Lightweight Morphology-Aware Encoding for Motion Learning

Ziyu Wu[1,2], Thomas Michel[2], Damien Rohmer[1]

[1]LIX, École polytechnique/CNRS, Institut Polytechnique de Paris, France
[2]Animaj, France

Abstract

We present a lightweight method for encoding, learning, and predicting 3D rigged character motion sequences that consider both the character's pose and morphology. Specifically, we introduce an enhanced skeletal embedding that extends the standard skeletal representation by incorporating the radius of proxy cylinders, which conveys geometric information about the character's morphology at each joint. This additional geometric data is represented using compact tokens designed to work seamlessly with transformer architectures. This simple yet effective representation demonstrated through three distinct tokenization strategies, maintains the efficiency of skeletal-based representations while enhancing the accuracy of motion sequence predictions across diverse morphologies. Notably, our method achieves these results despite being trained on a limited dataset, showcasing its potential for applications with scarce animation data.

CCS Concepts

• ***Computing methodologies*** → ***Animation;***

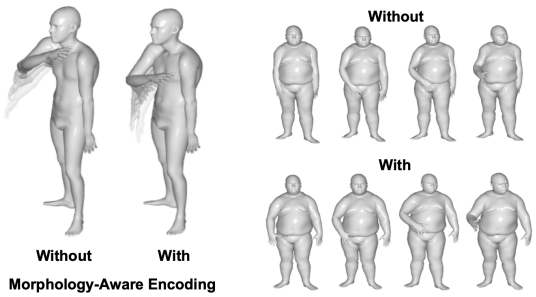

Figure 1: *Our Lightweight Morphology-Aware Encoding allows us to learn and infer intermediate skeletal motion without the need for the complete mesh while adapting to the character's shape, even for exaggerated forms. Left: A raising hand motion for a slim character is predicted more accurately than morphology-agnostic skeletal-based learning. Right: An upward motion of the arm of a large character learns to avoid its big belly.*

1. Introduction

Character animation is a cornerstone of modern media, including applications in animation cinema, video games, and virtual reality, where the demands for high quality generation are constantly increasing. Production pipelines typically rely on rigged characters composed of a hierarchical skeleton and a skin mesh attached through *skinning weights*, which are animated manually via keyframe techniques. While recent advancements in deep learning have enabled significant progress in automatically generating animations by leveraging high-quality datasets created by artists or motion capture, these methods are rarely adopted in professional production contexts. This gap arises because efficient data-driven motion generation approaches are often developed using a compact skeleton-based representation associated with a standard character shape. Such representations fail to capture the broad range of character morphologies encountered in the entertainment industry, including exaggeratedly thin or overweight characters. This limitation leads to two main effects: (i) the learned motion features do not account for subtle motion variations caused by weight and flexibility differences related to morphology. (ii) Transferring motion learned from one character to another can result in animation artifacts such as self-collisions. In contrast, mesh-based representations can fully capture character morphology while learning motion, but these representations are computationally and memory-intensive, making them impractical for real-world applications.

We propose in this work to combine the skeleton representation with a lightweight geometrical approximation of the mesh morphology to encode and learn the relationship between motion and character morphology. This approach maintains a compact, flexible, scalable, and efficient representation while leveraging information in standard character rig. Specifically, we enrich the traditional skeleton encoding by integrating a localized cylindrical approximation of the character's shape, with a radius associated with each joint associated along skinning weights. Our contribution involves the introduction and evaluation of three different encoding strategies for embedding this geometric information using transformer architectures. We show that the addition of such light geometrical

DOI: 10.2312/egs.20251048

information helps to improve the general accuracy for motion sequences compared to the use of skeleton-only representations and can qualitatively improve the generated motion, typically limiting self-collision for large characters in taking into account their morphology. The approach remains almost as computationally efficient and memory-light as skeleton-based representations, providing a scalable solution for application in 3D animation production.

2. Related Works

Learning motion sequences for 3D characters has become a very active area of research thanks to the development of deep learning techniques that have proven effective in learning motion through the 6 degrees of freedom of the joints in an animated skeleton [LZLL18]. These methods have been successfully applied to a variety of tasks, including retargeting, motion prediction, and motion generation [ZCP*24], and we refer to Mourot [MHC*22] for a more exhaustive listing. While skeletons can vary in bone length [LTIJ24], they are limited to representing only the pose of the character and cannot fully capture the character's morphology in relation to the space occupied by the real body parts, as defined by the character's skin. As a result, these approaches are highly efficient for encoding complex motion for characters with standard proportions, but still struggle to adapt to cartoon-like characters with exaggerated proportions, which restrict the range of motion in the limbs relative to their pose. Recent works like [ZLY*24] have primarily focused on skeleton or pose-based representations for tasks such as cross-domain motion retargeting.

Conversely, mesh representation defines a character's surface geometry using vertices and faces allowing point cloud interpolation. Applying learning frameworks at the level of the mesh vertices excels in capturing fine surface details making it adaptable to various designs. However, high-resolution meshes require significant memory and computational resources, posing challenges in real-time applications or when integrating with structured machine learning models like Transformers. Interestingly, intermediate approaches leveraging, for instance, spectral representation [LDLD23], have been explored, but remain an order of magnitude more costly compared to skeleton-based methods. Recent advancements, however, explore integrating shape information, as seen in works like [YLJ*24] which models dense geometric interactions, and [VCH*21] that focuses on contact preservation by considering character geometry. Furthermore, methods such as [LYS*22] operate without a skeleton, directly manipulating the mesh for pose transfer, and [ZCX*24] incorporates visual information from rendered characters to enhance motion retargeting.

Our approach is orthogonal to mesh-representation, but rather relies on the pre-existing relation between the skeleton and the skin mesh via the notion of skinning weights. We then enrich the skeletal representation with the notion of an approximated proxy representation of the character shape rather than full vertex information, therefore remaining close and compatible with skeletal-based approaches, while also leveraging lightweight geometrical information unlike purely skeleton-based methods.

3. Morphology-Aware Encoding

This section introduces our approach to enhance 3D character representation in skeletal animation by leveraging cylindrical mesh approximations aligned with the underlying skeleton structure. We consider a standard rigged character as input and assign each mesh point to the bone with the highest skinning weight. For each bone, we then build a proxy cylinder that approximates the local shape of the character. The cylinder's axis is aligned with the bone's direction, and its radius is set by the maximum distance between the mesh vertices assigned to the bone and the corresponding bone segment. As illustrated in Figure 2, this cylindrical mesh approximation represents the character as a set of cylinders that, possibly overlapping, completely wrap the character's body.

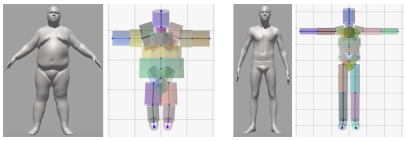

Figure 2: *Cylinder-based Mesh Approximation. For each character, left: original mesh, right: skeleton and cylinder.*

This Skeleton-Cylinder Approximation Model is then used to learn and predict motion sequences where the character morphology, which are represented by the cylinder radius and embedded in the learning-based architecture as an external token. The model use local Euler rotation angles at each joint as control values. In our application case, we aim to predict the middle frames of a motion sequence as shown in Figure 3. The entire sequence is used as input to the network, but the 25%-75% portion is masked, leaving the model aware only of the first and last 25% of the sequence. We adopted this specific 25-50-25 masking strategy to rigorously assess the model's capability to generate coherent and continuous motion sequence that seamlessly transitions between two given segments, which aligns with our intended application. While alternative masking strategies, such as randomly masking frames, could facilitate the development of a more generalized model for animated character motions, such an exploration falls beyond the scope of this study. Instead, our study specifically centers on evaluating the model's proficiency in the targeted in-betweening task.

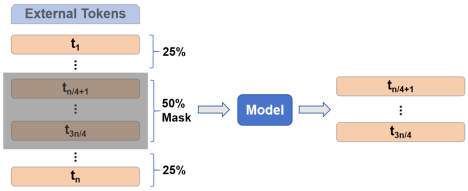

Figure 3: *Model Framework.*

We rely on a Transformer Encoder architecture to use the morphological information in encoding as fixed tokens, the bone lengths, and cylinder radii. We then propose and study three different tokenization strategies: (i) *One-token*: Aggregates all bone lengths and cylinder radii into a single token, simpler approach with low granularity. (ii) *Bone-token*: Encodes each bone's length and radius as separate tokens, providing finer skeletal detail and improving model specificity. (iii) *Time-Bone-token*: Combines time step control variables with corresponding bone length and radius

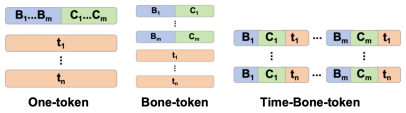

Figure 4: *Different Token Designs.*

into individual tokens for each time step, enhancing temporal coherence and motion prediction accuracy.

Figure 4 illustrates these tokenization approaches, where B_i represents the length of the i^{th} bone, C_i the radius of the i^{th} bone's cylinder, and t_j the j^{th} timestep in the animation sequence. In the experiment section, we will evaluate and discuss tokens containing only B (without radius information) and tokens containing B and C (with radius information) to verify their effectiveness.

4. Experiments

4.1. Dataset and Experiment Settings

We use the AMASS dataset [MGT*19] in considering identical motion sequences across multiple characters. We selected 6 distinct characters with different morphologies and extracted 289 motions sequences from the dataset, each spanning 40 time steps. Our framework focused on an in-betweening task, wherein the middle 50% of each motion sequence (corresponding to time steps 10–30) was masked. The model was tasked with predicting these missing frames using the initial and final 25% of the sequence as input. The data split, based on motion sequence across 6 characters, were the training set as 80% and the testing set as 20%. This dataset construction, comprising only 289 motion sequences in total, intentionally simulates a realistic production scenario where only a small number of characters have existing animations available for training. Consequently, while our results may not directly match the performance of state-of-the-art models that leverage significantly larger datasets, they demonstrate the potential of our approach to achieve promising results even with very limited training data. This highlights the model's ability to learn effectively in data-scarce environments, which is also a key contribution of this study.

Given the hierarchical nature of skeletal structures, errors in local transformations can accumulate with node depth (e.g. hand error accumulate with shoulder, elbow errors). To address this, the model employs an Inverse Linear Weights on loss function, defined as $\frac{1}{1+depth}$. This loss assigns higher weights to shallower nodes, effectively reducing depth-related error accumulation while maintaining accuracy for deeper nodes. After comparison with the uniform weights, we found this design can effectively reduce the overall error and bring better animation prediction results.

The model uses the token structure to encode skeletal and cylindrical information. Key configurations include 3 transformer encoder layers with a hidden dimension of 256, 8 attention heads, an L1 loss function with inverse linear weights, a learning rate of 0.001, and 150 training epochs. The experiments were conducted on a high-performance computing server equipped with one NVIDIA L4 GPU and an Intel(R) Xeon(R) CPU at 2.20 GHz. It should be noted that in order to ensure the accuracy of the results, the results in this section are the average of 3 runs.

4.2. Results

In this section, we evaluate the impact of different token designs on the performance of the Skeleton-Cylinder Approximation Model, using L1 loss with Inverse Linear Weights. This experiment aimed to assess how skeletal and morphology information structured as tokens influence motion sequence prediction accuracy. Results are shown in Figure 5 and Table 1. The improvement highlights the value of adding radius as morphology information in enhancing geometric accuracy, especially with improvements of more than 20% being achieved at most. Importantly, the less than 1.5% additional runtime increase from including radius information is negligible across all models, making it a practical enhancement.

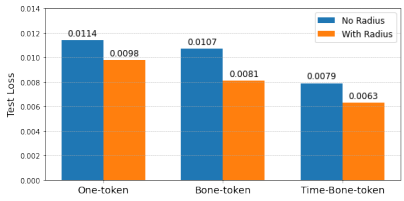

Figure 5: *Results on Different Token Models.*

Table 1: *Loss and Runtime (per sequence of 40 frames) with relative comparison to the no Radius variant.*

	Loss (Decrement)	Runtime
One-token - no Radius	0.0114 (—)	1.953s
One-token - with Radius	0.0098 (14.03%)	1.958s
Bone-token - no Radius	0.0107 (—)	1.951s
Bone-token - with Radius	0.0081 (24.30%)	1.960s
Time-Bone-token - no Radius	0.0079 (—)	4.211s
Time-Bone-token - with Radius	0.0063 (20.25%)	4.230s

When comparing different token designs, the Bone-token model reduced loss by 17.34% compared to One-token, while the Time-Bone-token achieved the highest improvement, reducing loss by 35.71% and 22.22% compared to One-token and Bone-token respectively. However, this accuracy comes at a runtime cost, with the Time-Bone-token requiring 116.03% more time than One-token and 115.81% more than Bone-token. Despite this, the approximately 20% improvement in loss demonstrates its suitability for accuracy-critical applications, while the Bone-token model offers a balanced trade-off between accuracy and computational efficiency, making it a better choice for resource-constrained scenarios.

To further illustrate the impact of including radius information, Figure 6 provides a qualitative comparison of motion sequence predictions for sample sequences using the Time-Bone-token. The model with radius information produces smoother and more accurate predictions. In contrast, the model without radius information demonstrates significant issues, with collisions between mesh and uncoordinated movements observed. These highlight the model's inability to account for geometric constraints without the radius information, leading to less realistic and physically implausible motion predictions. Also in Figure 1, we observe self-collisions between the arm and body. Interestingly, in the baseline prediction for the slim character, we observe issues such as collisions and oscillations, where the hand moves excessively in intermediate

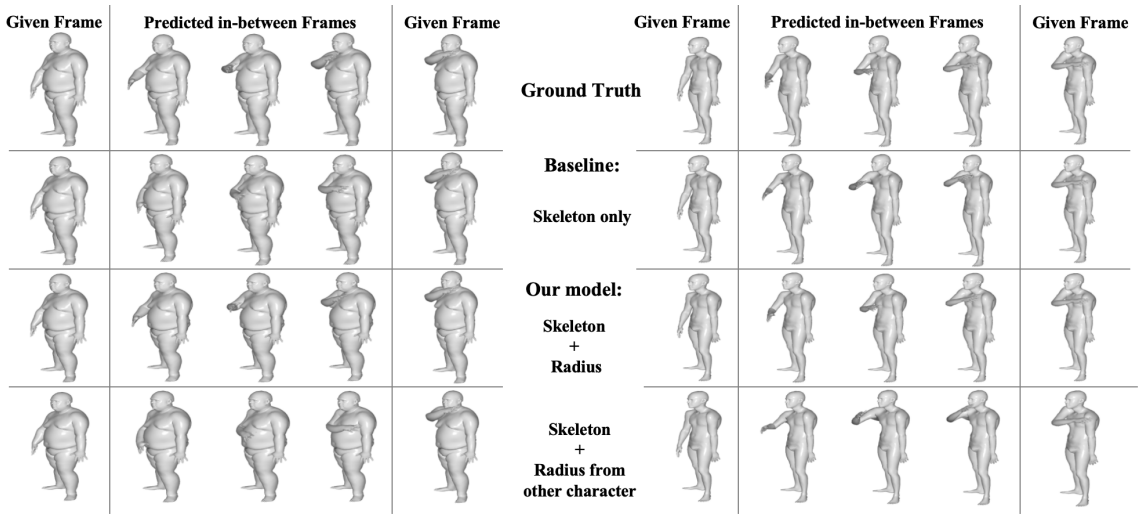

Figure 6: *Results for two different characters on the same animation sequence: Raise hand.*

frames compared to its position at the end of the given sequence. This counterintuitive behavior could originate from the model being trained on a dataset containing characters of varying shapes. Without shape-specific information, the model may generate predictions that oscillate between extremes observed in the training data, leading to unrealistic motion. Additionally, we tested the robustness of our approach in using different cylinder radii. For instance, Figure 6-bottom row illustrates the case where the cylinder radius is switched between the two characters. As expected, the bulky character adopts the motion style of the slimmer one, and vice versa.

5. Conclusion

This study presented a lightweight Morphology-Aware Encoding model, effectively integrating skeletal information and geometric approximations to enhance motion sequence prediction in 3D animation. Our experiments demonstrated that incorporating radius information and employing finer-grained token representations, such as the Time-Bone-token approach, significantly improved the model's ability to capture both temporal and spatial dependencies, as evidenced by both qualitative and quantitative evaluations. A key advantage of our approach lies in its inherent simplicity, facilitating straightforward implementation across diverse contexts. Importantly, our model exhibits remarkable effectiveness even when trained on small datasets, as demonstrated in our experiments with a limited number of sequences and characters. This makes it particularly well-suited for realistic production environments where training data for specific or unique characters, such as those prevalent in cartoon production, might be scarce. Future work will focus on refining token structures for enhanced efficiency, exploring the incorporation of additional skeletal parameters, and investigating more detailed geometric primitives beyond cylinders for surface approximation, further solidifying the model's capabilities in data-constrained scenarios.

References

[LDLD23] LEMEUNIER C., DENIS F., LAVOUE G., DUPONT F.: SpecTrHuMS: Spectral transformer for human mesh sequence learning. *Computers and Graphics 115* (2023). 2

[LTIJ24] LOVANSHI M., TIWARI V., INGLE R., JAIN S.: 3d skeleton-based non-autoregressive human motion prediction using encoder-decoder attention-based model. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024). 2

[LYS*22] LIAO Z., YANG J., SAITO J., PONS-MOLL G., ZHOU Y.: Skeleton-free pose transfer for stylized 3d characters. In *European Conference on Computer Vision* (2022), Springer, pp. 640–656. 2

[LZLL18] LI C., ZHANG Z., LEE W. S., LEE G. H.: Convolutional sequence to sequence model for human dynamics. *CVPR* (2018). 2

[MGT*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision* (Oct. 2019), pp. 5442–5451. 3

[MHC*22] MOUROT L., HOYET L., CLERC F. L., SCHNITZLER F., HELLIER P.: A survey on deep learning for skeleton-based human animation. *Eurographics STAR, CGF 41* (2022). 2

[VCH*21] VILLEGAS R., CEYLAN D., HERTZMANN A., YANG J., SAITO J.: Contact-aware retargeting of skinned motion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9700–9709. doi:10.1109/ICCV48922.2021.00958. 2

[YLJ*24] YE Z., LIU J.-W., JIA J., SUN S., SHOU M. Z.: Skinned motion retargeting with dense geometric interaction perception. In *Advances in Neural Information Processing Systems* (2024), Globerson A., Mackey L., Belgrave D., Fan A., Paquet U., Tomczak J., Zhang C., (Eds.), vol. 37, Curran Associates, Inc., pp. 125907–125934. 2

[ZCP*24] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI 46*, 6 (2024). 2

[ZCX*24] ZHANG H., CHEN Z., XU H., HAO L., WU X., XU S., ZHANG Z., WANG Y., XIONG R.: Semantics-aware motion retargeting with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 2155–2164. 2

[ZLY*24] ZHAO Q., LI P., YIFAN W., OLGA S.-H., WETZSTEIN G.: Pose-to-motion: Cross-domain motion retargeting with pose prior. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15170. 2