions. We compare our method with modeling body and hand motions separately. As shown in Table 5, our proposed holistic modeling strategy outperforms the separately modeling strategy (details in Appendix J). We also verify the rationality of the separate modeling between body-hand and facial motions in Table 6.

4.4. Ablation on Hierarchical Representations (RQ2)

We compare the reconstruction result of our H^2VQ with the Vanilla VQ (512 or 1024 codebook size) and RVQ methods on three datasets in Table 7. We take the commonly used MPJPE metric (Gower, 1975; Lin et al., 2023b; Chen et al., 2023b) to evaluate the reconstruction performance. As can be seen in Table 7, increasing the size of codebook naïvely is almost in vain or even worse for motion reconstruction. The hierarchical modeling strategy improves the reconstruction performance significantly when learning informative low-dimensional representations ($\sim 32.7\% \downarrow$ MPJPE *v.s.* VQ). Moreover, our H^2VQ is better than RVQ in reconstructing whole-body motions, with gains mainly coming from the modeling of body and hand discrete codes explicitly. When verifying the key insight of our hierarchical modeling on body-only datasets, in contrast to HumanML3D only including body-part motions, we compare the Vanilla VQ-VAE with the RVQ technique to verify our motivation in Appendix G. Additionally, we also explore how the scaling of the codebook size affects the performance of all quantization methods. Table 3 shows that the naïve scaling of the codebook is almost in vain for vector quantization, and the carefully designed H^2VQ reduces errors by about 20%.

4.5. Text-motion Aligned Model As a Prior (RQ3)

Here, we evaluate our core design of introducing a pretrained text-motion aligned model as a prior. Ablation results in Table 8 show that our introduced motion-aware prior benefits the alignment between the generated motions and texts. Visualization results in Appendix H show that our key design significantly helps capture the motion dynamic clues, especially on sequentiality, directions, and dynamics. We provide more empirical evidence to support the claim. We measure the text-similarity between two input cases. The similarity between "walking in a clockwise circle" and "walking in a counter-clockwise circle": is 0.98 (CLIP) vs.

	Motion-X			GRAB			HumanML3D
	All↓	Body ↓	Hand↓	All ↓	Body ↓	Hand ↓	Body ↓
Vanilla VQ (512)	140.66	92.20	46.45	78.23	38.29	31.48	77.21
Vanilla VQ (1024)	139.33	91.77	46.40	76.01	37.34	29.89	71.34
RVQ (512×2)	110.94	73.97	40.01	62.94	31.12	27.28	**63.05**
H^2VQ (512×2)	**92.97**	**62.34**	**37.20**	**46.74**	**24.33**	**24.59**	-

Table 7: Comparison of the motion reconstruction errors (MPJPE in mm) of different quantization methods on Motion-X, GRAB, and HumanML3D. Our H^2VQ shows significant improvements.

embedding	supervision	FID ↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$			Matching-score ↓	TMA-Matching-score ↓
			Top1 ↑	Top2 ↑	Top3 ↑	Top1 ↑	Top2 ↑	Top3 ↑		
GT		0.002	0.500	0.708	0.814	0.407	0.578	0.673	2.888	0.768
CLIP	✗	**1.086**	0.405	0.588	0.695	0.345	0.490	0.573	3.917	0.844
TMA	✗	1.290	0.416	0.596	0.699	0.395	0.550	0.637	3.950	0.815
TMA	✓	1.174	**0.416**	**0.603**	**0.703**	**0.399**	**0.555**	**0.638**	**3.894**	**0.809**

Table 8: Ablation on a pre-trained text-motion-aligned model for motion generation on Motion-X. Both TMA embedding and text-motion alignment supervision help generate text-aligned motions.

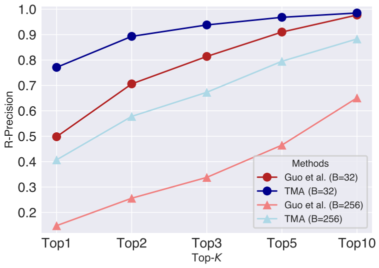


Figure 7: Comparison with existing metrics on Motion-X. Existing evaluation metrics (Guo et al., 2022) are illustrated in red, and ours are in green. The $B = 32$ and $B = 256$ settings for retrieval are denoted as "—●—" and "—▲—" respectively.

0.81 (TMA). The margin shows the sensitivity of TMA in different motion directions.

4.6. Analysis on Evaluation Metrics (RQ4)

We answer RQ4 from two aspects. (1) **Our *TMA-R-Precision*$^{(B)}$ and *TMA-Matching-score*$^{(B)}$ metrics are more accurate than the *R-Precision*$^{(B)}$ and *Matching-score* metrics (Guo et al., 2022)**. As shown in Figure 7, our TMA (in blue) shows stronger retrieval ability than Guo et al. (2022)'s retriever (in red) on both $B = 32$ and $B = 256$ settings. Moreover, Guo et al. (2022)'s retriever shows a larger retrieval gap than TMA when changing $B = 32$ to $B = 256$. Therefore, TMA can evaluate text-motion alignment more accurately than Guo et al. (2022). (2) **$B = 256$ is a more challenging retrieval setting than the $B = 32$ setting.** Retrieving text from motion in 32 text candidates is much easier than 256 candidates. As shown in Figure 7, when changing $B = 32$ to $B = 256$, the retrieval accuracy of both retrievers declines due to the increased size of the retrieval set, which makes the evaluation protocols more challenging. Overall, with a higher upper limit of retrieval capabilities, *TMA-R-Precision*$^{(256)}$ can better evaluate the performance of different methods on text-motion alignment. Additionally, *TMA-R-Precision*$^{(B)}$ and *TMA-Matching-score* are also more accurate and challenging metrics on the body-only dataset (HumanML3D). More details and numerical comparisons are in Appendix I.

4.7. Generalization Ability on Text Descriptions

For the "a man sleeps" cases, as shown in Figure 8, there is no case describing the sleeping motion in the training dataset directly. The successful case is similar to "the person is lying on the ground.", which we think is more reasonable than baselines. The result is mainly because sleeping and lying are similar at the semantic level, due to the good language prior of pre-trained TMA. We additionally provide some videos to verify the generalization ability of different texts in the supplementary video from different aspects: (i) different subject descriptions, like "the guy", and "the woman"; (ii) the robustness of different tenses, like simple or continuous tense. The results show good robustness and generalization ability on diverse or zero-shot texts.

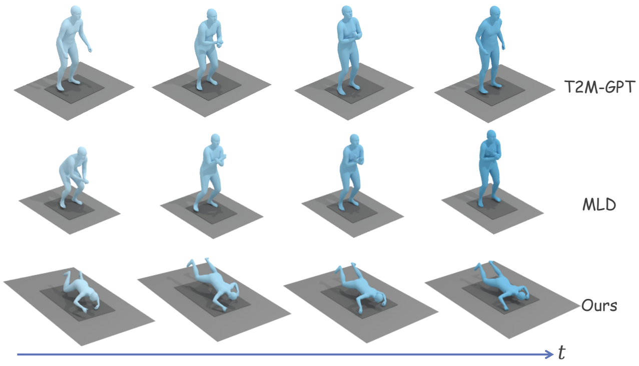


Figure 8: Zero-shot capability comparisons. Given the "a man sleeps", HumanTOMATO generates better text-aligned motions.

5. Conclusion

This work studies the problem of text-driven whole-body motion generation. We carefully clarify the existing challenges in generating vivid text-aligned whole-body motion on motion reconstruction and text-motion alignment. To tackle the challenges, two main technical contributions are proposed: (1) a Holistic Hierarchical VQ-VAE (H^2VQ) and a Hierarchical-GPT for fine-grained body and hand motion reconstruction and generation, and (2) a pre-trained text-motion-alignment model to help generate text-aligned motion. We conduct comprehensive experiments and ablations to verify the superiority and effectiveness of the proposed solution on both Motion-X and HumanML3D datasets. Our experimental results show that HumanTOMATO can generate vivid text-aligned whole-body motion. The limitations are discussed in Appendix K. Our future work mainly focuses on designing more efficient algorithms (Dai et al., 2024a) and scaling the motion-text data pairs via captioning motions (Chen et al., 2024) automatically.

Impact Statement

On the one hand, we explore the whole-body motion generation task and leverage the large-scale whole-body mocap dataset Motion-X to pre-train a motion-text-aligned prior. These could be a foundation for the field-related research community. Besides, based on the motion reconstruction via the proposed discrete latent compression scheme of human motions and large-scale motion data training, the pre-trained HumanTOMATO can provide motion prior, like VPoser (Pavlakos et al., 2019). It can also benefit Motion Capture models (Lin et al., 2023c; Yang et al., 2023) denoising and reducing the impact of noisy annotation. On the other hand, expressive, text-controllable, and high-quality motion generation can be implemented for many practical application scenarios, such as motion generation for games and animations, robotics, and motion interaction. In terms of ethical considerations and potential implications for society, the Motion-X dataset is mainly captured from Internet videos. Consequently, the motions we generate might be a bit similar to those online videos, potentially raising copyright concerns. This work primarily concentrates on developing algorithms and generating motions without portraits, without aiming to discuss these issues in depth.

Acknowledgement

The author team would like to acknowledge Haotian Zheng (HKU), Chenlai Qian (SEU), Jiale Liu (PSU), Wenhao Yang (NJU), and Xiaobo Xia (USYD) for their help in the writing and discussion. We would also like to acknowledge Zeyi Lin and the SwanHub engineering team for providing technical support for the demonstration. The reviews and the Area chairs help use improve the quality of this work significantly. The work is partially supported by National Key R&D Program of China under grant No.2022ZD0116004, by the Young Scientists Fund of the National Natural Science Foundation of China under grant No.62106154, by the Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, and by Shenzhen Science and Technology Program JCYJ20220818103001002 and ZDSYS20211021111415025.

References

Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D., and Chen, B. Skeleton-aware networks for deep motion retargeting. *TOG*, 39(4):62, 2020.

Ahn, H., Ha, T., Choi, Y., Yoo, H., and Oh, S. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, 2018.

Ahuja, C. and Morency, L.-P. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019.

Ao, T., Gao, Q., Lou, Y., Chen, B., and Liu, L. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM TOG*, 41(6): 1–19, 2022.

Barnes, C., Rizvi, S., and Nasrabadi, N. Advances in residual vector quantization: a review. *IEEE TIP*, 5(2):226–262, 1996. doi: 10.1109/83.480761.

Cai, Y., Wang, Y., Zhu, Y., Cham, T.-J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *ICCV*, 2021.

Chen, L.-H., Zhang, J., Li, Y., Pang, Y., Xia, X., and Liu, T. Humanmac: Masked motion completion for human motion prediction. *ICCV*, 2023a.

Chen, L.-H., Lu, S., Zeng, A., Zhang, H., Wang, B., Zhang, R., and Zhang, L. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024.

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

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

Dabral, R., Mughal, M. H., Golyanik, V., and Theobalt, C. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, pp. 9760–9770, 2023.

Dai, W., Chen, L.-H., Wang, J., Liu, J., Dai, B., and Tang, Y. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024a.

Dai, W., Chen, L.-H., Wang, J., Liu, J., Dai, B., and Tang, Y. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024b.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Fan, Y., Lin, Z., Saito, J., Wang, W., and Komura, T. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pp. 18770–18780, 2022.

Floridi, L. and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020.

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

Gower, J. C. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

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

Guo, C., Mu, Y., Javed, M. G., Wang, S., and Cheng, L. Momask: Generative masked modeling of 3d human motions. *arXiv preprint arXiv:2312.00063*, 2023.

Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., Elgharib, M., and Theobalt, C. Learning speech-driven 3d conversational gestures from video. In *ACM IVA*, pp. 101–108, 2021.

Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., and Liu, Z. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM SIGGRAPH*, 2022.

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

Kim, J., Kim, J., and Choi, S. Flame: Free-form language-based motion synthesis & editing. In *AAAI*, volume 37, pp. 8255–8263, 2023.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2013.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, T., Moon, G., and Lee, K. M. Multiact: Long-term 3d human motion generation from multiple action labels. In *AAAI*, volume 37, pp. 1231–1239, 2023.

Li, H., Zhang, S., Li, X., Su, L., Huang, H., Jin, D., Chen, L., Huang, J., and Yoo, J. Detectornet: Transformer-enhanced spatial temporal graph neural network for traffic prediction. In *ACM SIGSPATIAL*, pp. 133–136, 2021a.

Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., and Lu, C. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pp. 3383–3393, 2021b.

Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., and Li, X. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *ICCV*, pp. 10234–10243, 2023a.

Li, S., Zhuang, S., Song, W., Zhang, X., Chen, H., and Hao, A. Sequential texts driven cohesive motions synthesis with natural transitions. In *ICCV*, pp. 9498–9508, October 2023b.

Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., and Chen, C.-w. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *CVPR*, pp. 23222–23231, 2023a.

Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., and Zhang, L. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2023b.

Lin, J., Zeng, A., Wang, H., Zhang, L., and Li, Y. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pp. 21159–21168, 2023c.

Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., and Zheng, B. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, pp. 612–630. Springer, 2022.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 2019.

Liu, Y., Chen, C., and Yi, L. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *ICLR*, 2019.

Lu, Q., Zhang, Y., Lu, M., and Roychowdhury, V. Action-conditioned on-demand motion generation. In *ACM MM*, pp. 2249–2257, 2022.

Lucas, T., Baradel, F., Weinzaepfel, P., and Rogez, G. Posegpt: Quantization-based 3d human motion generation and forecasting. In *ECCV*, pp. 417–435. Springer, 2022.

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

Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., and Richard, A. From audio to photoreal embodiment: Synthesizing humans in conversations. *arXiv preprint arXiv:2401.01885*, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., and Jiang, H. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023.

Petrovich, M., Black, M. J., and Varol, G. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, 2022.

Petrovich, M., Black, M. J., and Varol, G. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. *ICCV*, 2023.

Plappert, M., Mandery, C., and Asfour, T. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. Babel: bodies, action and behavior with english labels. In *CVPR*, 2021.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., and Sheikh, Y. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *ICCV*, pp. 1173–1182, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C., and Liu, Z. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, pp. 11050–11059, 2022.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *NeurIPS*, 33:16857–16867, 2020.

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

Tevet, G., Gordon, B., Hertz, A., Bermano, A. H., and Cohen-Or, D. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022.

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

Tseng, J., Castellon, R., and Liu, K. Edge: Editable dance generation from music. In *CVPR*, pp. 448–458, 2023.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *NeruIPS*, 30, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 30, 2017.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Wang, D., Deng, Y., Yin, Z., Shum, H.-Y., and Wang, B. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, pp. 17979–17989, 2023b.

Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., and Dai, B. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, pp. 20460–20469, 2022a.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2023c.

Wang, Z., Yu, P., Zhao, Y., Zhang, R., Zhou, Y., Yuan, J., and Chen, C. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *AAAI*, 2020.

Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., and Huang, S. Humanise: Language-conditioned human motion generation in 3d scenes. *NuerIPS*, 35:14959–14971, 2022b.

Xie, Y., Jampani, V., Zhong, L., Sun, D., and Jiang, H. Omnicontrol: Control any joint at any time for human motion generation. *ICLR*, 2024.

Xu, S., Wang, Y.-X., and Gui, L.-Y. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, pp. 251–269. Springer, 2022.

Xu, S., Li, Z., Wang, Y.-X., and Gui, L.-Y. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. *ICCV*, 2023a.

Xu, S., Wang, Y.-X., and Gui, L.-Y. Stochastic multi-person 3d motion forecasting. In *ICLR*, 2023b.

Yan, S., Li, Z., Xiong, Y., Yan, H., and Lin, D. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*, 2019.

Yang, J., Zeng, A., Liu, S., Li, F., Zhang, R., and Zhang, L. Explicit box detection unifies end-to-end multi-person pose estimation. In *ICLR*, 2023.

Yao, H., Song, Z., Zhou, Y., Ao, T., Chen, B., and Liu, L. Moconvq: Unified physics-based motion control via scalable discrete representations. *arXiv preprint arXiv:2310.10198*, 2023.

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., and Black, M. J. Generating holistic 3d human motion from speech. In *CVPR*, 2023.

Yu, P., Zhao, Y., Li, C., Yuan, J., and Chen, C. Structure-aware human-action generation. In *ECCV*, 2020.

Yu, Z., Yin, Z., Zhou, D., Wang, D., Wong, F., and Wang, B. Talking head generation with probabilistic audio-to-visual diffusion priors. In *ICCV*, pp. 7645–7655, 2023.

Yuan, Y., Song, J., Iqbal, U., Vahdat, A., and Kautz, J. Physdiff: Physics-guided human motion diffusion model. *ICCV*, 2023.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *TASLP*, 30:495–507, 2021.

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

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

Zhang, Y., Black, M. J., and Tang, S. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020.

Zhao, R., Su, H., and Ji, Q. Bayesian adversarial human motion synthesis. In *CVPR*, 2020.

Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., and Gao, S. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *ICCV*, pp. 20807–20817, October 2023.

Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., and Liu, L. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2023.

Zhou, Z. and Wang, B. Ude: A unified driving engine for human motion generation. In *CVPR*, pp. 5632–5641, 2023.

Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., and Wang, Y. Human motion generation: A survey. *arXiv preprint arXiv:2307.10894*, 2023.

HumanTOMATO: Text-aligned Whole-body Motion Generation

Supplementary Material

Contents

A Related Work **2**
- A.1 Human Motion Generation. 2
- A.2 Text-driven Motion Generation. 2

B Implementation Details **3**
- B.1 Motion Representation 3
- B.2 Implementation Details of Hierarchical Motion VQ-VAE 3
- B.3 Implementation Details of the Hierarchical-GPT. 3
- B.4 Facial Motion Generator 4
- B.5 Compared Facial Generator Baselines 4

C Algorithm Flow of H^{2}VQ and comparison with Residual Vector Quantization **6**
- C.1 Training and Inference of H^{2}VQ 6
- C.2 Comparsion with Residual Vector Quantization (RVQ) 6

D Details about Text-motion Alignment Pre-training **7**
- D.1 Training Details 7
- D.2 Evaluation of the Alignment Model on retrieval tasks 7
- D.3 Retrieval Ability Comparison (TMA *v.s.* TEMOS) 10

E Failure Cases of Baselines **11**

F More Details on Main Results (RQ1) **12**
- F.1 Quantitative Comparison 12
- F.2 Qualitative Comparison 13

G Comparison on Different Vector Quantization Methods (RQ2) **16**
- G.1 Ablation on Different Quantization Methods 16
- G.2 Comparisons on different Codebook Sizes 19

H Text-motion Aligned Model As A Prior (RQ3) **20**
- H.1 Quantitative Results on HumanML3D 20
- H.2 Pre-trained Text-motion Aligned Model as a Prior 20

I Details on the Evaluation Metrics (RQ4) **23**

J Can We Generate Whole-body Motions by Parts Separately? **24**

K Limitation **25**

A. Related Work

A.1. Human Motion Generation.

Generating human motions (Zhu et al., 2023) can be divided into two main types according to inputs: motion synthesis (1) without any conditions (Yan et al., 2019; Zhao et al., 2020; Zhang et al., 2020; Cai et al., 2021) and (2) with some given conditions, such as text, audio, music, and interactive scenes (Ahn et al., 2018; Petrovich et al., 2022; Zhang et al., 2022; Chen et al., 2023b; Guo et al., 2022; Ahuja & Morency, 2019; Ghosh et al., 2021; Zhang et al., 2023; Lee et al., 2023; Yi et al., 2023; Wang et al., 2022a; Zhou & Wang, 2023; Wang et al., 2022b; Xu et al., 2022; Tseng et al., 2023; Siyao et al., 2022; Liu et al., 2022; Xu et al., 2023b; Dabral et al., 2023; Guo et al., 2023; Xie et al., 2024; Zhou et al., 2023; Zhi et al., 2023; Peng et al., 2023; Liu et al., 2023; Dai et al., 2024b). The second type will be more challenging and applicable due to either extracting and understanding motion and conditions or cross-modality alignment. To generate diverse, natural, high-quality human motions, many generative models have been explored by Wang et al. (2020); Hong et al. (2022); Yu et al. (2020); Zhang et al. (2023). Recently, diffusion-based models significantly improved the motion generation performance and diversity (Chen et al., 2023b; Tevet et al., 2023; Zhang et al., 2022; Wang et al., 2023c; Chen et al., 2023a; Xu et al., 2023a; Li et al., 2023a;b) with stable training. However, as human motion is a kind of high-dimensional spatio-temporal signal (Li et al., 2021a), these methods are still hard to tackle the motion data easily. Chen et al. (2023b); Zhang et al. (2023) learn low-dimensional motion latent in an encoding-decoding fashion, like VAE and VQ-VAE, in the first stage. Then, text-aligned motion latent representations could be easier to learn in the second stage. For holistic human motion generation with facial expressions and hand gestures, co-speech expression generation and gesture generation from human speech is also an arising topic in this area (Habibie et al., 2021; Yi et al., 2023). Specifically, TalkSHOW (Yi et al., 2023) takes the first attempt for face, hand, and body motion modeling via separate models since the facial expressions (*e.g.*, lip movement) are strongly correlated with the speech signals (Yu et al., 2023; Wang et al., 2023b), but the body and gesture motions are many-to-many mappings. Bearing the difficulties in jointly modeling the whole-body motions and the lack of whole-body data, there are no existing methods to explore text-driven whole-body motion generation.

A.2. Text-driven Motion Generation.

Text plays an important role in controlling human motion generation since it can describe the actions, directions, and dynamic body-part clues via a natural interaction way. Based on existing action recognition and motion capture datasets (Plappert et al., 2016; Mahmood et al., 2019; Liu et al., 2019; Punnakkal et al., 2021; Guo et al., 2022), text-driven motion generation has achieved rapid progress in recent years. The input text went from the original single-action category to multiple actions and arbitrary natural language (Ahn et al., 2018; Lee et al., 2023; Lu et al., 2022; Petrovich et al., 2022; Kim et al., 2023). The generated motions also range from upper-body motions to full-body motions (additionally with global trajectories and lower-body motions) and from short-time actions to long-term motions (Ahuja & Morency, 2019; Chen et al., 2023b; Zhang et al., 2023). Early attempts (Tevet et al., 2022; Guo et al., 2022) heavily rely on the given motion-text datasets, making the generated motion hard to generalize. For open-vocabulary motion generation, some works try to introduce large-scale pre-trained models (e.g., CLIP (Radford et al., 2021), and LLMs (Floridi & Chiriatti, 2020)) to make the text encoding powerful (Lucas et al., 2022; Jiang et al., 2023; Tevet et al., 2022; Hong et al., 2022; Lin et al., 2023a). However, existing methods suffer from two main issues. First, text-driven holistic motion generation is under-explored, while coherent hand gestures and facial expressions are essential to whole-body motions. Second, the distribution of motion is quite different from images, making CLIP prior weak in text-motion alignment, while LLMs only have textual priors. That is to say, previous efforts have not thoroughly explored motion-text alignment. Accordingly, modeling whole-body motion and exploring how to use motion-text-aligned priors are urgent for the community.

B. Implementation Details

B.1. Motion Representation

The raw motion representation consists of two parts (Aberman et al., 2020), static part (joints offsets) and dynamic part (joint movements) respectively. We further define motion generation tasks as generating diverse and vivid joint movements based on a uniform skeleton. We follow Guo et al. (2022) (*i.e.* H3D-Format) to randomly select a skeleton as a target skeleton, including body and hand joints, and retarget each motion sequence to it. As all motions share the same skeleton, in this way, we set the local joint offsets for all motions to be unchanged. As a pose can be decomposed as twist and swing (Li et al., 2021b), vanilla inverse kinematic (IK) algorithms will ignore the twist rotation, which will lead to the wrong supervision of joint movements. To verify whether rotation regularization helps motion generation and reconstruction, we take motion reconstruction as a pretext task. For motion reconstruction, we take a transformer-based VAE (Chen et al., 2023b) and convolution-based VQ-VAE (Zhang et al., 2023) as the architecture to evaluate the motion reconstruction performance on HumanML3D. As shown in Table 9, motion without rotation information reduces the reconstruction error. Besides, the results in Table 9 show that velocity is beneficial to motion reconstruction.

As discussed in the main paper (Section 4.2), for body-hand motion representations, we take the H3D-Format (Guo et al., 2022) as a basis and expand the body-only representation to holistic body-hand motion representation. Specifically, the *i-th* frame pose is defined by a tuple of root angular velocity ($\dot{r}^a \in \mathbb{R}$) along Y-axis, root linear velocities ($\dot{r}^x, \dot{r}^z \in \mathbb{R}$) on XZ-plane, root height $r^y \in \mathbb{R}$, local joints positions ($\mathbf{j}^p \in \mathbb{R}^{3N-1}$), and velocities ($\mathbf{j}^v \in \mathbb{R}^{3N}$), where N denotes the number of joints. For face motion representations, we follow Flame Format (Kim et al., 2023) and use $\mathbf{f} \in \mathbb{R}^{50}$ to represent the face expression. Thus, we represent the whole-body motion at frame i as $\mathbf{m}_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, \dot{r}^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{f}\}$.

	Input Format	MPJPE	PA-MPJPE	ACCEL.
VAE	H3D-format	49.51	39.47	7.131
	w/o rotation, w/o velocity	51.43	39.47	7.27
	w/o rotation	**46.84**	**36.16**	**6.603**
VQ-VAE	H3D-format	78.24	45.77	8.757
	w/o rotation, w/o velocity	76.34	39.98	8.622
	w/o rotation	**68.86**	**39.97**	**8.274**

Table 9: Ablation study of different motion representations on the Humanml3D dataset.

B.2. Implementation Details of Hierarchical Motion VQ-VAE

We take `Conv1d`$(\cdot)$ with skip connection as the basic module for both the body encoder and the hand encoder and downsample the feature from the body part by $2\times$ and the feature from the hand part by $4\times$, respectively. In detail, at each down-sampled timestamp, the number of body tokens is 2, and the number of hand tokens is 4. Therefore, we predict two body tokens at each down-sampled timestamp. The codebook size for both hand quantizer and body quantizer is set to 512×512. That is to say, $K = 512$ and the dimension of each code is 512. We take the AdamW (Loshchilov & Hutter, 2019) as an optimizer with a fixed learning rate 1×10^{-4}, batch size of 256, and exponential moving constant $\lambda = 0.99$. The α in H^{2}VQ loss $\mathcal{L}$ (Eqn. 6) is set as 0.02. The body-hand decoder upsamples the feature by $2\times$. All upsampling operation in the decoder is the nearest upsampling with a scaling factor of $2\times$. Training the H^{2}VQ takes about 8 hours on $4\times$ NVIDIA Tesla A100-80GB GPUs.

B.3. Implementation Details of the Hierarchical-GPT.

We employ 18 transformer layers with a dimension of 1024 and 16 heads. Since the design of different downsampling rates between two codebooks, we simply concat the tokens from pre-trained H^{2}VQ stage and set the maximum length of the code index sequence as 149.

Training. We combine the tokens from the hand codebook $\mathcal{C}_1$ and the body codebook $\mathcal{C}_2$ and feed them into the transformer, in which we employ a causal mask with $mask_{i,j} = -\infty \times \mathbf{1}(i < j) + \mathbf{1}(i \geq j)$, where $\mathbf{1}(\cdot)$ is the indicator function, to prevent information leakage from the following tokens. We employ the CLIP-ViT-L-14 model and pre-trained TMA Text encoder as the text encoder to encode the text, respectively, and freeze them in training. All the trainings are conducted on $4\times$ NVIDIA Tesla A100-80GB GPUs and cost 60 hours.

Inference. When performing inference, we feed the text encoder with raw texts and get the text embedding. Our Hierarchical-GPT predicts motion tokens in an auto-regressive fashion with the start token of text embedding. All our tests and inferences are conducted on $1\times$ NVIDIA Tesla A100-80GB GPU.

B.4. Facial Motion Generator

As discussed in the main paper (Section 3.3), similar to the body-hand motion generation pipeline, we take a two-stage strategy to generate facial motions. In the first stage, extended from our hierarchical modeling of body-hand motion, we train a Residual-VQVAE (RVQ) for facial motion compression and reconstruction. As the facial motion cannot be disentangled as body and hand motions for H^{2}VQ, we take the 2-layer RVQ as our quantizer. In the second stage, we learn a facial GPT for facial motion token prediction. For each down-sampling timestamp, we predict the first-level tokens first and then the second level, which is a kind of coarse-to-fine generation fashion (Wang et al., 2023a). With the encoded facial motion $\mathbf{m}^F = [\mathbf{m}_1^F, \mathbf{m}_2^F, \cdots, \mathbf{m}_L^F]$, we have corresponding code indices, denoted as $\mathrm{I}^1 = [\mathrm{I}_1^1, I_2^1, \cdots, \mathrm{I}_{L/r}^1]$ and $\mathrm{I}^2 = [\mathrm{I}_1^2, \mathrm{I}_2^2, \cdots, \mathrm{I}_{L/r}^2]$, where $r = 4$ denotes the down-sampling rate, which is used to convert the input motion sequence to discrete motion tokens. Note that all superscripts refer to RVQ levels. Therefore, as shown in Figure 2(b), the code indices prediction can be formulated as an auto-regressive prediction problem:

$$\begin{aligned} P(\mathrm{I}_{1,2,\cdots,L/r}^{F-\mathtt{i},F-\mathtt{ii}} \mid \mathbf{t}) &= \prod_{s=1}^{L/r} P(\mathrm{I}_s^{F-\mathtt{i},F-\mathtt{ii}} \mid \mathrm{I}_{<s}^{F-\mathtt{i},F-\mathtt{ii}}, \mathbf{t}) \\ &= \prod_{s=1}^{L/r} P(\mathrm{I}_s \mid \mathrm{I}_{<s}^{F-\mathtt{i},F-\mathtt{ii}}, \mathbf{t}) \cdot P(\mathrm{I}_s^{F-\mathtt{ii}} \mid \mathrm{I}_s^{F-\mathtt{i}}, \mathrm{I}_{<s}^{F-\mathtt{i},F-\mathtt{ii}}, \mathbf{t}), \end{aligned} \tag{7}$$

where we first predict the first level token index and then predict the second level at each down-sampled timestamp s.

B.5. Compared Facial Generator Baselines

We introduce VAE-based and diffusion-based facial motion generation baselines for compassion with our facial motion generator. The details are as follows.

B.5.1. Facial cVAE

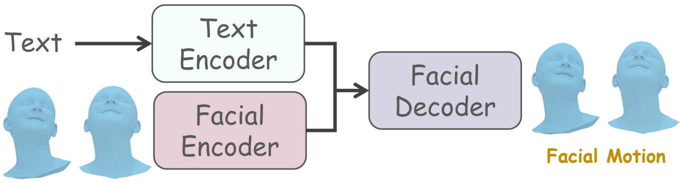


Figure 9: Facial cVAE Motion Generator.

We take a text-conditioned facial VAE (cVAE) (Petrovich et al., 2022) as a comparison. As shown in Figure 9, the facial VAE consists of three components. (1) A facial encoder. The Facial is a 6-layer transformer. The input facial motion is concatenated with a μ_F token and a Σ_F token. (2) A text encoder. The Facial is composed of a pre-trained DistllBERT (Sanh et al., 2019) and a 6-layer transformer. The input DistillBERT feature is concatenated with a μ_T token and a Σ_T token. (3) A facial decoder. The 6-layer transformer-based facial decoder generates facial motions from the z_F or z_T vector, which can be sampled from Gaussian distribution $\mathcal{N}(\mu_F, \Sigma_F)$ or $\mathcal{N}(\mu_T, \Sigma_T)$ via re-parameterizing trick (Kingma & Welling, 2013). The training loss consists of three components. (1) facial motion reconstruction loss:

$$\mathcal{L}_{rec} = \mathtt{SmoothL1}(\mathbf{m}^F, \hat{\mathbf{m}}^F),$$

where $\mathbf{m}^F, \hat{\mathbf{m}}^F$ are facial motions and reconstructed facial motions and $\mathtt{SmoothL1}(\cdot)$ is the SmoothL1-Loss. (2) KL Loss:

$$\begin{aligned} \mathcal{L}_{KL} =& \mathtt{KL}(\mathcal{N}(\mu_F, \Sigma_F), \mathcal{N}(\mathbf{0}, \mathbf{I})) + \mathtt{KL}(\mathcal{N}(\mu_T, \Sigma_T), \mathcal{N}(\mathbf{0}, \mathbf{I})) \\ &+ \mathtt{KL}(\mathcal{N}(\mu_F, \Sigma_F), \mathcal{N}(\mu_T, \Sigma_T)) + \mathtt{KL}(\mathcal{N}(\mu_T, \Sigma_T), \mathcal{N}(\mu_F, \Sigma_F)), \end{aligned}$$

where $\texttt{KL}(\cdot)$ is the Kullback-Leibler divergence function and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian distribution. (3) Cross-modal embedding similarity loss:

$$\mathcal{L}_E = \texttt{SmoothL1}(z_F, z_T).$$

The overall training loss is $\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_E$, where $\lambda_1 = \lambda_2 = 1 \times 10^{-5}$. In the inference stage, the text encoder encodes text embedding z_T first and then feeds it into the facial decoder to obtain the facial motions.

B.5.2. Diffusion-based Facial Motion Genrator

We take a text-conditioned facial motion diffusion model (Tevet et al., 2023) as a comparison. In this part, we simplify the $x = \mathbf{m}^F$ to represent facial motions. As shown in Figure 10, the diffusion model takes a text embedding as a condition and then concatenates the embedding and motion linear embedding x_t at timestamp t together to the transformer encoder. In each iteration, the output of the transformer encoder will be projected by a linear layer to predict the x_0. Note that the diffusion model here is used to predict x_0 but not noise. After $T = 1000$ step denoising, the model will return the generated facial motion. All parameter settings follow Tevet et al. (2023).

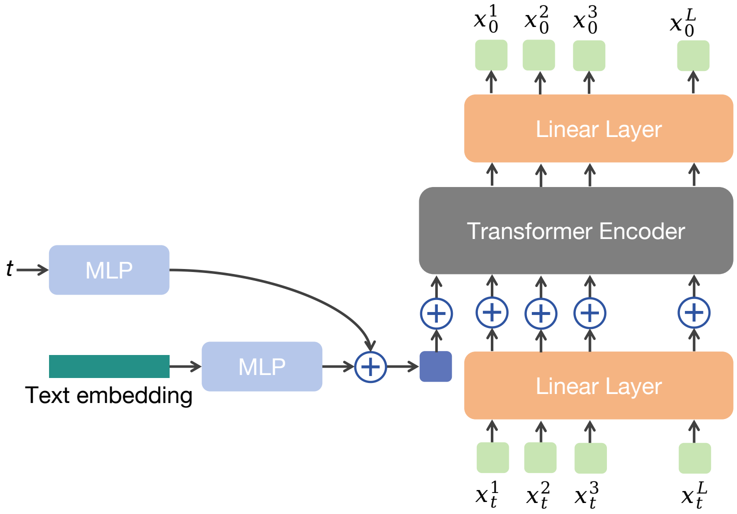


Figure 10: Diffusion-based Facial Motion Generator.

C. Algorithm Flow of H^{2}VQ and comparison with Residual Vector Quantization

C.1. Training and Inference of H^{2}VQ

In the main paper, we introduce the training and inference details in Section 3.2. For reading convenience, we provide the training and inference procedure of our Holistic Hierarchical VQ-VAE (H^{2}VQ-VAE) in Algorithm 1 and Algorithm 2.

Algorithm 1: Training procedure of Holistic Hierarchical VQ-VAE (H^{2}VQ-VAE)

```
Input: The initialized hand codebook $\mathcal{C}_1$, body codebook $\mathcal{C}_2$, hand quantizer $\mathcal{Q}^H(\cdot;\mathcal{C}^H)$, body quantizer $\mathcal{Q}^B(\cdot;\mathcal{C}^B)$
    ($|\mathcal{C}^H| = |\mathcal{C}^B| = K$), H$^2$VQ-VAE, the input motion $\mathbf{m}$, the optimization iterations $I_{max}$.
Output: The optimized H$^2$VQ-VAE network $\Theta$, codebooks $\mathcal{C}^H, \mathcal{C}^B$.
for $I = 0, 1, \ldots, I_{\max}$ do
    $\mathbf{z}^H = \texttt{Enc}_H(\mathbf{m}^H)$;
    $\mathbf{z}^B = \texttt{Enc}_B(\mathbf{m}^B)$;
    $\hat{\mathbf{z}}^H = \mathcal{Q}^H(\mathbf{z}^H;\mathcal{C}^H)$;
    $\hat{\mathbf{z}}^B = \mathcal{Q}^B(\texttt{Conv1d}(\texttt{Concat}(\texttt{Transform}(\hat{\mathbf{z}}^H), \mathbf{z}^B));\mathcal{C}^B)$;
    $\hat{\mathbf{m}} = \texttt{Dec}(\hat{\mathbf{z}}^B, \hat{\mathbf{z}}^H)$;
    $\Theta = \Theta - \nabla_\Theta \|\mathbf{m} - \texttt{Dec}(\hat{\mathbf{m}})\|_2^2 + \alpha\left(\|\mathbf{z}^H - \texttt{sg}(\hat{\mathbf{z}}^H)\|_2^2 + \|\mathbf{z}^B - \texttt{sg}(\hat{\mathbf{z}}^B)\|_2^2\right)$;
    Optimize two codebooks $\mathcal{C}^H, \mathcal{C}^B$ via EMA and Code Reset;
return H$^2$VQ-VAE network $\Theta$.
```

Algorithm 2: Inference procedure of Holistic Hierarchical VQ-VAE (H^{2}VQ-VAE)

```
Input: The pre-trained H$^2$VQ-VAE network $\Theta$, body and hand code indices sequence $\mathrm{I}^B = [\mathrm{I}_1^B, I_2^B, \cdots, \mathrm{I}_{L/r}^B, \texttt{End}]$
    and $\mathrm{I}^H = [\mathrm{I}_1^H, \mathrm{I}_2^H, \cdots, \mathrm{I}_{L/r}^H, \texttt{End}]$, codebook $\mathcal{C}^H = \{k, \mathbf{e}_1(k)\}_{k\in[K]}, \mathcal{C}^B = \{k, \mathbf{e}_2(k)\}_{k\in[K]}$.
Output: the noise prediction network $\boldsymbol{\epsilon_\theta}$.
$\hat{\mathbf{z}}^H \leftarrow$ Query $\mathcal{C}^H$ with $\mathrm{I}^H$;
$\hat{\mathbf{z}}^B \leftarrow$ Query $\mathcal{C}^B$ with $\mathrm{I}^B$;
return motion $\texttt{Dec}(\hat{\mathbf{z}}^B, \hat{\mathbf{z}}^H)$.
```

C.2. Comparsion with Residual Vector Quantization (RVQ)

As can be seen in Appendix C.1, our H^{2}VQ consists of two codebooks $\mathcal{C}^H$ and $\mathcal{C}^B$ with size K. The intuitive design insight is that the space of our code combination is $\mathcal{O}(K^2)$. However, scaling the size of the codebook to $2K$ only has the vector space of size $\mathcal{O}(K)$. Therefore, our H^{2}VQ enjoys the scaling of latent code size with low memory cost. An alternative way to scale the codebook size efficiently is the 2-level Residual Vector Quantization (RVQ) technique. As shown in Algorithm 3, RVQ quantized the residual error vectors recurrently in each level, which is also a hierarchical modeling strategy. However, RVQ does not model the hand and body motions explicitly, which makes it cannot reconstruct the whole-body motions better than H^{2}VQ. For more details, please refer to Zeghidour et al. (2021). The experimental comparisons are in Appendix G.

Algorithm 3: Residual Vector Quantization (RVQ)

```
Input: The output of the encoder $\mathbf{z} = \texttt{Enc}(\mathbf{m})$, $N_q$-level quantizers $\mathcal{Q}_i(\cdot)$ ($i = 1, 2, \cdots, N_q$).
Output: Quantized vector $\hat{\mathbf{z}}$.
$\hat{\mathbf{z}} = 0$;
$\mathbf{res} = \mathbf{z}$;
for $i = 1, 2, \ldots, N_q$ do
    $\hat{\mathbf{z}} += \mathcal{Q}_i(\mathbf{res})$;
    $\mathbf{res} -= \mathcal{Q}_i(\mathbf{res})$;
return $\hat{\mathbf{z}}$.
```

D. Details about Text-motion Alignment Pre-training

In this section, we will detail the training details of the TMA model and evaluate our pre-trained alignment model. Our trained TMA model demo is in the supplementary video.

D.1. Training Details

Here, we detail the training procedure on how to train a text-whole-body-motion alignment model. Recall a text-to-motion model, TEMOS (Petrovich et al., 2022), the VAE-based architecture consists of a motion encode, a text encoder, and a motion decoder. The training objective in TEMOS is the weighted sum of $\mathcal{L}_T = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_E\mathcal{L}_E$, where the three loss items are reconstruction loss, Kullback-Leibler (KL) divergence loss, and cross-modal embedding similarity loss respectively. Additionally, like Petrovich et al. (2023), we introduce an InfoNCE (Oord et al., 2018) loss term $\mathcal{L}_{NCE}$ into the optimization objective for learning text-motion-aligned representations. The InfoNCE loss aims to align pairwise text-motion embeddings and pull the negative motion-text pairs in the batch away like Radford et al. (2021). Therefore, the final training objective is

$$\min \mathcal{L}_T + \lambda_{NCE}\mathcal{L}_{NCE},$$

where all hyper-parameters are $\lambda_{KL} = 1 \times 10^{-5}, \lambda_E = 1 \times 10^{-5}, \lambda_{NCE} = 1 \times 10^{-1}$ respectively.

Note that, in a batch, different motion samples might be similar or even repetitive. Therefore, we will filter the similar negative samples in the InfoNCE loss. In other words, two motions with similar text descriptions (similarity higher than 0.85) will not be treated as negative samples. Technically, a pre-trained language model will calculate the similarity between two text descriptions $s_{i,j} = \langle \mathbf{t}_i, \mathbf{t}_j \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. Different from Petrovich et al. (2023) choosing MPNet[1] (Song et al., 2020) as the pre-trained language model, we take the Sentence-BERT (*aka* sBERT[2]) (Reimers & Gurevych, 2019) as the pre-trained language model, which is more accurate than MPNet.

To compare the accuracy of evaluating the similarity among sentences, we present a case study composed of 10 sentence samples in Example 1.

Example 1. Here, we present the 10 sentence samples used for evaluating sBERT and MPNet.

```
[
    0: ‘A human walking backwards.’,
    1: ‘A person is walking backwards.’,
    2: ‘Someone walks in a circle counterclockwise’,
    3: ‘A person walks a full counter-clockwise circle.’,
    4: ‘A human performs a tight 90° curve to the right.’,
    5: ‘A person walks a quarter circle clockwise with 4 steps.’,
    6: ‘human goes backwards starting with left’,
    7: ‘A person walks backwards.’,
    8: ‘a person walks in a circle to the left side.’,
    9: ‘trump’
]
```

We calculate the cosine similarity of these 10 sentences with sBERT and MPNet, respectively. As shown in Figure 11, sBERT reflects the sentence similarity more accurately than MPNet. For two sentences with very similar semantics, like `‘A human walking backwards.’` and `‘A person is walking backwards.’`, the similarity provided by sBERT is 0.958, while MPNet is 0.893. For two sentences completely unrelated, like `‘A human walking backwards.’` and `‘trump’`, the similarity provided by sBERT is 0.132, while MPNet is 0.758. In this case, the `‘trump’` example is not a motion description. sBERT clearly distinguishes it from other sentences, but MPNet cannot distinguish them significantly. Therefore, the sBERT is more discriminative than MPNet in negative filtering.

D.2. Evaluation of the Alignment Model on retrieval tasks

We take the Recall@K as the main metric to evaluate the retrieval performance to evaluate the performance of the TMA model. We evaluate both motion-to-text (M2T) and text-to-motion (T2M) retrieval performance with four main protocols. (A) *Retrieving in the full test test.* (B) *Retrieving in the full test test with a sBERT-score threshold (set ϵ as 0.9).* As some sentences have similar semantics, like “`A man is walking straight.`” and “`The person walks`

[1] https://huggingface.co/microsoft/mpnet-base.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

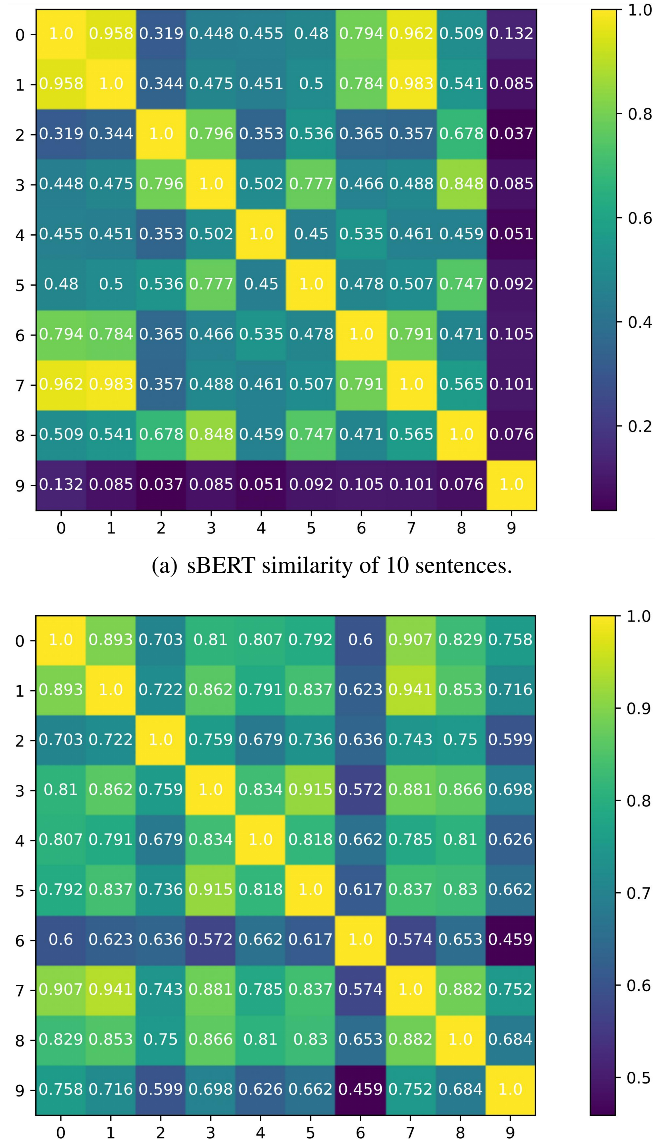
`forward.`”, we treat these retrieval results as positive targets if the retrieved text has a sBERT similarity higher than $\epsilon = 0.9$ with GT text. (C) *Retrieving in the* 256*-size sub-test set.* The 256-size retrieving set consists of one GT result and 255 negative results. (D) *Retrieving in the* 32*-size sub-test set.* Similar to Protocol C, the 32-size retrieving set consists of one GT result and 31 negative results. The T2M and M2T retrieval evaluation results on Motion-X are shown in Table 10. The T2M and M2T retrieval evaluation results on HumanML3D are shown in Table 11. The comparison with other text-motion retrieval methods on the protocol (C) and protocol (D) is shown in Appendix I. **The retrieval web demo on both body-only and whole-body datasets is shown in supplementary and will be public.**

	T2M					M2T				
	Recall@1	Recall@2	Recall@3	Recall@5	Recall@10	Recall@1	Recall@2	Recall@3	Recall@5	Recall@10
Protocol A	0.051	0.098	0.131	0.192	0.301	0.066	0.118	0.163	0.233	0.350
Protocol B	0.089	0.152	0.194	0.273	0.401	0.169	0.205	0.238	0.298	0.395
Protocol C	0.445	0.609	0.700	0.799	0.883	0.407	0.578	0.673	0.795	0.883
Protocol D	0.716	0.854	0.907	0.946	0.977	0.771	0.893	0.938	0.968	0.985

Table 10: Recall@K (T2M and M2T) of GT motions and texts on the Motion-X dataset.

	T2M					M2T				
	Recall@1	Recall@2	Recall@3	Recall@5	Recall@10	Recall@1	Recall@2	Recall@3	Recall@5	Recall@10
Protocol A	0.065	0.117	0.155	0.227	0.339	0.057	0.106	0.144	0.205	0.322
Protocol B	0.204	0.282	0.326	0.404	0.510	0.102	0.151	0.199	0.263	0.373
Protocol C	0.359	0.523	0.630	0.729	0.842	0.365	0.527	0.625	0.731	0.838
Protocol D	0.774	0.896	0.937	0.968	0.985	0.711	0.853	0.905	0.947	0.977

Table 11: Recall@K (T2M and M2T) of GT motions and texts on the HumanML3D dataset.



	0	1	2	3	4	5	6	7	8	9
0	1.0	0.958	0.319	0.448	0.455	0.48	0.794	0.962	0.509	0.132
1	0.958	1.0	0.344	0.475	0.451	0.5	0.784	0.983	0.541	0.085
2	0.319	0.344	1.0	0.796	0.353	0.536	0.365	0.357	0.678	0.037
3	0.448	0.475	0.796	1.0	0.502	0.777	0.466	0.488	0.848	0.085
4	0.455	0.451	0.353	0.502	1.0	0.45	0.535	0.461	0.459	0.051
5	0.48	0.5	0.536	0.777	0.45	1.0	0.478	0.507	0.747	0.092
6	0.794	0.784	0.365	0.466	0.535	0.478	1.0	0.791	0.471	0.105
7	0.962	0.983	0.357	0.488	0.461	0.507	0.791	1.0	0.565	0.101
8	0.509	0.541	0.678	0.848	0.459	0.747	0.471	0.565	1.0	0.076
9	0.132	0.085	0.037	0.085	0.051	0.092	0.105	0.101	0.076	1.0

(a) sBERT similarity of 10 sentences.

	0	1	2	3	4	5	6	7	8	9
0	1.0	0.893	0.703	0.81	0.807	0.792	0.6	0.907	0.829	0.758
1	0.893	1.0	0.722	0.862	0.791	0.837	0.623	0.941	0.853	0.716
2	0.703	0.722	1.0	0.759	0.679	0.736	0.636	0.743	0.75	0.599
3	0.81	0.862	0.759	1.0	0.834	0.915	0.572	0.881	0.866	0.698
4	0.807	0.791	0.679	0.834	1.0	0.818	0.662	0.785	0.81	0.626
5	0.792	0.837	0.736	0.915	0.818	1.0	0.617	0.837	0.83	0.662
6	0.6	0.623	0.636	0.572	0.662	0.617	1.0	0.574	0.653	0.459
7	0.907	0.941	0.743	0.881	0.785	0.837	0.574	1.0	0.882	0.752
8	0.829	0.853	0.75	0.866	0.81	0.83	0.653	0.882	1.0	0.684
9	0.758	0.716	0.599	0.698	0.626	0.662	0.459	0.752	0.684	1.0

(b) MPNet similarity of 10 sentences.

Figure 11: Sentences similarity comparison between the sBERT and MPNet .

D.3. Retrieval Ability Comparison (TMA *v.s.* TEMOS)

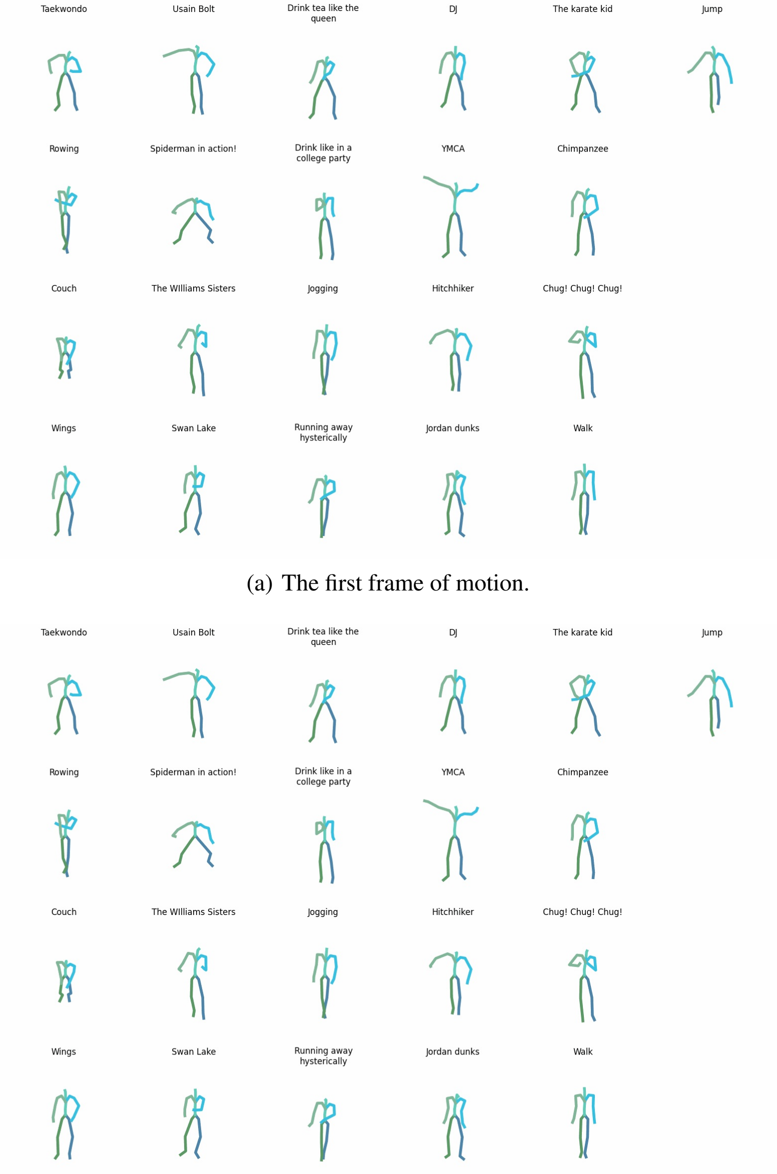
To verify the good alignment of TMA, we compare its retrieval ability with TEMOS Petrovich et al. (2022). As shown in Table 12, the TMA enjoys a good alignment between texts and motions by the contrastive training objective, which makes it with a larger margin than TEMOS in retrieval. This good retrieval ability provides a better alignment of two modalities, and provide a better motion-text alignment for motion generation.

Protocol	Model	T2M					M2T				
		Recall@1	Recall@2	Recall@3	Recall@5	Recall@10	Recall@1	Recall@2	Recall@3	Recall@5	Recall@10
A	TEMOS	0.034	0.062	0.082	0.117	0.178	0.034	0.067	0.096	0.137	0.204
A	TMA	**0.051**	**0.098**	**0.131**	**0.192**	**0.301**	**0.066**	**0.118**	**0.163**	**0.233**	**0.350**
B	TEMOS	0.115	0.163	0.190	0.237	0.308	0.112	0.135	0.155	0.190	0.246
B	TMA	**0.089**	**0.152**	**0.194**	**0.273**	**0.401**	**0.169**	**0.205**	**0.238**	**0.298**	**0.395**
C	TEMOS	0.233	0.341	0.412	0.492	0.601	0.263	0.360	0.426	0.504	0.614
C	TMA	**0.445**	**0.609**	**0.700**	**0.799**	**0.883**	**0.407**	**0.578**	**0.673**	**0.795**	**0.883**
D	TEMOS	0.502	0.641	0.717	0.802	0.897	0.528	0.654	0.725	0.807	0.907
D	TMA	**0.716**	**0.854**	**0.907**	**0.946**	**0.977**	**0.771**	**0.893**	**0.938**	**0.968**	**0.985**

Table 12: The Recall@K (T2M and M2T) of GT motions and texts on the Motion-X dataset (TMA *v.s.* TEMOS).

E. Failure Cases of Baselines

As discussed in Section 3.5, previous methods shown in Figure 4 will fail in some scenarios. We discuss the fashion of "Supervision by an image-text aligned prior explicitly" (Figure 4a) here. As there was no strong text-motion-aligned pre-trained model, MotionCLIP (Tevet et al., 2022) renders the generated motions as images and then supervises the alignment between text embeddings and image embeddings with the CLIP model. This supervision will cause the generated motion to be *over-smoothing*, even *stillness*. We show some over-smoothing cases[3,4] of MotionCLIP on here. As shown in Figure 12, there is almost no change at all between the first frame (Figure 12(a)) and the final frame (Figure 12(b)) of motion.



(a) The first frame of motion.

(b) The final frame of motion.

Figure 12: Visualization of MotionCLIP generated results. The first frame and the final frame of motions are shown in the figure.

[3] https://github.com/GuyTevet/MotionCLIP/issues/5.

[4] https://github.com/GuyTevet/MotionCLIP/issues/15.

F. More Details on Main Results (RQ1)

F.1. Quantitative Comparison

In the main paper, we report the metrics of our HumanTOMATO and related baselines in Table 1. We repeat the evaluation 5 times and report the $\texttt{mean}^{\pm \texttt{std}}$ results in Table 13 and Table 14. Experimental results show our strength than baseline on generation quality, text-motion alignment, and diversity.

	FID↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$		
		Top1↑	Top2↑	Top3↑	Top1↑	Top2↑	Top3↑
GT	-	0.500$^{\pm 0.002}$	0.708$^{\pm 0.002}$	0.814$^{\pm 0.002}$	0.407$^{\pm 0.003}$	0.578$^{\pm 0.004}$	0.673$^{\pm 0.003}$
TEMOS	9.147$^{\pm 0.002}$	0.279$^{\pm 0.001}$	0.442$^{\pm 0.005}$	0.555$^{\pm 0.001}$	0.258$^{\pm 0.000}$	0.389$^{\pm 0.002}$	0.444 $^{\pm 0.000}$
T2M-GPT	1.366$^{\pm 0.059}$	0.368$^{\pm 0.005}$	0.553$^{\pm 0.003}$	0.655$^{\pm 0.007}$	0.310$^{\pm 0.001}$	0.446$^{\pm 0.007}$	0.527$^{\pm 0.014}$
MotionDiffuse	**1.129**$^{\pm 0.034}$	0.391$^{\pm 0.024}$	0.587$^{\pm 0.023}$	0.695$^{\pm 0.016}$	0.368$^{\pm 0.003}$	0.493$^{\pm 0.002}$	0.584$^{\pm 0.009}$
MDM	3.800$^{\pm 0.020}$	0.352$^{\pm 0.003}$	0.547$^{\pm 0.002}$	0.634$^{\pm 0.004}$	0.310$^{\pm 0.004}$	0.430$^{\pm 0.007}$	0.530$^{\pm 0.014}$
MLD	3.407$^{\pm 0.020}$	0.385$^{\pm 0.002}$	0.571$^{\pm 0.001}$	0.683$^{\pm 0.001}$	0.333$^{\pm 0.004}$	0.477$^{\pm 0.003}$	0.561$^{\pm 0.001}$
HumanTOMATO	1.174$^{\pm 0.015}$	**0.416**$^{\pm 0.009}$	**0.603**$^{\pm 0.007}$	**0.703**$^{\pm 0.007}$	**0.399**$^{\pm 0.000}$	**0.555**$^{\pm 0.005}$	**0.638**$^{\pm 0.004}$

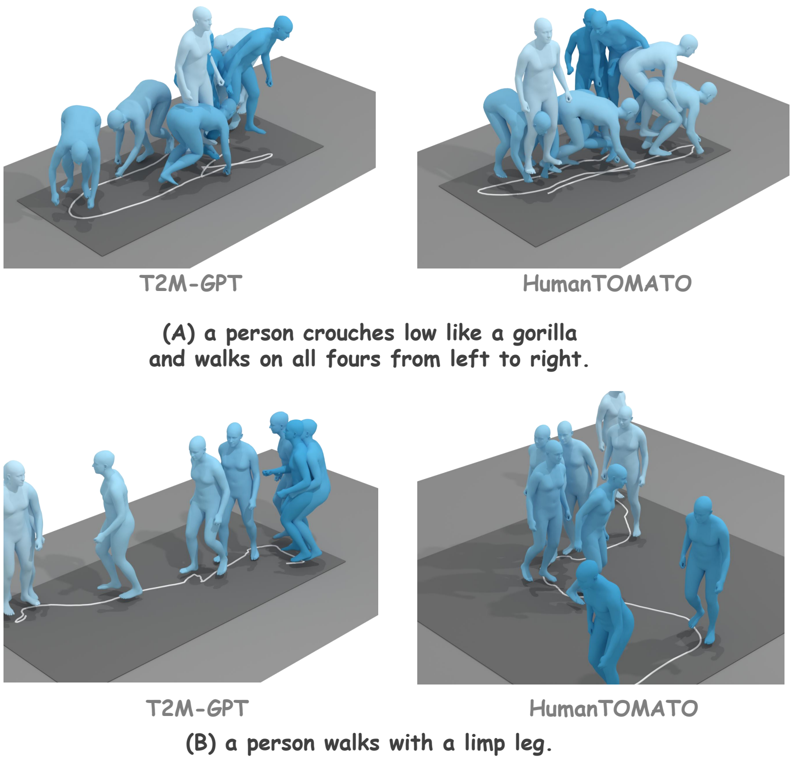
Table 13: Quantitative Comparison on the Motion-X dataset (FID, TMA-R-Precision$^{(256)}$, and R-Precision$^{(32)}$ metrics).

	Matching Score↓	TMA-Matching Score↓	MModality↑	Diversity↑
GT	2.888 $^{\pm 0.006}$	0.768$^{\pm 0.000}$	-	11.087 $^{\pm 0.271}$
TEMOS	5.482$^{\pm 0.008}$	0.928$^{\pm 0.003}$	1.195$^{\pm 0.045}$	9.764$^{\pm 0.239}$
T2M-GPT	4.316$^{\pm 0.053}$	0.881$^{\pm 0.004}$	2.356$^{\pm 0.093}$	10.753$^{\pm 0.063}$
MotionDiffuse	3.950$^{\pm 0.035}$	0.829$^{\pm 0.003}$	1.654$^{\pm 0.071}$	10.580$^{\pm 0.170}$
MDM	4.050$^{\pm 0.023}$	0.840$^{\pm 0.004}$	**2.530**$^{\pm 0.041}$	**11.400**$^{\pm 0.370}$
MLD	3.901$^{\pm 0.011}$	0.883 $^{\pm 0.002}$	2.448$^{\pm 0.034}$	10.420$^{\pm 0.234}$
HumanTOMATO	**3.894**$^{\pm 0.008}$	**0.809**$^{\pm 0.002}$	1.732$^{\pm 0.194}$	10.812$^{\pm 0.034}$

Table 14: Quantitative Comparison on the Motion-X dataset (Matching Score, TMA-Matching Score, and MModality metrics).

F.2. Qualitative Comparison

In the main paper, we compare our method with baseline methods with key-frame sequence visualization. We provide more comparison in Figure 13. In Figure F, the lighter colors represent earlier snapshots. As can be seen, T2M-GPT lacks temporal sensitivity and will generate motions that do not match the text description. In contrast, our method will enjoy these scenarios well and generate vivid motions well aligned with texts.

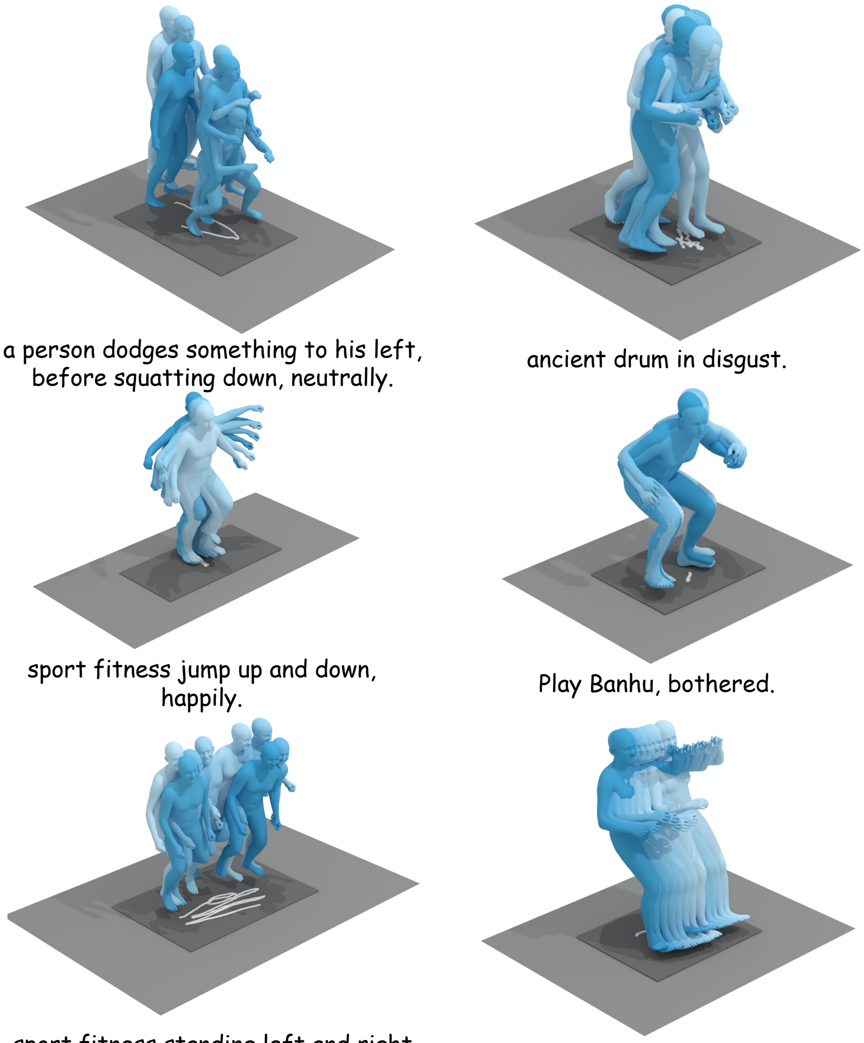


(A) a person crouches low like a gorilla and walks on all fours from left to right.

(B) a person walks with a limp leg.

Figure 13: Qualitative comparison with T2M-GPT.

Additionally, we visualize more generated results of HumanTOMATO in Figure 14 and Figure 15, which show our good generation performance.



a person dodges something to his left, before squatting down, neutrally.

ancient drum in disgust.

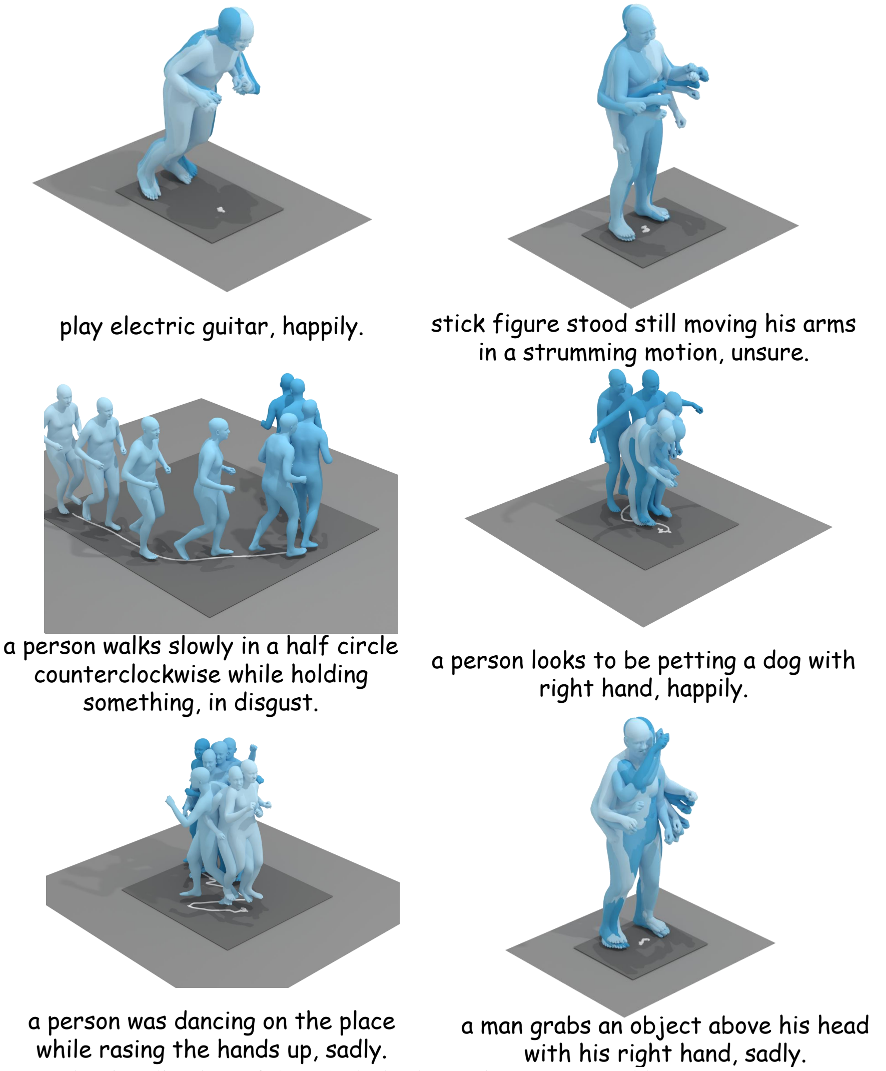
sport fitness jump up and down, happily.

Play Banhu, bothered.

sport fitness standing left and right leg swing, happily.

Play Big ruan,sadly.

Figure 14: Visualization of the whole-body motions generated by HumanTOMATO (1).



play electric guitar, happily.

stick figure stood still moving his arms in a strumming motion, unsure.

a person walks slowly in a half circle counterclockwise while holding something, in disgust.

a person looks to be petting a dog with right hand, happily.

a person was dancing on the place while rasing the hands up, sadly.

a man grabs an object above his head with his right hand, sadly.

Figure 15: Visualization of the whole-body motions generated by HumanTOMATO (2).

G. Comparison on Different Vector Quantization Methods (RQ2)

G.1. Ablation on Different Quantization Methods

In the main paper (Section 4.4), we report the MPJPE for evaluating the reconstruction error of Vanilla VQ, RVQ, and H^{2}VQ respectively. Although HumanML3D only includes body-part motions, we compare the Vinilla VQ-VAE with the RVQ technique to verify our motivation on hierarchical motion modeling, whose results are shown in Table 17. Additionally, as shown in Table 15 and Table 16, we provide more evaluation metrics on PA-MPJPE and Acceleration error (Accel.) (Gower, 1975; Lin et al., 2023b; Chen et al., 2023b). to evaluate the reconstruction quality. Evaluation results show that naïvely increasing the codebook size is almost in vain, and hierarchical modeling is effective for action modeling. Besides, our H^{2}VQ is a better design on whole-body motions than RVQ.

	MPJPE			PA-MPJPE			Accel.		
	All↓	Body ↓	Hand↓	All ↓	Body ↓	Hand ↓	All ↓	Body ↓	Hand↓
Vanilla VQ (512)	140.66	92.20	46.45	58.23	47.72	17.03	23.73	19.99	26.46
Vanilla VQ (1024)	139.33	91.77	46.40	57.30	46.79	17.01	23.54	19.71	26.35
RVQ	110.94	73.97	40.01	40.63	35.84	14.46	21.22	17.76	23.75
H^{2}VQ	**92.97**	**62.34**	**37.20**	**34.21**	**30.76**	**14.05**	**18.95**	**16.53**	**20.72**

Table 15: Different vector quantization methods on Motion-X.

	MPJPE			PA-MPJPE			Accel.		
	All↓	Body ↓	Hand↓	All ↓	Body ↓	Hand ↓	All ↓	Body ↓	Hand↓
Vanilla VQ (512)	78.23	38.29	31.48	35.32	21.75	14.51	11.01	7.32	13.71
Vanilla VQ (1024)	76.01	37.34	29.89	33.42	20.92	14.14	10.70	7.23	13.25
RVQ	62.94	31.12	27.28	25.61	15.96	**13.06**	**8.80**	6.67	**10.37**
H^{2}VQ	**46.74**	**24.33**	**24.59**	**22.00**	**13.95**	13.48	10.11	**6.05**	13.09

Table 16: Different vector quantization methods on GRAB.

	MPJPE (Body)↓	PA-MPJPE (Body) ↓	Accel. (Body)↓
Vanilla VQ (512)	77.209	45.53	8.36
Vanilla VQ (1024)	71.34	40.75	7.59
RVQ	**63.05**	**30.99**	**6.46**

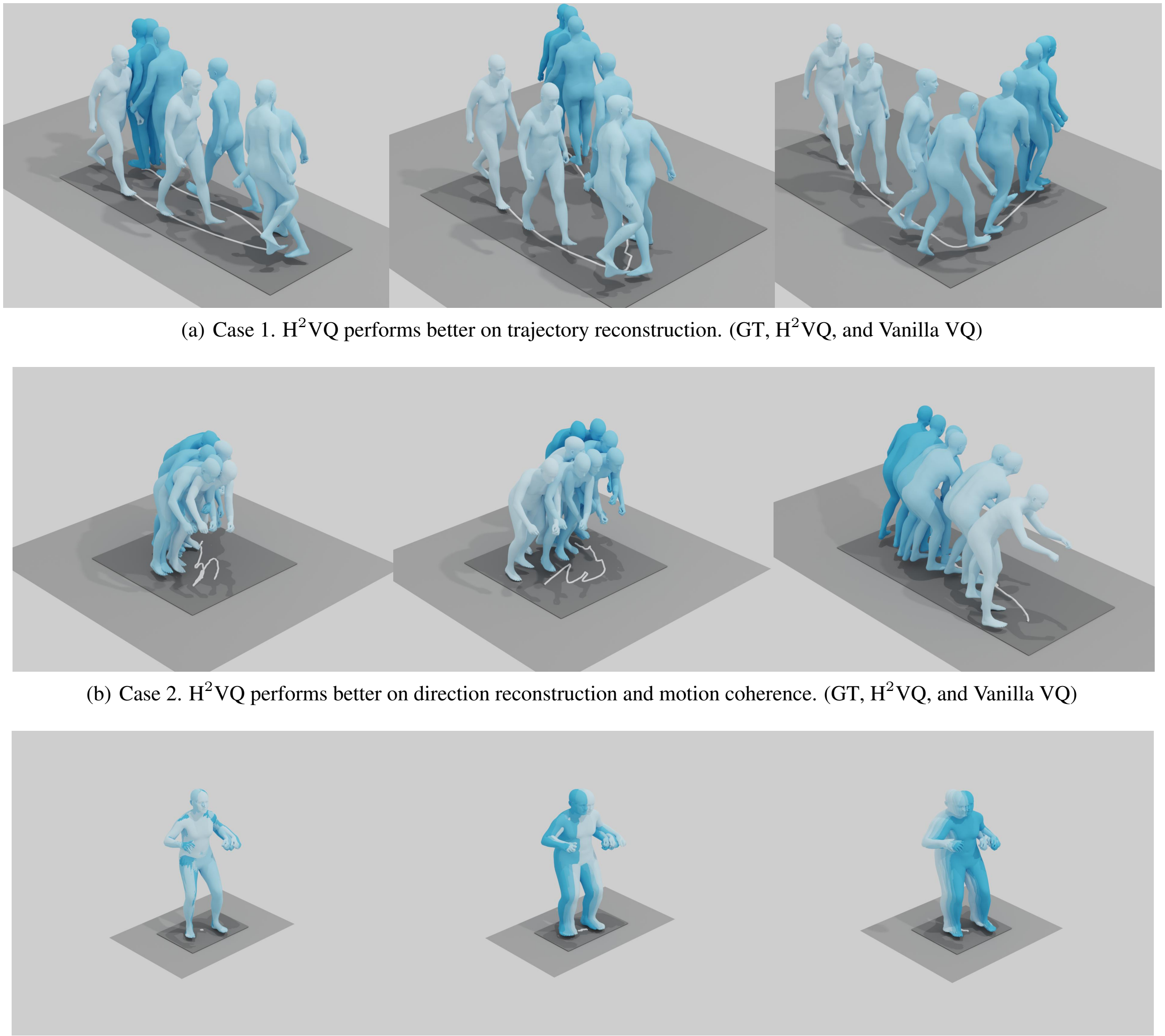
Table 17: Different vector quantization methods on HumanML3D.

We additionally discuss how the H^{2}VQ helps the motion generation from the aspect of motion quality and text-motion alignment. We take the T2M-GPT as the baseline and compare it to the hierarchical reconstruction setting. The difference between the two settings is with or without the H^{2}VQ method. As shown in Table 18, the H^{2}VQ helps both motion generation and text-motion alignment significantly.

	FID↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$			TMA-Matching Score ↓	Matching Score ↓
		Top1↑	Top2↑	Top3↑	Top1↑	Top2↑	Top3↑		
GT	-	0.500	0.708	0.814	0.407	0.578	0.673	0.768	2.888
T2M-GPT w/o H^{2}VQ	1.366	0.368	0.553	0.655	0.310	0.446	0.527	0.881	4.316
T2M-GPT w/ H^{2}VQ	**1.086**	**0.405**	**0.588**	**0.695**	**0.345**	**0.490**	**0.573**	**0.844**	**3.917**

Table 18: The ablation on how can H^{2}VQ help the whole-body motion generation on T2M-GPT.

We show more visualization results here. Our method excels in two perspectives, body-part reconstruction and hand-part reconstruction. On the one hand, From 16(a), our method H^{2}VQ in the middle column achieves a significantly higher level of accuracy in reconstructing global translation. From 16(b), our method could perform better on movement direction reconstruction and motion coherence. From 16(c), our method could reconstruct motion more precisely than other methods even with minor motion movements. On the other hand, because of our decoupled design, our method performs better on hand movement and pose reconstruction. As shown in 17, ours (in blue) can precisely reconstruct the GT hand pose (in green), while the Vanilla VQ-VAE method fails in most of these cases, which demonstrates the superiority of our design.



(a) Case 1. H^{2}VQ performs better on trajectory reconstruction. (GT, H^{2}VQ, and Vanilla VQ)

(b) Case 2. H^{2}VQ performs better on direction reconstruction and motion coherence. (GT, H^{2}VQ, and Vanilla VQ)

(c) Case 3. H^{2}VQ performs better on reconstructing motions with low amplitude. (GT, H^{2}VQ, and Vanilla VQ)

Figure 16: Visualization of motion reconstruction on the Motion-X dataset (body motion reconstruction perspective). From the left to right are GT, H^{2}VQ, and Vanilla VQ, respectively.

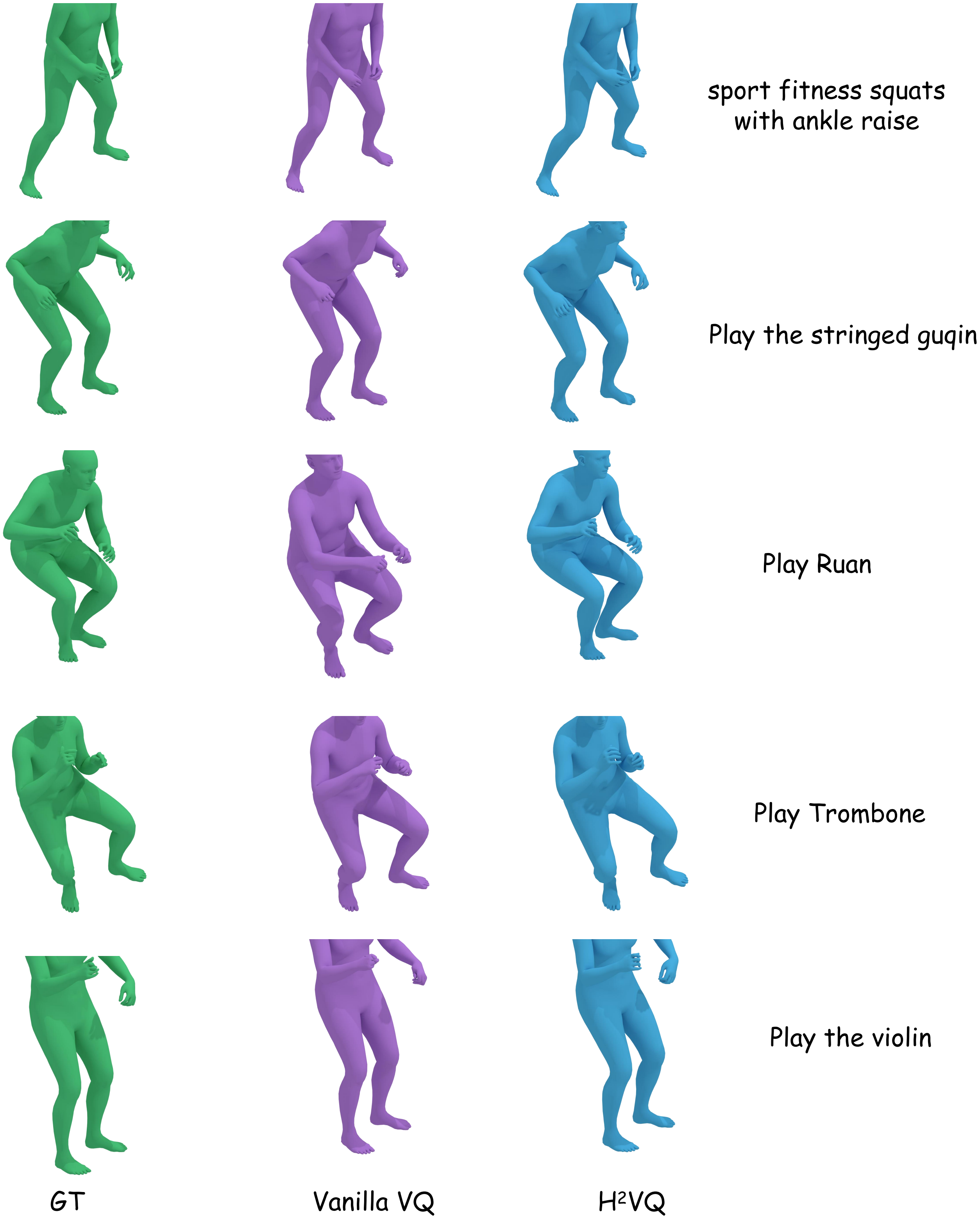


Figure 17: Visualization of motion reconstruction on the Motion-X dataset (hands motion reconstruction perspective). From the left to right are GT, Vanilla VQ, and H^{2}VQ, respectively.

G.2. Comparisons on different Codebook Sizes

We discuss how much the scaling of codebook size benefits the generation results. We perform the comparison on Vanilla VQ, RVQ, and H^{2}VQ. As shown in Figure 18, H^{2}VQ performs best among the three quantization methods. When doubling the codebook size, the final reconstruction error (MPJPE) reduces marginally. This verifies that scaling of codebook size in VQ-VAE is almost in vain. This observation supports the basic intuition on the designing of H^{2}VQ.

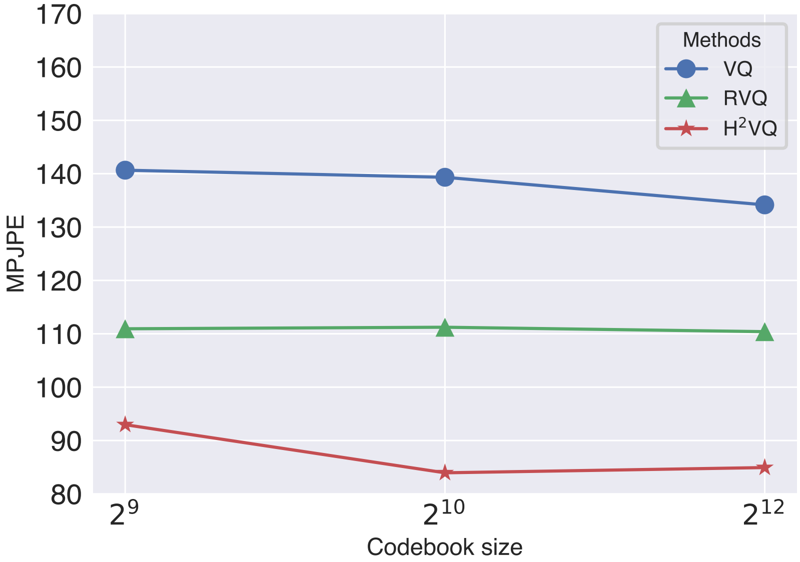


Figure 18: The ablation on the codebook size. Reconstruction results of GT, Vanilla VQ, and H^{2}VQ are presented respectively.

H. Text-motion Aligned Model As A Prior (RQ3)

H.1. Quantitative Results on HumanML3D

In the main paper, we verify that the pre-trained text-motion-aligned model provides a strong prior to text-aligned whole-body motion generation. Additionally, the text-motion-aligned prior not only benefits the whole-body motion generation but also helps the text-motion alignment in body-only motion generation. We take the T2M-GPT as baseline (line 1 in the Table 19), and we ablate whether the TMA language embedding and text-motion-alignment supervision help to generate the text-aligned body-only motions. As shown in Table 19, our experiments on HumanML3D show that both the motion-aware language prior and the text-motion-alignment supervision help to generate higher quality and text-aligned motions (on FID and TMA-R-Precision$^{(256)}$).

embedding	supervision	FID ↓	TMA-R-Precision$^{(256)}$			R-Precision$^{(256)}$			Matching-score ↓	TMA-Matching-score↓
			Top1 ↑	Top2 ↑	Top3 ↑	Top1 ↑	Top2 ↑	Top3 ↑		
CLIP	✗	0.474	0.082	0.129	0.168	0.169	0.259	0.341	3.155	1.322
TMA	✗	0.326	0.147	0.206	0.269	0.177	0.281	0.396	2.915	1.285
TMA	✓	**0.312**	**0.159**	**0.223**	**0.276**	**0.184**	**0.292**	**0.396**	**2.906**	**1.282**

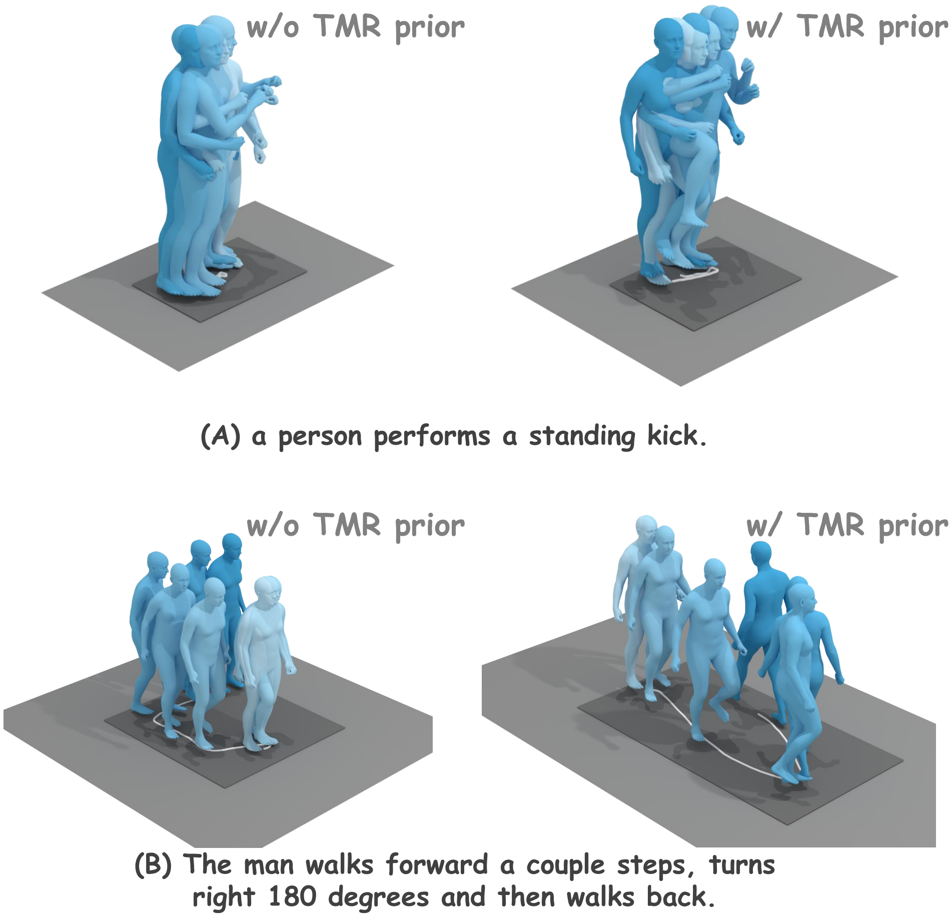
Table 19: Abaltion on how pre-trained text-motion aligned model helps to generate the text-aligned body-only motion (on HumanML3D).

H.2. Pre-trained Text-motion Aligned Model as a Prior

We test on the Motion-X dataset first to explore whether our text-motion-aligned text encoder helps the generated motions align well with the given text. As shown in Figure 19(a), the model with our design performs the "`kick`" motion. As shown in Figure 19(b) and Figure 19(c), HumanTOMATO learning with motion-aware language prior has a better understanding of motion trajectory and temporal relations.

We test some cases in the wild to explore whether our text-motion-aligned text encoder helps the generated motions align well with the given text. We show some cases for comparison in Figure 20. In Figure 20(a), if T2M-GPT learns without motion-aware language prior, the person walks in a quarter of counter-clockwise circle. The model with motion-aware language prior will generate the motion well aligned with the given text on direction and trajectory. For the second case in Figure 20(b), our design helps the model to generate motions much better in the motion direction. For the third case, our method is better aligned with text on the caption "`back`" and does not switch the left or right backward direction.

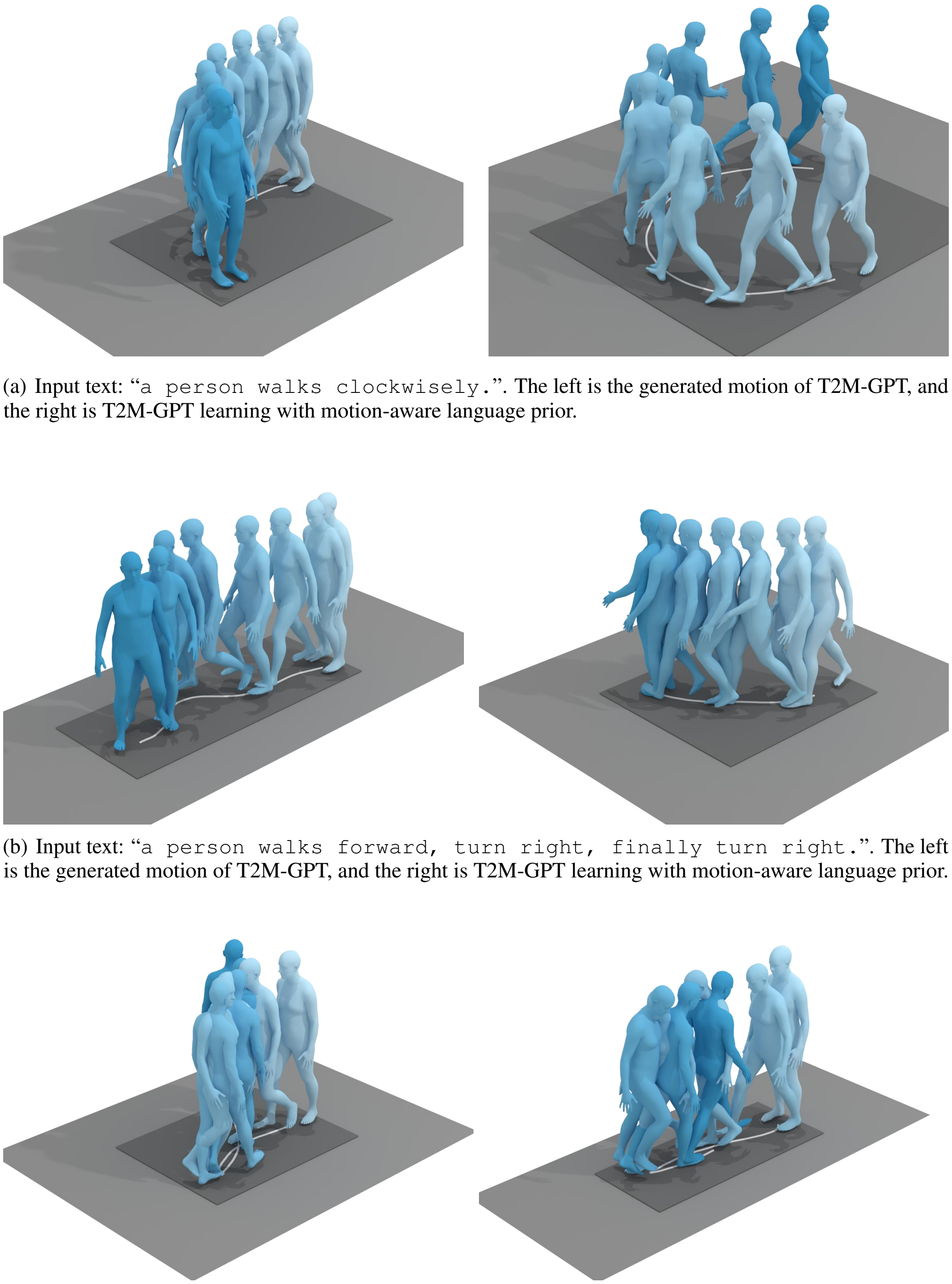
In summary, as claimed in Section 3.5, our method can understand the motion dynamic clues better on sequentiality, directions, and dynamics.



(A) a person performs a standing kick.

(B) The man walks forward a couple steps, turns right 180 degrees and then walks back.

Figure 19: Visualization on our HumanTOMATO, learning without (left) or with (right) motion-aware language prior. The left is the generated motion of HumanTOMATO without language prior, and the right is HumanTOMATO.



(a) Input text: “a person walks clockwisely.”. The left is the generated motion of T2M-GPT, and the right is T2M-GPT learning with motion-aware language prior.

(b) Input text: “a person walks forward, turn right, finally turn right.”. The left is the generated motion of T2M-GPT, and the right is T2M-GPT learning with motion-aware language prior.

(c) Input text: “A person walks forward and then turns back.”. The left is the generated motion of T2M-GPT, and the right is T2M-GPT learning with language prior.

Figure 20: Visualization on T2M-GPT, learning without (left) or with (right) motion-aware language prior. The left is the generated motion of T2M-GPT, and the right is T2M-GPT learning with language prior.

I. Details on the Evaluation Metrics (RQ4)

In Section 4.6, we analyze why the proposed evaluation metrics of alignment between generated motions and given texts are more accurate and challenging on the Motion-X dataset. Here, we provide more comparisons on both body-only and whole-body datasets to verify the universality of the proposed metrics, all of which are calculated 3 times to calculate the mean and standard value ($mean^{\pm std}$). The comparison is shown in Table 20 and Table 21. We also visualize the comparison on the HumanML3D dataset in Figure 21. Similar to the conclusion in Section 4.6, our metrics are more accurate and challenging than Guo et al. (2022)'s in the following two aspects. (1) *TMA-R-Precision*$^{(B)}$ and *TMA-Matching-score*$^{(B)}$ metrics are more accurate than Guo et al. (2022)'s *R-Precision*$^{(B)}$ and *Matching-score* metrics. (2) $B = 256$ is a more challenging retrieval setting than the $B = 32$ setting.

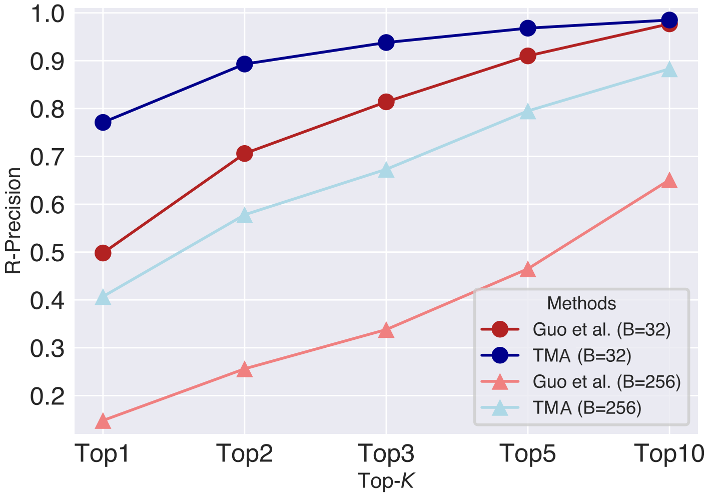


Figure 21: Comparison with existing metrics on HumanML3D. Existing evaluation metrics (Guo et al., 2022) are illustrated in red and ours are in blue. The $B = 32$ and $B = 256$ settings for retrieval are denoted as "—●—" and "—▲—" respectively.

	Top1	Top2	Top3	Top5	Top10
Guo et al. (2022) $B = 32$	$0.498^{\pm.006}$	$0.706^{\pm.005}$	$0.814^{\pm.003}$	$0.910^{\pm.003}$	$0.977^{\pm.001}$
TMA $B = 32$	**0.771**$^{\pm.001}$	**0.893**$^{\pm.003}$	**0.938**$^{\pm.002}$	**0.968**$^{\pm.001}$	**0.985**$^{\pm.000}$
Guo et al. (2022) $B = 32$	$0.148^{\pm.002}$	$0.256^{\pm.004}$	$0.338^{\pm.004}$	$0.465^{\pm.003}$	$0.651^{\pm.002}$
TMA $B = 256$	**0.407**$^{\pm.003}$	**0.578**$^{\pm.004}$	**0.673**$^{\pm.003}$	**0.795**$^{\pm.001}$	**0.883**$^{\pm.001}$

Table 20: R-Precision of GT motions and texts on the Motion-X dataset.

	Top1	Top2	Top3	Top5	Top10
Guo et al. (2022) $B = 32$	$0.511^{\pm.003}$	$0.705^{\pm.002}$	$0.795^{\pm.003}$	$0.887^{\pm.003}$	$0.964^{\pm.003}$
TMA $B = 32$	**0.711**$^{\pm.005}$	**0.853**$^{\pm.001}$	**0.905**$^{\pm.002}$	**0.947**$^{\pm.001}$	**0.977**$^{\pm.001}$
Guo et al. (2022) $B = 256$	$0.167^{\pm.002}$	$0.279^{\pm.002}$	$0.368^{\pm.003}$	$0.490^{\pm.004}$	$0.659^{\pm.003}$
TMA $B = 256$	**0.365**$^{\pm.003}$	**0.527**$^{\pm.002}$	**0.625**$^{\pm.004}$	**0.731**$^{\pm.003}$	**0.838**$^{\pm.002}$

Table 21: R-Precision of GT motions and texts on the HumanML3D dataset.

J. Can We Generate Whole-body Motions by Parts Separately?

In this section, we will discuss whether we can generate whole-body motions by parts separately. To answer this question, we provide an ablation on whether to model them separately in Table 22. In Table 22, the "Modeling Separately" means modeling the hand and body motion separately.

	FID↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$		
		Top1↑	Top2↑	Top3↑	Top1↑	Top2↑	Top3↑
GT	-	$0.500^{\pm 0.002}$	$0.708^{\pm 0.002}$	$0.814^{\pm 0.002}$	$0.407^{\pm 0.003}$	$0.578^{\pm 0.004}$	$0.673^{\pm 0.003}$
Modeling Separately	$2.209^{\pm 0.047}$	$0.359^{\pm 0.002}$	$0.551^{\pm 0.003}$	$0.666^{\pm 0.002}$	$0.306^{\pm 0.003}$	$0.459^{\pm 0.002}$	$0.552^{\pm 0.002}$
HumanTOMATO	$\mathbf{1.174}^{\pm 0.015}$	$\mathbf{0.416}^{\pm 0.009}$	$\mathbf{0.603}^{\pm 0.007}$	$\mathbf{0.703}^{\pm 0.007}$	$\mathbf{0.399}^{\pm 0.000}$	$\mathbf{0.555}^{\pm 0.005}$	$\mathbf{0.638}^{\pm 0.004}$

Table 22: Abalation of modeling strategy on the Motion-X dataset (FID, TMA-R-Precision$^{(256)}$, and R-Precision$^{(32)}$ metrics).

As shown in Table 22, modeling body and hands separately will result in a large performance loss in whole-body motion generation. As a result, we take the H^{2}VQ and Hierarchical-GPT as the technical design choice.

K. Limitation

Although this work makes great progress on the novel task, and the significant improvement of motion reconstruction and text-aligned generation, it still has some shortcomings. Most text2motion efforts proposed by the community are hard to generate physically plausible motions (Yuan et al., 2023). Generating physically plausible motions requires post-processing in a simulation environment, which is left as our future work. First, the natural textual description utilization for whole-body motion generation needs to be further explored. This work simply uses the sequential semantic descriptions following previous works without frame-level or fine-grained whole-body descriptions. Second, the face generation lacks a unified generation scheme. Due to the limited holistic facial expression data and face motion descriptions (e.g., only commonly used emotion here), a simple generator is not the best design choice. As rich data comes, a unified framework could be future work. Additionally, we will unify more text-motion-pairwise data for training a better motion generation model.