https://doi.org/10.2352/EI.2026.38.12.GENAI-176

Expanding Motion Representations for Complementary Human Movement Modeling in Motion Language Models

Lawrence Amadi [1] ***Chih-Hsien Chou*** ***Ning Lu***
Futurewei Technologies; Austin, USA

Abstract

Human motion language models (Human-MLMs) have shown that text, speech, and motion can be brought into a shared generative framework. Most MLMs, however, encode body motion primarily through joint-centric SMPL features defined with respect to a fixed reference pose. While effective, this view compresses rich human movement into a single vocabulary and can under-represent local structures and movement co-dependencies of body part biomechanics. In this paper, we present ***JaBOR*** *(Joint-aligned Bone Orientation Representation), a hierarchical bone-centric motion representation designed as a general plug-in for Human-MLMs. JaBOR complements conventional joint-based motion features by expressing each bone in a locally rotation-normalized coordinate frame defined by four structurally selected keypoints:* ***Pivot, Free, Axis,*** *and* ***Plane****. Combined with compositional body-part tokenization and multi-representation training, JaBOR exposes LLMs to richer motion vocabularies without changing the high-level goal of existing MLM pipelines.*

We evaluate JaBOR in a preliminary motion tokenization-level study by integrating it into three Human-MLM pipelines: MotionGPT, MotionGPT3, and LoM. On the reported benchmarks, JaBOR improves Fréchet Inception Distance (FID) from 0.232 to 0.211 for MotionGPT and from 0.208 to 0.181 for MotionGPT3, while preserving diversity in two of the three reported settings. These results support the view that bone-based motion representation is a useful complementary signal for motion-language modeling. We position exercise form coaching and broader multimodal motion translation as promising future directions enabled by this representation.

Introduction

Human motion can be encoded as a language-like signal. It is structured, compositional, temporally ordered, and tightly coupled with speech, intent, and context. This observation has motivated a new line of research on **Human Motion Language Models** that unify motion understanding and generation with natural language modeling techniques [1, 2, 3, 4]. Language models such as MotionGPT treat motion as a foreign language and learn motion token vocabularies that can be aligned with text prompts [1]. More recent models extend this paradigm to richer multimodal settings, including speech, gesture, facial expression, and part-aware generation [2, 3, 4]. Despite this progress, existing Human-MLMs still inherit a largely **joint-centric** view of motion. In common SMPL-based pipelines, motion is encoded through joint rotations, joint positions, root-joint motion parameters, and related quantities measured relative to a fixed *T–pose* or *A–pose* reference. This representation is natural for animation and rendering [6, 7], and it has supported progress in text-to-motion generation [1, 8]. However, as Fig. 1 demonstrates, it is not obvious that the same representation is optimal for **autoregressive motion-language modeling**, where LLMs must compress the very broad range of human movements into a finite token vocabulary. Joint-only descriptions may obscure local bone structure, weaken body part-level editability, and underutilize alternate sensor-aligned viewpoints that are valuable for embodied reasoning.

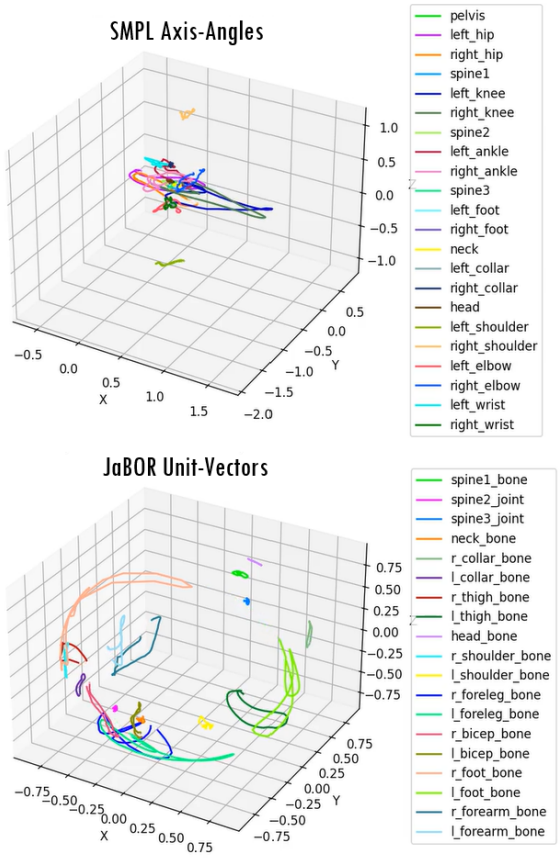


Figure 1. *Body part motion patterns of SMPL joint-based motion representation compared to the proposed JaBOR bone-based motion representation. In JaBOR, body parts have visibly unique movement patterns at different sections of the unit sphere and the values are neatly normalized to* $[-1, 1]$.

This paper investigates a simple but important question: **can Human-MLMs benefit from learning motion from both joints and bones?** We answer this question at the representation level. Rather than proposing a new end-to-end backbone, we introduce

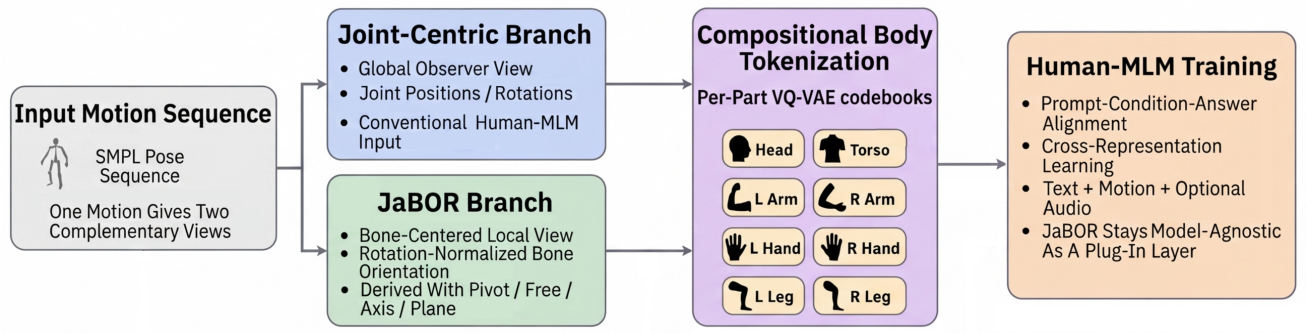


Figure 2. *Dual-motion vocabulary pipeline. Motion sequence is converted to (i) conventional joint-centric tokens and (ii) JaBOR bone-centric tokens. Per-body-part VQ-VAEs learn codebooks, and the Human-MLM is trained with prompt-condition-answer alignment across text, motion, and, when available, audio.*

JaBOR, a **Joint-aligned Bone Orientation Representation** designed to plug into existing Human-MLMs. JaBOR describes each bone in a local, rotation-normalized coordinate system obtained from the surrounding kinematic structure of the body pose skeleton. In contrast to joint parameters defined only with respect to a static *T–pose* or *A–pose* reference, JaBOR is explicitly pose-to-pose **local**, structurally **hierarchical**, temporally **dynamic**, and **bone-centric**. Our design follows two principles. First, the representation should preserve compatibility with the SMPL ecosystem, which remains the dominant source of standardized 3D human motion for motion-language research [5, 7]. Second, the representation should expose frame-to-frame local motion structure in a form that can be discretized and composed at the level of body parts. To this end, JaBOR is integrated with **compositional body-part tokenization**, inspired by recent part-aware models [3, 4], and used as a plug-in motion vocabulary that can be aligned with joint-based tokens during training.

The scope of this paper is intentionally focused. We report preliminary quantitative results on the quality of tokenized motion. We do not claim downstream-task gains beyond the numbers currently available, and we introduce **exercise form coaching** as an interesting future Human-MLM downstream task rather than a validated benchmark in the present study.

Our contributions are as follows:

1. We introduce **JaBOR**, a hierarchical local bone-orientation representation derived from SMPL pose structure and defined by *Pivot-Free-Axis-Plane* keypoints selections.
2. We present JaBOR as a **motion representation plug-in** for Human-MLMs rather than a model-specific architecture.
3. We combine JaBOR with **compositional body-part tokenization** and multi-representation training to expand the motion vocabulary available to language models.
4. We report preliminary evidence that JaBOR improves motion-token quality on popular Human-MLMs, including **9.1% lower FID** on MotionGPT and **13.0% lower FID** on MotionGPT3.

Related Work

Human motion language models. Recent work has reframed human motion as a modality that can be aligned with diffusion [13, 14] or language models [9]. T2M-GPT demonstrated that discrete motion codebooks and autoregressive language modeling are competitive for text-to-motion generation [8]. MotionGPT extended this idea into a unified motion-language model, treating human motion as a foreign language and supporting multiple motion-related tasks within a prompt-based framework [1]. MotionGPT3 further argued that discrete text and continuous motion should be modeled with partially separated streams to reduce cross-modal interference, yielding a bimodal architecture for motion understanding and generation [2]. A parallel line of work has broadened the modality space. EMAGE introduced holistic co-speech gesture generation with body, face, and hands, together with the BEAT2 dataset [4]. The Language of Motion (LoM) unified verbal and non-verbal signals using multimodal language modeling and emphasized compositional body-part generation and editing [3]. These works collectively motivate richer motion representations other than conventional joint-based motion encoding.

Motion representation for generation and understanding. Much of the literature still operates on SMPL-derived joint parameters, joint coordinates, root motion signals, or reduced joint rotations [5, 6, 7]. This representation is effective and standardized, especially through datasets such as AMASS [7] and HumanML3D [10]. However, it also reflects a graphics-oriented viewpoint: motion is often described as deviations from a fixed *A–pose* or *T–pose* reference, which simplifies rendering and rigging but does not necessarily provide the most expressive language for autoregressive modeling. Our work is motivated by the idea that **bones**, not only joints, offer an additional and complementary unit of motion semantics. Bone-centric features expose local structure and may better support body part composition, fine editing, and translation across sensor domains. In this sense, JaBOR should be viewed as a representation-level complement to joint-centric Human-MLMs rather than a replacement for SMPL.

Discrete tokenization and compositionality. Discrete latent modeling through vector quantization has become a standard building block in motion-language models [1, 3, 4, 8, 11]. VQ-VAE-style tokenizers enable continuous motion sequences to be mapped into discrete vocabularies that are amenable to language-model pretraining [11]. In language-conditioned motion generation, this enables prompt-based alignment between text and motion codebooks. Recent part-aware models further decompose the body into substructures, allowing localized control and improved expressiveness [3, 4]. Our approach follows this trend and integrates JaBOR with body-part tokenization so that local bone structure becomes available at the vocabulary level.

Method

Our goal is not to replace existing Human-MLMs with a new backbone, but to **expand the motion vocabulary** available to them. Given an SMPL-based motion sequence, we construct a second representation in parallel to the standard joint-centric encoding. This new representation, JaBOR, is then tokenized and aligned with the original motion stream through multi-representation training. Figure 2 illustrates the high-level pipeline: one motion sequence produces both a joint-based token stream and a JaBOR-based token stream, and the Human-MLM is trained to model and align the motion vocabularies.

Joint-aligned Bone Orientation Representation

The core idea of the proposed Joint-aligned Bone Orientation Representation (JaBOR) is to describe a bone in a **local canonical frame** induced by nearby joints. For each target bone, we select four keypoints–(A) **Pivot:** the reference joint around which local alignment is defined, (B) **Free:** a child joint connected to the pivot, defining the target bone, (C) **Axis:** a neighboring joint connected to the pivot and used to define a stable axis, and (D) **Plane:** an additional neighboring joint that is used to resolve orientation within a plane. The joints selection follows the hierarchical SMPL kinematic tree. The **Free** keypoint is a child of the **Pivot**, while the **Axis** and **Plane** keypoints are chosen from joints that are siblings or ancestors of the Pivot. This gives each bone a locally meaningful structural context. We then perform a rigid alignment decoupled into three steps illustrated in Fig. 3:

1. **Translation:** move the pose (or quadruplet joints) so that the Pivot keypoint rest on the cartesian origin.
2. **Axis alignment:** rotate the pose so that the bone from Pivot to Axis lies on a canonical X or Y axis.
3. **Plane normalization:** rotate again so that the Pivot-Axis-Plane triplet lies in the canonical XY plane.

After these transformations, the target bone–line segment from Pivot to Free–becomes a **rotation-normalized local vector**. JaBOR encodes the orientation of this aligned free bone through its direction in the canonical frame. Intuitively, the representation answers the question: *once local body context has been normalized, which direction does this bone point?* This construction differs conceptually from fixed-reference SMPL joint rotations. Instead of describing a pose only as deviations from a global fixed reference pose, JaBOR describes each bone relative to its **local structural neighborhood**. The resulting representation is hierarchical with nearest bones dependencies, and better suited for root–to–leaf compositional body part and movement reasoning.

Why Bones Complement Joints

Joint-centric SMPL features are effective because they are standardized and directly compatible with rendering. However, they have two limitations for motion language modeling. First, they are primarily organized around **joint parameters with respect to a fixed reference pose**, which can under-emphasize local dependencies between adjacent bones. Second, they encourage the language model to learn one dominant motion vocabulary, even though the same motion can be described by complementary representations and different useful viewpoints. JaBOR addresses both issues. Because the representation is bone-centric and locally normalized, it more explicitly captures **relative structure**

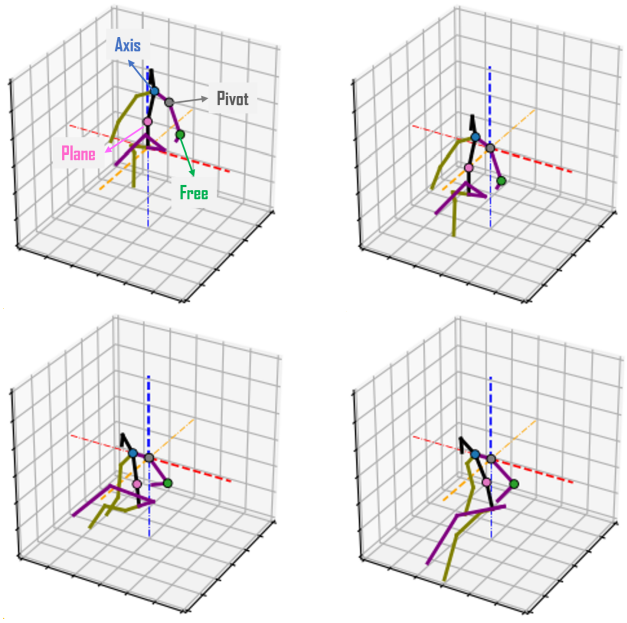


Figure 3. *A step-by-step demonstration of JaBOR transformation for left arm bone, as an example. Red, blue, and yellow dashed lines represent XYZ cartesian frame, respectively. Left shoulder, left elbow, neck, and spine joints are the selected Pivot, Free, Axis, and Plane keypoints, respectively. (1) Translate Pivot to the origin; (2) align the Axis (shoulder) bone to X-axis; (3) rotate the Pivot, Axis and Plane triplet into the XY-plane; and encode the unit vector direction of the Free (left arm) bone after orientation alignment.*

that is harder to read directly from SMPL joint rotations. And when used jointly with the standard SMPL joint representation during training, it acts as a **second language for motion**, exposing the model to a complementary vocabulary rather than forcing a single representation to serve all semantic roles.

Dual-motion Representations and Compositional Body-parts Tokenization

Our training recipe has two components.

Multi-representation pretraining. The Human-MLM is exposed to paired motion views of the same sequence: the existing joint-based representation and the new JaBOR representation. The objective is not merely reconstruction, but alignment between complementary vocabularies derived from the same motion.

Compositional motion tokenization. Inspired by LoM [3], we decompose the body into **five groups** or **eight parts**: head, torso, left and right hands, left and right arms, and left and right legs. For each group, a VQ-VAE learns a motion codebook. The language model is then trained and fine-tuned through **prompt-condition-answer** alignment across space, time, and representations. This allows the model to reason over localized body motion while also learning dependencies across parts and modalities.

The resulting system should be interpreted as a motion **representation plug-in layer** placed before the Human-MLM, not as a new model family. In principle, any motion-language framework that consumes tokenized or aligned motion representations can incorporate JaBOR.

Preliminary Results

Experimental Scope

This paper reports only verified results from preliminary training experiments. At this stage, our goal is to evaluate JaBOR as a **representation-level plug-in** rather than to claim a full state-of-the-art model. Accordingly, we report the impact of JaBOR on three human-motion language models–MotionGPT [1], MotionGPT3 [2], and LOM [3]–using two standard motion-generation metrics. **Fréchet Inception Distance (FID):** a fidelity measure based on the distance between generated and real motion embeddings. **Diversity (DIV):** a measure of how varied generated motions are across prompts or samples. We intentionally do not report other training and downstream-task metrics beyond these numbers, because they have not yet been fully verified in a form suitable for publication.

Motion-token quality results with and without JaBOR. Lower is better for FID; higher is better for Diversity.

Method	FID ↓	Diversity ↑
Real	0.002	9.503
MotionGPT	0.232	9.528
MotionGPT + JaBOR	**0.211**	**9.601**
MotionGPT3	0.208	**9.700**
MotionGPT3 + JaBOR	**0.181**	9.643

Diversity results of the motion tokens learned by the Language of Motion (LOM) model with and without JaBOR.

Method	Diversity ↑
The Language of Motion	15.165
The Language of Motion + JaBOR	**15.191**

Discussion

The most consistent signal in the current study is an improvement in **fidelity**. The use of JaBOR instead of SMPL-based motion representation reduces FID from **0.232 to 0.211** on MotionGPT and from **0.208 to 0.181** on MotionGPT3. These correspond to relative reductions of approximately **9.1%** and **13.0%**, respectively. This performance gain suggests that the bone-centric representation provides information that existing joint-based vocabularies do not fully capture. The diversity results are more nuanced. MotionGPT shows notable increase from **9.528 to 9.601**, while The Language of Motion shows marginal increase from **15.165 to 15.191**. MotionGPT3, however, shows a slight decrease from **9.700 to 9.643**. Taken together, these results suggest that JaBOR can improve fidelity while **largely preserving diversity**, but the diversity effect is not yet uniform across architectures and perhaps could be improved with better hyperparameter engineering.

This asymmetry is important. It indicates that JaBOR is not simply a generic regularizer that improves every metric in every setting. Rather, it is a complementary motion view whose effect depends on the baseline tokenizer and training pipeline. This is precisely why we frame this present paper as a **preliminary motion representation study for Human-MLMs** instead of a finished benchmark paper. Figure 4 shows the distinct rotation characteristics (or normalized orientation properties) of each bone capatured by JaBOR which makes for better compositional body part motion tokenization.

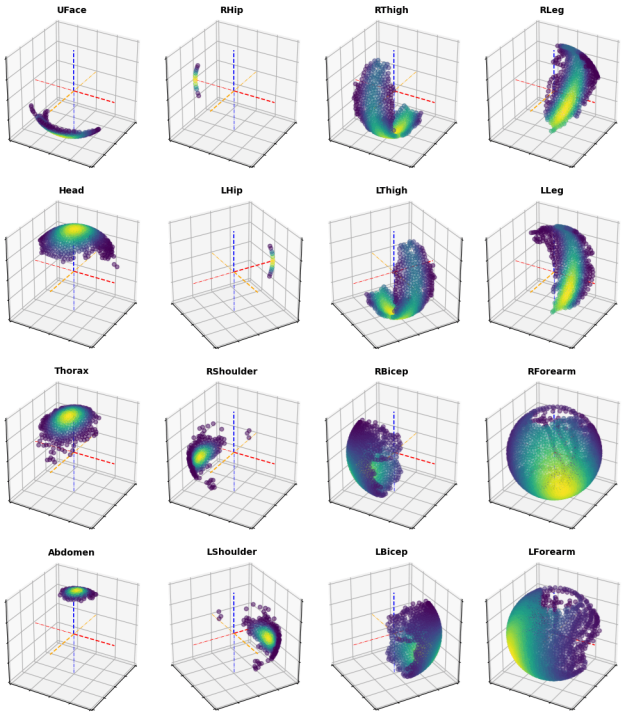


Figure 4. *Likelihoods of JaBOR bone orientations sampled from a subset of motion sequences from the AMASS dataset. Brighter colors indicate more plausible or frequent orientations of the given bone after rotation normalization by local-joints. The scatter plots show the direction of the bones' unit vectors from the cartesian origin to the unit-sphere surface.*

Limitations

We state some limitations of the current paper explicitly as stronger empirical support is still needed.

1. **No joint+bone ablation.** The current results show that bone-based JaBOR performs comparably to joint-based SMPL on one metric and outperforms it on another. But we do not yet evaluate the gains from the combined use of the complementary motion representations.
2. **No repeated-run significance analysis.** We report the verified preliminary metrics, but do not yet make claims about the statistical significance of performance gains or drops.
3. **No downstream evaluation in this paper.** Although the broader research targets text-to-motion generation, motion completion, question answering, gesture generation, emotion understanding, and exercise-related applications, those results are not yet mature enough to report here.
4. **No formal exercise coaching benchmark yet.** We therefore treat exercise form coaching as future work.

These limitations do not diminish the central observation of the current study: a carefully designed bone-centric motion representation produces a useful complementary foreign language encoding of human motion for Human-MLMs.

Exercise Coaching as a Future Direction

One motivation for richer motion representation is that real-world applications often require more than global motion plausibility. They require **localized reasoning** about how specific body parts move, coordinate, and deviate from desired form relative to other body parts. Exercise coaching is a natural example: good feedback often depends on the relationships between the torso, hips, legs, shoulders, arms and head rather than on a single global description of motion relative to a reference pose.

We do not present an exercise-coaching benchmark in this paper. Instead, we argue that JaBOR and compositional body-part tokenization provide a representation substrate that is well matched to such tasks. A model that can reason jointly over global joint motion and locally normalized bone motion may be better positioned to support future applications in Virtual AI coaching, physical rehabilitation, embodied tutoring, and body language feedback. Establishing this empirically will require new datasets, evaluation protocols, and likely human-centered studies.

Conclusion

We presented **JaBOR**, a hierarchical **Joint-aligned Bone Orientation Representation** for human motion language models. JaBOR is designed as a **motion encoding plug-in** that complements, rather than replaces, standard joint-centric SMPL motion encodings. By normalizing each bone relative to a structurally defined local frame and combining the bone-based view with compositional body-part tokenization, JaBOR expands the motion vocabulary available to Human-MLMs. The current evidence is preliminary but encouraging. Across reported baselines, JaBOR improves FID for MotionGPT and MotionGPT3 and preserves or slightly improves diversity in two of the three MLMs reported. These results support the broader hypothesis that motion-language models can benefit from exposure to **different languages of motion**, not only one joint-centric vocabulary.

Future work should strengthen this finding through controlled ablations, repeated trials, and task-level evaluations. In particular, we view cross-representation motion translation in bilingual motion language models and exercise form coaching as promising next steps. More broadly, this work suggests that representation design remains a central lever for improving the reasoning and generative capabilities of Human-MLMs.

References

[1] Biao Jiang, Xin Chen, Wei Liu, Jingbo Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations (ICLR)*, 2026.

[3] Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The Language of Motion: Unifying verbal and non-verbal language of 3D human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[4] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015.

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[8] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Xinshuang Zou, Minjing Ruhnke, Daniel Sun, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 2020.

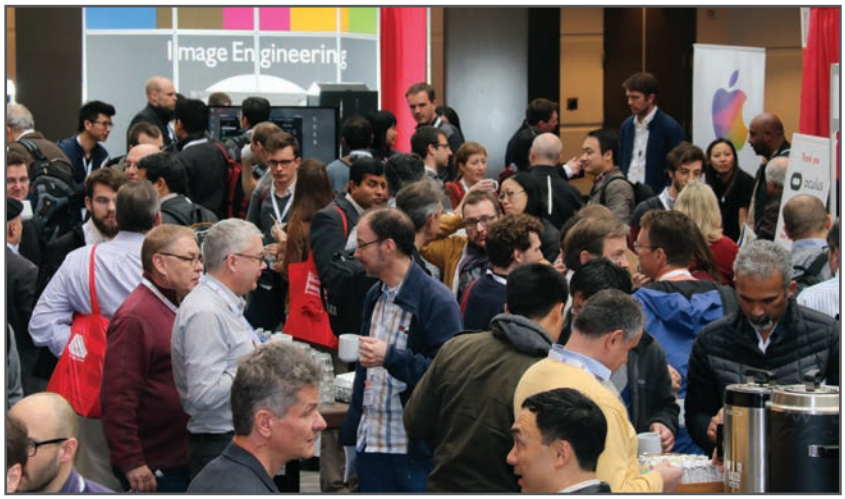
[13] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[14] eokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing. In *the Conference on Computer Vision and Pattern Recognition proceedings*, 2025.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org