Continuous Intermediate Token Learning With Implicit Motion Manifold for Keyframe Based Motion Interpolation

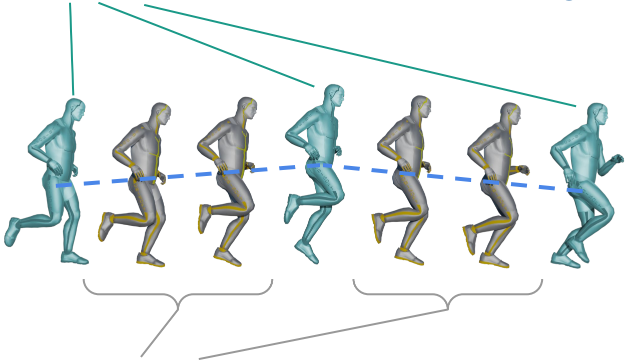
Clinton Mo
clmo6615@uni.sydney.edu.au | University of Sydney
(Kun Hu, Chengjiang Long, & Zhiyong Wang)

IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023
WED-PM-147
17-22 June 2023

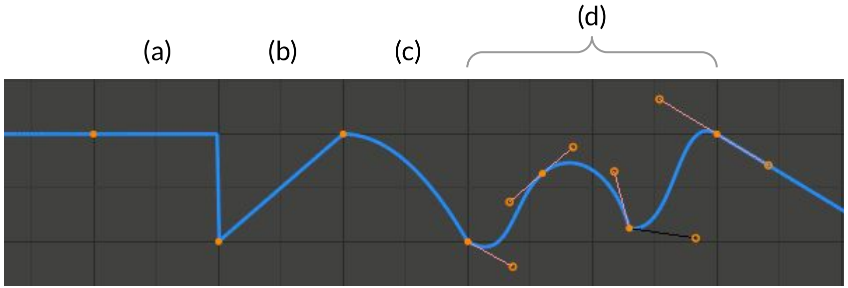
THE UNIVERSITY OF SYDNEY

Motion Interpolation in Animation Processes

Keyframes: The definitions and timings of motion details, in the form of **key poses**.



Interpolation: To fill the temporal gaps between keyframes



Common interpolation algorithms:

(a) Constant
(b) Linear interpolation (a.k.a. LERP)
(c) Quadratic/polynomial functions
(d) Bezier/spline functions

Refining Keyframed Motions

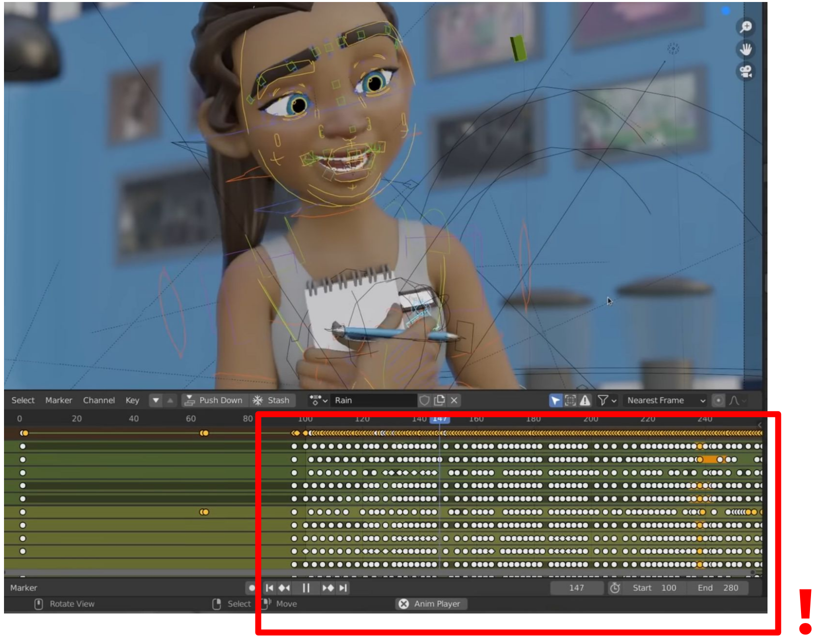
Motion features *only* exist when defined by a keyframe

+ Greater control over final motion

– Burden on animator manpower to **define all details**

– High cost for detail-heavy styles, notably **realistic** motions.

– **Destructive** keyframing process:
Early mistakes = **Start-over** for refining



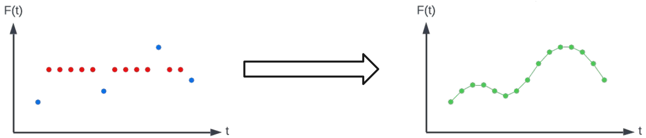
We can automate this process using **learned interpolation**

Transformer was built for Discrete Data

Existing masking approaches do not address the **continuous nature** of motion data

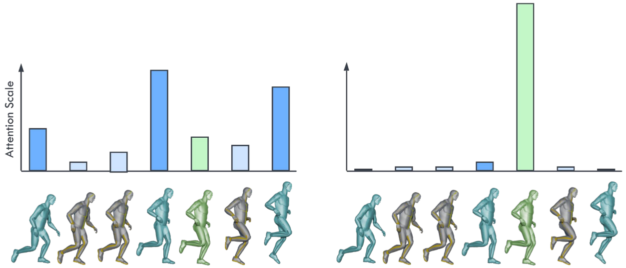
Monolithic masking[1]

- Discrete representation
- Poor predictive precision
  - Discontinuous input → Smooth continuous output



LERP masking [2][3]

+ **Continuous** representation, *however*
- Provides a trivial local minimum
- Over-reliance on LERP pattern



Expected attention scores   Actual attention scores

[1] Oreshkin, B. N., Valkanas, A., Harvey, F. G., Ménard, L. S., Bocquelet, F., & Coates, M. J. (2022). Motion Inbetweening via Deep Δ-Interpolator. arXiv preprint arXiv:2201.06701.

[2] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In CVPR'22.

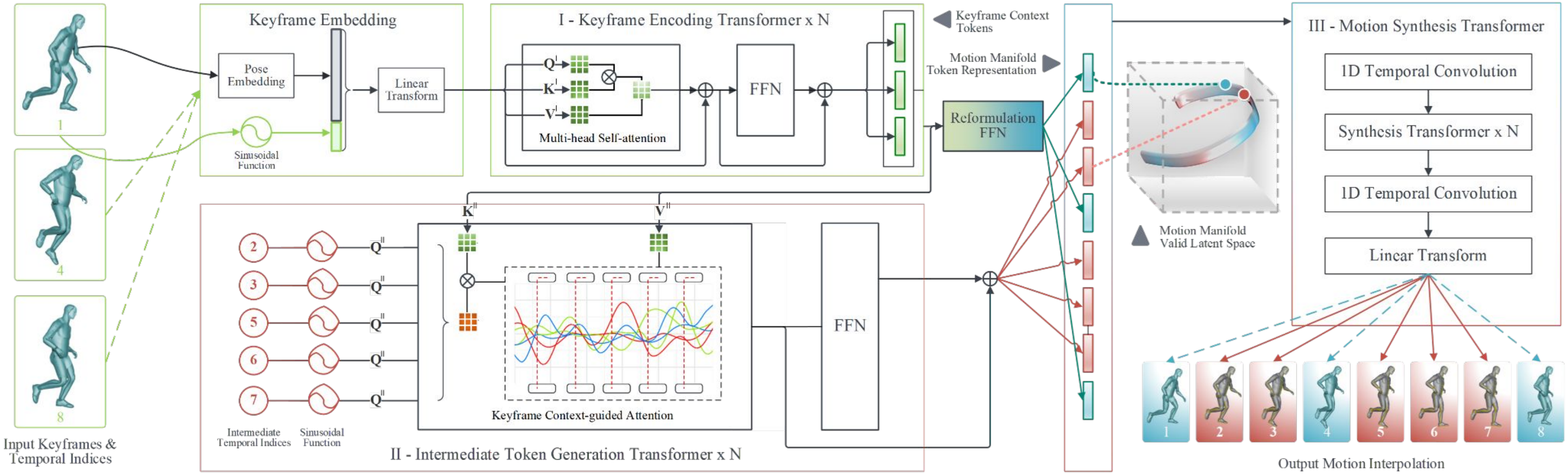
[3] Duan, Y., Lin, Y., Zou, Z., Yuan, Y., Qian, Z., & Zhang, B. (2022, June). A unified framework for real time motion completion. In AAAI'22.

Our approach: Implicit manifold learning

Manifold: Smooth high-dimensional surface representation for masks

Three-stage process:

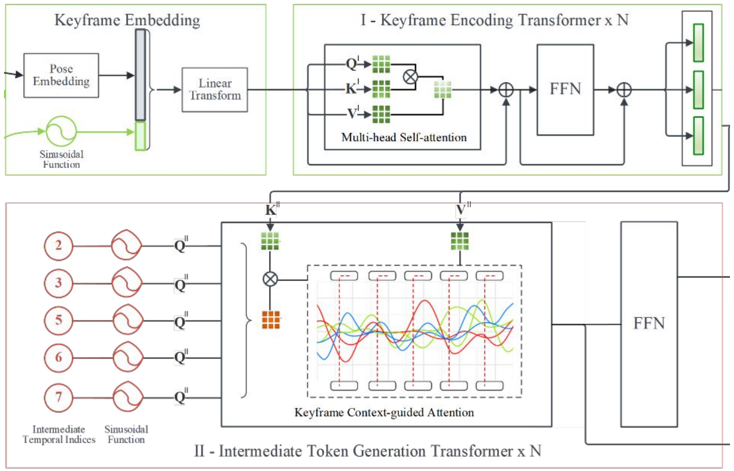
① Keyframe encoding → ② Initial manifold estimation → ③ Constrained manifold & refinement



Initial Manifold Estimation

Context-guided attention

- *Query*: Sinusoidal positions of intermediate frames
- *Key & Value*: Stage-I embeddings map a **latent subspace** on temporal axis
- Self-attention is removed to increase independent feature variety



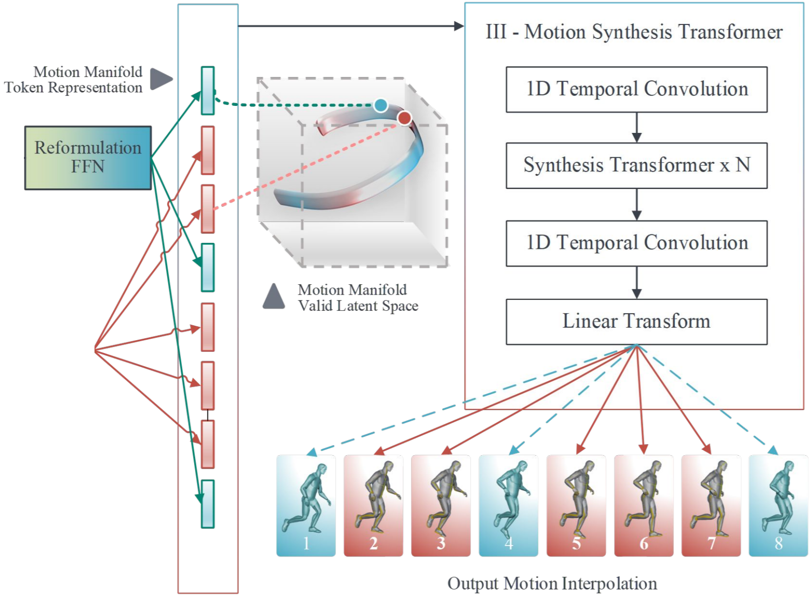
Manifold Control & Refinement

Reformulation FFN

- Transforms Stage-I encodings → equivalent Stage-II manifold elements
- Enables Stages I & II to **cooperate**

Motion Synthesis Transformer

- *Convolutional interactions* necessitate cohesive manifold behaviour
- Refinement of Stage-II motion details through *self-attention*

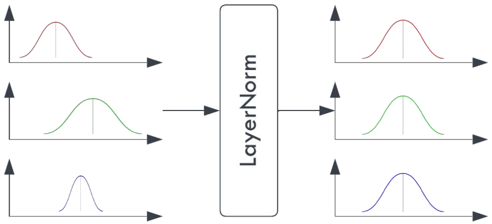


Sequence-level Re-centreing (Seq-RC)

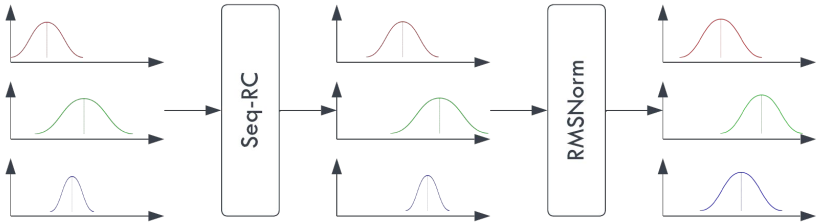
Transformer models normalise feature vectors using **LayerNorm**

- Token-wise normalisation disables **consistent feature magnitudes** in continuous sequences

Instead, we propose **sequence re-centreing + RMSNorm**[1] to preserve feature magnitudes



Layer Normalisation



Our method (Seq-RC)

[1] Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. NeurIPS 2019.

Consistent improvement across all metrics

Our method outperforms all SOTA approaches in all major metrics:

- **L2P**: Global joint position L2
- **L2Q**: Global joint quaternion rotation L2
- **NPSS**: Fourier-based visual similarity

	L2P			L2Q			NPSS		
KF interval	**5**	**15**	**30**	**5**	**15**	**30**	**5**	**15**	**30**
LERP	0.178	0.837	1.327	**0.146**	0.544	0.779	0.073	0.304	0.642
BERT[1]	0.277	0.886	1.331	0.222	0.584	0.803	0.092	0.280	0.602
$TG_{Complete}$[2]	0.299	0.854	1.391	0.244	0.608	0.923	0.136	0.401	0.628
MAE[3]	0.275	0.737	1.123	0.262	0.536	0.757	0.111	0.299	0.585
Δ-interpolator[4]	0.209	0.823	1.313	0.158	0.492	0.770	0.091	0.267	0.638
Our method	**0.151**	**0.557**	**0.940**	0.163	**0.455**	**0.677**	**0.052**	**0.216**	**0.450**

[1] Duan, Y., Lin, Y., Zou, Z., Yuan, Y., Qian, Z., & Zhang, B. (2022, June). A unified framework for real time motion completion. In AAAI'22.

[2] Harvey, F. G., Yurick, M., Nowrouzezahrai, D., & Pal, C. (2020). Robust motion in-betweening. ACM TOG'20.

[3] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In CVPR'22.

[4] Oreshkin, B. N., Valkanas, A., Harvey, F. G., Ménard, L. S., Bocquelet, F., & Coates, M. J. (2022). Motion Inbetweening via Deep Δ-Interpolator. arXiv preprint.

Summary

This research work claims three contributions:

1) Transformer-based **manifold learning** architecture for motion interpolation
2) **Sequence-level re-centreing** for continuous feature representations
3) **Comparative** & **ablation study** of our method vs SOTA & various settings