2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Text-to-Motion Generation for Diverse Human Body-Motion Simulation

Jingze Gong[1], Yusheng Wang[2], Jun Ota[2]

***Abstract*— For practical deployment of autonomous robots in human-existed environment, it is essential to train robot policies with simulation environments that reflect the diversity of human body features and motion behaviors. However, existing datasets often overlook this need, relying on simplified skeleton models that ignore variations in age, height, or body shape. Moreover, realistic scenarios representing such diversity are largely missing due to the high cost and complexity of data collection.**

In this work, we address these limitations by constructing a human motion dataset that captures a wide range of body types and age groups, using accurate and characterized body models. These detailed representations allow robots to better learn how physical attributes influence movement, thereby enhancing their responsiveness and safety during interaction. To further expand the dataset efficiently, we also explore data generation techniques that create diverse motion samples from limited inputs. Our approach enables the scalable construction of simulation environments that reflect human variability, offering a valuable resource for future robot policy learning.

I. Introduction

Fig. 1. Simulation of work motion in an industrial environment.

Human-robot collaboration is becoming increasingly vital in the modern industrial and household domains. Ensuring safety and adaptability in these interactions requires simulation environments (Fig. 1) that can replicate realistic human motions and physical characteristics. Especially in reinforcement learning (RL) for robots, such simulation environments are essential for training robots that must operate safely and intelligently alongside humans.

Current simulations face two critical limitations that hinder their effectiveness:

1) **Lack of Diverse Body-Motion Representation**: Most datasets use simplified skeleton models that fail to capture the effects of human body variability, such as differences in height, age, and body shape. This limits a robot's ability to generalize behaviors safely across different human collaborators.
2) **Low-Fidelity Human Motion Synthesis**: Motions are often hard-coded or replayed kinematically, lacking biomechanical plausibility. This reduces realism and leads to overfitting to deterministic trajectories rather than accommodating real-world human motion variability.

To address these challenges, we construct a new dataset capturing motion data from individuals across different age groups and body types. Our dataset uses precise, feature-rich body models that enable more informed learning of how physical characteristics influence motion.

Furthermore, we explore text-to-motion generation techniques to synthesize motion data, extending the diversity and volume of training examples without proportional increases in collection cost.

The contributions of this work are listed as follows.

- **Simulation Diversity**: Our data provides a wider range of human body and motion characteristics in simulation environments.
- **Rich Annotation for Analysis**: The inclusion of body descriptions such as age, height, and body shape enables in-depth analysis of how these attributes influence human motion.
- **Efficient Data Scaling**: Generative methods reduce reliance on labor-intensive data capture.

Overall, our work contributes a practical and scalable foundation for constructing safer, more realistic worker simulations to train collaborative robots.

II. Related Works

A. Human Motion Representation

Human motion is typically represented in two major ways. The first is the skeleton-based representation, where motion is abstracted as a sequence of joint positions and orientations. This approach simplifies computation but often overlooks fine-grained body dynamics and physical characteristics. The second, more detailed method is the SMPL model [15], which encodes both motion and body shape using parametric mesh representations. Unlike skeletons, SMPL allows for accurate rendering of individual body differences, making it more suitable for studies involving body-motion diversity.

B. Text-to-Motion Human Motion Generation

Text-to-motion generation[10], [29], [2], [20], [21], [14], [1], [17], [23] has advanced rapidly with the integration of

[1]J. Gong is with Precision Engineering, Graduate School of Engineering, The University of Tokyo, Japan. gong@race.t.u-tokyo.ac.jp
[2]Y. Wang, J. Ota are with Research into Artifacts, Center for Engineering, Graduate School of Engineering, The University of Tokyo, Japan. {wang,ota}@race.t.u-tokyo.ac.jp

large language models (e.g., LLAMA [24]) and generative techniques like diffusion [30], [11] or VAE-based methods [10], [22]. These models generate diverse motions conditioned on text prompts and are evaluated primarily on the realism and diversity of generated sequences. However, a critical gap remains: they do not incorporate human physical features into the generation process. As a result, motions are typically detached from the body-specific context in which they occur.

C. Human Motion Datasets

Motion datasets[3], [19], [27], [4], [16], [7], [31], [12], [28], [6] are essential in various tasks related to human. Text-to-motion datasets[3], [19], [27], [4], [16] have been central to recent research, providing paired motion and language annotations. However, they lack metadata on human body characteristics, limiting their utility for modeling personalized or physically grounded motions. Scene-aware datasets such as PROX [9] and SAMP [8] partially address scene context but do not integrate text and body descriptors jointly. Some studies attempt to fuse data from separate sources([32][13]), but this often results in unnatural human-scene interactions due to misalignment.

Our dataset bridges this gap by combining body-aware motion data, scene context, and text prompts. It enables exploration of how physical traits influence motion and supports data-efficient generation of diverse, realistic simulations for applications like human-robot collaboration.

III. Methods

A. Objective

Our goal is to construct a simulation environment that reflects the diversity of human bodies and motions to support safe and realistic human-robot collaboration. We develop a **body-characterized text-to-motion generation framework** that expands our dataset. This includes collecting and modeling human motion across a wide range of physical attributes (e.g., age, height, body shape) and integrating this knowledge into a generative pipeline guided by textual inputs.

B. Challenges

Data Collection: how to capture diverse body characteristics (e.g., age, height, size) and obtain representative data from specific groups such as elderly individuals.

Model Design: how to develop a model that is both efficient and realistic for robot policy learning, while effectively incorporating body attributes into motion generation.

C. Constructing the Training Dataset

To address the lack of body diversity in motion datasets, we built a novel dataset conditioned on **body characteristics** such as age, height, and build. It enables training motion generative models to produce more realistic, personalized motions for simulation environments.

We designed a standardized *data collection protocol* where participants of different age groups (young, middle-aged, elderly) performed predefined actions. Motions were captured via skeletal tracking and reconstructed into SMPL representations to retain both shape and pose.

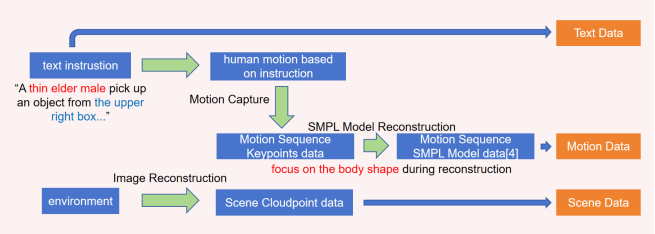


Fig. 2. Collection Process

Text prompts were normalized to include body descriptors (e.g., "a thin elderly man") and scene context, making the data suitable for text-to-motion tasks. As shown in Fig. 9, we standardized text prompts to explicitly include age, height, and body descriptors, and introduced wearable elder simulation equipment to mimic physical limitations in younger participants(Fig. **elder simulation equipment**), enhancing diversity.

To ensure consistency and usability, we established a labeling and alignment framework:

- Define height and weight categories for body descriptions.
- Standardize scene layouts and interaction types.
- Align all motion data in a unified format (e.g., SMPL+scene).

D. Generative Model Design & Training

To extend controllability and realism in simulation environments, we build a body-characterized text-to-motion generative model, guided by the architecture shown in Figure 3. Unlike traditional generation methods, our model emphasizes human physical diversity by incorporating age, body shape, and motion condition descriptions into the generation process.

Based on Motion Agent[26]'s structure, we adopt a two-stage design:

- **Tokenization via VQVAE[25]:** Raw motion sequences are splited into motion tokens(four frames per token) through a pretrained VQVAE encoder, enabling discrete motion representation.
- **Text-Driven Motion Prediction:** A large language model (LLM), enhanced by LoRA fine-tuning, learns to map text motion descriptions to token sequences. These tokens are decoded by the VQVAE decoder into full motion sequence.

This design allows the model to utilize information from pretrained LLM to generate motion sequences that reflect different body and motion characteristics, such as slower or more constrained movements for elders.

The loss function is defined as a cross-entropy loss over the token sequence:

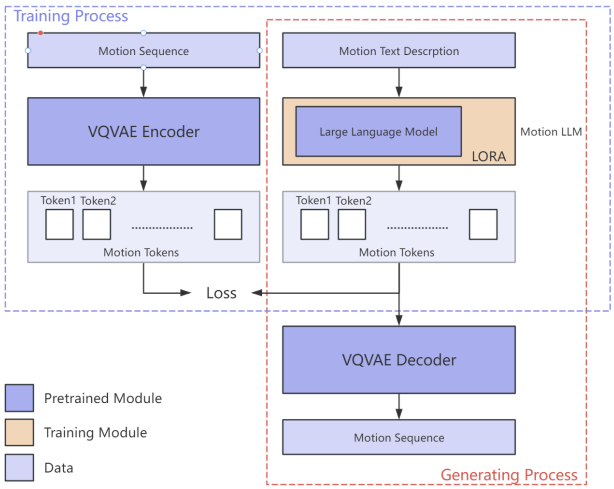


Fig. 3. Generative Model Framework

$$\mathscr{L} = -\sum_{t=1}^{T} \log P(\hat{z}_t = z_t \mid z_{<t}, \text{text}). \tag{1}$$

Here, z_t denotes the ground-truth token at timestep t from the VQVAE encoder, and $\hat{z}_t$ is the token predicted by the LoRA-tuned LLM. $P(\cdot)$ is the probability distribution over the motion token vocabulary, text represents the input natural language motion description, and T is the length of the token sequence.

This autoregressive formulation encourages the model to generate coherent and semantically accurate token sequences that are later decoded into realistic motion sequences by VQVAE decoder.

IV. Experiments

A. Constructing the Training Dataset

1) Simulation Environment Data Construction: We collect simulation environment samples that integrates **motion data**, **text instructions**, and **scene context**. Each sample consists of a *characterized motion sequence*, paired with a *3D scene representation* (point cloud). A normalized text prompt with age, body shape, and action descriptions guides both motion and scene configuration. The motion sequences are then converted to SMPL format and spatially aligned with the scene. As illustrated in Fig. 6, we use 12 motion caption camera for data collection.

B. Correlation Analysis

We evaluate the effectiveness of our dataset and model design through both quantitative analysis and qualitative observation of generated motions.

To examine how body features affect motion patterns, we computed correlation matrices across age groups and individuals from the collected motion sequences. Each sequence, originally a $f \times n \times 3$ array (frames × joints × spatial dimensions), was padded to a uniform length and then flattened into a 1D vector. The processed data were grouped by age and identity for correlation analysis across different groups. The correlation matrix was calculated as:

$$\text{Correlation}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}},$$

where X and Y are 1D vectors representing motion groups. We observed(Fig. 7):

- **Within Age Groups**: High correlation suggests consistent motion patterns within each age group.
- **Within Individuals Across Ages**: High correlation shows personal motion traits persist across life stages.
- **Between Age Groups**: Lower correlation highlights distinct motion patterns across age groups.

The results showed:

- **High intra-group similarity:** Motions within the same age group (e.g., elderly individuals) show strong correlation, indicating consistent motion patterns across different people of similar body conditions.
- **High intra-personal similarity:** Motions by the same person across different age simulations (e.g., using elder equipment) remain similar, verifying individual-specific motion styles.
- **Low inter-group similarity:** Motions across age groups display significantly lower correlation, confirming that body features meaningfully impact movement behavior.

C. Generative Model Design & Training

1) Baseline Model and Training Configuration: We employed a **LLaMA3-based text encoder [24]**, followed by a pretrained VQVAE-based decoder [26] that predicts motion sequence. More training details can be inferred from AppendixV-A.

D. Model Performance

Table I summarizes the performance of the baseline model trained under different VQVAE pretraining settings. The results show that VQVAEs pretrained on external datasets such as KIT [18] and HumanML3D [5] yield higher MPJPE and lower token accuracy when applied to our dataset. This indicates that those models struggle to generalize to our motion data, likely due to the increased complexity and the presence of diverse human body shapes in our dataset.

When the VQVAE is pretrained directly on our dataset, the model achieves significantly better results: MPJPE drops to 0.2057 and token accuracy rises to 0.9709. The high token accuracy demonstrates that the LoRA-finetuned large language model can effectively interpret and generate motion sequences grounded in our data, supporting the potential for large-scale text-to-motion synthesis.

However, the SMPL body shape MSE remains relatively high (0.4291), even under our best-performing setting. This suggests that while our model captures general motion effectively, the current VQVAE architecture still struggles to encode fine-grained body shape details. Future work should explore architecture improvements for more precise modeling of physical characteristics.

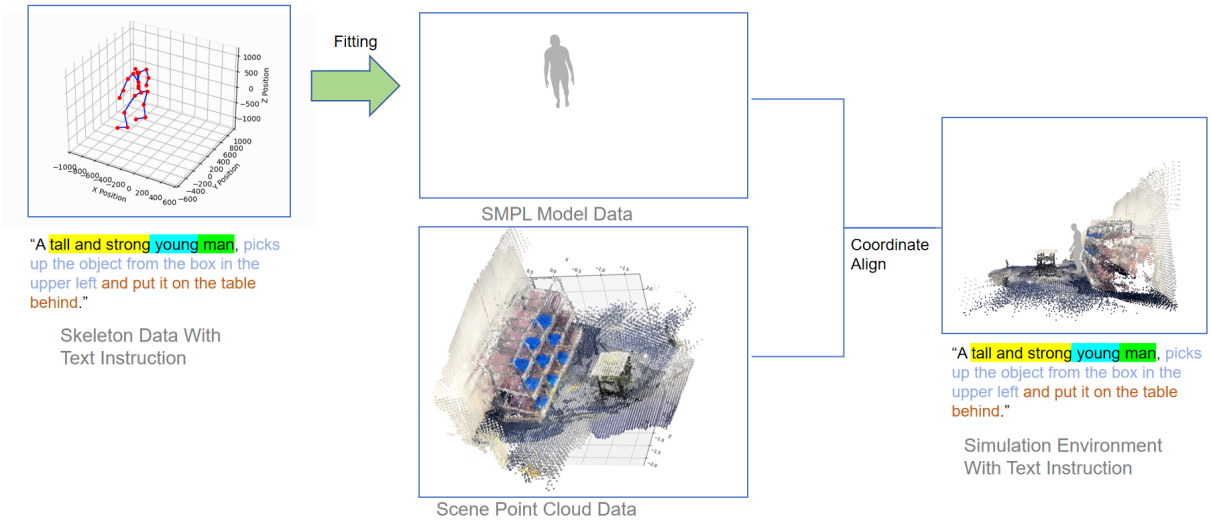


Fig. 4. Combining Motion and Scene into Simulation Environments

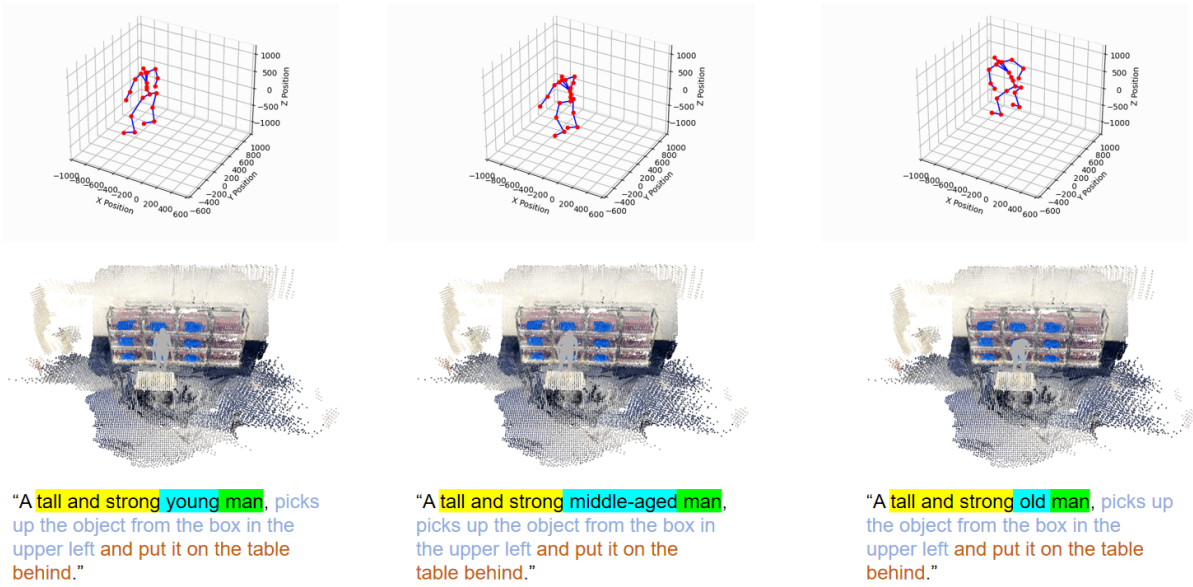


Fig. 5. Samples of Generated Motion Data

TABLE I
GENERATIVE MODEL TRAINING RESULTS

Setting	Epochs	MPJPE	SMPL Body Shape MSE	Token Accuracy
Baseline + our dataset + VQVAE Pretrained on KIT[18]	60	0.2276	0.4670	0.8177
Baseline + our dataset + VQVAE Pretrained on HumanML3D[5]	60	0.2253	0.4735	0.8205
Baseline + our dataset + VQVAE Pretrained on our dataset	60	**0.2057**	**0.4291**	**0.9709**

V. CONCLUSION

In this work, we address a critical gap in human-robot collaboration research: the lack of simulation environments that reflect diverse human body types and motions. We construct a motion dataset enriched with body-specific annotations such as age, height, and build, and propose a generation framework that leverages this data to produce realistic, controllable human motion.

Our analysis shows that models pretrained on existing datasets struggle to generalize to our body-aware data, showing the complexity and uniqueness of our dataset. By pretraining the motion encoder on our own data, we achieve significant improvements in motion accuracy and token prediction, proving that our dataset structure is compatible with scalable generation using large language models.

Despite these gains, our results also reveal limitations in the ability of current VQVAE architectures to capture

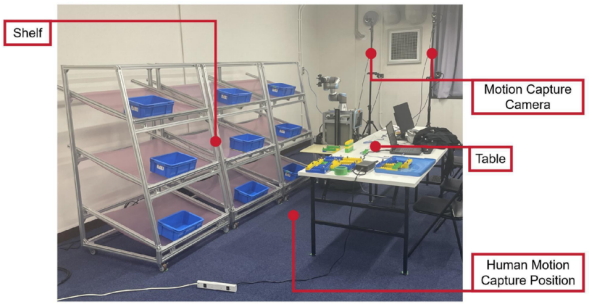
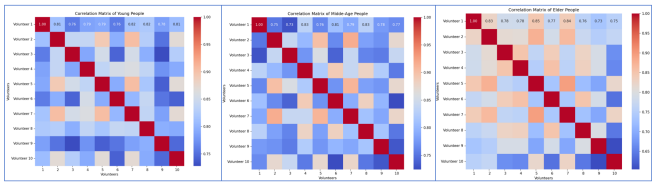
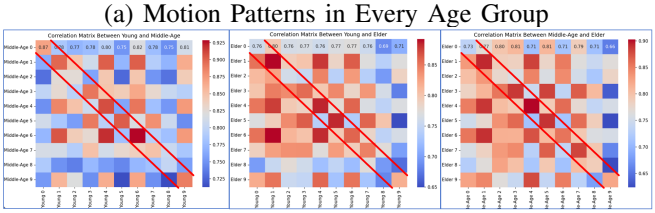


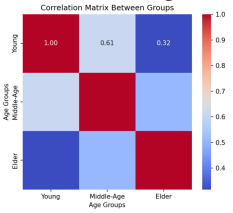
Fig. 6. Experimental Environment



(a) Motion Patterns in Every Age Group



(b) Motion Patterns Across Age Groups by person



(c) Motion Patterns Across Age Groups of all people

Fig. 7. Correlation Matrices for Motion Sequences

detailed body shape information, as reflected by the relatively high body shape MSE. This points to future directions in model refinement, especially for applications demanding precise physical modeling.

A. Future Work

Our dataset and generation framework establish a foundation for body-aware motion simulation. Two key directions remain for further exploration:

1) Modeling Body-Motion Relationships: Our motion data includes explicit body features (e.g., age, height, body type), adding complexity and realism. Future work may involve a dedicated module to better model how these attributes influence motion dynamics.

2) Application to Robot Learning: To assess practical impact and promote adoption, we plan to integrate our simulation environments into robot RL tasks. Initial experiments will explore how body-aware motion improves learning performance and safety in collaborative scenarios.

Overall, our dataset and methods offer a foundation for more realistic and various human body characteristics in

Scene	Age	Number	Motion sequence count
1	Younger	10 (8 Male 2 Female)	81
1	Elder (Younger with equipment)		81
1	Middle Age (Younger with half equipment)		81

Fig. 8. Collection Plan

	Body Description		Age	Gender
A	Thin	Tall	young	man
	Fat	Short	middle-age	woman
	Average size	Average height	old	

first action		second action		
pick up the object from the box	on the upper left	and from the box	on the upper left	then put them on the table behind
	on the left		on the left	
	on the lower left		on the lower left	
	in the middle upper		in the middle upper	
	in the middle		in the middle	
	in the middle bottom		in the middle bottom	
	on the upper right		on the upper right	
	on the right		on the right	
	on the lower right		on the lower right	

Fig. 9. Text Prompt Format

simulation environments, supporting safer and more adaptive robot training that accounts for human physical variability. We hope this work encourages broader adoption and experimentation in both academic and applied robotics settings.

APPENDIX

A. Generative Model Training Status

To benchmark our dataset, We fine-tuned the open-source **Motion Agent** model [26] on our dataset to benchmark performance and verify compatibility with existing architectures. The model was trained using the Adam optimizer, a batch size of 32, and a learning rate of 1×10^{-4} for 60 epochs on one NVIDIA H100 GPU.

To support effective training, we also include:

- **Train/Test Dataset Code:** Provided in Python, compatible with standard deep learning pipelines.
- **Motion Sequence Masking:** Handles variable-length sequences using binary masks.
- **Data Augmentation:** Improves diversity by splitting 120fps sequences into multiple 30fps versions.

1) Evaluation Metrics: We evaluate model performance using MPJPE (Mean Per Joint Position Error) and APE (Average Position Error), which measure how accurately the generated motions align with the intended text prompts and scene contexts. In addition, Token Accuracy is used to assess how well the large language model understands and generates motion tokens.

ACKNOWLEDGMENT

Special thanks are extended to the team involved in data collection, whose tools and models significantly contributed to this work.

REFERENCES

[1] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action composition for 3d humans," *2022 International Conference on 3D Vision (3DV)*, pp. 414–423, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252185316

Fig. 10. Equipment for Manipulating Elder's Behavior

[2] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, "Mofusion: A framework for denoising-diffusion-based motion synthesis," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9760–9770, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254408793

[3] J. de Lin, A. Zeng, S. Lu, Y.-Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *ArXiv*, vol. abs/2307.00818, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259316966

[4] M. N. Finean, L. Petrović, W. Merkt, I. Marković, and I. Havoutis, "Motion planning in dynamic environments using context-aware human trajectory prediction," *Robotics and Autonomous Systems*, vol. 166, p. 104450, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889023000891

[5] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5152–5161.

[6] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220870974

[7] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4316–4327, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232427806

[8] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black, "Stochastic scene-aware motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 374–11 384.

[9] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3d human pose ambiguities with 3d scene constraints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2282–2292.

[10] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] P. Jin, Y. Wu, Y. Fan, Z. Sun, W. Yang, and L. Yuan, "Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] R. Khirodkar, A. Bansal, L. Ma, R. A. Newcombe, M. Vo, and K. Kitani, "Egohumans: An egocentric 3d multi-human benchmark," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19 750–19 762, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258947228

[13] J. Kim, J. Kim, J. Na, and H. Joo, "Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions," in *Computer Vision and Pattern Recognition*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267035145

[14] T. J. D. Lee, G. Moon, and K. M. Lee, "Multiact: Long-term 3d human motion generation from multiple action labels," *ArXiv*, vol. abs/2212.05897, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254563762

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[16] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, K. Bailey, D. S. Fosas, C. K. Liu, Z. Liu, J. J. Engel, R. D. Nardi, and R. A. Newcombe, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *European Conference on Computer Vision*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270521958

[17] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 965–10 975, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233210075

[18] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big Data*, vol. 4, no. 4, p. 236–252, Dec. 2016. [Online]. Available: http://dx.doi.org/10.1089/big.2016.0028

[19] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "Babel: Bodies, action and behavior with english labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 722–731.

[20] X. Shi, C. Luo, J. Peng, H. Zhang, and Y. Sun, "Fg-mdm: Towards zero-shot human motion generation via chatgpt-refined descriptions," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265658847

[21] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247450907

[22] ——, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision*. Springer, 2022, pp. 358–374.

[23] R. Tian, Z. Zhao, W. Liu, H. Liu, W. Mao, Z. Zhao, and K. Yan, "Samp: A model inference toolkit of post-training quantization for text processing via self-adaptive mixed-precision," in *Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252367903

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[25] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:20282961

[26] Q. Wu, Y. Zhao, Y. Wang, X. Liu, Y.-W. Tai, and C.-K. Tang, "Motion-agent: A conversational framework for human motion generation with llms," *arXiv preprint arXiv:2405.17013*, 2024.

[27] M. Yan, Y. Zhang, S. Cai, S. Fan, X. Lin, Y. Dai, S. Shen, C. Wen, L. Xu, Y. Ma, and C. Wang, "Reli11d: A comprehensive multimodal human motion dataset and method," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2250–2262, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268733040

[28] D. Yang, J. Kang, L. Ma, J. Greer, Y. Ye, and S.-H. Lee, "Divatrack: Diverse bodies and motions from acceleration-enhanced three-point trackers," *Computer Graphics Forum*, vol. 43, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267657525

[29] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "Generating human motion from textual descriptions with discrete representations," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 730–14 740, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255942203

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

[31] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang, "Egobody: Human body shape and motion of interacting people from head-mounted devices," in *European Conference on Computer Vision*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:251040567

[32] S. Zhang, Y. Zhang, F. Bogo, M. Pollefeys, and S. Tang, "Learning motion priors for 4d human body capture in 3d scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 343–11 353.