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

Supplementary Material

Contents

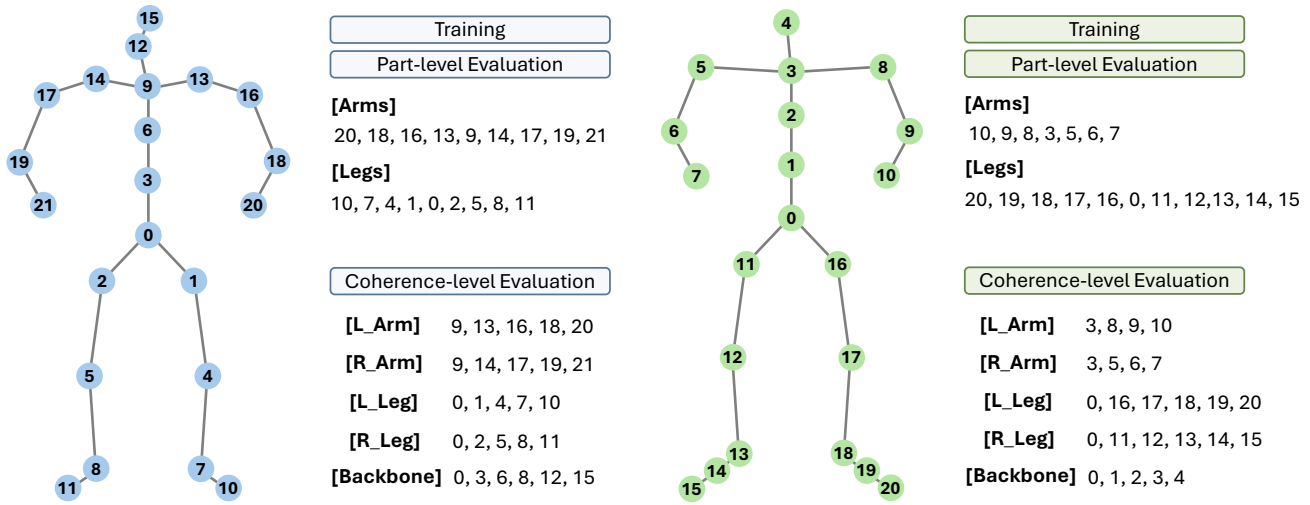
A. Implementation Details

A.1. Networks & Training

Temporal-aware VQ-VAE utilizes a codebook with 256 entries and a code dimension of 128. In the Local Temporal Enhancement (LTE) module, we set the window size to 8. For Global Temporal Enhancement (GTE), we adopt a 3-layer GCN, where each layer has a hidden dimension of 128. The part VQ-VAE uses a codebook with 128 entries and the same code dimension of 128. In its LTE module, the window size is set to 12, while the GTE configuration remains identical to that of the holistic VQ-VAE. For training, we employ the AdamW [10] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate follows a two-stage schedule: 2×10^{-4} for the first 200K iterations, and 1×10^{-5} for the remaining iterations. We use a batch size of 256 and set the approximation loss weight to $\mathcal{L}_{app} = 1.0$.

Part-aware Text Grounding module employs 4 MLPs (3-layer with ReLU activation) to diversely transform the text embeddings. This configuration was determined to yield optimal performance through extensive experimental validation, as detailed in Sec. E.4.1. For the contrastive learning objective, we set the temperature parameter to $\tau = 0.05$. In the overall loss function, the diversity loss $\mathcal{L}_{div}$ is weighted by $\lambda_{div} = 0.1$. The gating network for the Part Gate is implemented as a simple linear layer. For LLM-generated part text supervision, we apply the auxiliary loss $\mathcal{L}_{\text{aux}}$ with a weighting factor of $\lambda_{aux} = 0.1$.

Part-Guided Network constructs Part Guidance from three time steps of the part transformer, a design choice chosen based on extensive experimental validation (Sec. E.5.1), and employs a holistic transformer with 10 layers and a token dimension of 128, as well as a part transformer with 8 layers and the same token dimension of 128. In the **Holistic-Part Fusion** module, both the self-attention and cross-attention layers use 6 attention heads with a head dimension of 64. For training, we also use the AdamW optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The learning rate is set to 2×10^{-4} for the initial 100K iterations and then reduced to 5×10^{-6} for the remaining iterations. We use a batch size of 64 to accommodate the memory requirements of the transformer architecture, and all experiments are conducted on a single NVIDIA A5000 GPU.



(a) Part Division for HumanML3D

(b) Part Division for KIT-ML

Figure S1. Body part division for HumanML3D and KIT-ML.

A.2. Body Part Division

Fig. S1 illustrates how we construct the Legs and Arms for training and part-level evaluation on both the HumanML3D [3] and KIT-ML [13] datasets. It also shows how the five parts (Left Arm, Right Arm, Left Leg, Right Leg, Backbone) are defined for coherence-level evaluation.

B. Part-level Evaluation Metrics Details

B.1. Design Details

T2M [3] proposes a motion encoder for evaluating generated motions; however, this encoder is designed for assessing holistic motion. To enable part-level evaluation, we train this motion encoder separately on arms and legs data, resulting in specialized part motion encoders capable of evaluating motion quality at the part level. R-Precision, FID, and MM-Dist are computed following the same evaluation protocol as in T2M [3].

B.2. Quantitative Results & Ablation Studies

We demonstrate the generality of our metrics by reporting part-level evaluation results on additional models [15, 22] in Tab. S1. Furthermore, Tab. S2 reports the part-level evaluation results for the legs in our proposed methods.

Table S1. Quantitative comparison with **part-level evaluation** metrics on HumanML3D.

Method	Part	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
MoMask [5]	Arms	$0.452^{\pm.003}$	$0.761^{\pm.002}$	$0.175^{\pm.003}$	$3.440^{\pm.006}$
	Legs	$0.403^{\pm.003}$	$0.687^{\pm.003}$	$0.104^{\pm.003}$	$3.513^{\pm.009}$
AttT2M [22]	Arms	$0.431^{\pm.003}$	$0.738^{\pm.002}$	$0.264^{\pm.002}$	$3.489^{\pm.010}$
	Legs	$0.375^{\pm.003}$	$0.656^{\pm.003}$	$0.190^{\pm.003}$	$3.541^{\pm.011}$
ParCo [23]	Arms	$0.468^{\pm.003}$	$0.767^{\pm.003}$	$0.215^{\pm.003}$	$3.326^{\pm.008}$
	Legs	$0.407^{\pm.003}$	$0.699^{\pm.002}$	$0.118^{\pm.003}$	$3.482^{\pm.011}$
LGTM [15]	Arms	$0.436^{\pm.003}$	$0.745^{\pm.003}$	$0.398^{\pm.002}$	$3.421^{\pm.015}$
	Legs	$0.384^{\pm.003}$	$0.689^{\pm.003}$	$0.325^{\pm.002}$	$3.547^{\pm.018}$
Ours	Arms	$\mathbf{0.506}^{\pm.003}$	$\mathbf{0.802}^{\pm.002}$	$\mathbf{0.133}^{\pm.002}$	$\mathbf{3.079}^{\pm.005}$
	Legs	$\mathbf{0.463}^{\pm.003}$	$\mathbf{0.755}^{\pm.003}$	$\mathbf{0.078}^{\pm.003}$	$\mathbf{3.122}^{\pm.008}$

Table S2. Ablation studies with **part-level evaluation** metrics on HumanML3D.

Part	PG	PTG	HPF	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
Legs				$0.397^{\pm.003}$	$0.691^{\pm.003}$	$0.169^{\pm.003}$	$3.416^{\pm.012}$
	✓			$0.422^{\pm.003}$	$0.715^{\pm.003}$	$0.114^{\pm.003}$	$3.328^{\pm.011}$
	✓	✓		$0.456^{\pm.002}$	$0.744^{\pm.003}$	$0.095^{\pm.003}$	$3.175^{\pm.006}$
	✓	✓	✓	$\mathbf{0.463}^{\pm.003}$	$\mathbf{0.755}^{\pm.003}$	$\mathbf{0.078}^{\pm.003}$	$\mathbf{3.122}^{\pm.008}$

C. Coherence-level Evaluation Metrics Details

C.1. Design Details

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. To quantify the instanta-

neous motion strength of each part in a noise-robust manner, we compute the temporal-wise root mean square (RMS) velocity for each body part g, defined as:

$$\mathbf{x}_g(t) = \sqrt{\frac{1}{n_g} \sum_{j \in g} \|\hat{\mathbf{p}}_j(t) - \hat{\mathbf{p}}_j(t-1)\|^2} \tag{S1}$$

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 of length L frames with stride $L/2$ frames—local time intervals that allow adaptive normalization to account for varying motion dynamics throughout the sequence. This standardization (zero mean, unit variance) ensures that the subsequent cross-correlation reflects only the relative phase and shape similarity of motion signals rather than their absolute magnitudes:

$$\bar{x}_g^{(w)} = \frac{1}{|\mathcal{T}_w|} \sum_{t \in \mathcal{T}_w} x_g(t), \quad \sigma_g^{(w)} = \sqrt{\frac{1}{|\mathcal{T}_w|} \sum_{t \in \mathcal{T}_w} (x_g(t) - \bar{x}_g^{(w)})^2}$$
$$s_g^{(w)}(t) = \frac{x_g(t) - \bar{x}_g^{(w)}}{\sigma_g^{(w)} + \varepsilon}, \quad t \in \mathcal{T}_w \tag{S2}$$

where $\mathcal{T}_w$ is the set of frame indices in the w-th sliding window, $\bar{x}_g^{(w)}$ and $\sigma_g^{(w)}$ denote the local mean and standard deviation of the part-wise RMS velocity, and ε is a small constant for numerical stability. For each part pair (g, h), we compute the pairwise cross-correlation:

$$r_{g,h}^{(w)}(\tau) = \frac{\sum_{t \in \mathcal{T}_w} s_g^{(w)}(t) s_h^{(w)}(t+\tau)}{\sqrt{\sum_{t \in \mathcal{T}_w} (s_g^{(w)}(t))^2} \sqrt{\sum_{t \in \mathcal{T}_w} (s_h^{(w)}(t))^2}} \tag{S3}$$

where $\tau \in [-\tau_{\max}, \tau_{\max}]$ represents the temporal lag (time shift in frames) between the motion signals of parts g and h. By computing cross-correlation across multiple lags, we can detect phase-shifted synchrony—coordinated movements even when body parts move with natural timing offsets—for instance, in walking, arm motion naturally leads or lags leg motion. We aggregate correlation values across all lags using a softmax-weighted average, retain only positive correlations (in-phase movements), and apply an exponential penalty based on the expected lag magnitude to suppress spurious correlations from unrelated movements or excessive delays:

$$R_{g,h}^{(w)} = \mathbb{E}_\tau \left[r_{g,h}^{(w)}(\tau) \right], \quad \langle |\tau| \rangle_w = \mathbb{E}_\tau \left[|\tau| \right] \tag{S4}$$

$$\tilde{R}_{g,h}^{(w)} = \max(0, R_{g,h}^{(w)}) \cdot \exp\left(-\frac{\langle |\tau| \rangle_w}{\kappa} \right) \tag{S5}$$

where $\mathbb{E}_\tau[\cdot] = \frac{\sum_\tau (\cdot) \exp(r_{g,h}^{(w)}(\tau)/\sigma)}{\sum_\tau \exp(r_{g,h}^{(w)}(\tau)/\sigma)}$ denotes the softmax-weighted expectation over temporal lags with temperature σ, $\langle |\tau| \rangle_w$ measures the expected absolute lag (average timing offset between coordinated movements), and κ controls the strength of the delay penalty. The temporal coherence score is then computed as:

$$S_{\text{temporal}} = \frac{1}{W} \sum_{w=1}^{W} \frac{1}{|\mathcal{P}|} \sum_{(g,h) \in \mathcal{P}} \tilde{R}_{g,h}^{(w)} \tag{S6}$$

where W is the total number of sliding windows, $\mathcal{P}$ denotes the set of all body part pairs, yielding a measure of overall rhythmic coordination.

Spatial Coherence score evaluates the physical plausibility of spatial relationships within each frame. To capture the overall spatial configuration of each part while being robust to joint-level noise, we define the representative position of part g as the average position of its joints:

$$\mathbf{c}_g(t) = \frac{1}{n_g} \sum_{j \in g} \hat{\mathbf{p}}_j(t) \tag{S7}$$

Using these representative points, we measure inter-part distances $d_{g,h}(t) = \|\mathbf{c}_g(t) - \mathbf{c}_h(t)\|$ to assess global spatial relationships and angles relative to the torso to evaluate local anatomical plausibility and articulation consistency:

$$\mathbf{u}_g(t) = \frac{\Delta \hat{\mathbf{p}}_g(t)}{\|\Delta \hat{\mathbf{p}}_g(t)\|}, \quad \theta_g(t) = \arccos(\mathbf{u}_g(t) \cdot \mathbf{u}_{\text{TR}}(t)) \tag{S8}$$

where $\Delta \hat{\mathbf{p}}_g(t) = \hat{\mathbf{p}}_{\text{end}}(t) - \mathbf{c}_g(t)$ is the part direction vector from $\mathbf{c}_g(t)$ to the end joint of part g, $\mathbf{u}_g(t)$ and $\mathbf{u}_{\text{TR}}(t)$ are the normalized part and torso orientation vectors, and $\theta_g(t)$ is the articulation angle.

Using statistics computed from the HumanML3D [3], we normalize the distance and angle measurements to obtain z-scores that quantify how much each frame deviates from typical human motion patterns. The normalized scores are then converted to spatial consistency scores using Gaussian kernels:

$$z_{g,h}^{(d)}(t) = \frac{d_{g,h}(t) - \mu_{g,h}^{(d)}}{\sigma_{g,h}^{(d)} + \varepsilon}, \quad z_g^{(\theta)}(t) = \frac{\theta_g(t) - \mu_g^{(\theta)}}{\sigma_g^{(\theta)} + \varepsilon}$$
$$s_{g,h}^{(d)}(t) = \exp\left(-\frac{(z_{g,h}^{(d)}(t))^2}{\beta_d^2} \right), \quad s_g^{(\theta)}(t) = \exp\left(-\frac{(z_g^{(\theta)}(t))^2}{\beta_\theta^2} \right) \tag{S9}$$

where $\mu_{g,h}^{(d)}$ and $\sigma_{g,h}^{(d)}$ are the mean and standard deviation of inter-part distance $d_{g,h}$ in HumanML3D, $\mu_g^{(\theta)}$ and $\sigma_g^{(\theta)}$ are the statistics for articulation angle θ_g, ε is a small constant

Table S3. Ablation studies on σ and κ parameters.

Hyperparameters	Temporal Coherence ↑
$\sigma = 0.05, \kappa = 3$	0.92
$\sigma = 0.05, \kappa = 5$	0.93
$\sigma = 0.05, \kappa = 10$	0.93
$\sigma = 0.05, \kappa = 15$	0.91
$\sigma = 0.1, \kappa = 3$	0.95
$\boldsymbol{\sigma = 0.1, \kappa = 5}$	**0.96**
$\sigma = 0.1, \kappa = 10$	0.94
$\sigma = 0.1, \kappa = 15$	0.93
$\sigma = 0.2, \kappa = 3$	0.93
$\sigma = 0.2, \kappa = 5$	0.94
$\sigma = 0.2, \kappa = 10$	0.92
$\sigma = 0.2, \kappa = 15$	0.91
$\sigma = 0.5, \kappa = 3$	0.89
$\sigma = 0.5, \kappa = 5$	0.91
$\sigma = 0.5, \kappa = 10$	0.88
$\sigma = 0.5, \kappa = 15$	0.86

Table S4. Ablation studies on L and τ_{max} parameters.

Method	Temporal Coherence ↑
$L = 20, \tau_{\text{max}} = 10$	0.95
$\boldsymbol{L = 20, \tau_{\text{max}} = 15}$	**0.96**
$\boldsymbol{L = 20, \tau_{\text{max}} = 20}$	**0.96**
$L = 20, \tau_{\text{max}} = 30$	0.94
$L = 30, \tau_{\text{max}} = 10$	0.94
$L = 30, \tau_{\text{max}} = 15$	0.95
$L = 30, \tau_{\text{max}} = 20$	0.93
$L = 30, \tau_{\text{max}} = 30$	0.90
$L = 40, \tau_{\text{max}} = 10$	0.89
$L = 40, \tau_{\text{max}} = 15$	0.89
$L = 40, \tau_{\text{max}} = 20$	0.86
$L = 40, \tau_{\text{max}} = 30$	0.82

Table S5. Ablation studies on β_d and β_θ parameters.

β_d	β_θ	Spatial Coherence ↑
1.0	1.0	0.97
1.5	**1.5**	**0.99**
2.0	2.0	0.98
3.0	3.0	0.94

for numerical stability, and β_d and β_θ are bandwidth parameters controlling the sensitivity to deviations. The spatial coherence score is then computed by averaging these consistency scores across all frames:

Table S6. Quantitative comparison with **coherence-level (TC, SC) scores** on HumanML3D.

Method	Temporal Coherence (TC)↑	Spatial Coherence (SC)↑
Real motion	$0.96^{\pm.032}$	$0.99^{\pm.026}$
ParCo [23]	$0.49^{\pm.062}$	$0.59^{\pm.057}$
LGTM [15]	$0.51^{\pm.048}$	$0.65^{\pm.039}$
T2M-GPT [19]	$0.85^{\pm.050}$	$0.87^{\pm.036}$
MoMask [5]	$0.84^{\pm.047}$	$0.90^{\pm.044}$
Ours	$\mathbf{0.88}^{\pm.051}$	$\mathbf{0.92}^{\pm.041}$

Table S7. Quantitative comparison with **coherence-level (TC, SC) scores** on KIT-ML.

Method	Temporal Coherence (TC)↑	Spatial Coherence (SC)↑
Real motion	$0.87^{\pm.045}$	$0.91^{\pm.037}$
ParCo [23]	$0.32^{\pm.066}$	$0.44^{\pm.062}$
LGTM [15]	$0.40^{\pm.061}$	$0.49^{\pm.058}$
T2M-GPT [19]	$0.77^{\pm.057}$	$0.80^{\pm.049}$
MoMask [5]	$0.79^{\pm.042}$	$0.79^{\pm.065}$
Ours	$\mathbf{0.80}^{\pm.048}$	$\mathbf{0.81}^{\pm.054}$

$$S_{\text{spatial}} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{(g,h)\in\mathcal{P}} s_{g,h}^{(d)}(t) + \sum_{g\in\mathcal{G}} s_{g}^{(\theta)}(t)\right] \quad \text{(S10)}$$

where T is the total number of frames, and $\mathcal{G}$ represents the set of all body parts.

C.2. Hyperparameter Analysis

We experimented with various combinations of hyperparameters in our coherence-level evaluation metrics (TC, SC) to find optimal values against ground truth motions from HumanML3D. Results for Temporal Coherence are reported in Tab. S3 and Tab. S4, while results for Spatial Coherence are reported in Tab. S5.

C.3. Quantitative Results

We demonstrate the generality of our metrics by reporting coherence-level evaluation results on additional models [15, 19] in Tab. S6. Tab. S7 presents the results on an additional dataset, KIT-ML [13], following the same evaluation process as HumanML3D [3]. The optimal hyperparameters for KIT-ML are $(\sigma = 0.05,\ \kappa = 5)$, $(L = 20,\ \tau_{\text{max}} = 15)$, and $(\beta_d = \beta_\theta = 1.5)$.

C.4. Ablation Studies

We provide ablation study results for the proposed components on coherence-level evaluation in Tab. S8. The results demonstrate that our ParTY, which leverages a Part-Guided Network, successfully preserves coherence while enhancing part expressiveness, compared to conventional part-wise

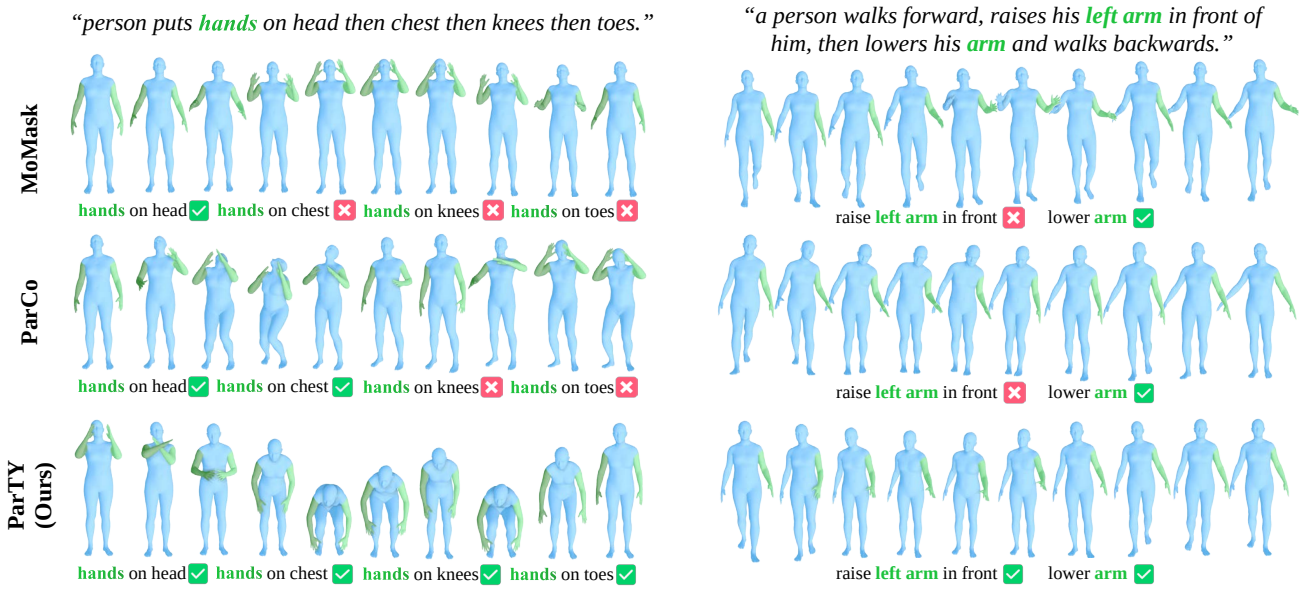


Figure S2. Qualitative comparison on part-text alignment on HumanML3D.

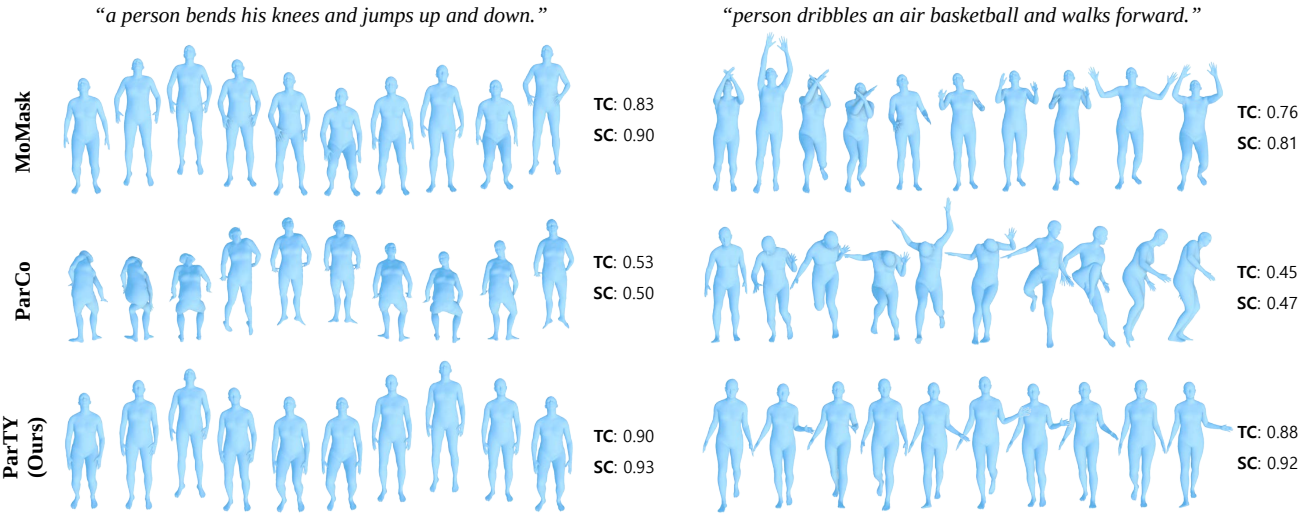


Figure S3. Qualitative comparison on coherence on HumanML3D.

Table S8. Ablation studies with **coherence-level evaluation** metrics on HumanML3D.

PG	PTG	HPF	Temporal Coherence↑	Spatial Coherence↑
			$0.43^{\pm.056}$	$0.54^{\pm.044}$
✓			$0.81^{\pm.048}$	$0.86^{\pm.029}$
✓	✓		$0.83^{\pm.045}$	$0.89^{\pm.034}$
✓	✓	✓	$\mathbf{0.88}^{\pm.051}$	$\mathbf{0.92}^{\pm.041}$

approaches that independently generate part motions and simply integrate them.

D. Additional Qualitative Results

D.1. Part-Text Alignment & Coherence

Fig. S2 and Fig. S3 show additional samples for part-text alignment and coherence, respectively. ParTY demonstrates strong performance in both aspects.

D.2. Long & Complex

Fig. S4 shows that ParTY accurately handles each action throughout long and complex motions (over 170 frames).

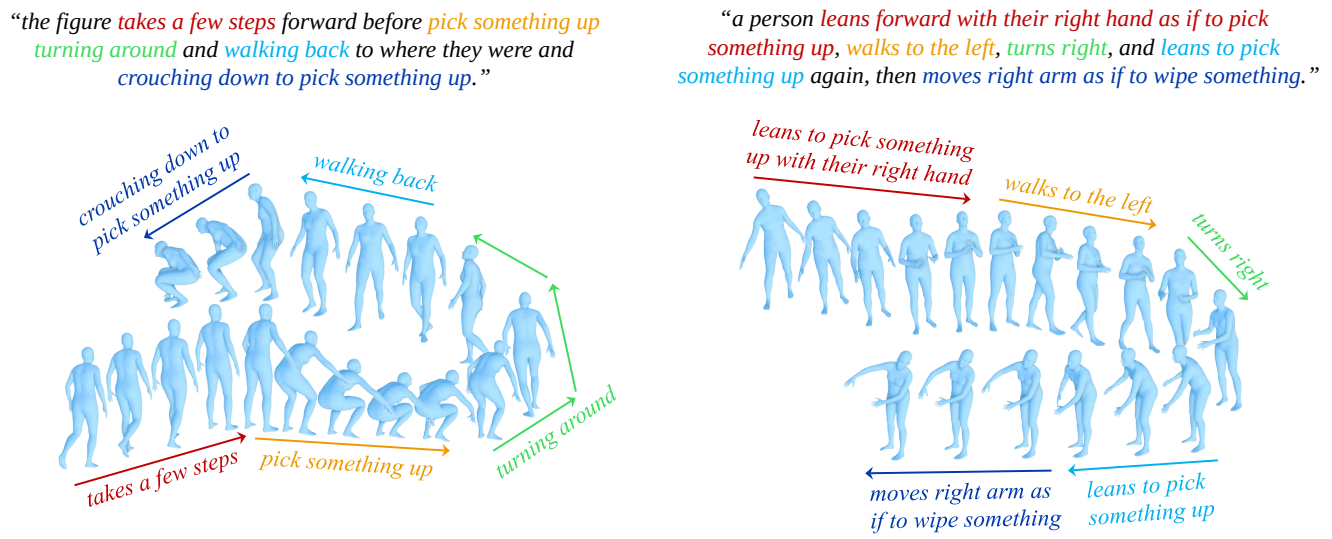


Figure S4. Qualitative comparison on long and complex motions on HumanML3D.

Table S9. Comprehensive ablation studies of the proposed component. **Bold** indicates the best result, while underlined refers the second-best. The right arrow → indicates that closer values to ground truth are preferred.

PG	PTG	HPF	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
			Top-1	Top-2	Top-3				
			$0.494^{\pm.003}$	$0.686^{\pm.003}$	$0.780^{\pm.003}$	$0.158^{\pm.005}$	$3.087^{\pm.008}$	$9.722^{\pm.067}$	$1.445^{\pm.051}$
✓			$0.520^{\pm.002}$	$0.711^{\pm.003}$	$0.802^{\pm.003}$	$0.086^{\pm.003}$	$2.913^{\pm.010}$	$9.688^{\pm.045}$	$1.841^{\pm.079}$
		✓	$0.506^{\pm.003}$	$0.695^{\pm.003}$	$0.791^{\pm.003}$	$0.132^{\pm.003}$	$3.019^{\pm.013}$	$9.710^{\pm.056}$	$1.860^{\pm.064}$
✓	✓		$0.545^{\pm.003}$	$0.734^{\pm.003}$	$0.828^{\pm.003}$	$0.051^{\pm.003}$	$2.799^{\pm.010}$	$9.556^{\pm.042}$	$1.973^{\pm.068}$
✓		✓	$0.529^{\pm.002}$	$0.721^{\pm.003}$	$0.813^{\pm.003}$	$0.063^{\pm.002}$	$2.858^{\pm.008}$	$9.528^{\pm.059}$	$2.101^{\pm.071}$
	✓	✓	$0.538^{\pm.003}$	$0.729^{\pm.003}$	$0.822^{\pm.003}$	$0.075^{\pm.002}$	$2.826^{\pm.011}$	$\mathbf{9.517}^{\pm.048}$	$2.076^{\pm.052}$
✓	✓	✓	$\mathbf{0.550}^{\pm.003}$	$\mathbf{0.744}^{\pm.003}$	$\mathbf{0.836}^{\pm.003}$	$\mathbf{0.035}^{\pm.002}$	$\mathbf{2.779}^{\pm.006}$	$9.534^{\pm.066}$	$\mathbf{2.155}^{\pm.046}$

Table S10. Comparison of computational complexity. Part-level evaluation reports the average performance of arms and legs.

Method	#Params	AIT	Part-level Evaluation (Avg.)		
			R-Precision (Top-3)↑	FID↓	MM-Dist↓
T2M-GPT [19]	247.5M	277ms	$0.706^{\pm.003}$	$0.198^{\pm.003}$	$3.534^{\pm.008}$
ParCo [23]	168.4M	643ms	$0.733^{\pm.003}$	$0.166^{\pm.003}$	$3.404^{\pm.010}$
MoMask [5]	44.8M	80ms	$0.724^{\pm.003}$	$0.139^{\pm.003}$	$3.476^{\pm.007}$
Ours	78.3M	158ms	$0.778^{\pm.003}$	$0.105^{\pm.003}$	$3.100^{\pm.007}$

E. Additional Quantitative Results

All experiments were performed on the HumanML3D [3] dataset, except for Tab. S18.

E.1. Detailed Results & Ablation Studies

We provide quantitative comparisons including additional models in Tab. S18. Detailed ablation study results for the proposed components are presented in Tab. S9. Note that PTG is fed into the part transformer and is thus related to its output; when used alone without PG and HPF, this output cannot be reflected. Therefore, this case was excluded from the ablation study.

E.2. Computational Complexity

We report computational complexity, including the number of parameters and average inference time (AIT), in Tab. S10. Notably, ParTY reduces the parameter count by more than 2× and improves AIT by over 4× compared to ParCo [23], a conventional part-wise method, while also enhancing part expressiveness. Although ParTY has slightly higher complexity than MoMask due to the use of part transformers, it achieves substantially higher part expressiveness, representing a reasonable trade-off.

Table S11. Porting Temporal-aware VQ-VAE to ParCo [23]. AIT is averaged over 100 samples on an RTX A5000 GPU.

Method	Window size	Reconstruction			Generation	
		# Params	FID↓	MPJPE↓	FID↓	AIT
ParCo	4	6.35M	0.021	0.108	0.109	65ms
+ Ours	4	7.93M	0.005 (+76%)	0.034 (+68%)	0.077 (+29%)	
ParCo	8	4.42M	0.047	0.166	0.172	33ms
+ Ours	8	6.01M	0.009 (+81%)	0.041 (+75%)	0.085 (+50%)	(+49%)
ParCo	12	3.59M	0.090	0.243	0.255	23ms
+ Ours	12	5.18M	0.017 (+80%)	0.060 (+75%)	0.098 (+61%)	(+64%)

Table S12. Ablation studies on proposed components in Temporal-aware VQ-VAE.

LTE	GTE	FID ↓	MPJPE ↓
		$0.038^{\pm.000}$	$0.121^{\pm.000}$
✓		$0.018^{\pm.000}$	$0.049^{\pm.000}$
	✓	$0.020^{\pm.000}$	$0.071^{\pm.000}$
✓	✓	$0.007^{\pm.000}$	$0.019^{\pm.000}$

Table S13. Quantitative results according to the number of embeddings in PTG.

# of Emb.	R-Precision ↑			FID ↓	MM-Dist ↓
	Top-1	Top-2	Top-3		
1	$0.531^{\pm.003}$	$0.722^{\pm.003}$	$0.815^{\pm.003}$	$0.060^{\pm.003}$	$2.842^{\pm.011}$
2	$0.538^{\pm.003}$	$0.731^{\pm.002}$	$0.827^{\pm.003}$	$0.052^{\pm.003}$	$2.825^{\pm.015}$
3	$0.541^{\pm.002}$	$0.735^{\pm.003}$	$0.829^{\pm.003}$	$0.047^{\pm.003}$	$2.811^{\pm.007}$
4	$\mathbf{0.550}^{\pm.003}$	$\mathbf{0.744}^{\pm.003}$	$\mathbf{0.836}^{\pm.003}$	$\underline{0.035}^{\pm.002}$	$\mathbf{2.779}^{\pm.006}$
5	$\underline{0.546}^{\pm.003}$	$\underline{0.739}^{\pm.002}$	$\underline{0.833}^{\pm.003}$	$\mathbf{0.033}^{\pm.002}$	$\underline{2.804}^{\pm.011}$
6	$0.531^{\pm.003}$	$0.727^{\pm.003}$	$0.819^{\pm.003}$	$0.045^{\pm.003}$	$2.878^{\pm.014}$

E.3. Temporal-aware VQ-VAE

E.3.1. Porting to Additional Model

To demonstrate the generality of our Temporal-aware VQ-VAE, we report the results of porting it to an additional model, ParCo [23], in Tab. S11.

E.3.2. Ablation Studies

We conducted ablation studies on the proposed components (LTE, GTE) in Temporal-aware VQ-VAE and report the reconstruction performance in Tab. S12.

E.4. Part-aware Text Grounding

E.4.1. Analysis of the Number of Embeddings

We conducted experiments under various conditions to find the optimal number of embeddings K in Part-aware Text Grounding (PTG). As shown in Tab. S13, we tested cases with 1, 2, 3, 4, 5, and 6 embeddings. The results showed that 4 embeddings achieved the best R-Precision performance, while 5 views showed slightly lower R-Precision but the highest FID performance. From 1 to 4 embeddings, we observed a general improvement in performance, indicating that adequate representation through a sufficient number of embeddings is necessary for properly aligning text descriptions to each part. However, the performance decline with 6 embeddings suggests that simply increasing the number of embeddings is not beneficial. This can be attributed to potential role overlap between embeddings and increased complexity in the Part Gate's weighting process when the number of embeddings exceeds a certain threshold.

Table S14. Ablation studies on auxilary loss.

Method	R-Precision ↑			FID ↓	MM-Dist ↓
	Top-1	Top-2	Top-3		
Ours (Full model)	$0.550^{\pm.003}$	$0.744^{\pm.003}$	$0.836^{\pm.003}$	$0.035^{\pm.002}$	$2.779^{\pm.006}$
w/o aux. loss	$0.543^{\pm.003}$	$0.739^{\pm.003}$	$0.830^{\pm.003}$	$0.042^{\pm.003}$	$2.805^{\pm.011}$

Table S15. Ablation studies on auxilary loss with **part-level evaluation** metrics.

Method	Part	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
Ours (Full model)	Arms	$0.506^{\pm.003}$	$0.802^{\pm.002}$	$0.133^{\pm.002}$	$3.079^{\pm.005}$
	Legs	$0.463^{\pm.003}$	$0.755^{\pm.003}$	$0.078^{\pm.003}$	$3.122^{\pm.008}$
w/o aux. loss	Arms	$0.491^{\pm.003}$	$0.793^{\pm.002}$	$0.148^{\pm.003}$	$3.098^{\pm.008}$
	Legs	$0.444^{\pm.002}$	$0.732^{\pm.003}$	$0.092^{\pm.003}$	$3.167^{\pm.008}$

E.4.2. Ablation Studies on Auxilary Loss

Tab. S14 reports ablation study results on the auxiliary loss, and Tab. S15 presents the corresponding results for part-level evaluation. These results demonstrate that using LLM-generated part text as auxiliary supervision provides modest improvements to text-part alignment.

E.4.3. LLM Prompt Details

We provide the complete prompt used to generate part text descriptions with an LLM in Fig. S5. We employ `gemini-2.5-flash` as our LLM, which takes a holistic motion description as input and generates corresponding part-level descriptions for arms and legs. When a part description indicates "No significant movement," we use the holistic text embedding as auxiliary supervision for that part instead of the LLM-generated part text. The **Examples** in Fig. S5 were constructed as follows:

- **Example 1:**
  Holistic Description: "A person waves with their right hand while standing still."
  Part-level Descriptions:
  – Arms: Right arm raises up and waves side to side.
  – Legs: No significant movement.
- **Example 2:**
  Holistic Description: "A person runs forward quickly."
  Part-level Descriptions:
  – Arms: Both arms swing back and forth rhythmically to support the running motion.

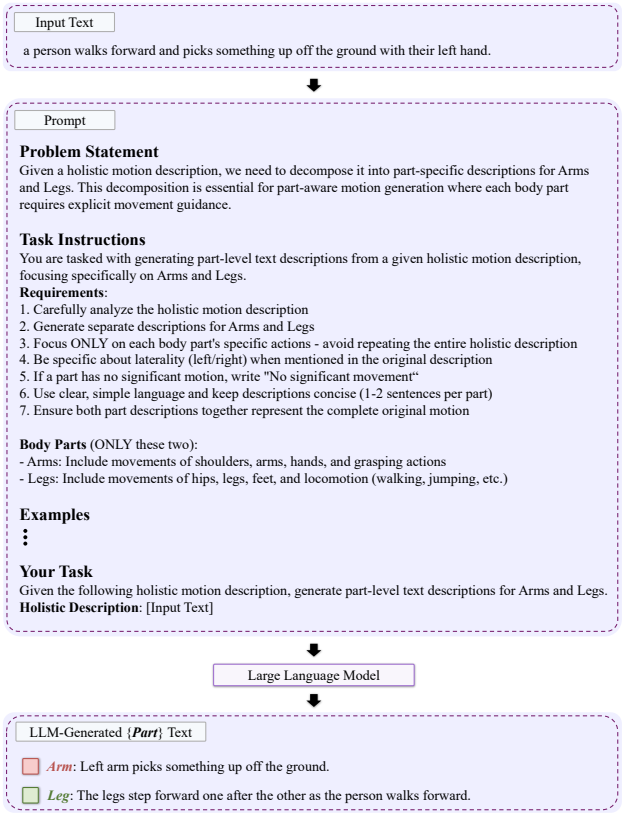


Figure S5. LLM prompt details.

  - Legs: The legs alternate rapidly, pushing off the ground to propel the body forward in a running motion.
- **Example 3:**
  Holistic Description: "A person jumps up and raises both arms above their head."
  Part-level Descriptions:
  - Arms: Both arms lift upward and extend above the head.
  - Legs: The legs push off the ground forcefully to jump, then land back down.
- **Example 4:**
  Holistic Description: "A person sits down on a chair."
  Part-level Descriptions:
  - Arms: No significant movement.
  - Legs: The legs bend at the knees and lower the body into a sitting position.
- **Example 5:**
  Holistic Description: "A person kicks a ball with their right foot."
  Part-level Descriptions:
  - Arms: No significant movement.
  - Legs: The right leg swings forward to kick, while the left leg supports the body's weight.

Table S16. Quantitative results with multiple LLM-generated part texts.

Types	R-Precision ↑			FID ↓	MM-Dist ↓
	Top-1	Top-2	Top-3		
LLM_1	$0.549^{\pm.003}$	$0.744^{\pm.002}$	$\mathbf{0.836}^{\pm.003}$	$0.035^{\pm.002}$	$\mathbf{2.777}^{\pm.009}$
LLM_2	$\mathbf{0.551}^{\pm.003}$	$0.743^{\pm.003}$	$0.835^{\pm.003}$	$0.038^{\pm.003}$	$2.792^{\pm.009}$
LLM_3	$0.550^{\pm.003}$	$0.744^{\pm.003}$	$\mathbf{0.836}^{\pm.002}$	$0.035^{\pm.002}$	$2.779^{\pm.006}$
LLM_4	$0.550^{\pm.003}$	$\mathbf{0.745}^{\pm.003}$	$\mathbf{0.836}^{\pm.003}$	$0.036^{\pm.003}$	$2.788^{\pm.007}$
LLM_5	$0.548^{\pm.003}$	$0.743^{\pm.003}$	$\mathbf{0.836}^{\pm.002}$	$\mathbf{0.032}^{\pm.003}$	$2.790^{\pm.011}$

E.4.4. Quality Evaluation of Generated Text

We generated part text descriptions multiple times using the same prompt and measured performance using them as auxiliary supervision, with results reported in Tab. S16. Although the LLM-generated text varies slightly across generations, the performance remains stable with minimal fluctuation due to the small weight of the auxiliary loss.

E.4.5. Group of Text Descriptions

We provide the group of text descriptions mentioned in manuscript Fig. 6 related to Part-aware Text Grounding. This group consists of 30 *"squat"*-related text descriptions, constructed as follows: we extracted all text descriptions containing *"squat"* from the entire HumanML3D dataset, encoded them into features using the CLIP [14] text encoder, and selected the text description *"a person squats down"* along with 29 text descriptions whose features are closest to this anchor description, resulting in a total of 30 descriptions. The following is the list of text descriptions:

1. *"a person squats down."*
2. *"a man is doing squats."*
3. *"a person is squatting down."*
4. *"a person is squatting down."*
5. *"the person is doing squats."*
6. *"someone is squatting down."*
7. *"a person squatting, raises arms."*
8. *"a person slightly squats down."*
9. *"a person lifting weights or squatting."*
10. *"a person performs a single squat."*
11. *"a person does an exercise squat."*
12. *"a person raises his hands, squats."*
13. *"a person in squat position while extending elbows."*
14. *"a person is squatting while moving their hands."*
15. *"a person squats down and then stands back up."*
16. *"a person squats down and holds out their arms."*
17. *"a squatting person raises their arms upwards from their sides."*
18. *"a person holds something above their shoulders and squats slightly."*
19. *"a person squats down and puts their hands above their head."*
20. *"a person does a squat and raises both arms over its head."*

Table S17. Quantitative results for different window sizes in PG.

Window size	R-Precision ↑			FID ↓	MM-Dist ↓
	Top-1	Top-2	Top-3		
1	$0.545^{\pm.003}$	$0.735^{\pm.003}$	$0.829^{\pm.003}$	$0.051^{\pm.003}$	$2.822^{\pm.012}$
2	$0.547^{\pm.003}$	$0.740^{\pm.003}$	$0.833^{\pm.002}$	$0.040^{\pm.003}$	$2.801^{\pm.008}$
3	$0.550^{\pm.003}$	$\mathbf{0.744}^{\pm.003}$	$\mathbf{0.836}^{\pm.002}$	$\mathbf{0.035}^{\pm.002}$	$2.779^{\pm.006}$
4	$\mathbf{0.551}^{\pm.003}$	$0.741^{\pm.003}$	$0.835^{\pm.003}$	$0.036^{\pm.002}$	$\mathbf{2.754}^{\pm.009}$
5	$0.548^{\pm.003}$	$0.740^{\pm.002}$	$0.833^{\pm.003}$	$0.039^{\pm.003}$	$2.793^{\pm.007}$

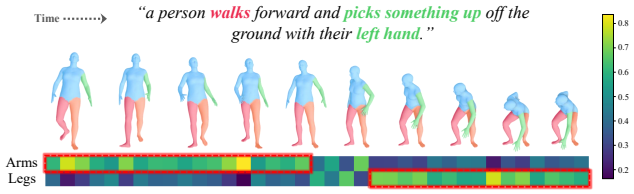


Figure S6. Visualization of cross attention map of HPF.

21. *"a figure squatting whilst hold both arms out towards the front."*
22. *"the person did a squat and they raised both arms forward."*
23. *"person squats down repeatedly with arms raised up to hold weights."*
24. *"a person squats down and moves their hands around at head level."*
25. *"person is standing in a slight squat position with hands resting on thighs."*
26. *"person is squatting and raises both arms up straight out and then down."*
27. *"a person squats down and puts their hands up to their face while squatting."*
28. *"the person puts their hand on their face then squats down like they're going underwater."*
29. *"a person raised their arms above their head, and squatted."*
30. *"a person squats by bending both knees and both elbows and moving arms above legs without touching them."*

E.5. Part Guidance

E.5.1. Analysis of the Size of Window

We evaluate quantitative performance using Part Guidance generated with different window sizes and report the results in Tab. S17. Performance increases progressively from window size 1 to 3, then slightly decreases from window size 4 onward. This suggests that a window size of 3 provides the optimal amount of future part information for the model to utilize effectively, while window sizes of 4 or larger introduce additional complexity that negatively impacts performance.

In this study, we conduct a user study to compare human motion results generated by multiple models. Each result consists of a video of the motion and frame-by-frame images extracted from it. Your role is to examine these results and rate them according to their visual quality. When evaluating, please focus on the following three criteria:

Part-Text Alignment: Evaluate how well the motion performs the actions described in the text description for each body part. There should be no left/right confusion.

Temporal Coherence of Part Motions: Evaluate how temporally synchronized the timing of part motions is. For example, during a walking motion, when the left arm moves forward, the right leg should move forward as well, which would be considered temporally coherent timing of each part motion.

Spatial Coherence of Part Motions: In each frame, evaluate how physically (spatially) coherent the part motions are within the full-body motion. For example, if an arm or leg bends unnaturally, it is not physically correct, and thus would be considered to have low spatial coherence of part motions.

Please note that you should exclude facial movements (expressions, etc.) and hand movements from your evaluation.

After reviewing results, please rank them from the most natural and high-quality result (5) to the most unnatural or low-quality result (1). Thank you for your participation!

Figure S7. Guidelines for user study.

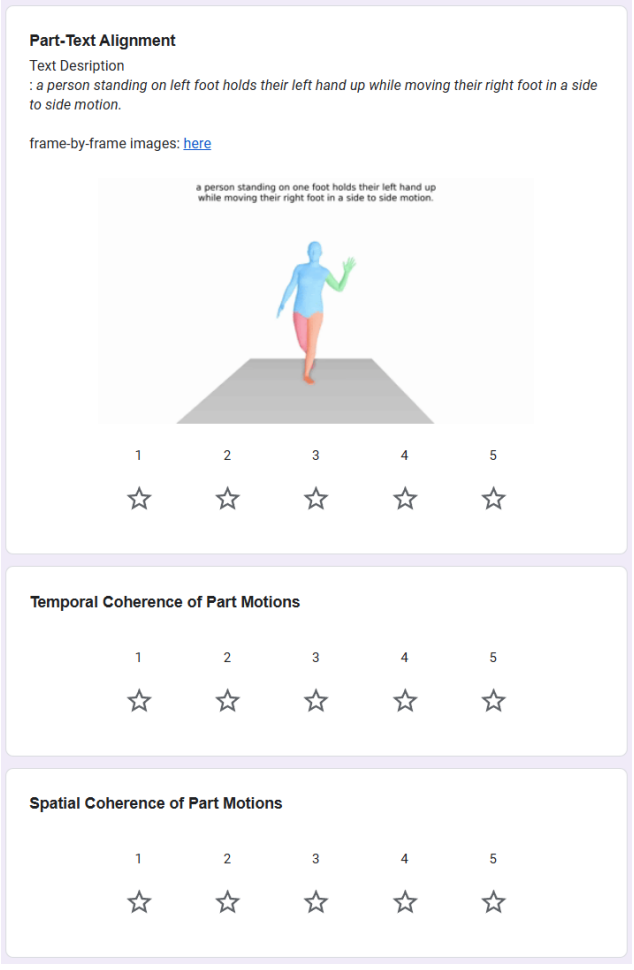
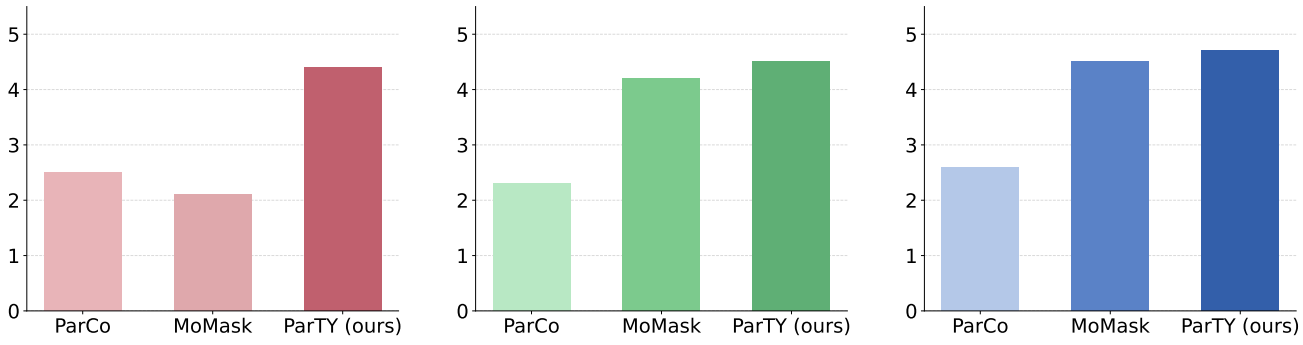


Figure S8. User evaluation interface for the user study.



(a) Part-Text Alignment (b) Temporal Coherence of Part Motions (c) Spatial Coherence of Part Motions

Figure S9. User study results on HumanML3D dataset. Each bar represents the average score on a scale from 1 to 5.

E.6. Holistic-Part Fusion

E.6.1. Additional Visualization of Attention Map

We provide additional sample of cross-attention map visualization from the Holistic-Part Fusion process in Fig. S6.

F. User Study Details

We conducted a user study to validate our proposed model and metrics. As described in Fig. S7, we provided participants with guidelines for evaluating (1) Part-Text Alignment, (2) Temporal Coherence of Part Motion, and (3) Spatial Coherence of Part Motion. Participants assessed motions generated by MoMask [5], ParCo [23], and ParTY (ours), assigning scores for each metric as shown in Fig. S8.

Fig. S9 presents the results of our user study with 50 participants. The results in (a) confirm that ParTY achieves superior part expressiveness compared to other methods, consistent with human visual perception. Furthermore, (b) and (c) demonstrate that ParTY maintains stable coherence and that our proposed coherence-level metrics effectively capture aspects of motion quality that align with human judgment.

Table S18. 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	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
	TM2T [4]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
	T2M [3]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
	MDM [16]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
	MLD [1]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
	T2M-GPT [19]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	AttT2M [22]	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
	ParCo [23]	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$0.109^{\pm.005}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$	$1.382^{\pm.060}$
	ReMoDiffuse [20]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
	MMM [12]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$	$1.164^{\pm.041}$
	SALAD [6]	$\mathbf{0.581}^{\pm.003}$	$\mathbf{0.769}^{\pm.003}$	$\mathbf{0.857}^{\pm.002}$	$0.076^{\pm.002}$	$\mathbf{2.649}^{\pm.009}$	$9.696^{\pm.096}$	$1.751^{\pm.062}$
	BAD [8]	$0.517^{\pm.002}$	$0.713^{\pm.003}$	$0.808^{\pm.003}$	$0.065^{\pm.003}$	$2.901^{\pm.008}$	$9.694^{\pm.068}$	$1.194^{\pm.044}$
	BAMM [11]	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$2.919^{\pm.008}$	$9.717^{\pm.089}$	$1.687^{\pm.051}$
	MoMask [5]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
	Light-T2M [18]	$0.511^{\pm.003}$	$0.699^{\pm.002}$	$0.795^{\pm.002}$	$0.040^{\pm.002}$	$3.002^{\pm.008}$	-	$1.670^{\pm.061}$
	MoGenTS [17]	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$0.033^{\pm.001}$	$2.867^{\pm.006}$	$9.570^{\pm.077}$	-
	LAMP [9]	$0.557^{\pm.003}$	$0.751^{\pm.002}$	$0.843^{\pm.001}$	$0.032^{\pm.002}$	$2.759^{\pm.007}$	$9.571^{\pm.069}$	-
	DisCoRD [2]	$0.524^{\pm.003}$	$0.715^{\pm.003}$	$0.809^{\pm.002}$	$0.032^{\pm.002}$	$2.938^{\pm.010}$	-	$1.288^{\pm.043}$
	BiPO [7]	$0.523^{\pm.003}$	$0.714^{\pm.002}$	$0.809^{\pm.002}$	$0.030^{\pm.002}$	$2.880^{\pm.009}$	$9.556^{\pm.076}$	$1.374^{\pm.047}$
	Motion Anything [21]	$0.546^{\pm.003}$	$0.735^{\pm.002}$	$0.829^{\pm.002}$	$\mathbf{0.028}^{\pm.005}$	$2.859^{\pm.010}$	$\mathbf{9.521}^{\pm.083}$	$2.705^{\pm.068}$
	ParTY (Ours)	$0.550^{\pm.003}$	$0.744^{\pm.003}$	$0.836^{\pm.003}$	$0.035^{\pm.002}$	$2.779^{\pm.006}$	$9.534^{\pm.066}$	$2.155^{\pm.046}$
KIT-ML	Real motion	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
	TM2T [4]	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$3.599^{\pm.153}$	$4.591^{\pm.026}$	$9.473^{\pm.117}$	$\mathbf{3.292}^{\pm.081}$
	T2M [3]	$0.370^{\pm.005}$	$0.569^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$3.401^{\pm.008}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$
	MDM [16]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.190^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
	MLD [1]	$0.390^{\pm.003}$	$0.609^{\pm.003}$	$0.734^{\pm.002}$	$0.404^{\pm.013}$	$3.204^{\pm.010}$	$10.80^{\pm.082}$	$2.192^{\pm.079}$
	T2M-GPT [19]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
	AttT2M [22]	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$0.870^{\pm.039}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
	ParCo [23]	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$0.453^{\pm.027}$	$2.820^{\pm.028}$	$10.95^{\pm.094}$	$1.245^{\pm.022}$
	ReMoDiffuse [20]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$2.814^{\pm.012}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$
	MMM [12]	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$0.316^{\pm.028}$	$2.977^{\pm.019}$	$10.91^{\pm.101}$	$1.232^{\pm.039}$
	SALAD [6]	$0.477^{\pm.006}$	$\mathbf{0.711}^{\pm.005}$	$\mathbf{0.828}^{\pm.005}$	$0.296^{\pm.012}$	$\mathbf{2.585}^{\pm.016}$	$\mathbf{11.097}^{\pm.095}$	$1.004^{\pm.040}$
	BAD [8]	$0.417^{\pm.006}$	$0.631^{\pm.006}$	$0.750^{\pm.006}$	$0.221^{\pm.012}$	$2.941^{\pm.025}$	$11.00^{\pm.100}$	$1.170^{\pm.047}$
	BAMM [11]	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$2.723^{\pm.026}$	$11.01^{\pm.094}$	$1.609^{\pm.065}$
	MoMask [5]	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	-	$1.131^{\pm.043}$
	Light-T2M [18]	$0.444^{\pm.006}$	$0.670^{\pm.007}$	$0.794^{\pm.005}$	$0.161^{\pm.009}$	$2.746^{\pm.016}$	-	$1.005^{\pm.036}$
	MoGenTS [17]	$0.445^{\pm.006}$	$0.671^{\pm.006}$	$0.797^{\pm.005}$	$0.143^{\pm.004}$	$2.711^{\pm.024}$	$10.92^{\pm.090}$	-
	LAMP [9]	$\mathbf{0.479}^{\pm.006}$	$0.691^{\pm.005}$	$0.826^{\pm.005}$	$0.141^{\pm.013}$	$2.704^{\pm.018}$	$10.929^{\pm.101}$	-
	DisCoRD [2]	$0.434^{\pm.007}$	$0.657^{\pm.005}$	$0.775^{\pm.004}$	$0.169^{\pm.010}$	$2.792^{\pm.015}$	-	$1.266^{\pm.046}$
	BiPO [7]	$0.444^{\pm.005}$	$0.674^{\pm.006}$	$0.803^{\pm.005}$	$0.164^{\pm.008}$	$2.658^{\pm.015}$	$10.833^{\pm.111}$	$1.098^{\pm.047}$
	Motion Anything [21]	$0.449^{\pm.007}$	$0.678^{\pm.004}$	$0.802^{\pm.006}$	$\mathbf{0.131}^{\pm.003}$	$2.705^{\pm.024}$	$10.94^{\pm.098}$	$1.374^{\pm.069}$
	ParTY (Ours)	$0.449^{\pm.006}$	$0.680^{\pm.007}$	$0.804^{\pm.006}$	$0.155^{\pm.014}$	$2.694^{\pm.030}$	$11.21^{\pm.082}$	$1.166^{\pm.049}$

References

[1] 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. 11

[2] Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. Discord: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14602–14612, 2025. 11

[3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 3, 4, 6, 11

[4] 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. 11

[5] 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. 2, 4, 6, 10, 11

[6] Seokhyeon Hong et al. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 11

[7] Seong-Eun Hong, Soobin Lim, Juyeong Hwang, Minwook Chang, and Hyeongyeop Kang. Bipo: Bidirectional partial occlusion network for text-to-motion synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–42, 2026. 11

[8] Seyed Rohollah Hosseyni et al. Bad: Bidirectional auto-regressive diffusion for text-to-motion generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. 11

[9] Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*, 2024. 11

[10] Ilya Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] 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. 11

[12] 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. 11

[13] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2, 4

[14] 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. 8

[15] 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. 2, 4

[16] 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. 11

[17] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. *Advances in Neural Information Processing Systems*, 37:130739–130763, 2024. 11

[18] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9797–9805, 2025. 11

[19] 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. 4, 6, 11

[20] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 11

[21] Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. Motion anything: Any to motion generation. *arXiv preprint arXiv:2503.06955*, 2025. 11

[22] 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. 2, 11

[23] 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. 2, 4, 6, 7, 10, 11