CVPR 2026

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

KunHo Heo, SuYeon Kim, Yonghyun Gwon, Youngbin Kim, MyeongAh Cho

Paper: https://arxiv.org/pdf/2603.09611

Project Page: https://visualsciencelab-khu.github.io/ParTY_project/

Code: https://github.com/VisualScienceLab-KHU/ParTY

Backgrounds

Text-driven Human Motion Generation

Input

"*Walking forward and steps over an object, and then continue walking.*"

Text Description

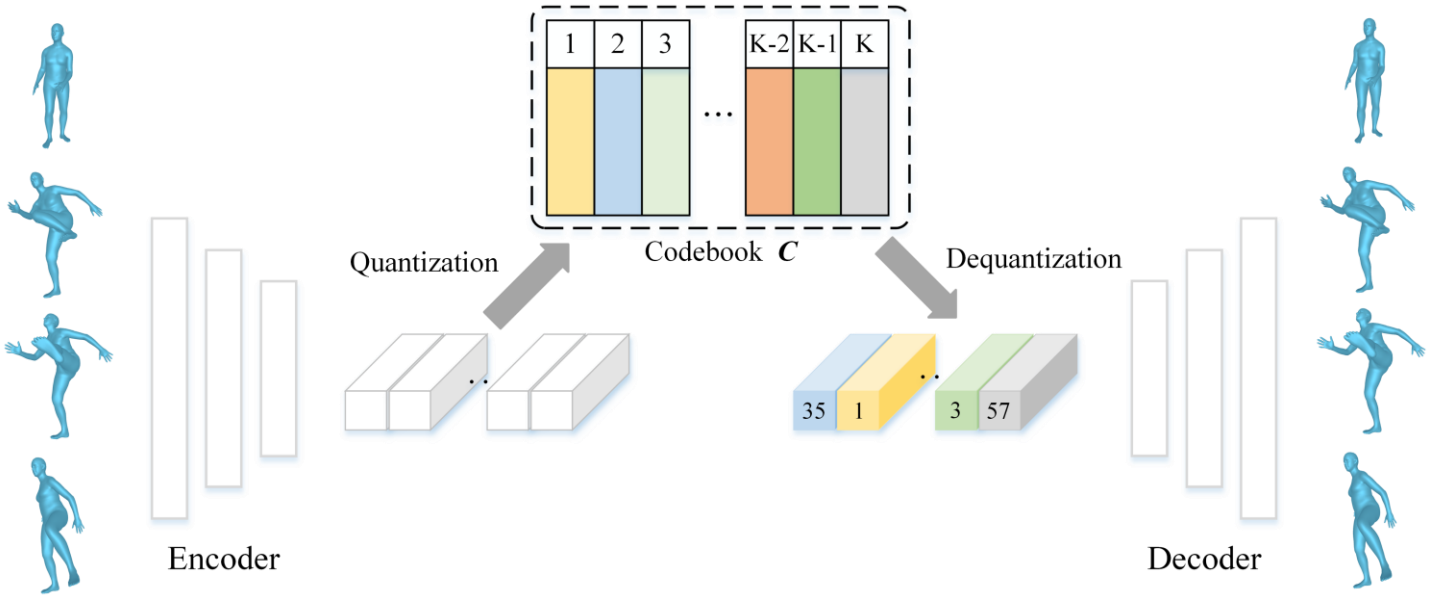
Output

3D Motion Sequence

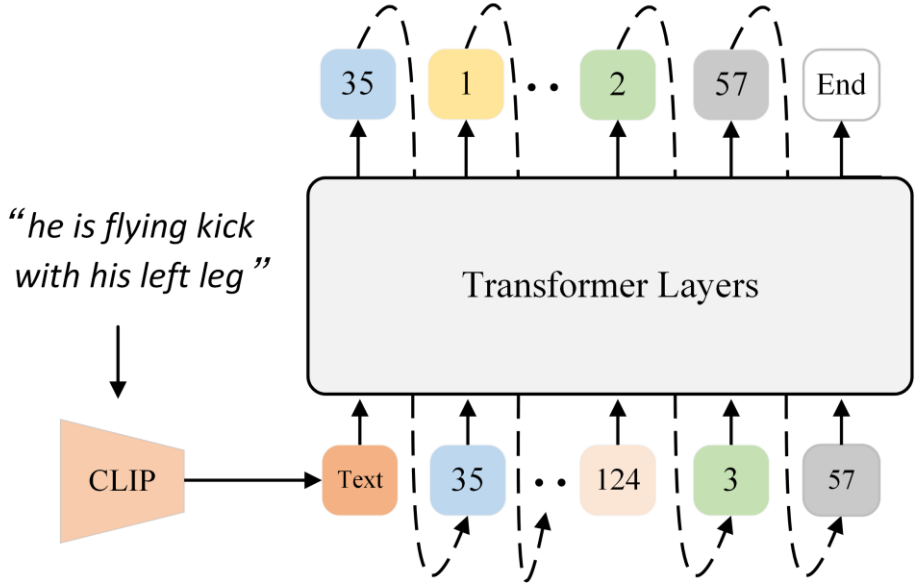
Previous Works

T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations

S1: Motion Quantization



S2: Text-to-Motion Token Prediction

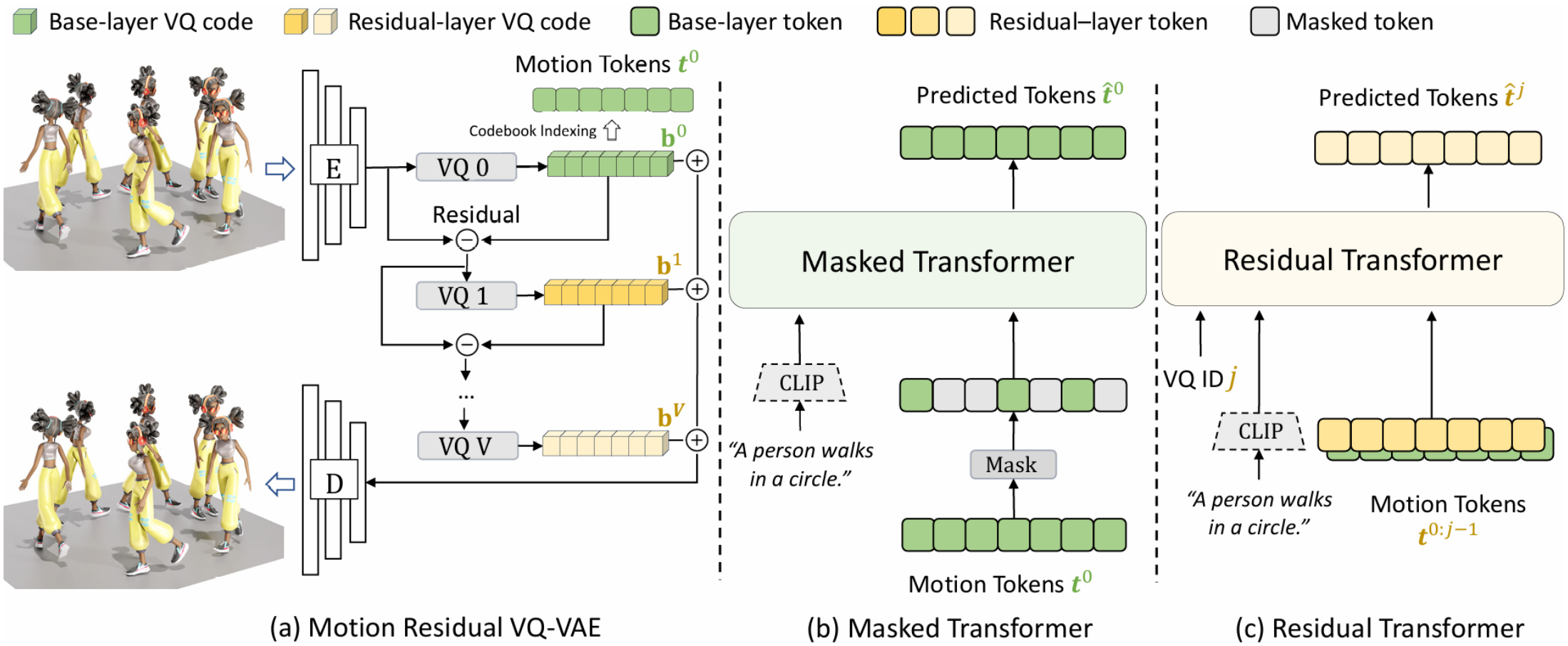


- Proposal of a "Motion Quantization + Text-to-Motion Token Prediction" architectures.
- They employed ***VQ-VAE*** for **S1**, and ***Transformer*** for **S2**.

Zhang, Jianrong, et al. "Generating human motion from textual descriptions with discrete representations." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

Previous Works

MoMask: Generative Masked Modeling of 3d Human Motions



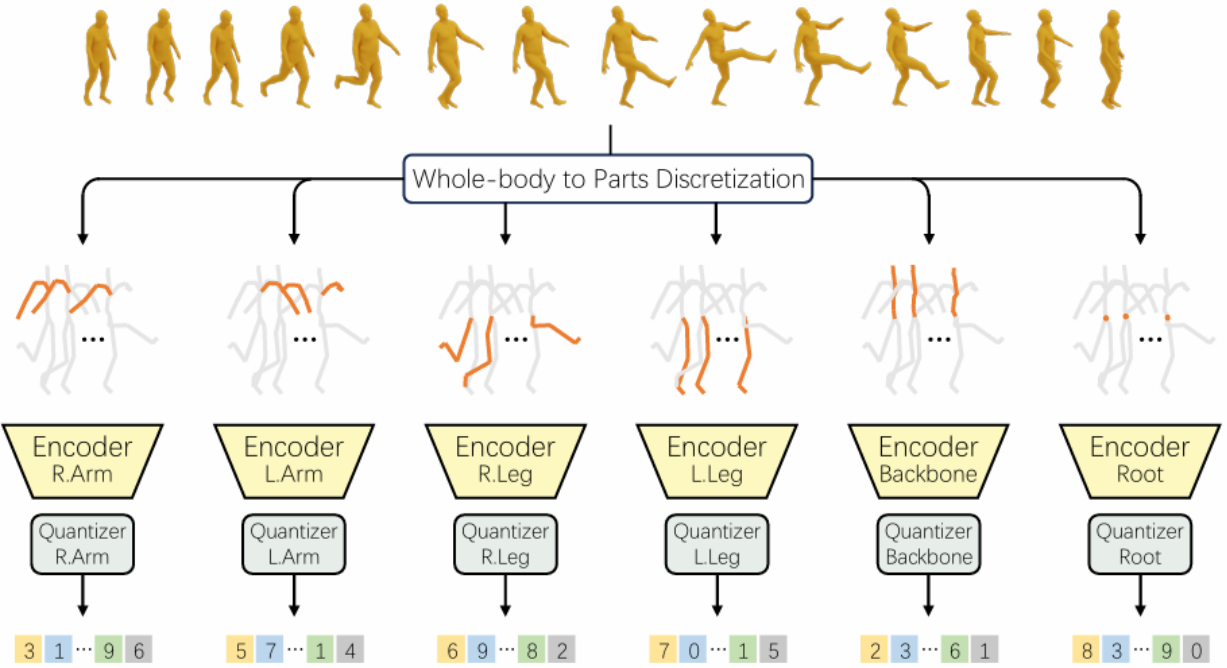
- Proposal of a **Residual Motion Quantization** and **Masked Transformer**.

Guo, Chuan, et al. "Momask: Generative masked modeling of 3d human motions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

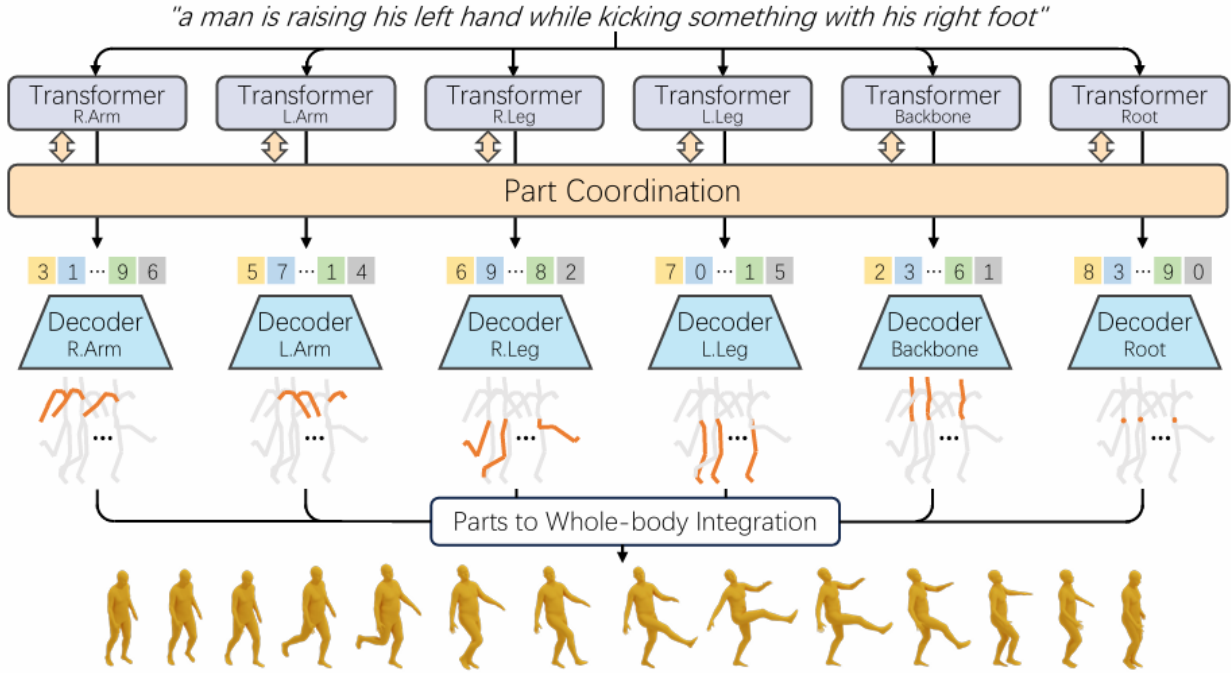
Previous Works

ParCo: Part Coordinating Text-to-Motion Synthesis

S1: Part-aware Motion Quantization



S2: Part-aware Motion Token Prediction



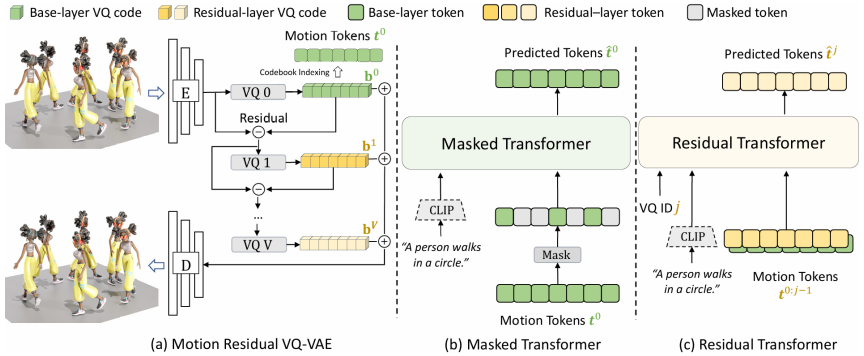
- Subdivided **S1** and **S2** processes by **body part**.
- **Part Coordination** added in **S2** enables interaction between parts.

Zou, Qiran, et al. "Parco: Part-coordinating text-to-motion synthesis." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

Previous Works

Holistic & Part-wise Methods

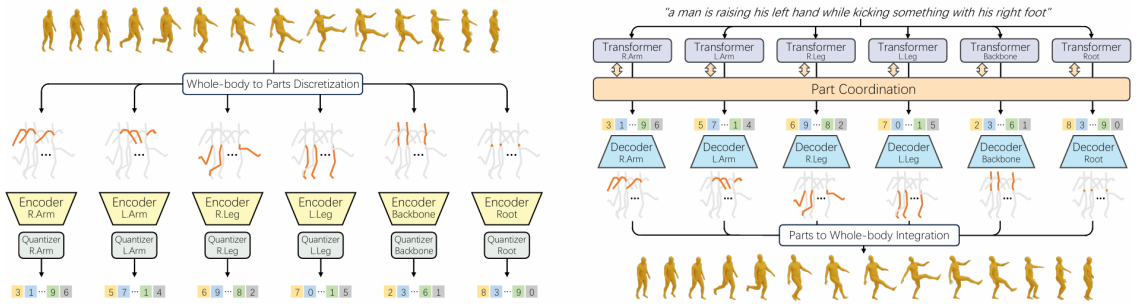
Holistic Method



Representative: ***MoMask***

- Generating **full-body** motion directly from text

Part-wise Method



Representative: ***ParCo***

- Split the body into anatomical parts and independently generate motions **separately for each part**.

Guo, Chuan, et al. "Momask: Generative masked modeling of 3d human motions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
Zou, Qiran, et al. "Parco: Part-coordinating text-to-motion synthesis." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

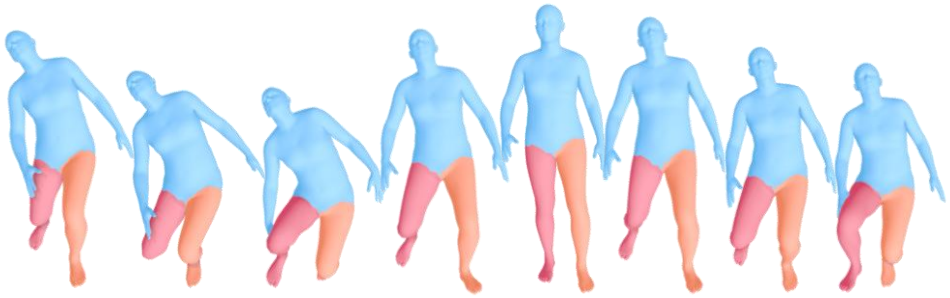
Problem Definition

Trade-Off: Text-Part Alignment & Inter-Part Coherence

"a person lunges forwards on their right leg, stands, and lunges on their left leg."

Holistic (MoMask)

✗ lunge on right leg ✗ lunge on left leg

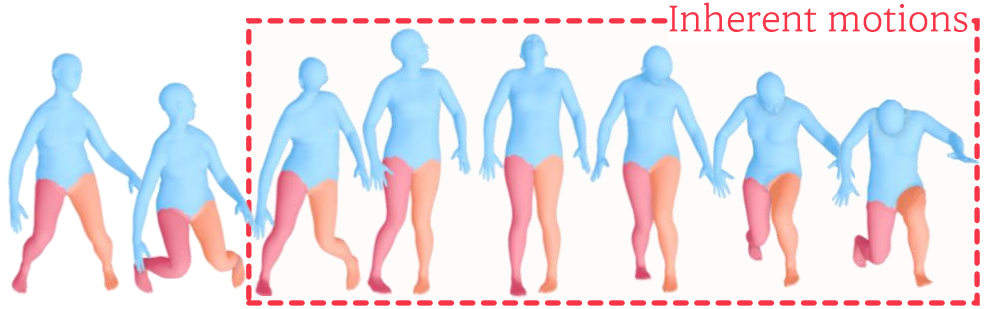


Text-Part Alignment ☹ Inter-Part Coherence ☺

Trade-Off

Part-wise (ParCo)

✗ lunge on right leg ✓ lunge on left leg

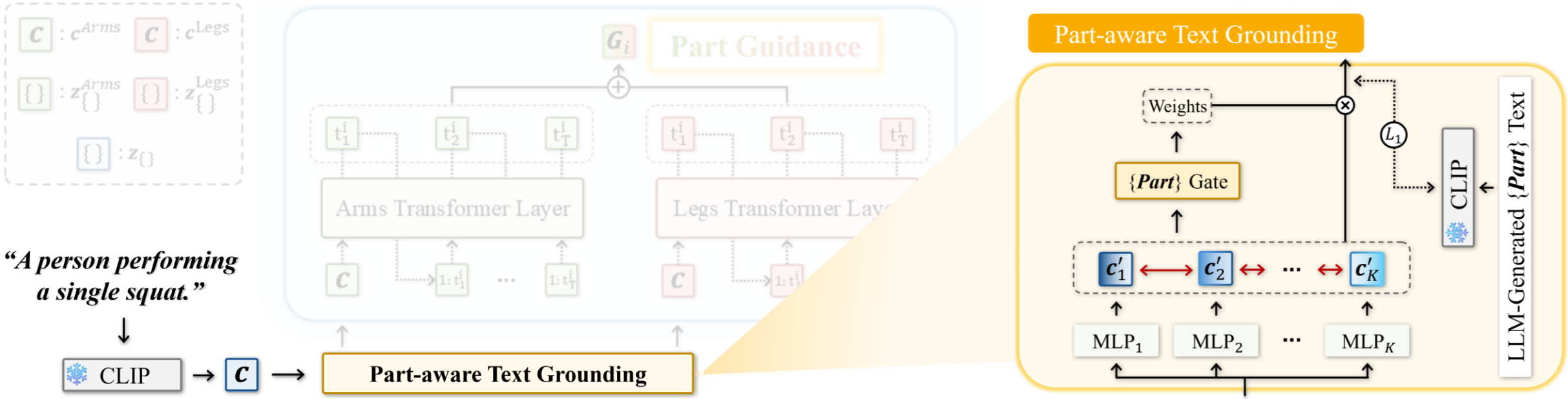


Text-Part Alignment ☺ Inter-Part Coherence ☹

- *Holistic* methods maintain **Inter-Part Coherence** well but limited **Text-Part Alignment**.
- In contrast, *Part-wise* methods show enhanced **Text-Part Alignment** but compromised **Inter-Part Coherence** as a trade-off.

Method Proposal

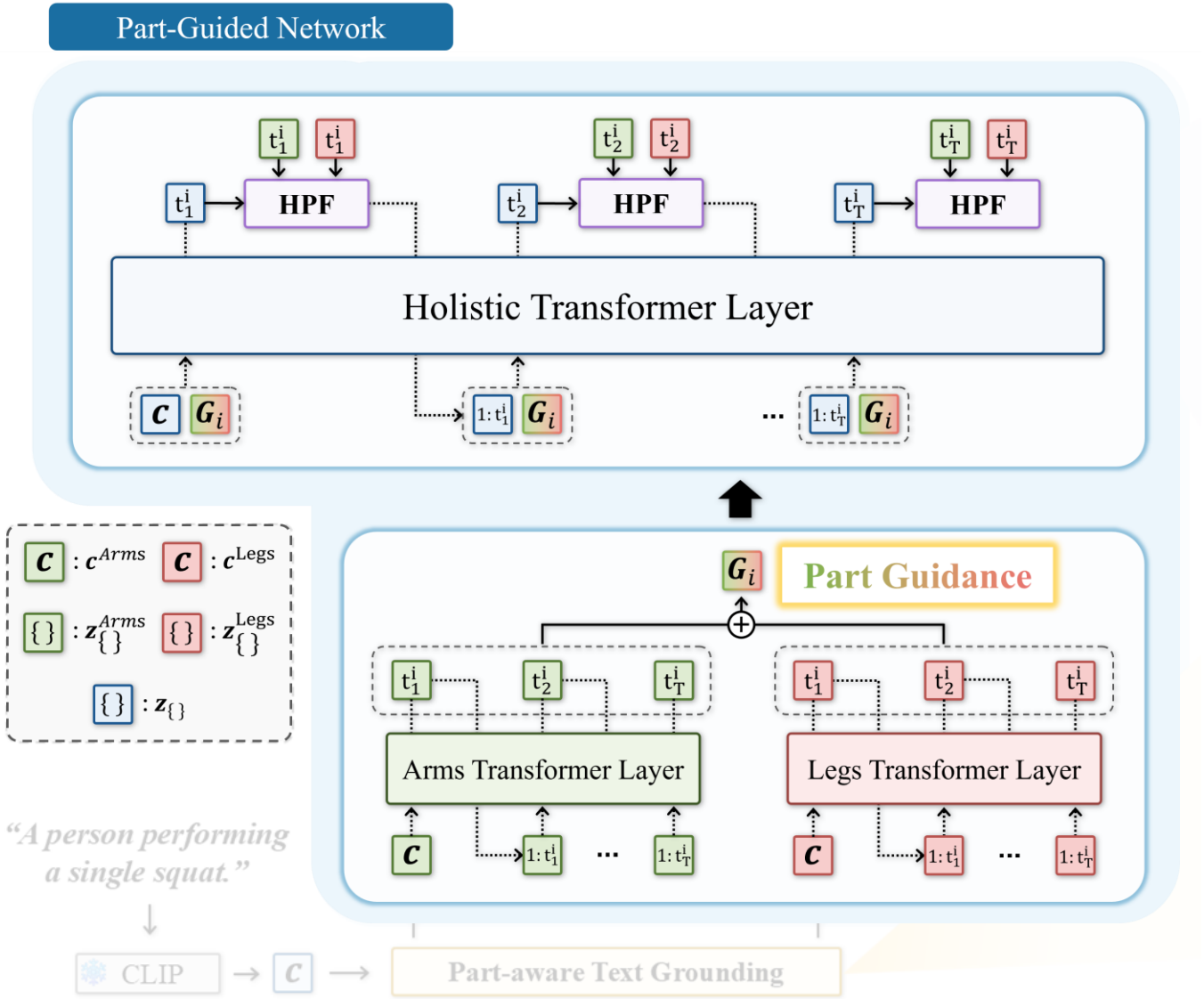
Part-aware Text Grounding



- **Part-aware Text Grounding (PTG)** 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.

Method Proposal

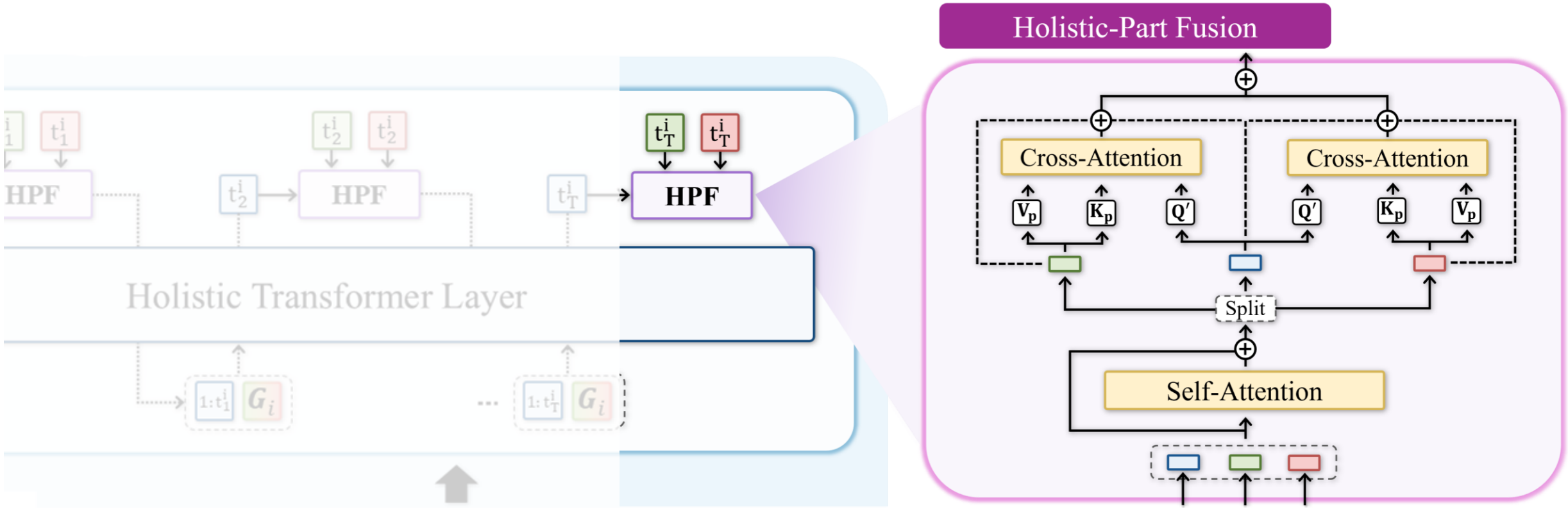
Part-Guided Network



- **Part-Guided Network** is 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.

Method Proposal

Holistic-Part Fusion



- During holistic motion generation, a **Holistic-Part Fusion (HPF)** is also employed, which directly fuses holistic and part motions, allowing part motion information to be incorporated throughout the process.

Experimental Results

Dataset & Evaluation Metrics

Dataset: HumanML3D & KIT-ML

- HumanML3D (14,616motions, 44,970 texts) and KIT-ML (3,911 motions, 6,278 texts).

Example of HumanML3D.

A person shakes an item with his left hand.

The person is leaving at someone with his left hand.

A person waves his left hand repeatedly above his head.

Guo, Chuan, et al. "Generating diverse and natural 3d human motions from text." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
Wald, Johanna, et al. "Rio: 3d object instance re-localization in changing indoor environments." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Experimental Results

Dataset & Evaluation Metrics

Evaluation Metrics

- **FID (Fréchet Inception Distance)**
  - Similarity between the **distribution** of generated motions and real motions.
- **R-Precision**
  - The accuracy of **semantic** matching between text descriptions and motion sequences.
- **Multi-Modal Distance (MM-Dist)**
  - **Semantic similarity** between text descriptions and motion sequences.
- **Multimodality**
  - **Diversity** of motion sequences generated from the same text description.

Experimental Results

Dataset & Evaluation Metrics

Proposed Evaluation Metrics: Part-level

- Extend conventional evaluation metrics to the part-level (arms, legs)
  - Part FID
  - Part R-Precision
  - Part MM-Distance

Proposed Evaluation Metrics: Coherence-level

- **Temporal Coherence (TC)**
  - How well different body parts remain **temporally coordinated** over time, including natural phase shifts such as arm–leg timing offsets during walking.
- **Spatial Coherence (SC)**
  - How **physically plausible** each frame is by checking whether inter-part distances and part-to-torso angles stay close to natural human pose statistics.

Experimental Results

Quantitative Comparison

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	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
	T2M-GPT	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	ParCo	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$0.109^{\pm.005}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$	$1.382^{\pm.060}$
	MMM	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$	$1.164^{\pm.041}$
	BAMM	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$2.919^{\pm.008}$	$9.717^{\pm.089}$	$1.687^{\pm.051}$
	MoMask	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
	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	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.190^{\pm.022}$	$10.85^{\pm.109}$	$\mathbf{1.907}^{\pm.214}$
	T2M-GPT	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
	ParCo	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$0.453^{\pm.027}$	$2.820^{\pm.028}$	$10.95^{\pm.094}$	$1.245^{\pm.022}$
	MMM	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$0.316^{\pm.028}$	$2.977^{\pm.019}$	$10.91^{\pm.101}$	$1.232^{\pm.039}$
	BAMM	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$2.723^{\pm.026}$	$\mathbf{11.01}^{\pm.094}$	$1.609^{\pm.065}$
	MoMask	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	-	$1.131^{\pm.043}$
	ParTY (Ours)	$\mathbf{0.449}^{\pm.006}$	$\mathbf{0.680}^{\pm.007}$	$\mathbf{0.804}^{\pm.006}$	$\mathbf{0.155}^{\pm.014}$	$\mathbf{2.694}^{\pm.030}$	$11.21^{\pm.082}$	$1.166^{\pm.049}$

Experimental Results

Quantitative Comparison

Part-level Evaluation

Method	Part	R-Precision (Top-1)↑	R-Precision (Top-3)↑	FID↓	MM-Dist↓
MoMask	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	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	$\mathbf{0.463}^{\pm.003}$	$\mathbf{0.755}^{\pm.003}$	$\mathbf{0.078}^{\pm.003}$	$\mathbf{3.122}^{\pm.008}$

Coherence-level Evaluation

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

Experimental Results

Quantitative Comparison

Ablation Studies

*PG: Part-Guidance

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

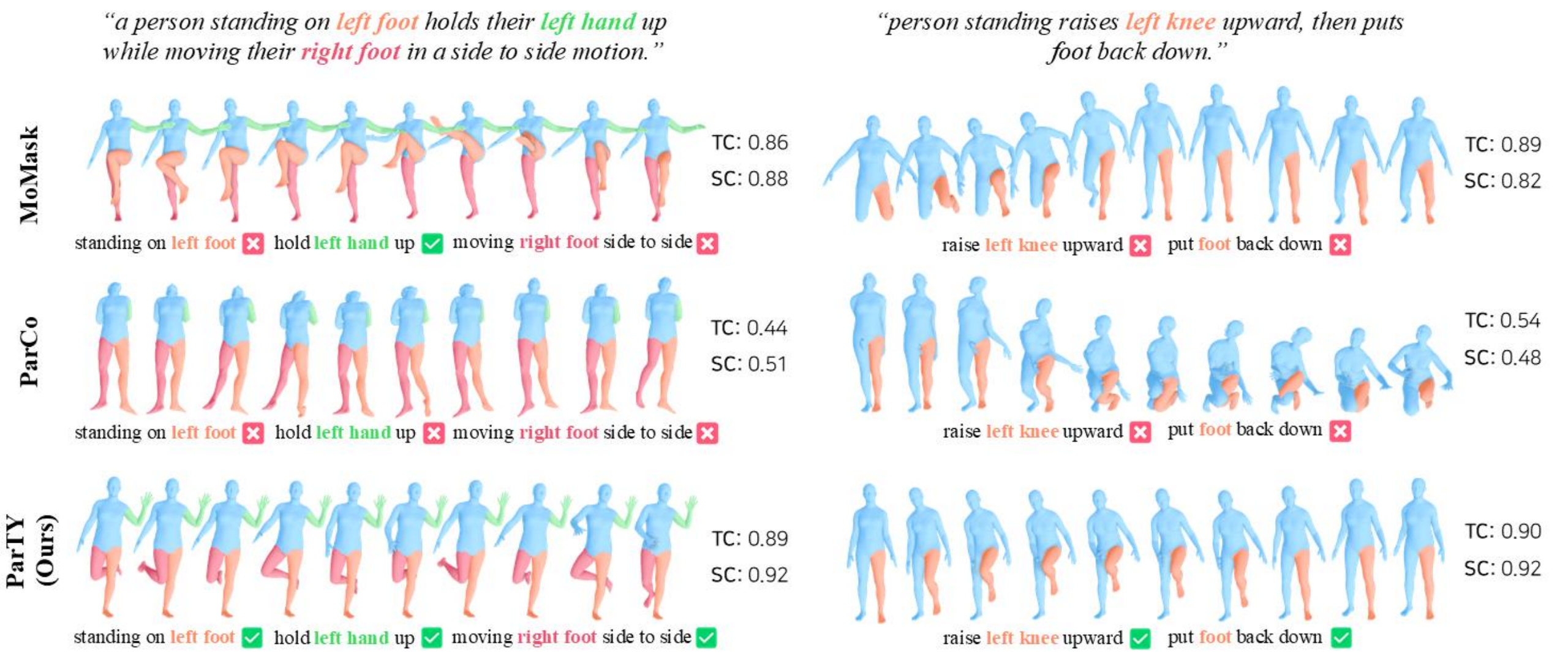
Ablation Studies with Part-level metrics

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

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

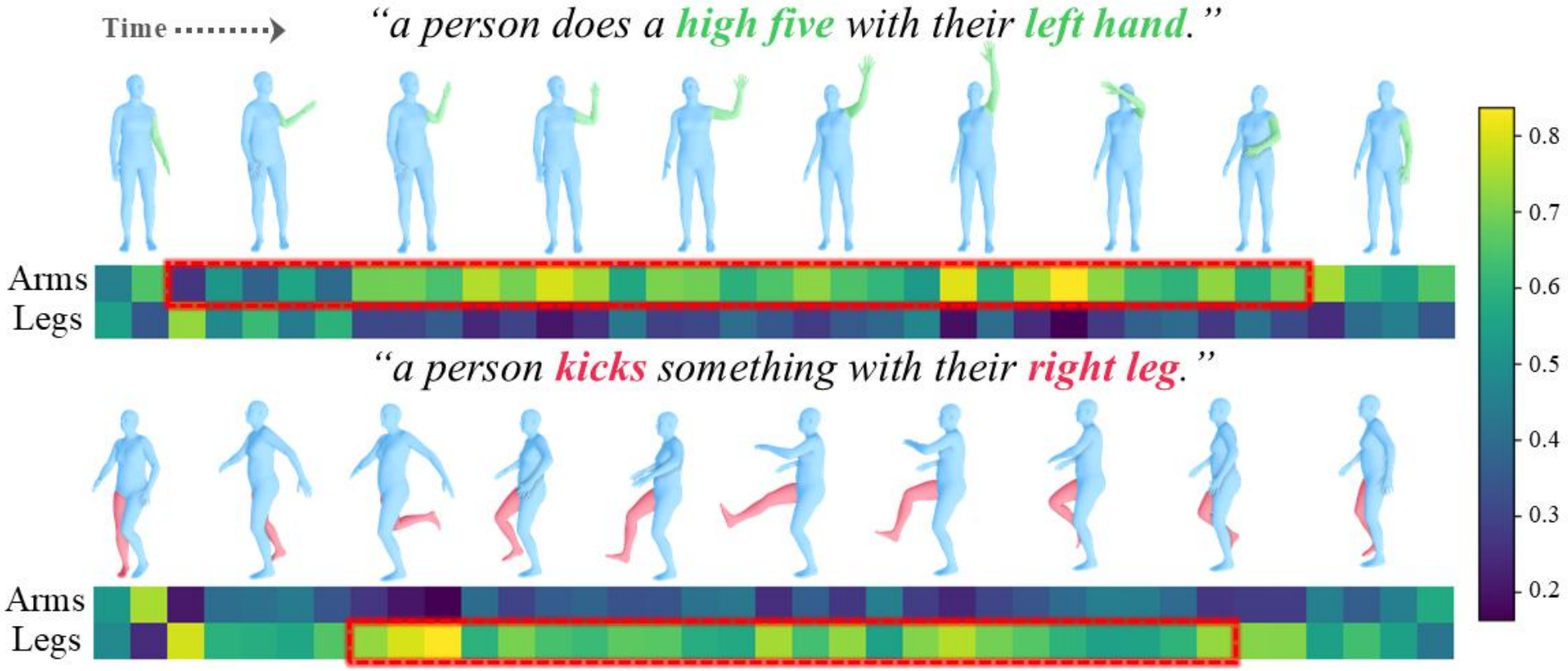
Experimental Results

Qualitative Comparison



Experimental Results

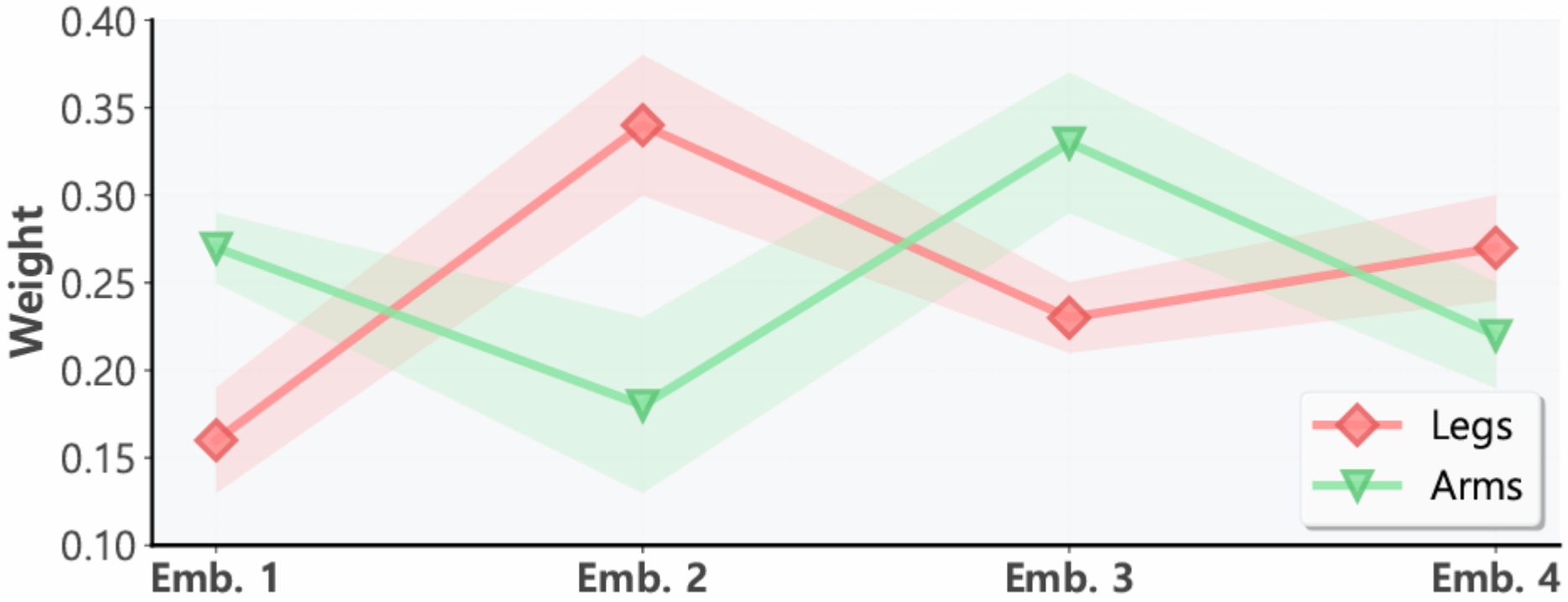
Qualitative Comparison



- Visualization of cross attention map of **HPF**. Rows correspond to body parts and columns represent temporal frames

Experimental Results

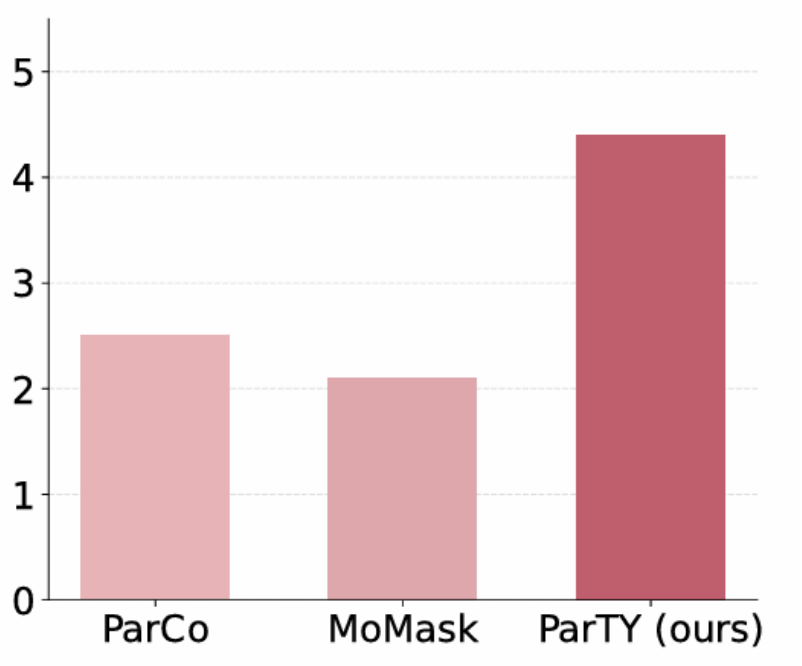
Qualitative Comparison



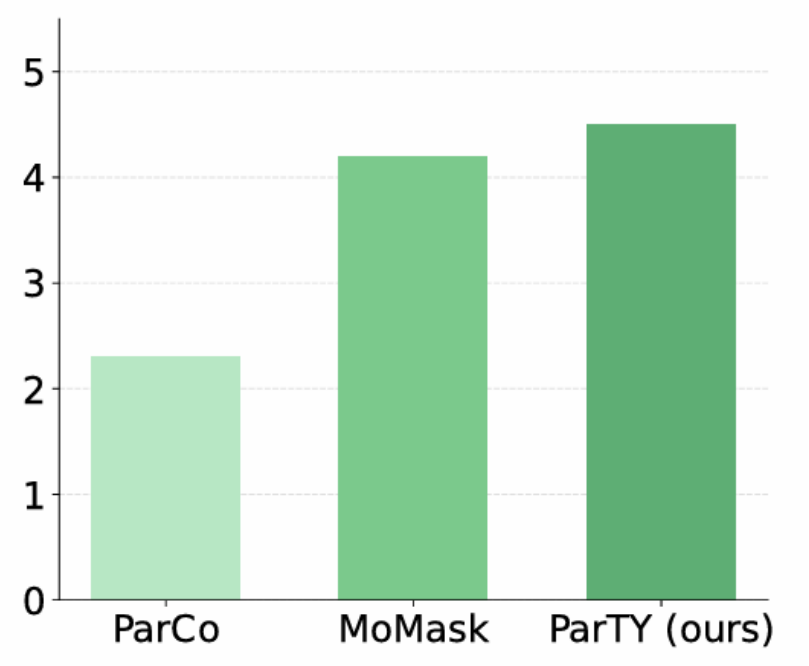
- Embedding selection ratios in **PTG**. Mean and standard deviation of weights are computed over semantically similar text descriptions that share common motion patterns.

Experimental Results

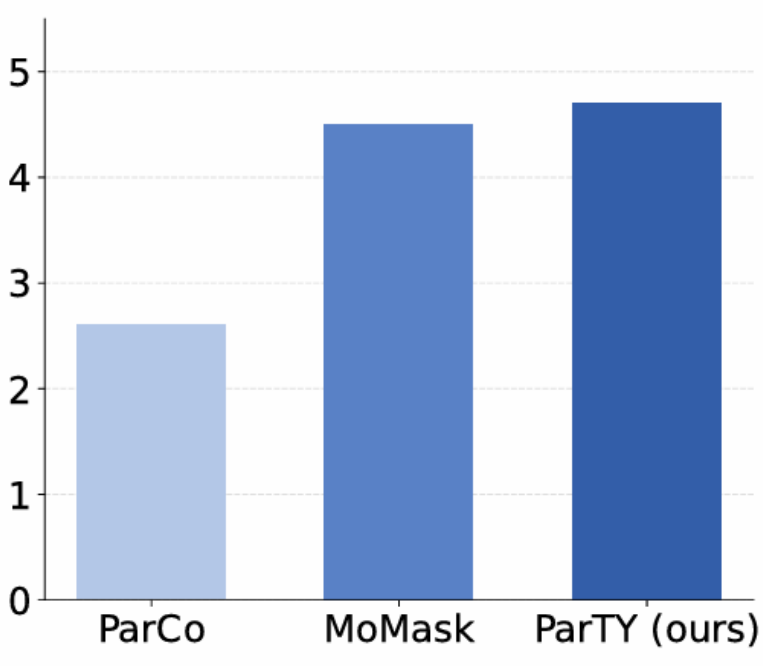
User Study



(a) Part-Text Alignment



(b) Temporal Coherence of Part Motions



(c) Spatial Coherence of Part Motions

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

Conclusion

Limitations & Future Works

Lack of physical plausibility in proposed metrics

- In practice, there are cases where TC/SC scores remain high even when physical realism is reduced due to contact/physics-related artifacts (e.g., foot sliding). There is room to introduce a module that can address these issues based on our Part-Guided structure.

Ambiguity in Part Decomposition

- Since we divided the full body into only two parts (Arms and Legs), there is room to extend this to a finer-grained five-part decomposition (R Leg / L Leg / Backbone / R Arm / L Arm).