

Generating Attribute-Aware Human Motions from Textual Prompt

Xinghan Wang¹, Kun Xu¹, Fei Li², Cao Sheng², Jiazhong Yu², Yadong Mu^{1*}

¹Peking University, China

²China Tower Corporation Limited
xinghan.wang@pku.edu.cn, myd@pku.edu.cn

Abstract

Text-driven human motion generation has recently attracted considerable attention, allowing models to generate human motions based on textual descriptions. However, current methods neglect the influence of human attributes—such as age, gender, weight, and height—which are key factors shaping human motion patterns. This work represents a pilot exploration for bridging this gap. We conceptualize each motion as comprising both attribute information and action semantics, where textual descriptions align exclusively with action semantics. To achieve this, a new framework inspired by Structural Causal Models is proposed to decouple action semantics from human attributes, enabling text-to-semantics prediction and attribute-controlled generation. The resulting model is capable of generating attribute-aware motion aligned with the user’s text and attribute inputs. For evaluation, we introduce a comprehensive dataset containing attribute annotations for text-motion pairs, setting the first benchmark for attribute-aware motion generation. Extensive experiments validate our model’s effectiveness.

Introduction

In recent years, text-driven human motion generation has gained significant attention (Guo et al. 2022a; Zhang et al. 2022; Chen et al. 2023) in both research community and industry owing to its potential applications in video games, filmmaking, virtual reality, and robotics. The goal is to generate realistic human motion sequences based on given textual descriptions. However, human motion patterns are significantly influenced by human attributes such as age, gender, weight, and height. As illustrated in Figure 1, individuals with different attributes demonstrate distinct motion patterns. For instance, the movements of an elderly person and a teenager differ substantially. Therefore, incorporating human attributes into motion generation is crucial for producing realistic and context-aware movements. Also, users may need to generate motions that correspond to specific human attributes, demanding text-to-motion generation controlled by attributes. Despite this, none of the existing methods have addressed this issue.

Incorporating human attributes into text-driven motion generation faces two major challenges. The first is that a

*Corresponding author.

"The person performed a jump by bending at the knees and then performing a maximal-level jump."

Gender: Male
Age: 65
Weight: 64kg
Height: 176cm



"The person performed a jump by bending at the knees and then performing a maximal-level jump."

Gender: Female
Age: 23
Weight: 59kg
Height: 167cm



Figure 1: Data samples from HumanAttr dataset with text prompts and human attributes. Notice that the motion patterns of subjects with different attributes vary significantly.

motion sequence consists of both *action semantics* (e.g., walking, running) and *human attributes* (e.g., gender, age), whereas textual descriptions usually focus solely on the action semantics aspect. Existing text-to-motion frameworks typically align text and motion in a shared space without distinguishing between semantic content and human attribute information, which may hinder the alignment. This necessitates a new framework that can separate action semantics from human attributes, ensuring that texts are accurately aligned with semantic content.

The second challenge lies in the absence of a large-scale text-to-motion dataset that includes annotations for a broad spectrum of human attributes. Existing datasets either lack human attribute annotations entirely (Guo et al. 2022a; Lin et al. 2024; Punnakkal et al. 2021; Liang et al. 2024b; Chen et al. 2024), encompass subjects with only a small range of distinct attributes (Plappert, Mandery, and Asfour 2016; Jang et al. 2020; Troje 2002; Ma et al. 2024; Kim et al. 2021), or have very limited scale (University and of Singapore 2011; Leightley et al. 2015). For example, the KIT dataset (Plappert, Mandery, and Asfour 2016) includes attribute annotations for 4k motion sequences, but 90% of the 55 subjects are aged between 18 and 45. This data gap impedes the development of attribute-aware motion generation

Dataset	Subjects	Motions	Minutes	Age Range
BMLmovi (Troje 2002)	86	1,801	161.8	[17, 33]
EKUT (Plappert, Mandery, and Asfour 2016)	4	348	30.1	[24, 44]
ETRI-Activity3D (Jang et al. 2020)	100	3,727	691.8	[21, 88]
ETRI-LivingLab (Kim et al. 2021)	50	600	160.5	[64, 80]
K3Da (Leightley et al. 2015)	54	236	34.5	[18, 81]
Kinder-Gator (Aloba et al. 2018)	20	362	34.9	[5, 32]
KIT (Plappert, Mandery, and Asfour 2016)	55	4,231	463.1	[15, 55]
Nymeria (Ma et al. 2024)	264	6,850	552.4	[18, 50]
SFU (University and of Singapore 2011)	7	44	6.4	[18, 30]
Total	640	18,199	2,135.4	[5, 88]

Table 1: Statistics of the collected sub-datasets. The statistics are gathered after data filtering.

models capable of generalizing to diverse textual descriptions and human attributes.

This work represents a pilot exploration that incorporates human attributes to generate high-quality 3D human motion aligned with prompted text and attribute controls. To address the above-mentioned semantics-attribute coupling in human motions, we introduce **AttrMoGen** (**Attribute-controlled Motion Generator**). It comprises a Semantic-Attribute Decoupling VQVAE (Decoup-VQVAE) that extracts attribute-free semantic tokens from human motions, and a Semantics Generative Transformer that predicts these semantic tokens from textual prompts. To eliminate attribute information from the semantic tokens, we leverage Structural Causal Model (SCM) (Pearl and Mackenzie 2018), formulating the problem as causal factor disentanglement. In particular, action semantics S are causal factors determining target tokens Y , while human attributes A are non-causal for Y but crucial for constructing raw motion X . The causal model is trained using a causal information bottleneck, which disentangles S from A by limiting their mutual information $I(S; A)$. Additionally, a bottleneck term $-\lambda I(X; S)$ is used to restrict information flow from X to S , ensuring that S captures only the essential semantics. During inference, the Semantics Generative Transformer predicts semantic tokens from text input, which are combined with attributes input to generate the motion via the Decoup-VQVAE decoder.

We compiled a comprehensive dataset called **HumanAttr**, which includes texts and attribute annotations for each motion sequence. HumanAttr integrates data from multiple sources, resulting in a collection of 18.2k motion sequences from 640 subjects with a wide range of attributes. Example data entries are presented in Figure 1. Each attribute label consists of the subject’s age and gender. A portion of the data ($\sim 74\%$) also contains weight and height information, but lacks age diversity thus are not utilized in the main experiments. An additional experiment on this subset that contains weight and height is included in supplementary materials for reference. Extensive experiments are conducted on the dataset, evaluating a range of current representative text-to-motion models, including our proposed AttrMoGen model. The results demonstrate its ability to generate realistic motion that aligns with texts and attributes.

Related Work

Human Motion Generation. Generating realistic human motion that aligns with specific control signals has been a

longstanding problem. A prominent focus is text-to-motion generation. Early works (Tevet et al. 2022a; Petrovich, Black, and Varol 2022, 2023; Yu, Tanaka, and Fujiwara 2024) primarily focused on establishing a joint embedding space for motion sequences and texts. Recent development has been dominated by autoregressive models and diffusion models. The former (Guo et al. 2022a; Huang et al. 2025; Lu et al. 2023) represent motions with discrete tokens and implement text-to-motion generation via next token prediction, while diffusion models (Jin et al. 2024; Yuan et al. 2023; Zhang et al. 2024b; Wang et al. 2023; Karunratanakul et al. 2023; Chen et al. 2023; Xie et al. 2024b,a; Liang et al. 2024a; Huang et al. 2024) perform denoising in the motion space or latent space to generate natural and smooth human motions. Recently, some studies (Zhong et al. 2024; Guo et al. 2024b; Kim et al. 2024; Song et al. 2024) explore style-based motion models, emphasizing how an action is performed (*e.g.*, proud, depressed, angry). Yet, style refers to subjective intent behind an action, while attributes determine the objective inherent biomechanics (*e.g.*, child and the elderly naturally differ in stride length, joint range, and expressiveness).

Text-motion Datasets. With the proliferation of mocap-based databases (CMU Graphics Lab 2000; Shahrudy et al. 2016; Liu et al. 2019; Mandery et al. 2015; Mahmood et al. 2019), numerous text-motion datasets have been proposed. Conventional ones like KIT and HumanML3D provide motion with corresponding sequence-level textual descriptions, primarily developed for text-to-motion task. Subsequent works have advanced the field by incorporating multi-person scenarios (Liang et al. 2024b; Fang et al. 2024), addressing long motion sequences (Li et al. 2023; Han et al. 2024), enhancing the quality of text descriptions (Tang et al. 2023; Chen et al. 2024), adding full-body text descriptions (Lin et al. 2024), and providing frame-level or segment-level annotations (Punnakkal et al. 2021; Wang, Kang, and Mu 2024). Several motion datasets (University and of Singapore 2011; Plappert, Mandery, and Asfour 2016; Troje 2002; Ma et al. 2024; Jang et al. 2020; Kim et al. 2021) provide the attribute information of each subject, yet are limited in data scale or attribute range. For instance, 90% of the subjects in KIT are aged between 18 and 45. This indicates the lack of large-scale collections that encompass a wide range of human attributes.

The HumanAttr Dataset

Dataset Collection. Existing benchmark datasets for human motion generation such as HumanML3D and KIT primarily consist of motion sequences and corresponding text descriptions, lacking the annotations for human attributes. To address this gap, we compile a comprehensive dataset that includes data entries of (motion, text, attribute). The attribute labels contain age and gender for each data entry, and a portion of the data ($\sim 74\%$) also have weight and height information. The dataset, named HumanAttr, integrates data from various sources, annotated with diverse human attributes and textual descriptions. As illustrated in Table 1, HumanAttr consists of nine sub-datasets that encompass a variety of subjects with varying attributes. In total, it contains 18.2k

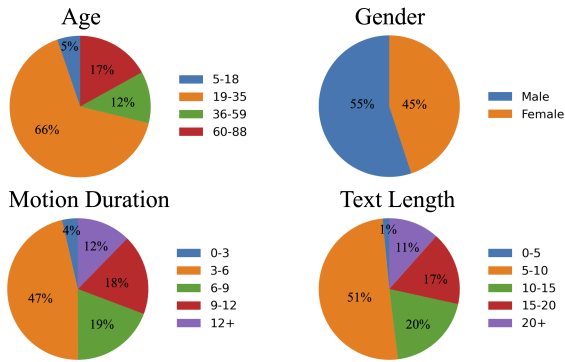


Figure 2: Statistics of age, gender, motion duration (in seconds), and text length (in words) of the HumanAttr dataset.

motion sequences from 640 subjects, with ages ranging from 5 to 88 years and balanced gender. Figure 2 presents the distribution of age, gender, motion duration, and text length within the dataset.

Dataset Processing. This section provides a brief introduction to the pipeline for processing HumanAttr. Initially, data from different sources are converted into a 3D coordinates representation of $L \times V \times 3$, where L denotes the sequence length and V is the number of markers. Since the number of markers varies across different data sources, we adopt the method from (Li et al. 2024) and use gradient descent optimization to fit the marker data into unified parameterized SMPL model (Loper et al. 2023). To make the dataset compatible with previous work in the field, processing protocol from HumanML3D (Guo et al. 2022a) is then applied, resulting in a 263-dimensional input vector.

However, some of the motion data exhibited serious jittering issues caused by the Kinect system. To address this, additional processing is applied to specific sub-datasets. Motions with excessive jitter or outlier coordinates are discarded, and a 1-D Gaussian denoising is applied to enhance motion quality. The dataset is further augmented by mirroring the motion data following the approach in HumanML3D to increase the training samples and improve model generalization. For text annotations, all sub-datasets include either textual descriptions or action category labels. Action labels are expanded into full text descriptions using the information from the original database, ensuring each motion entry has a meaningful text.

Our Proposed Method

Our goal is to generate 3D human motion that aligns with both a given textual description and human attributes control. The core observation is that human motion patterns can be decomposed into two factors: action semantics and human attributes. Since textual descriptions primarily focus on the semantics, existing methods that directly align motion with text annotations may hinder the alignment. Moreover, current methods lack the capability to generate attribute-aware motion based on attribute control input.

To tackle this challenge, we propose an architecture composed of two main components. The overall architecture is

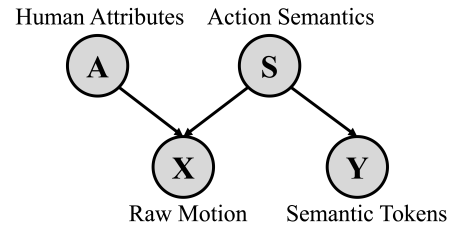


Figure 3: Structural Causal Model for our Decoup-VQVAE. Our objective is to learn an encoder capable of decoupling the Y -causative action semantics S from raw motion X .

illustrated in Figure 4. The first component is a Semantic-attribute Decoupling VQVAE (Decoup-VQVAE), inspired by Structural Causal Model (SCM) (Pearl and Mackenzie 2018). Specifically, the encoder removes attribute information from the raw motion with the help of a causal information bottleneck, resulting in attribute-free semantic tokens via vector quantization. The decoder can then reconstruct the original motion using both the semantic tokens and attribute labels. The second component is a Semantics Generative Transformer, which predicts semantic tokens from text by a generative transformer. During inference, semantic tokens are generated from textual input and combined with user-defined attributes to generate attribute-aware human motion.

Semantic-attribute Decoupling VQVAE

To obtain attribute-free semantic tokens, we employed Structural Causal Model (SCM) to formulate the problem as causal factor disentanglement. Specifically, let X represent the raw motion and Y represent the desired semantic tokens. The action semantics S are the causal factors that determine Y . Human attributes, denoted as A , are non-causal factors that are independent of Y , but essential in constructing the raw motion X . This leads to our causal formulation, with the SCM illustrated in Figure 3. Our aim is to learn an encoder $S = f(X)$ that can decouple the Y -causative action semantics S from raw motion X , and a decoder $\hat{X} = g(S, A)$ that can reconstruct the motion X from S, A . Ideally, f should clearly separate S from A , retaining only the essential semantic information of the raw motion. Inspired by (Zhang et al. 2024a; Hua et al. 2022), we introduce a causal information bottleneck (CIB) to tackle this challenge. Formally, The CIB objective function is defined as

$$CIB(X, Y, S, A) = I(X; S, A) + I(Y; S) - I(S; A) - \lambda I(X; S), \quad (1)$$

where $I(\cdot)$ denotes the mutual information and λ is the bottleneck weight. (Zhang et al. 2024a) proves that maximizing the CIB objective is equivalent to maximizing the mutual information $I(X, Y; S, A)$ between the latent factors S and A and the variables X and Y with an information bottleneck. The role of each term and our customized implementation will be demonstrated below.

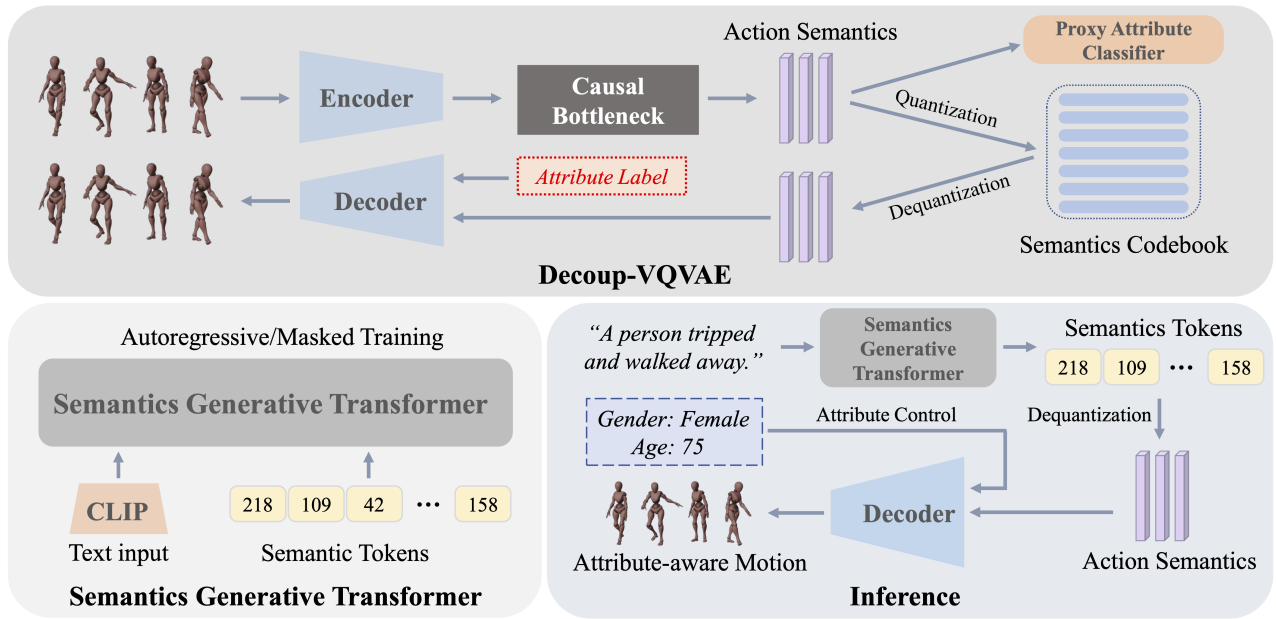


Figure 4: Overall architecture of our proposed AttrMoGen. The encoder of Decoup-VQVAE uses a causal information bottleneck to decouple action semantics from human attributes, producing attribute-free semantic tokens. The decoder then reconstructs motion from these semantic tokens and attribute labels. The Semantics Generative Transformer predicts semantic tokens from textual input, which are subsequently combined with attribute inputs to generate attribute-aware human motions during inference.

Decoupling. The term $-I(S; A)$ restricts the mutual information between S and A , which plays a crucial role in disentangling the causal factor S from the non-causal factor A . In our case, note that the mutual information has the following upper bound:

$$I(S; A) = H(A) - H(A|S) \quad (2)$$

$$\leq \log|A| - \mathbb{E}_{s \sim p(S)} H(A|S = s), \quad (3)$$

where H denotes the entropy and $|A|$ denotes the total size of the attribute space. Since the attribute set is finite, $\log|A|$ is a constant. Given a batch of samples $\{s_i\}_{i=1}^B$, the above upper bound of mutual information can be estimated as:

$$\hat{I}(S; A) = \text{const} - \sum_{i=1}^B H(A|S = s_i) \quad (4)$$

$$= \text{const} + \sum_{i=1}^B \sum_{a \in \mathcal{A}} p(a|s_i) \log p(a|s_i), \quad (5)$$

where \mathcal{A} is the attribute label set. The problem reduces to modeling the conditional distribution $p(a|s_i)$. To this end, a proxy attribute classification network h is introduced, whose goal is to classify the human attribute A from the semantic embedding S . The output score of h then indicates the conditional probability: $p(A|s_i) = h(s_i)$, which enables the computation of the following loss function through Equation 5:

$$\mathcal{L}_{entropy} = - \sum_{i=1}^B H(A|S = s_i). \quad (6)$$

Minimizing $\mathcal{L}_{entropy}$ reduces the mutual information $I(S; A)$ between S and A , eliminating attribute information from S and achieving effective decoupling. During training, the proxy attribute classifier h is updated alternately with the encoder f and decoder g , supervised by the cross-entropy loss \mathcal{L}_{CE} with ground truth attribute labels.

Bottleneck. The information bottleneck term $-\lambda I(X; S)$ limits the information flow from X to S , ensuring that S contains only essential semantics without excessive irrelevant information, fostering a robust and compact representation (Tishby, Pereira, and Bialek 2000). The mutual information $I(X; S)$ can be measured using the Kullback-Leibler (KL) divergence:

$$I(X; S) = \mathbb{E}_{x \sim p(X)} \mathcal{D}_{KL}(p(S|x) \| p(S)) \quad (7)$$

$$= \mathbb{E}_{x \sim p(X)} \mathcal{D}_{KL}(p(S|x) \| \mathbb{E}_{x' \sim p(X)} p(S|x')) \quad (8)$$

$$\leq \mathbb{E}_{x \sim p(X)} \mathbb{E}_{x' \sim p(X)} \mathcal{D}_{KL}(p(S|x) \| p(S|x')). \quad (9)$$

The inequality holds according to Jensen's inequality as the KL divergence is convex *w.r.t.* the second argument, and a rigorous proof is provided in the supplementary. The objective is basically alleviating the impact of X on the posterior distribution of S . To achieve this, we minimize the following objective:

$$D(X; X^-) = \mathcal{D}_{KL}(p(S|X) \| p(S|X^-)), \quad (10)$$

where X^- is the counterfactual motion that shares the same semantics with the original X but has different attributes. Note that our decoder g can generate motion based on specific semantics and any given attribute control, therefore X^-

Algorithm 1: Optimization of Decoup-VQVAE

for each training iteration **do**
 1. Sample a batch $\{(X_i, A_i)\}_{i=1}^B$ from dataset.
 2. Calculate semantic embedding $S_i = f(X_i)$.
 3. Randomize the attribute label A_i into A_i^- .
 4. Obtain counterfactual motion $X_i^- = g(S_i, A_i^-)$.
 5. Update encoder $f(S|X)$ and decoder $g(X|S, A)$ by minimizing $\mathcal{L}_{vqvae} + \alpha\mathcal{L}_{entropy} + \lambda\mathcal{L}_{bottleneck}$.
 6. Update proxy attribute classifier $h(A|S)$ by minimizing \mathcal{L}_{CE} .
end for

can be derived through

$$X^- = g(S, A^-), \quad (11)$$

where A^- is the counterfactual attribute obtained by randomizing the original attribute A . Suppose \mathbf{X} is a batch of samples and \mathbf{X}^- is their corresponding counterfactuals. As the encoder $S = f(X)$ serves as an estimator for the distribution $p(S|X)$, the KL-divergence in Equation 10 can be minimized by aligning $f(\mathbf{X})$ and $f(\mathbf{X}^-)$. To achieve this, we define the following similarity matrix analogous to (Lv et al. 2022):

$$\tilde{D}(\mathbf{X}, \mathbf{X}^-)_{ij} = \text{Cosine}(\mathbf{S}[:, i], \mathbf{S}^-[:, j]), \quad (12)$$

where $\mathbf{S} = f(\mathbf{X}) \in \mathbb{R}^{B \times D}$, $\mathbf{S}^- = f(\mathbf{X}^-) \in \mathbb{R}^{B \times D}$, $\tilde{D} \in \mathbb{R}^{D \times D}$, *Cosine* denotes cosine similarity. This brings the following loss function, which enforces \mathbf{S} and \mathbf{S}^- to be closer while preserving independence between channels, fostering feature space diversity:

$$\mathcal{L}_{bottleneck} = \|\tilde{D}(\mathbf{X}, \mathbf{X}^-) - \mathbf{I}\|_F^2, \quad (13)$$

where \mathbf{I} denotes the identity matrix. A detailed illustration of this workflow is in Algorithm 1.

Reconstruction. The remaining terms, $I(X; S, A) + I(Y; S)$, ensure that the information contained in S and A are sufficient to reconstruct X , while also guarantee that S is informative enough to deduce Y . Note the following lower bound derived by (Zhang et al. 2024a):

$$I(X; S, A) \geq \mathbb{E}_{p(X, S, A)} \log p_\theta(X|S, A), \quad (14)$$

where p_θ denotes the decoder g . This implies that $I(X; S, A)$ can be addressed using a reconstruction loss

$$\mathcal{L}_{rec} = |X - \hat{X}|_F^2 = |X - g(S, A)|_F^2. \quad (15)$$

The second term $I(Y; S)$ ensures the deduction from the causal factor S to the target semantic token Y , which is, by definition, the quantization process from the semantic embedding to the codebook index. Note that $I(Y; S) = H(Y) - H(Y|S)$, where $H(Y)$ reflects the full exploitation of the codebook and $-H(Y|S)$ demands precise indexing from S to Y . This can be addressed by the loss $\mathcal{L}_{commit} + \mathcal{L}_{embed}$ from VQVAE (Van Den Oord, Vinyals et al. 2017). In summary, the overall loss function is formulated as

$$\mathcal{L}_{overall} = \mathcal{L}_{vqvae} + \alpha\mathcal{L}_{entropy} + \lambda\mathcal{L}_{bottleneck}, \quad (16)$$

where $\mathcal{L}_{vqvae} = \mathcal{L}_{rec} + \mathcal{L}_{commit} + \mathcal{L}_{embed}$ is the conventional VQVAE loss and α, λ are weight coefficients.

Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow
	Top-1	Top-2	Top-3		
MoMask	0.685 \pm 0.003	0.864 \pm 0.002	0.925 \pm 0.001	0.245 \pm 0.009	2.602 \pm 0.009
w/ attr test	0.603 \pm 0.002	0.781 \pm 0.002	0.853 \pm 0.002	0.957 \pm 0.019	3.815 \pm 0.014
w/ attr train	0.689 \pm 0.002	0.872 \pm 0.002	0.933 \pm 0.002	0.203 \pm 0.007	2.518 \pm 0.012
w/o entropy	0.686 \pm 0.003	0.867 \pm 0.002	0.930 \pm 0.001	0.489 \pm 0.008	2.523 \pm 0.014
w/o bottleneck	0.686 \pm 0.003	0.870 \pm 0.002	0.933 \pm 0.001	0.184 \pm 0.006	2.486 \pm 0.011
$\lambda = 0.25$	0.698 \pm 0.002	0.881 \pm 0.002	0.942 \pm 0.002	0.098 \pm 0.004	2.326 \pm 0.012
$\lambda = 1$	0.701 \pm 0.002	0.881 \pm 0.002	0.941 \pm 0.001	0.088 \pm 0.003	2.332 \pm 0.007
$\alpha = 0.005$	0.691 \pm 0.002	0.869 \pm 0.001	0.931 \pm 0.001	0.236 \pm 0.007	2.481 \pm 0.016
$\alpha = 0.02$	0.697 \pm 0.002	0.878 \pm 0.002	0.939 \pm 0.001	0.139 \pm 0.004	2.364 \pm 0.012
AttrMoGen	0.705 \pm 0.002	0.882 \pm 0.002	0.940 \pm 0.001	0.089 \pm 0.003	2.266 \pm 0.012

Table 2: Results of ablation studies. “ \uparrow ” denotes that higher is better. “ \downarrow ” denotes that lower is better. The default settings for AttrMogen are $\lambda = 0.5$ and $\alpha = 0.01$.

Semantics Generative Transformer

To predict semantic tokens from text, a Semantics Generative Transformer is employed, whose architecture can be either a GPT-like autoregressive model (Zhang et al. 2023; Lu et al. 2023) or a masked transformer (Guo et al. 2024a). We adopt MoMask (Guo et al. 2024a) due to its strong performance and efficiency. During training, the model aims to predict randomly masked semantic tokens conditioned on the CLIP (Radford et al. 2021) features of the textual input, following a BERT-like scheme (Devlin 2018). At inference time, semantic tokens are first predicted from the text using the Semantics Generative Transformer. These tokens, combined with attribute inputs (e.g., age, gender), are then used to generate the motion via the Decoup-VQVAE decoder. Figure 4 presents a schematic illustration.

Experiments

Implementation and Evaluation Details. For the model architecture, MoMask (Guo et al. 2024a) is adopted due to its superior performance and efficiency. To ensure a fair comparison, all experimental settings and hyperparameters strictly follow those in MoMask. The weight coefficients α, λ of Decoup-VQVAE in Equation 16 are set to 0.01, 0.5 empirically. For attribute control, we categorize the attributes into discrete labels to accommodate sub-datasets that only provide age groups (e.g., Nymeria) and enable in-depth analyses based on categorization. Gender is labeled 0 (male) or 1 (female); age is grouped as 0 (5–18), 1 (19–35), 2 (36–59), and 3 (60–88). One-hot encodings of these labels are fused via MLPs, then concatenated with semantic embeddings for decoding. For the evaluation metrics, we follow the standard protocol (Guo et al. 2022a) and report the following metrics: FID, R-Precision, Multimodal Distance (MM-Dist), Diversity and Multimodality.

Ablation Studies. A straightforward way to incorporate human attributes is to directly supplement attribute information into the text prompts and then employ conventional text-to-motion frameworks. Take the original MoMask model as an example, we conduct two key ablations: (1) human attributes are added to text prompts during only test time but not during training (‘w/ attr test’). (2) human attributes are added to text prompts during both training and testing (‘w/

Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Top-1	Top-2	Top-3				
Real motion	0.750 \pm .0.002	0.913 \pm .0.002	0.964 \pm .0.001	0.001 \pm .0.000	1.793 \pm .0.004	19.259 \pm .0.152	-
TM2T (Guo et al. 2022b)	0.479 \pm .0.002	0.686 \pm .0.002	0.800 \pm .0.002	1.491 \pm .0.033	3.787 \pm .0.014	18.598 \pm .0.142	2.841 \pm .0.085
T2M (Guo et al. 2022a)	0.592 \pm .0.002	0.778 \pm .0.002	0.859 \pm .0.001	1.909 \pm .0.031	3.827 \pm .0.018	18.856 \pm .0.217	2.627 \pm .0.133
MotionDiffuse (Zhang et al. 2022)	0.670 \pm .0.002	0.857 \pm .0.002	0.928 \pm .0.001	0.416 \pm .0.011	2.704 \pm .0.015	18.968 \pm .0.150	2.435 \pm .0.119
MDM (Tevet et al. 2022b)	0.330 \pm .0.006	0.479 \pm .0.009	0.565 \pm .0.007	1.581 \pm .0.361	9.482 \pm .0.105	16.294 \pm .0.296	4.853 \pm .0.084
T2M-GPT (Zhang et al. 2023)	0.661 \pm .0.002	0.824 \pm .0.002	0.889 \pm .0.002	0.279 \pm .0.017	3.203 \pm .0.022	19.184 \pm .0.190	3.170 \pm .0.162
MLD (Chen et al. 2023)	0.660 \pm .0.002	0.850 \pm .0.001	0.922 \pm .0.001	0.349 \pm .0.011	2.673 \pm .0.013	19.576 \pm .0.239	1.938 \pm .0.086
GenMoStyle (Guo et al. 2024b)	0.680 \pm .0.001	0.861 \pm .0.002	0.925 \pm .0.001	0.332 \pm .0.007	2.649 \pm .0.012	19.118 \pm .0.206	1.588 \pm .0.074
MoMask (Guo et al. 2024a)	0.685 \pm .0.003	0.864 \pm .0.002	0.925 \pm .0.001	0.245 \pm .0.009	2.602 \pm .0.009	18.981 \pm .0.132	1.438 \pm .0.084
AttrMoGen	0.705 \pm .0.002	0.882 \pm .0.002	0.940 \pm .0.001	0.089 \pm .0.003	2.266 \pm .0.012	19.268 \pm .0.212	1.250 \pm .0.076

Table 3: Performance of representative existing text-to-motion methods as well as our proposed AttrMoGen which incorporates attribute information on the HumanAttr test set. “ \uparrow ” denotes that higher is better. “ \downarrow ” denotes that lower is better. “ \rightarrow ” denotes that results are better when closer to the real motion.

Attribute	Group	Method	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	
			Top-1	Top-2	Top-3			Generated / Real	
Age	5-18	MoMask	0.511 \pm .0.010	0.733 \pm .0.010	0.835 \pm .0.008	0.736 \pm .0.094	2.630 \pm .0.042	14.144 \pm .0.229 / 14.552 \pm .0.200	
		AttrMoGen	0.544 \pm .0.007	0.752 \pm .0.009	0.842 \pm .0.005	0.778 \pm .0.079	2.487 \pm .0.038	14.500 \pm .0.179 / 14.552 \pm .0.200	
	19-35	MoMask	0.684 \pm .0.004	0.858 \pm .0.002	0.918 \pm .0.002	0.240 \pm .0.011	2.634 \pm .0.019	18.571 \pm .0.133 / 18.836 \pm .0.118	
		AttrMoGen	0.702 \pm .0.003	0.876 \pm .0.002	0.932 \pm .0.002	0.090 \pm .0.003	2.349 \pm .0.012	18.708 \pm .0.205 / 18.836 \pm .0.118	
	36-59	MoMask	0.445 \pm .0.007	0.643 \pm .0.005	0.744 \pm .0.006	1.944 \pm .0.092	3.762 \pm .0.037	14.939 \pm .0.129 / 14.508 \pm .0.156	
		AttrMoGen	0.483 \pm .0.007	0.677 \pm .0.008	0.781 \pm .0.006	0.669 \pm .0.039	3.185 \pm .0.035	14.873 \pm .0.114 / 14.508 \pm .0.156	
60-88	MoMask	0.375 \pm .0.004	0.597 \pm .0.003	0.731 \pm .0.003	0.145 \pm .0.014	1.702 \pm .0.022	12.175 \pm .0.131 / 11.953 \pm .0.154		
	AttrMoGen	0.393 \pm .0.003	0.618 \pm .0.003	0.755 \pm .0.004	0.097 \pm .0.009	1.209 \pm .0.017	11.915 \pm .0.142 / 11.953 \pm .0.154		
Gender	male	MoMask	0.672 \pm .0.003	0.851 \pm .0.002	0.914 \pm .0.001	0.277 \pm .0.010	2.710 \pm .0.016	18.664 \pm .0.194 / 19.092 \pm .0.172	
		AttrMoGen	0.701 \pm .0.003	0.877 \pm .0.002	0.934 \pm .0.002	0.097 \pm .0.006	2.340 \pm .0.020	18.910 \pm .0.179 / 19.092 \pm .0.172	
	female	MoMask	0.681 \pm .0.003	0.871 \pm .0.003	0.934 \pm .0.002	0.246 \pm .0.010	2.468 \pm .0.012	19.232 \pm .0.194 / 19.433 \pm .0.223	
		AttrMoGen	0.694 \pm .0.004	0.879 \pm .0.003	0.941 \pm .0.002	0.118 \pm .0.006	2.176 \pm .0.018	19.615 \pm .0.148 / 19.433 \pm .0.223	

Table 4: Performance comparison on different attribute groups. “ \uparrow ” denotes that higher is better. “ \downarrow ” denotes that lower is better. “ \rightarrow ” denotes that results are better when closer to the real motion.

attr train’). The results are illustrated in Table 2. As shown, directly providing attributes through text prompts during inference fails to generate high-quality motions. Including attribute information in the training texts enhances MoMask’s performance, but remains less effective than our proposed attribute-control scheme.

In addition, we analyze how each component impacts the performance of our proposed model, as well as the influence of the hyperparameters α , λ in Equation 16. For clarity, AttrMoGen without $\mathcal{L}_{entropy}$ and $\mathcal{L}_{bottleneck}$ are denoted as ‘w/o entropy’ and ‘w/o bottleneck’ in the table respectively. The results are presented in Table 2.

Quantitative Evaluations. A number of representative text-to-motion methods are evaluated on the HumanAttr dataset, including approaches based on VAE, autoregressive models, masked transformers, and diffusion. Additionally, we adopt a style-controlled motion generation model GenMoStyle (Guo et al. 2024b) as a baseline by substituting style labels with attribute labels and using text prompts to generate content codes during inference. All experiments are repeated 20 times, and the mean values along with a 95% confidence interval are reported. The overall performance is

summarized in Table 3. As shown, our method outperforms competing methods in R-precision, FID, MM-Dist, and Diversity. Notably, AttrMoGen advances the FID of MoMask from 0.245 to 0.089 and the MM-Dist from 2.602 to 2.266, demonstrating that incorporating human attributes can significantly enhance the quality of generated motion. Table 4 further compares the performance of MoMask and AttrMoGen within each attribute group, showing that AttrMoGen consistently outperforms MoMask across all groups.

The effectiveness of attribute control is validated as follows. First, we pretrain an attribute classifier on the training set. We then use this classifier to classify the attributes of generated motions. For an ideal generation model, the predicted attributes should match the input controls. We adopt two testing protocols: (a) generating motion from (text, attribute) pairs drawn from the original test set, to evaluate performance on the original data distribution; and (b) generating motion from text with randomly assigned attributes, to assess generalization beyond the original dataset. Table 5 shows that AttrMoGen demonstrates higher classification accuracy, indicating that it successfully generates motions aligned with user-specified attributes.

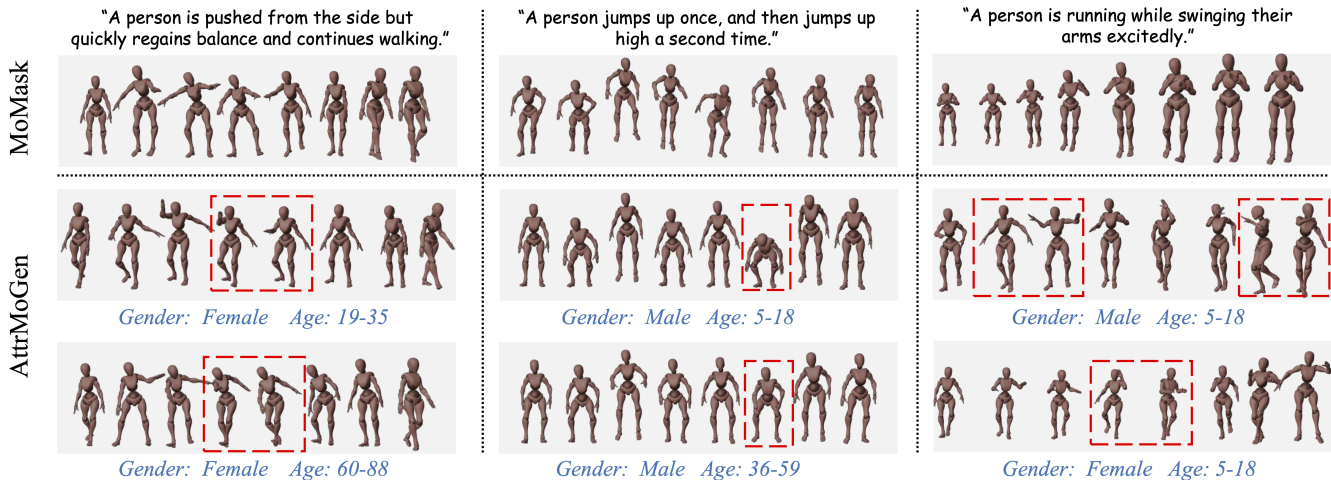


Figure 5: Visualization of generated motions of MoMask and AttrMoGen. As shown, subjects of different attributes exhibit variations in the extent and patterns of movements.

Attribute	Group	True		Random	
		MoMask	AttrMoGen	MoMask	AttrMoGen
Gender	male	0.747 \pm 0.002	0.992 \pm 0.001	0.614 \pm 0.004	0.992 \pm 0.001
	female	0.546 \pm 0.003	0.985 \pm 0.001	0.383 \pm 0.004	0.983 \pm 0.001
	avg.	0.657 \pm 0.002	0.989 \pm 0.001	0.515 \pm 0.004	0.988 \pm 0.001
Age	5-18	0.314 \pm 0.009	0.422 \pm 0.012	0.098 \pm 0.005	0.148 \pm 0.007
	19-35	0.694 \pm 0.002	0.728 \pm 0.003	0.621 \pm 0.003	0.661 \pm 0.003
	36-59	0.409 \pm 0.009	0.439 \pm 0.009	0.106 \pm 0.007	0.102 \pm 0.005
	60-88	0.556 \pm 0.005	0.787 \pm 0.005	0.162 \pm 0.005	0.387 \pm 0.009
	avg.	0.619 \pm 0.002	0.691 \pm 0.002	0.460 \pm 0.002	0.527 \pm 0.002

Table 5: Attribute classification accuracy of generated motion controlled by true and random attribute labels.

Visualizations. To illustrate AttrMoGen’s ability of attribute control, we visualized the generated motions with different attribute inputs for the same text inputs, alongside MoMask’s results for comparison. As shown in Figure 5, the extent and patterns of movements vary across different attributes, which validate AttrMoGen’s effectiveness in producing attribute-aware motions.

Since the impact of gender on motion is hard to observe visually, we design a t-SNE visualization in Figure 6. First, motion features extracted by the VQVAE of MoMask and AttrMoGen are visualized. Unlike MoMask that showcases distinct gender-based clustering, Decoup-VQVAE removes attribute cues, resulting in no observable gender clusters. Then we feed MoMask- and AttrMoGen-generated motions into a pretrained feature extractor and visualize the features. Here, AttrMoGen exhibits stronger gender-based clustering than MoMask, indicating better alignment of generated motions with the input attributes.

Conclusion

This work explores the novel direction of integrating human attributes into text-to-motion generation. The core design of our framework is a decoupling encoder inspired by Structural Causal Model, which separates action semantics

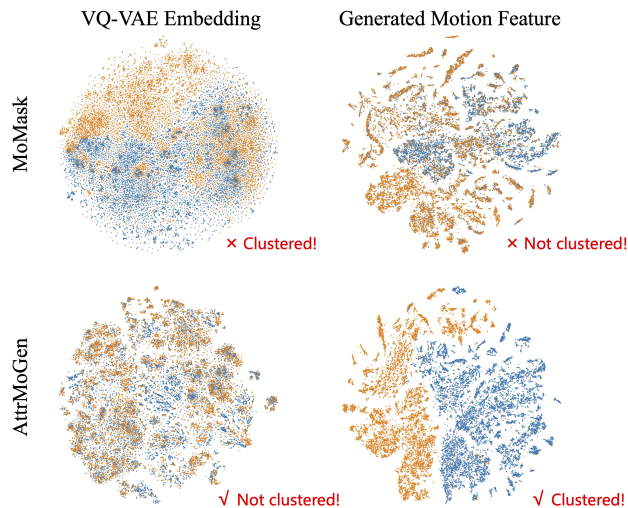


Figure 6: t-SNE visualization with colors representing male and female. The VQVAE embeddings of AttrMoGen exhibit no gender-based clustering, indicating effective removal of gender cues from the motion. Meanwhile, with attributes control input, features of the generated motion by AttrMoGen displays distinct gender-based clusters, demonstrating better alignment with attributes input. Best viewed in color.

from human attributes, enabling text-to-semantics prediction and attribute-controlled generation. We also present a text-to-motion dataset with attribute annotations called HumanAttr, which features subjects with diverse attributes. Extensive evaluations demonstrate the potential of generating realistic human motions based on text and attribute control.

Acknowledgements

The research is supported by a grant from Bytedance (No.CT20250811106734).

References

- Aloba, A.; Flores, G.; Woodward, J.; Shaw, A.; Castonguay, A.; Cuba, I.; Dong, Y.; Jain, E.; and Anthony, L. 2018. Kinder-Gator: The UF Kinect Database of Child and Adult Motion. In *Eurographics (Short Papers)*, 13–16.
- Chen, L.-H.; Lu, S.; Zeng, A.; Zhang, H.; Wang, B.; Zhang, R.; and Zhang, L. 2024. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *arXiv preprint arXiv:2405.20340*.
- Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your commands via motion diffusion in latent space. In *CVPR*, 18000–18010.
- CMU Graphics Lab. 2000. CMU Graphics Lab Motion Capture Database. <http://mocap.cs.cmu.edu/>. Accessed: 2024-11-01.
- Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fang, Q.; Fan, Y.; Li, Y.; Dong, J.; Wu, D.; Zhang, W.; and Chen, K. 2024. Capturing Closely Interacted Two-Person Motions with Reaction Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 655–665.
- Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024a. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.
- Guo, C.; Mu, Y.; Zuo, X.; Dai, P.; Yan, Y.; Lu, J.; and Cheng, L. 2024b. Generative Human Motion Stylization in Latent Space. *arXiv preprint arXiv:2401.13505*.
- Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *CVPR*, 5152–5161.
- Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts.
- Han, B.; Peng, H.; Dong, M.; Ren, Y.; Shen, Y.; and Xu, C. 2024. AMD: Autoregressive Motion Diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2022–2030.
- Hua, H.; Yan, J.; Fang, X.; Huang, W.; Yin, H.; and Ge, W. 2022. Causal information bottleneck boosts adversarial robustness of deep neural network. *arXiv preprint arXiv:2210.14229*.
- Huang, Y.; Wan, W.; Yang, Y.; Callison-Burch, C.; Yatskar, M.; and Liu, L. 2025. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, 180–196. Springer.
- Huang, Y.; Yang, H.; Luo, C.; Wang, Y.; Xu, S.; Zhang, Z.; Zhang, M.; and Peng, J. 2024. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 224–232.
- Jang, J.; Kim, D.; Park, C.; Jang, M.; Lee, J.; and Kim, J. 2020. ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10990–10997. IEEE.
- Jin, P.; Wu, Y.; Fan, Y.; Sun, Z.; Yang, W.; and Yuan, L. 2024. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems*, 36.
- Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2151–2162.
- Kim, B.; Kim, J.; Chang, H. J.; and Choi, J. Y. 2024. MoST: Motion Style Transformer between Diverse Action Contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1705–1714.
- Kim, D.; Lee, I.; Kim, D.; and Lee, S. 2021. Action recognition using close-up of maximum activation and eTri-activity3d livinglab dataset. *Sensors*, 21(20): 6774.
- Leightley, D.; Yap, M. H.; Coulson, J.; Barnouin, Y.; and McPhee, J. S. 2015. Benchmarking human motion analysis using kinect one: An open source dataset. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–7. IEEE.
- Li, S.; Zhuang, S.; Song, W.; Zhang, X.; Chen, H.; and Hao, A. 2023. Sequential texts driven cohesive motions synthesis with natural transitions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9498–9508.
- Li, Y.-L.; Wu, X.; Liu, X.; Wang, Z.; Dou, Y.; Ji, Y.; Zhang, J.; Li, Y.; Lu, X.; Tan, J.; et al. 2024. From isolated islands to pangea: Unifying semantic space for human action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16582–16592.
- Liang, H.; Bao, J.; Zhang, R.; Ren, S.; Xu, Y.; Yang, S.; Chen, X.; Yu, J.; and Xu, L. 2024a. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 482–493.
- Liang, H.; Zhang, W.; Li, W.; Yu, J.; and Xu, L. 2024b. Inter-gen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 1–21.
- Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2024. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36.
- Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*.
- Lu, S.; Chen, L.-H.; Zeng, A.; Lin, J.; Zhang, R.; Zhang, L.; and Shum, H.-Y. 2023. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*.

- Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8046–8056.
- Ma, L.; Ye, Y.; Hong, F.; Guzov, V.; Jiang, Y.; Postyeni, R.; Pesqueira, L.; Gamino, A.; Baiyya, V.; Kim, H. J.; et al. 2024. Nymeria: A Massive Collection of Multi-modal Egocentric Daily Motion in the Wild. *arXiv preprint arXiv:2406.09905*.
- Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes.
- Mandery, C.; Terlemez, Ö.; Do, M.; Vahrenkamp, N.; and Asfour, T. 2015. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, 329–336. IEEE.
- Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions.
- Petrovich, M.; Black, M. J.; and Varol, G. 2023. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 9488–9497.
- Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT motion-language dataset. *Big data*.
- Punnakkal, A. R.; Chandrasekaran, A.; Athanasiou, N.; Quiros-Ramirez, A.; and Black, M. J. 2021. BABEL: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 722–731.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.
- Song, W.; Jin, X.; Li, S.; Chen, C.; Hao, A.; Hou, X.; Li, N.; and Qin, H. 2024. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821–830.
- Tang, Y.; Liu, J.; Liu, A.; Yang, B.; Dai, W.; Rao, Y.; Lu, J.; Zhou, J.; and Li, X. 2023. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22106–22117.
- Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022a. Motionclip: Exposing human motion generation to clip space.
- Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022b. Human motion diffusion model. In *ICLR*.
- Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.
- Troje, N. F. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5): 2–2.
- University, S. F.; and of Singapore, N. U. 2011. SFU Motion Capture Database.
- Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Wang, X.; Kang, Z.; and Mu, Y. 2024. Text-controlled Motion Mamba: Text-Instructed Temporal Grounding of Human Motion. *arXiv preprint arXiv:2404.11375*.
- Wang, Y.; Leng, Z.; Li, F. W.; Wu, S.-C.; and Liang, X. 2023. Fg-T2M: Fine-Grained Text-Driven Human Motion Generation via Diffusion Model.
- Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2024a. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.
- Xie, Z.; Wu, Y.; Gao, X.; Sun, Z.; Yang, W.; and Liang, X. 2024b. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6252–6260.
- Yu, Q.; Tanaka, M.; and Fujiwara, K. 2024. Exploring Vision Transformers for 3D Human Motion-Language Models with Motion Patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 937–946.
- Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; and Kautz, J. 2023. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16010–16021.
- Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14730–14740.
- Zhang, M.; Bi, K.; Chen, W.; Chen, Q.; Guo, J.; and Cheng, X. 2024a. CausalDiff: Causality-Inspired Disentanglement via Diffusion Model for Adversarial Defense. *arXiv preprint arXiv:2410.23091*.
- Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.
- Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2024b. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36.
- Zhong, L.; Xie, Y.; Jampani, V.; Sun, D.; and Jiang, H. 2024. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, 405–421. Springer.