ParTY: Part-Guidance for Expressive Text-to-Motion Synthesis

KunHo Heo SuYeon Kim Yonghyun Gwon Youngbin Kim MyeongAh Cho†
Kyung Hee University
{hkh7710, spoiuy3, mathewgwon, youngbean, maycho}@khu.ac.kr
https://visualsciencelab-khu.github.io/ParTY_project/

Abstract

*Text-to-motion synthesis aims to generate natural and expressive human motions from textual descriptions. While existing approaches primarily focus on generating holistic motions from text descriptions, they struggle to accurately reflect actions involving specific body parts. Recent part-wise motion generation methods attempt to resolve this but face two critical limitations: (i) they lack explicit mechanisms for aligning textual semantics with individual body parts, and (ii) they often generate incoherent full-body motions due to integrating independently generated part motions. To overcome these issues and resolve the fundamental trade-off in existing methods, we propose **ParTY**, a novel framework that enhances part expressiveness while generating coherent full-body motions. ParTY comprises: **(1) Part-Guided Network**, which first generates part motions to obtain part guidance, then uses it to generate holistic motions; **(2) Part-aware Text Grounding**, which diversely transforms text embeddings and appropriately aligns them with each body part; and **(3) Holistic-Part Fusion**, which adaptively fuses holistic motions and part motions. Extensive experiments, including **part-level** and **coherence-level** evaluations, demonstrate that ParTY achieves substantial improvements over previous methods.*

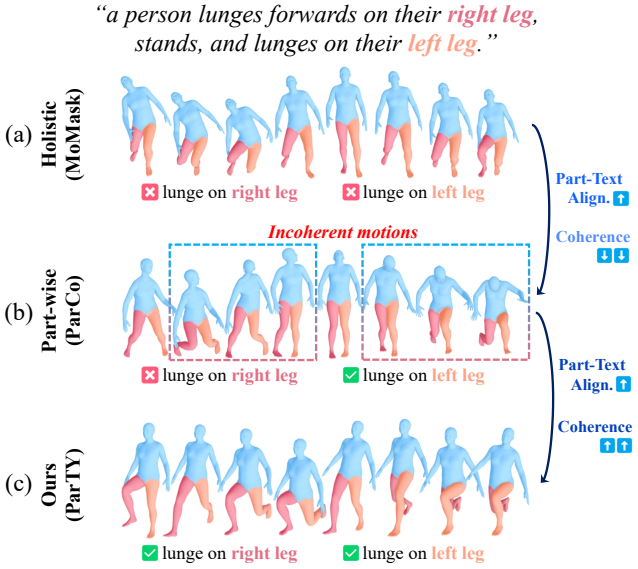

Figure 1. (a) Holistic methods maintain coherence well but limited part-text alignment. In contrast, (b) Part-wise methods show enhanced part-text alignment (e.g., correctly performing the left leg lunge) but compromised coherence as a trade-off (e.g., neck distortion and misaligned arm and leg movements). (c) Our ParTY resolves this trade-off by achieving superior performance in both part-text alignment and coherence.

1. Introduction

Text-to-motion synthesis aims to generate human motions from textual descriptions, with applications in animation [15], virtual reality [10], video games [22, 36], and robotics [3, 17, 18]. Recent architectures [12, 13, 25, 26, 37, 39] have improved semantic alignment with the input text, and visual fidelity. However, most adopt a holistic generation approach: synthesizing full-body motion directly from text. While this strategy generates globally coherent motion, it fundamentally lacks the capacity to model part-specific semantics by treating the body as a single entity. As a result, fine-grained part actions in the text are often misrepresented or overlooked, as shown in Fig. 1 (a).

To address this limitation, part-wise methods [30, 40] have emerged, which split the body into anatomical parts and independently generate motions separately for each part. This decomposition provides explicit part-level control and opens up significant potential for improved part-specific expressiveness compared to holistic approaches. However, existing part-wise methods have yet to fully realize this potential and face two critical challenges: **(i) Insufficient text-to-part semantic alignment**: Existing methods miss fine-grained, part-relevant cues in text, so motions fail to reflect intended part-level behaviors. **(ii) Lack of**

† Corresponding author

inter-part coherence: Since each part motion is generated independently and then simply combined without consideration of global consistency, the resulting full-body motion often lacks overall coherence, as illustrated by the part-wise method in Fig. 1 (b).

To tackle these challenges and bridge holistic and part-wise methods, we propose **ParTY**, a novel framework that resolves the inherent trade-off between part-specific expressiveness and full-body coherence. For fine-grained ***text-to-part alignment***, we introduce **Part-aware Text Grounding** module: it transforms a single sentence embedding into multiple diverse embeddings and dynamically selects appropriate embeddings for each body part. During the selection process, it leverages auxiliary text information about each part generated by an LLM, which is used only during training. This enables fine-grained text-part alignment over previous methods.

To address the lack of ***coherence*** among independently generated part motions, we propose a **Part-Guided Network**, a dual-generation framework that first generates part motions and then uses them as guidance to generate holistic motions, rather than generating each part independently and combining them. Specifically, part motions are generated for several time steps to create part guidance, which conditions the holistic motion generation by providing future part-level information. During holistic motion generation, a **Holistic-Part Fusion** is also employed, which directly fuses holistic and part motions, allowing part motion information to be incorporated throughout the process. These approaches enable the generation of coherent movements across the entire body.

While our method improves part-specific expressiveness and full-body coherence, evaluating these improvements remains challenging. Conventional metrics [10] operate exclusively at the holistic-level, making them incapable of accurately assessing part-level semantic alignment. Moreover, no metric exists to directly evaluate motion coherence across the full-body. To address these issues, we propose new evaluation protocols: **part-level metrics** that expand [10]'s approach for evaluating part-specific expressiveness and **temporal and spatial coherence metrics**. Extensive experiments including these evaluations and Fig. 1 (c) demonstrate that our **ParTY** effectively combines the advantages of both holistic and part-wise methods, leveraging expressive part motions while maintaining coherence.

Our contributions are summarized as follows:

- We introduce a **Part-Guided Network** that addresses the *coherence* problem inherent to part-wise methods by first generating part motions and then using them as guidance for holistic motion generation, with a **Holistic-Part Fusion** module that adaptively fuses both motion representations to maintain coordination.
- We propose **Part-aware Text Grounding**, which enhances fine-grained *text-to-part alignment* by diversely transforming text embeddings and appropriately aligning them with each body part, leveraging LLM-generated part descriptions as auxiliary information.
- Through extensive experiments, we demonstrate that our method achieves state-of-the-art performance on conventional metrics. Furthermore, using our newly proposed **part-level and coherence-level evaluation metrics**, we validate and analyze the effectiveness of our approach in improving part expressiveness and motion coherence.

2. Related Works

Text-to-Motion Generation. Text-to-motion synthesis aims to translate textual descriptions into realistic human movements, offering intuitive control through natural language [5, 28]. Unlike traditional methods relying on action classes [9, 21, 23] or audio signals [19, 20, 33], text-driven approaches provide expressive control capabilities. Early approaches established cross-modal alignment through joint embedding spaces [1, 2], followed by probabilistic frameworks [4, 24] that captured the one-to-many relationship between text and motion. Discrete representation learning emerged with [11] enhancing cross-modal understanding, while [14, 37] combined VQ-VAE [34] with transformers for autoregressive generation. Recent diffusion-based methods [6, 7, 32, 35, 38] dramatically improve motion quality through iterative denoising processes, though often with computational overhead. Temporal context modeling has been explored through masked modeling strategies [12] and hybrid approaches [25, 26]. Despite these advances, most approaches treat human motion as a monolithic entity rather than a coordinated system of interrelated parts, struggling to capture nuanced relationships between textual descriptions and specific body movements that require precise inter-part coordination.

Part-wise Text-to-Motion Generation. Part-wise approaches treat the human body as a system of coordinated parts rather than a monolithic entity, aiming to enhance expressiveness and control over individual body components. Early work included SCA [8], which divided the body into upper and lower segments with independent networks, demonstrating potential for specialized control but struggling with coordination. Additionally, AttT2M [39] introduced body-part attention-based encoders, though its single decoder limited nuanced control. More recently, ParCo [40] used separate VQ-VAEs for each body part with token-sharing for coordination, but did not leverage part-specific text descriptions. LGTM [30] decomposed text descriptions into part-specific prompts using an LLM, but this extraction of only part-related text loses the overall context of the sentence. Due to these limitations, although these approaches achieve some improvement in part expressiveness, they still exhibit insufficient part-level detail. Moreover, the simple

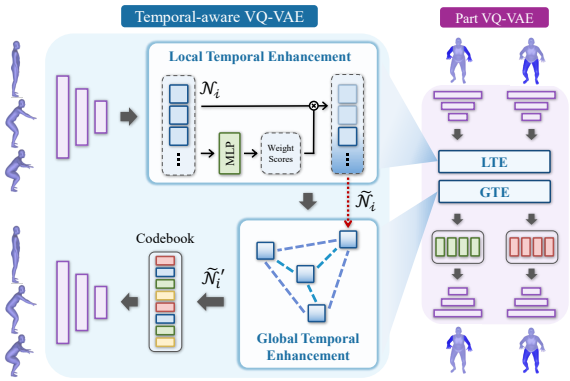

Figure 2. Architecture of the Temporal-aware VQ-VAE. Part VQ-VAE follows an identical architecture, where the sole distinction lies in processing part-level rather than full-body motion data.

integration of independently generated parts inevitably results in incoherent motions. Our method addresses these limitations, enhancing both aspects simultaneously.

3. Method

Our method consists of two stages. First, we quantize motion sequences into codebooks for both the full-body and parts (arms and legs), as shown in Fig. 2. Second, we train holistic and part transformers using these codebooks to predict codebook sequences that match the text description, as shown in Fig. 3. During inference, the predicted codebook sequences are decoded by the pre-trained VQ-VAE decoder from the first stage to reconstruct the motions. Further network details can be found in the supplementary material.

3.1. Temporal-aware VQ-VAE

Recent studies have increasingly adopted VQ-VAE [34] to quantize sequential motions into discrete codebooks [12, 13, 25, 26, 37, 40]. However, compressing motion sequences—where temporal flow is crucial—into discrete codebooks through fixed-size windows inherently causes temporal information loss. While reducing the window size can mitigate this loss by increasing codebook entries, this introduces a trade-off with model size without truly resolving the problem. This constitutes a fundamental limitation of VQ-VAE-based approaches. To address this limitation without increasing model complexity, we propose **Temporal-aware VQ-VAE** that enhances both local and global temporal information to enable codebook quantization while preserving temporal details.

Our goal is to quantize motion sequences into codebooks while preserving temporal information. To achieve this, we enhance local and global temporal information before quantization, as shown in Fig. 2. As motion sequences are encoded frame-by-frame through the encoder, **Local Temporal Enhancement (LTE)** bundles frame-level features with a window size of w, resulting in feature groups $\{\mathcal{N}_i\}_{i=1}^{t/w}$ where t denotes total frames. For each group, we compute feature weights via an MLP_i (a 3-layer MLP with ReLU activation) and perform weighted summation to produce an enhanced group-level feature $\{\tilde{\mathcal{N}}_i\}_{i=1}^{t/w}$, formulated as:

$$\tilde{\mathcal{N}}_i = \sum_{j=1}^{w} \alpha_{ij} \cdot f_{ij}, \quad f_{ij} \in \mathcal{N}_i \tag{1}$$

where f_{ij} is the j-th feature in group $\mathcal{N}_i$, and α_{ij} is the weight computed by applying softmax to MLP_i outputs, ensuring $\sum_{j=1}^{w} \alpha_{ij} = 1$.

In **Global Temporal Enhancement (GTE)**, we employ a Graph Convolutional Network [16] to preserve global temporal dependencies. We define the nodes of the GCN as the group-level features $\{\tilde{\mathcal{N}}_i\}_{i=1}^{t/w}$ and update each node by capturing relationships among them:

$$\tilde{\mathcal{N}}'_i = \text{GELU}\left(\sum_{k=1}^{t/w} \hat{A}_{ik}(\tilde{\mathcal{N}}_k W)\right) \tag{2}$$

where $\tilde{\mathcal{N}}'_i$ is the updated i-th node feature, $\tilde{\mathcal{N}}_k$ is the k-th node feature, $\hat{A}_{ik}$ is the normalized adjacency matrix, and W is a learnable weight matrix. Finally, each $\tilde{\mathcal{N}}'_i$ is mapped to a single codebook entry through quantization. By preserving temporal information, our approach encodes longer motion sequences into single codebook entries with reduced information loss, thereby decreasing model size and inference time. Related analysis is provided in Tab. 4 and Sec. 4.4.

Training. Our VQ-VAE is optimized by $\mathcal{L}_{vq} = \mathcal{L}_{rec} + \lambda_{app} \cdot \mathcal{L}_{app}$, where $\mathcal{L}_{rec}$ is the L_1 reconstruction loss between decoded and ground truth joint positions, and $\mathcal{L}_{app}$ is the L_2 approximation loss between quantized codebook vectors and the encoded vectors, which encourages the encoder to produce features close to learned codebook entries.

3.2. Part-aware Text Grounding

After completing the motion quantization stage, in the second stage, we proceed to motion generation using transformers. Before feeding text embedding into each part's transformer, we apply Part-aware Text Grounding (PTG) to generate text embeddings tailored to each body part. As shown in Fig. 3, we first obtain text embedding $\mathbf{c}$ from the text description through CLIP [29]. This embedding is fed into K distinct MLPs, producing transformed embeddings $\mathbf{c}'_n = \text{MLP}_n(\mathbf{c})_{n\in\{1,\dots,K\}}$. To ensure these embeddings maintain semantic consistency while achieving diversity, we employ contrastive learning where each $\mathbf{c}'_n$ treats $\mathbf{c}$ as a positive anchor and $\{\mathbf{c}'_m\}_{m\neq n}^{K}$ as negatives. The text

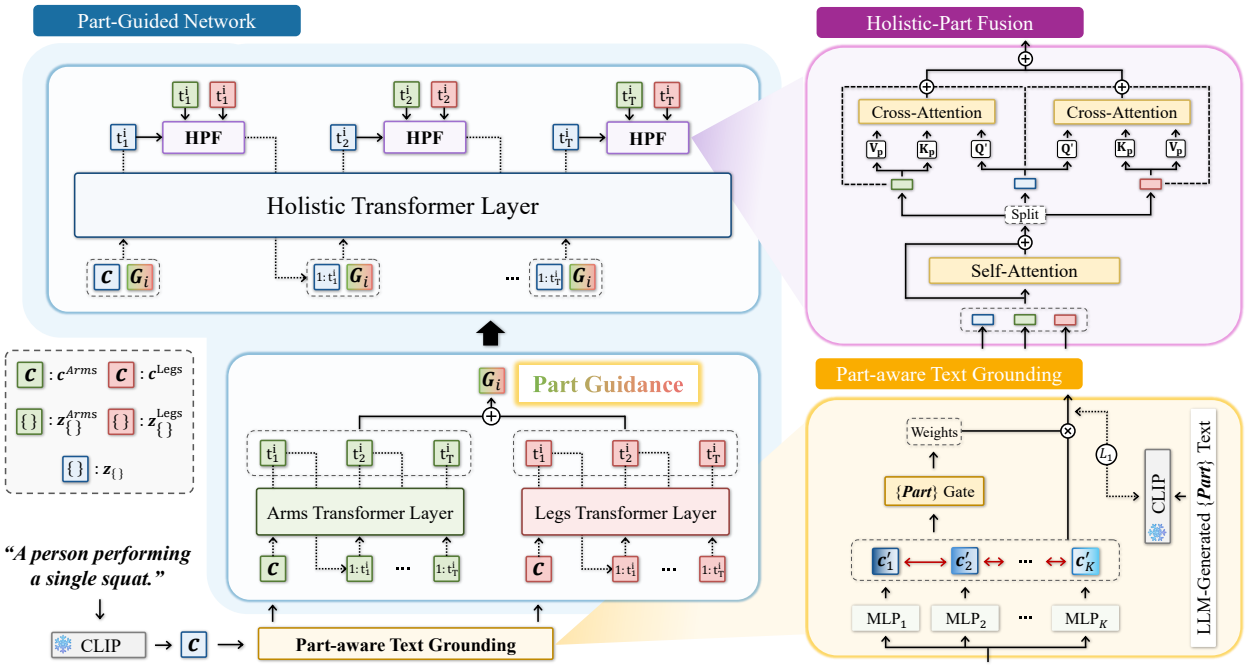

Figure 3. Overview of ParTY. Text embeddings are processed through Part-aware Text Grounding, then part transformers generate Part Guidance for the holistic transformer to generate motion tokens, with Holistic-Part Fusion applied during generation. The notation {***Part***} indicates that the process is performed for both arms and legs.

diversity loss is:

$$\mathcal{L}_{\text{div}} = \frac{1}{K}\sum_{n=1}^{K}\mathcal{L}^{(n)}$$
$$\mathcal{L}^{(n)} = -\log\frac{\exp(\text{s}(\mathbf{c}'_n,\mathbf{c})/\tau)}{\exp(\text{s}(\mathbf{c}'_n,\mathbf{c})/\tau)+\sum_{m\neq n}^{K}\exp(\text{s}(\mathbf{c}'_n,\mathbf{c}'_m)/\tau)} \quad (3)$$

where $\text{s}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter. This encourages each MLP to explore different semantic aspects while preserving core meaning. After generating diverse embeddings, the {***Part***} Gate—a gating network dedicated to arms and legs—dynamically selects appropriate embeddings through adaptive weighting. To improve part-specific selection, we use LLM-Generated {***Part***} Text as auxiliary supervision: detailed motion descriptions for each part are generated from the original text (e.g., given *"a person walks forward and picks something up off the ground with their left hand,"* the LLM generates *"Left arm picks something up off the ground"* for the arm and *"The legs step forward one after the other"* for the legs), embedded via CLIP, and used to compute an auxiliary L_1 loss $\mathcal{L}_{\text{aux}}$ that aligns PTG outputs with these part-specific embeddings. Notably, since LLM-generated texts are only used during training, our method remains efficient at inference time.

3.3. Part-Guided Network

To achieve expressive motion generation, we propose a Part-Guided Network that first generates part motion tokens for a certain number of time steps and leverages them as part guidance for generating holistic motion tokens. Following PTG, the precisely aligned text embeddings are fed into their respective part transformers, which autoregressively generate part motion tokens. The generation proceeds in cycles: in the i-th cycle, each part transformer generates T consecutive tokens for time steps $t_i = \{t_1^i, t_2^i, \ldots, t_T^i\}$, and the generated tokens from each part are fused to create the i-th Part Guidance $\mathbf{G}_i$:

$$\mathbf{G}_i = \sum_{t\in t_i}\mathbf{z}_t^{\text{fuse}} \quad (4)$$
$$\mathbf{z}_t^{\text{fuse}} = \text{MLP}(\mathbf{z}_t^{\text{Arms}} + \mathbf{z}_t^{\text{Legs}}) \quad (5)$$
$$\mathbf{z}_t^{p} = f_p(\mathbf{z}_{1:t-1}^{p}, \mathbf{c}^{p}), \quad p \in \{\text{Arms}, \text{Legs}\} \quad (6)$$

where f_p denotes part transformer that autoregressively generates motion token $\mathbf{z}_t^p$ conditioned on previous tokens $\mathbf{z}_{1:t-1}^p$ and part-specific text embedding $\mathbf{c}^p$ from PTG.

Next, the i-th Part Guidance is fed into the holistic transformer at each step during the generation of holistic motion tokens for time steps in t_i. The holistic motion token $\mathbf{z}_t$ generated at time step $t \in t_i$ by the holistic transformer is

formulated as:

$$\mathbf{z}_t = f(\mathbf{z}_{1:t-1}, \mathbf{c}, \mathbf{G}_i) \tag{7}$$

$$\mathbf{z}_{1:t-1} = \texttt{HPF}(\mathbf{z}_{1:t-1}, \mathbf{z}^{\text{Arms}}_{1:t-1}, \mathbf{z}^{\text{Legs}}_{1:t-1}) \tag{8}$$

where $\mathbf{z}_{1:t-1}$ denotes the previous holistic motion tokens refined by the Holistic-Part Fusion (HPF), $\mathbf{c}$ is original text embedding without processing through the PTG module. Overall, this process of the i-th generation cycle—generating part motion tokens for T steps and subsequently generating holistic motion tokens for T steps—continues cyclically until the holistic transformer reaches the end token.

Holistic-Part Fusion. To ensure coherent movements across the entire body, we continuously integrate part motion information into the holistic transformer through the Holistic-Part Fusion (HPF) during holistic motion token generation. HPF first concatenates the holistic tokens $\mathbf{z}_{1:t-1}$ with arm tokens $\mathbf{z}^{\text{Arms}}_{1:t-1}$ and leg tokens $\mathbf{z}^{\text{Legs}}_{1:t-1}$, then performs self-attention using the standard scaled dot-product attention mechanism $\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$, where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are linear projections of the concatenated tokens $[\mathbf{z}_{1:t-1}; \mathbf{z}^{\text{Arms}}_{1:t-1}; \mathbf{z}^{\text{Legs}}_{1:t-1}]$. The attended output is then split back into separate sequences $\tilde{\mathbf{z}}_{1:t-1}$, $\tilde{\mathbf{z}}^{\text{Arms}}_{1:t-1}$, and $\tilde{\mathbf{z}}^{\text{Legs}}_{1:t-1}$ using split tokens, which are learnable vectors that serve as separators between different tokens. We apply cross-attention operations with $\tilde{\mathbf{z}}_{1:t-1}$ as query and each part token as key/value:

$$\mathbf{z}^{\text{p}}_{\text{cross}} = \text{Attn}(\mathbf{Q}', \mathbf{K}_{\text{p}}, \mathbf{V}_{\text{p}}), \quad p \in \{\text{Arms}, \text{Legs}\} \tag{9}$$

where $\mathbf{Q}'$ and $(\mathbf{K}_{\text{p}}, \mathbf{V}_{\text{p}})$ are linear projections of $\tilde{\mathbf{z}}_{1:t-1}$ and $\tilde{\mathbf{z}}^{\text{p}}_{1:t-1}$, respectively. Finally, the two cross-attention outputs are added to produce the HPF output.

Training. To train both holistic and part transformers, we design separate loss functions for each. Let $d(\mathbf{z}|t)$ denote the conditional distribution of $\mathbf{z}$ at time step t. The holistic motion loss and part motion loss are defined as:

$$\mathcal{L}_{\text{hol}} = \mathbb{E}_{\mathbf{z}_t, t\sim d(\mathbf{z}_t,t)}[-\log d(\mathbf{z}_t|t)] \tag{10}$$

$$\mathcal{L}_{\text{part}} = \sum_{p\in\{\text{Arms,Legs}\}} \mathbb{E}_{\mathbf{z}^p_t, t\sim d(\mathbf{z}^p_t,t)}[-\log d(\mathbf{z}^p_t|t)] \tag{11}$$

The total loss is $\mathcal{L} = \mathcal{L}_{\text{hol}} + \mathcal{L}_{\text{part}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}$, where λ_{div} and λ_{aux} are weighting coefficients for the text diversity loss and part text auxiliary loss, respectively.

4. Experiments

We evaluate on HumanML3D [10] (14,616 motions, 44,970 texts) and KIT-ML [27] (3,911 motions, 6,278 texts), following the standard splits and pose representation from [10]. Additional quantitative and qualitative experiment results can be found in the supplementary material.

4.1. Evaluation Metrics

We adopt evaluation metrics from [10] using pre-trained encoders. R-Precision and MM-Dist measure text-motion alignment and semantic similarity in the feature space. FID evaluates motion quality through distributional differences between generated and real motions. We also report Diversity (variance across motion pairs) and Multimodality (variance for motions from the same text). Following prior work, we run each evaluation 20 times (5 times for Multimodality) and report averages with 95% confidence intervals.

4.1.1. Part-level Evaluation Metrics

To evaluate part-specific expressiveness, we extend conventional evaluation metrics to the part-level. We independently train arms and legs motion encoders using the T2M [10] encoder architecture, which is commonly used for evaluation in most studies [25, 26, 37, 39, 40]. With these trained part-specific encoders, we compute part-level R-Precision, FID, and MM-Dist by adapting full-body metrics to evaluate part-level performance.

4.1.2. Coherence-level Evaluation Metrics

To evaluate motion coherence at the frame-level, we introduce the Temporal Coherence (TC) and Spatial Coherence (SC) score, which evaluate both temporal and spatial consistency across body parts. A motion sequence is represented by j joints with 3D position $\hat{\mathbf{p}}_j(t)$ at time step t, partitioned into five body parts: left arm, right arm, left leg, right leg, and backbone.

Temporal Coherence score quantifies temporal coordination between body parts over time. For each body part g, we compute the temporal-wise RMS velocity:

$$\mathbf{x}_g(t) = \sqrt{\frac{1}{n_g}\sum_{j\in g}\|\hat{\mathbf{p}}_j(t) - \hat{\mathbf{p}}_j(t-1)\|^2} \tag{12}$$

where the sum is over all joints j belonging to part g, and n_g is the number of joints from part g. To compare motion patterns across parts with different movement intensities, we apply z-normalization within sliding windows w—local time intervals that allow adaptive normalization to account for varying motion dynamics throughout the sequence. We then measure part-wise correlations within each window through cross-correlation functions $r^{(w)}_{g,h}(\tau)$ across temporal lags τ. This allows detection of phase-shifted synchrony—for instance, in walking, arm motion naturally leads or lags leg motion. To aggregate these correlations, we compute a refined correlation score:

$$\tilde{R}^{(w)}_{g,h} = \max(0, \mathbb{E}_\tau[r^{(w)}_{g,h}(\tau)]) \cdot \exp\left(-\frac{\langle|\tau|\rangle_w}{\kappa}\right) \tag{13}$$

where $\mathbb{E}_\tau[\cdot]$ denotes softmax-weighted averaging over lags (emphasizing high-correlation lags), $\langle|\tau|\rangle_w$ is the expected absolute lag, and the exponential term penalizes excessive

Table 1. Quantitative comparison on HumanML3D and KIT-ML. **Bold** indicates the best result, while underlined refers the second-best. The right arrow → indicates that closer values to ground truth are preferred.

Datasets	Method	R Precision ↑			FID↓	MM-Dist↓	Diversity→	MultiModality↑
		Top 1	Top 2	Top 3				
HumanML3D	Real motion	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
	MDM [31]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
	T2M-GPT [37]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	ParCo [40]	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$0.109^{\pm.005}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$	$1.382^{\pm.060}$
	MMM [26]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$	$1.164^{\pm.041}$
	BAMM [25]	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$2.919^{\pm.008}$	$9.717^{\pm.089}$	$1.687^{\pm.051}$
	MoMask [12]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
	ParTY (Ours)	$\mathbf{0.550}^{\pm.003}$	$\mathbf{0.744}^{\pm.003}$	$\mathbf{0.836}^{\pm.003}$	$\mathbf{0.035}^{\pm.002}$	$\mathbf{2.779}^{\pm.006}$	$\mathbf{9.534}^{\pm.066}$	$2.155^{\pm.046}$
KIT-ML	Real motion	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
	MDM [31]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.190^{\pm.022}$	$10.85^{\pm.109}$	$\mathbf{1.907}^{\pm.214}$
	T2M-GPT [37]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
	ParCo [40]	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$0.453^{\pm.027}$	$2.820^{\pm.028}$	$10.95^{\pm.094}$	$1.245^{\pm.022}$
	MMM [26]	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$0.316^{\pm.028}$	$2.977^{\pm.019}$	$10.91^{\pm.101}$	$1.232^{\pm.039}$
	BAMM [25]	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$2.723^{\pm.026}$	$\mathbf{11.01}^{\pm.094}$	$1.609^{\pm.065}$
	MoMask [12]	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	-	$1.131^{\pm.043}$
	ParTY (Ours)	$\mathbf{0.449}^{\pm.006}$	$\mathbf{0.680}^{\pm.007}$	$\mathbf{0.804}^{\pm.006}$	$\mathbf{0.155}^{\pm.014}$	$\mathbf{2.694}^{\pm.030}$	$11.21^{\pm.082}$	$1.166^{\pm.049}$

delays to suppress spurious matches from unrelated movements. The temporal coherence score is then $S_{\text{temporal}} = \frac{1}{W|\mathcal{P}|}\sum_{w=1}^{W}\sum_{(g,h)\in\mathcal{P}}\tilde{R}_{g,h}^{(w)}$, where W is the number of windows and $\mathcal{P}$ is the set of all part pairs, yielding a measure of overall rhythmic coordination.

Spatial Coherence score evaluates the physical plausibility of spatial relationships within each frame. For each body part g, we compute its representative position as the centroid (average 3D position) of all joints belonging to that part: $\mathbf{c}_g(t) = \frac{1}{n_g}\sum_{j\in g}\hat{\mathbf{p}}_j(t)$. We then measure: (1) inter-part distances $d_{g,h}(t) = \|\mathbf{c}_g(t) - \mathbf{c}_h(t)\|$ between the centroids of part pairs (g, h), and (2) part-torso angles $\theta_g(t)$ measuring the angular alignment between each part and the torso. To define what constitutes *physically plausible* human poses, we estimate reference statistics—means and standard deviations—of these quantities over the HumanML3D dataset [10], forming empirical distributions of natural human geometry. Each per-frame measurement is normalized into z-scores $z_{g,h}^{(d)}$ and $z_g^{(\theta)}$, which are converted to consistency scores using Gaussian kernels:

$$s_{g,h}^{(d)}(t) = \exp\left(-\frac{(z_{g,h}^{(d)}(t))^2}{\beta_d^2}\right), \quad s_g^{(\theta)}(t) = \exp\left(-\frac{(z_g^{(\theta)}(t))^2}{\beta_\theta^2}\right) \tag{14}$$

where larger deviations from typical human geometry lead to exponentially lower consistency. The spatial coherence score is defined as:

$$S_{\text{spatial}} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{(g,h)\in\mathcal{P}} s_{g,h}^{(d)}(t) + \sum_{g\in\mathcal{G}} s_g^{(\theta)}(t)\right] \tag{15}$$

Table 2. Quantitative comparison with **part-level evaluation** metrics on HumanML3D.

Method	Part	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
MoMask [12]	Arms	$0.452^{\pm.003}$	$0.761^{\pm.002}$	$0.175^{\pm.003}$	$3.440^{\pm.006}$
	Legs	$0.403^{\pm.003}$	$0.687^{\pm.003}$	$0.104^{\pm.003}$	$3.513^{\pm.009}$
ParCo [40]	Arms	$0.468^{\pm.003}$	$0.767^{\pm.003}$	$0.215^{\pm.003}$	$3.326^{\pm.008}$
	Legs	$0.407^{\pm.003}$	$0.699^{\pm.002}$	$0.118^{\pm.003}$	$3.482^{\pm.011}$
Ours	Arms	$\mathbf{0.506}^{\pm.003}$	$\mathbf{0.802}^{\pm.002}$	$\mathbf{0.133}^{\pm.002}$	$\mathbf{3.079}^{\pm.005}$
	Legs	$\mathbf{0.463}^{\pm.003}$	$\mathbf{0.755}^{\pm.003}$	$\mathbf{0.078}^{\pm.003}$	$\mathbf{3.122}^{\pm.008}$

where T is the total number of frames, and $\mathcal{G}$ is the set of body parts, averaging consistency scores across all frames and spatial relationships. TC and SC provide a comprehensive assessment of motion quality by capturing both temporal coordination and spatial plausibility of generated human motions.

4.2. Quantitative Evaluation

As shown in Tab. 1, ParTY achieves state-of-the-art performance in R-Precision, MM-Dist, and FID metrics across both HumanML3D and KIT-ML datasets. The superior R-Precision and MM-Dist results demonstrate more precise text-motion alignment, while the leading FID scores confirm enhanced motion quality.

For part-level evaluation, as shown in Tab. 2, ParTY outperforms both the existing part-wise method ParCo [40] and the holistic method MoMask [12] across all body parts in R-Precision, MM-Dist, and FID metrics. These results demonstrate that ParTY successfully achieves expressive part-level motion generation, overcoming a fundamental

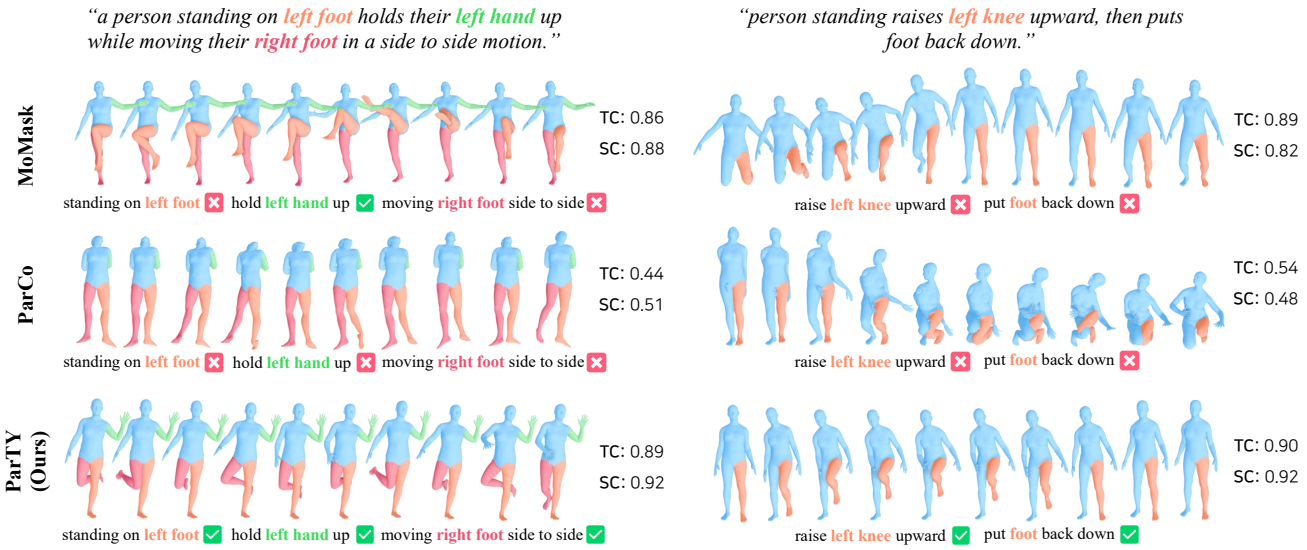

Figure 4. Qualitative comparison on HumanML3D. Colored text in the descriptions corresponds to the colored body parts in the generated motions, with coherence-level (TC, SC) scores displayed for each sample.

Table 3. Quantitative comparison with **coherence-level (TC, SC) scores** on HumanML3D. We run each evaluation 20 times and report averages with 95% confidence intervals.

Method	Temporal Coherence (TC)↑	Spatial Coherence (SC)↑
ParCo [40]	$0.49^{\pm.062}$	$0.59^{\pm.057}$
MoMask [12]	$0.84^{\pm.047}$	$0.90^{\pm.044}$
Ours	$\mathbf{0.88}^{\pm.051}$	$\mathbf{0.92}^{\pm.041}$

challenge common to both part-wise and holistic methods.

For coherence-level evaluation, we measure both temporal and spatial coherence scores and report the results in Tab. 3. Consistent with our problem statement, ParCo, a part-wise method, shows low scores across both coherence metrics, while MoMask, a holistic method, achieves substantially more stable scores. This validates that our proposed coherence-level metrics effectively capture the coherence deficiency in part-wise methods, which was our key observation. Notably, ParTY outperforms not only ParCo but also achieves marginally better scores than MoMask, demonstrating that our Part-Guided Network with Holistic-Part Fusion module successfully addresses the coherence issues inherent in part-wise methods.

4.3. Qualitative Evaluation

We conduct a visual comparison of motion generation results between our ParTY, ParCo [40], and MoMask [12]. As shown in Fig. 4, both MoMask and ParCo fail to accurately reflect part-specific descriptions, whereas our ParTY accurately captures all part descriptions. Moreover, in terms of motion coherence, the part-wise method ParCo exhibits significant issues such as neck distortion and misaligned upper and lower body orientations, resulting in low temporal and spatial coherence scores. In contrast, ParTY maintains coherent motions across all frames, achieving high temporal and spatial coherence scores. These results demonstrate that ParTY effectively resolves the existing trade-off between part-specific expressiveness and full-body coherence.

Table 4. Porting Temporal-aware VQ-VAE to MoMask [12]. Reconstruction evaluates VQ-VAE performance, while Generation evaluates final performance including the transformer. Mean Per Joint Position Error (MPJPE) measures positional accuracy, and Average Inference Time (AIT) is averaged over 100 samples on an RTX A5000 GPU.

Method	Window size	Reconstruction			Generation	
		# Params	FID↓	MPJPE↓	FID↓	AIT
MoMask	4	19.44M	0.020	0.030	0.045	80ms
+ Ours	4	19.86M	0.003 (+85%)	0.011 (+63%)	0.033 (+26%)	
MoMask	8	10.15M	0.042	0.055	0.094	43ms
+ Ours	8	10.58M	0.005 (+88%)	0.014 (+74%)	0.039 (+58%)	(+46%)
MoMask	12	7.67M	0.079	0.091	0.126	29ms
+ Ours	12	8.09M	0.011 (+86%)	0.023 (+75%)	0.042 (+67%)	(+64%)

4.4. Discussions

All experiments were performed on the HumanML3D [10] dataset. More detailed experimental results and analysis are provided in the supplementary material.

Cost Efficiency of Temporal-aware VQ-VAE. In VQ-VAE-based approaches, the window size—which defines how many consecutive frames are encoded into a single codebook entry—critically impacts both model performance and inference efficiency. To demonstrate the cost efficiency of proposed Temporal-aware VQ-VAE, we apply it to MoMask [12] and report parameters, reconstruction, and generation performance across different window sizes

Table 5. Ablation studies of the proposed components. PG indicates Part Guidance.

PG	PTG	HPF	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
			$0.494^{\pm.003}$	$0.780^{\pm.003}$	$0.158^{\pm.005}$	$3.087^{\pm.008}$
✓			$0.520^{\pm.002}$	$0.802^{\pm.003}$	$0.086^{\pm.003}$	$2.913^{\pm.010}$
✓	✓		$0.545^{\pm.003}$	$0.828^{\pm.003}$	$0.051^{\pm.003}$	$2.799^{\pm.008}$
✓	✓	✓	$\mathbf{0.550}^{\pm.003}$	$\mathbf{0.836}^{\pm.002}$	$\mathbf{0.035}^{\pm.002}$	$\mathbf{2.779}^{\pm.006}$

Table 6. Ablation studies with **part-level evaluation** metrics.

Part	PG	PTG	HPF	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
Arms				$0.433^{\pm.003}$	$0.736^{\pm.002}$	$0.232^{\pm.004}$	$3.347^{\pm.014}$
	✓			$0.470^{\pm.003}$	$0.769^{\pm.003}$	$0.166^{\pm.002}$	$3.251^{\pm.006}$
	✓	✓		$0.501^{\pm.003}$	$0.798^{\pm.003}$	$0.152^{\pm.002}$	$3.102^{\pm.007}$
	✓	✓	✓	$\mathbf{0.506}^{\pm.003}$	$\mathbf{0.802}^{\pm.002}$	$\mathbf{0.133}^{\pm.002}$	$\mathbf{3.079}^{\pm.005}$

in Tab. 4. First, when porting with the same window size of 4, the LTE and GTE modules add minimal parameters but yield notable performance improvements. For MoMask-only, increasing the window size substantially reduces both model parameters and average inference time (AIT), but at the cost of significant performance degradation due to mapping longer motion sequences to each codebook. In contrast, our Temporal-aware VQ-VAE effectively prevents this degradation by capturing and preserving critical temporal information. While our Part-Guided Network inevitably increases inference time due to separate part transformers, the Temporal-aware VQ-VAE enables larger window sizes without performance loss, compensating for the increased inference cost.

Analysis of Part-Guided Network. All transformers in our method autoregressively generate motion tokens, relying solely on information up to the current time step. Through Part Guidance (PG), previously generated part motion tokens are leveraged to guide holistic motion generation. The effectiveness of this approach is demonstrated in Tab. 5, where adding PG leads to notable improvements across all metrics, showing that part information serves as valuable guidance for holistic motion generation. Furthermore, as shown in Tab. 6, utilizing PG effectively incorporates part-level information, positively impacting part expressiveness. Additionally, Holistic-Part Fusion (HPF) enables the generation of more part-aware holistic motions by adaptively fusing holistic and part motions throughout the generation process. As shown in Fig. 5, for two text descriptions emphasizing different body parts, the attention scores are notably higher for the corresponding parts. This demonstrates that the HPF module effectively captures dynamic inter-part relationships and performs appropriate fusion by adaptively attending to task-relevant body parts throughout the motion sequence.

Analysis of Part-aware Text Grounding. As shown in

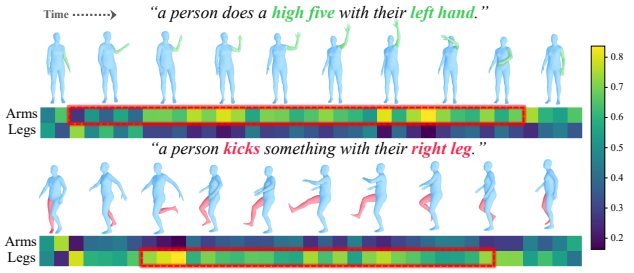

Figure 5. Visualization of cross attention map of HPF. Rows correspond to body parts and columns represent temporal frames. We visualize the normalized attention weights between the holistic motion token and each part motion token.

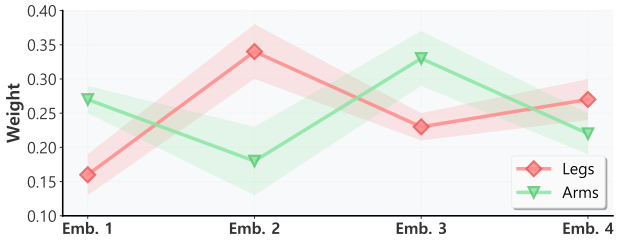

Figure 6. Embedding selection ratios in PTG. Mean and standard deviation of weights are computed over semantically similar text descriptions that share common motion patterns.

Tab. 5 and Tab. 6, PTG improves text-motion alignment metrics (R-Precision and MM-Dist) and part expressiveness, confirming that it enables more precise correspondence between textual descriptions and part motions. Fig. 6 demonstrates how different embeddings are weighted and selected for specific body parts given the same input text. We adopt 4 embeddings based on experimental validation showing optimal performance. Legs show high selection ratios for embedding 2, while arms favor embedding 3, confirming that PTG effectively transforms textual embeddings into multiple specialized representations and dynamically selects the most relevant features for each body part.

5. Conclusion

We present ParTY, a novel model for text-driven motion generation that addresses the fundamental trade-off in existing methods. Our approach enhances part-text alignment through part-specific selection of diverse text embeddings and maintains coherent motions via a part-guided generation framework. Moreover, we introduce part-level and coherence-level evaluation metrics to comprehensively validate our model. Extensive experiments on HumanML3D and KIT-ML demonstrate that ParTY achieves state-of-the-art performance, outperforming both holistic and part-wise methods. By bridging the gap between expressive part motion and full-body motion coherence, ParTY establishes a new standard for text-to-motion generation.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT)(RS-2024-00456589).

References

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

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

[3] André Antakli, Erik Hermann, Ingo Zinnikus, Han Du, and Klaus Fischer. Intelligent distributed human motion simulation in human-robot collaboration environments. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 319–324, 2018. 1

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2

[5] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021. 2

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[7] Xuehao Gao, Yang Yang, Zhenyu Xie, Shaoyi Du, Zhongqian Sun, and Yang Wu. Guess: Gradually enriching synthesis for text-driven human motion generation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[8] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 2

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

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 2, 5, 6, 7

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 2, 3, 6, 7

[13] Seyed Rohollah Hosseyni et al. Bad: Bidirectional autoregressive diffusion for text-to-motion generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. 1, 3

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2

[15] Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus Magnor. High-fidelity neural human motion transfer from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1541–1550, 2021. 1

[16] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[17] Hema Koppula and Ashutosh Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *International conference on machine learning*, pages 792–800. PMLR, 2013. 1

[18] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[19] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 2

[20] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 2

[21] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022. 2

[22] Dennis Majoe, Lars Widmer, and Juerg Gutknecht. Enhanced motion interaction for multimedia applications. In *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, pages 13–19, 2009. 1

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

[24] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[25] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024. 1, 2, 3, 5, 6

[26] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 1, 2, 3, 5, 6

[27] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. 5

[28] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 2

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

[30] Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. Lgtm: Local-to-global text-driven human motion diffusion model. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 1, 2

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 6

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[33] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[35] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[36] Mohammed Yeasin, Ediz Polat, and Rajeev Sharma. A multiobject tracking framework for interactive multimedia applications. *IEEE transactions on multimedia*, 6(3):398–405, 2004. 1

[37] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 1, 2, 3, 5, 6

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

[39] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 1, 2, 5

[40] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. *arXiv preprint arXiv:2403.18512*, 2024. 1, 2, 3, 5, 6, 7