Continuous Intermediate Token Learning with Implicit Motion Manifold for Keyframe Based Motion Interpolation – Supplemental Material –

Clinton A. Mo[1], Kun Hu[1,*], Chengjiang Long[2], Zhiyong Wang[1]
[1]School of Computer Science, The University of Sydney, NSW 2006, Australia
[2]Meta Reality Labs, Burlingame, CA, USA
clmo6615@uni.sydney.edu.au, {kun.hu, zhiyong.wang}@sydney.edu.au, clong1@meta.com

Abstract

The supplementary material provides additional ablation studies on the two datasets: LaFan1 and CMU datasets. More detailed experimental results are discussed regarding the effectiveness of Stage-II and our initialisation strategy, the robustness of the proposed method for noisy latent representations, as well as the usage of global position & rotation loss terms. Moreover, we discuss the capability of using our method for the scenario of non-uniformly distributed keyframes. Note that we could not include the material in the main part of the paper due to the space limit.

1. Intermediate token generation stage

As shown in Table 1, removing Stage II leads to worse performance on both LaFan1 and CMU datasets. In this case, keyframe guidance is not adopted to the tokenization of missing frames. This results in a tendency that the intermediate tokens distribute differently with the keyframe representations in the latent space.

Table 1. Ablation study by removing stage II in place of MAE-styled monolithic tokens.

Dataset	LaFAN1			CMU		
Method	L2P	L2Q	NPSS	L2P	L2Q	NPSS
w/ Stage II	**0.3790**	**0.3951**	**0.1677**	**0.2226**	**0.3157**	**0.1625**
w/o Stage II	0.4681	0.4808	0.2003	0.3823	0.4304	0.2762

2. Stage-II with LERP initialisation

As shown in Table 2, LERP initialisation for Stage-II is not as good as the strategy devised in the proposed method. In the context of motion interpolation, the inclusion of an LERP reference generally leads to local optimums, even though it can provide initially reasonable results.

Table 2. Comparison between initial query tokens of Stage II in our architecture.

Dataset	LaFAN1			CMU		
Stage II Initialization	L2P	L2Q	NPSS	L2P	L2Q	NPSS
LERP+Temporal Indices	0.7518	0.6577	0.2891	0.3521	0.4275	0.2670
Temporal Indices	**0.3790**	**0.3951**	**0.1677**	**0.2226**	**0.3157**	**0.1625**

3. The stochasticity of human body motion

We have investigated stochasticity by adding Gaussian noises with $\sigma = 0.005$ to the latent space and the evaluation metrics are listed in Table 3. It can be observed that the proposed method is generally robust to Gaussian noises. Specifically, in $\text{TG}_{\text{complete}}$ [2], due to the unknown future keyframe, stochasticity is important for adapting to unknown scenarios after the next known keyframe. In our method. we address animation workflows where all keyframes are known, and thus the stochasticity is not necessary.

Table 3. Impact of stochasticity in the latent space.

Dataset	LaFAN1			CMU		
Stage III input	L2P	L2Q	NPSS	L2P	L2Q	NPSS
w/ Gaussian noise $\sigma = 0.005$ noise	0.4218	0.4169	0.2301	0.2993	0.3892	0.2311
w/o noise	**0.3790**	**0.3951**	**0.1677**	**0.2226**	**0.3157**	**0.1625**

4. Global position & rotation losses

Global objectives allow the model to balance local pose feature accuracy when it would lead to more effective global joint positions and rotations. This becomes a more prominent issue with joints that are further down the skeletal hierarchy. Furthermore, since we evaluate the performance in the global space, i.e. L2P and L2Q, optimising global fea-

*Corresponding author.

Table 4. Performance of our method (left) and the masked auto-encoder method (right) with uniformly and randomly distributed keyframes, tested with the LaFAN1 dataset.

Keyframe distribution	*Our method - Random*			*Our method - Uniform*			Keyframe distribution	*MAE - Random*			*MAE - Uniform*		
# of keyframes	L2P	L2Q	NPSS	L2P	L2Q	NPSS	# of keyframes	L2P	L2Q	NPSS	L2P	L2Q	NPSS
5	1.1668	0.7473	1.0356	0.7630	0.5870	0.4118	5	1.4231	0.8969	1.3666	1.0161	0.7410	0.6108
7	0.7949	0.5889	0.6307	0.5007	0.4436	0.2306	7	1.0640	0.7423	0.8633	0.7770	0.5887	0.4238
9	0.5885	0.4797	0.4053	0.3782	0.3584	0.1625	9	0.8564	0.6375	0.6093	0.6615	0.4991	0.3555
11	0.4755	0.4100	0.3076	0.2993	0.2990	0.1228	11	0.7388	0.5642	0.5158	0.5514	0.4325	0.2803
13	0.4067	0.3632	0.2514	0.2438	0.2545	0.0990	13	0.6637	0.5151	0.4490	0.4794	0.3800	0.2416
15	0.3491	0.3291	0.2088	0.2103	0.2247	0.0864	15	0.5922	0.4700	0.3869	0.4182	0.3415	0.2164

Table 5. Comparison between different loss setting when removing global positional or rotational loss, i.e. L_{FK_p} and L_{FK_q}.

Dataset	LaFAN1			CMU		
Method	L2P	L2Q	NPSS	L2P	L2Q	NPSS
Our method	**0.3790**	**0.3951**	**0.1677**	**0.2226**	**0.3157**	**0.1625**
w/o global position loss	0.6179	0.4950	0.2411	0.3391	0.3891	0.2368
w/o global rotation loss	0.5897	0.5216	0.2644	0.3150	0.4547	0.2525

tures becomes more necessary. The benefit of global loss functions is shown in Table 5.

5. Non-uniformly distributed keyframes

Our method can be used with arbitrarily defined keyframes in variable-length motions. Table 4 shows a comparison between randomly and uniformly distributed keyframe settings on LaFAN1 dataset. The metrics were evaluated with our method and MAE method [3]. Since random keyframes generally have higher entropy for representation, they lead to lower performances. All results of our main paper's experiments, as well as other results in the supplementary material, are conducted under uniform sampling for consistency.

6. More visualisations

Figure 1 - 3 illustrates three additional interpolation examples for the comparison of the LERP, BERT [1], Δ-interpolator [4], $\text{TG}_{\text{complete}}$ [2], MAE [3] and our proposed method with their L2P, L2Q and NPSS metrics. Specifically, the three examples are for motions of *getting up*, *sport with walking*, and *avoiding obstacles*.

7. Video demo

A video demo containing qualitative comparisons for different methods can be found at: https://youtu.be/V2DOMoqEBvQ.

References

[1] Yinglin Duan, Yue Lin, Zhengxia Zou, Yi Yuan, Zhehui Qian, and Bohan Zhang. A unified framework for real time motion completion. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 4459–4467, 2022. 2

[2] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics*, 39(4):60–1, 2020. 1, 2

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[4] Boris N Oreshkin, Antonios Valkanas, Félix G Harvey, Louis-Simon Ménard, Florent Bocquelet, and Mark J Coates. Motion inbetweening via deep δ-interpolator. *arXiv preprint arXiv:2201.06701*, 2022. 2

Figure 1. Illustration of the interpolation on a motion sequence of *getting up*.

Figure 2. Illustration of the interpolation on a motion sequence of *sport with walking*.

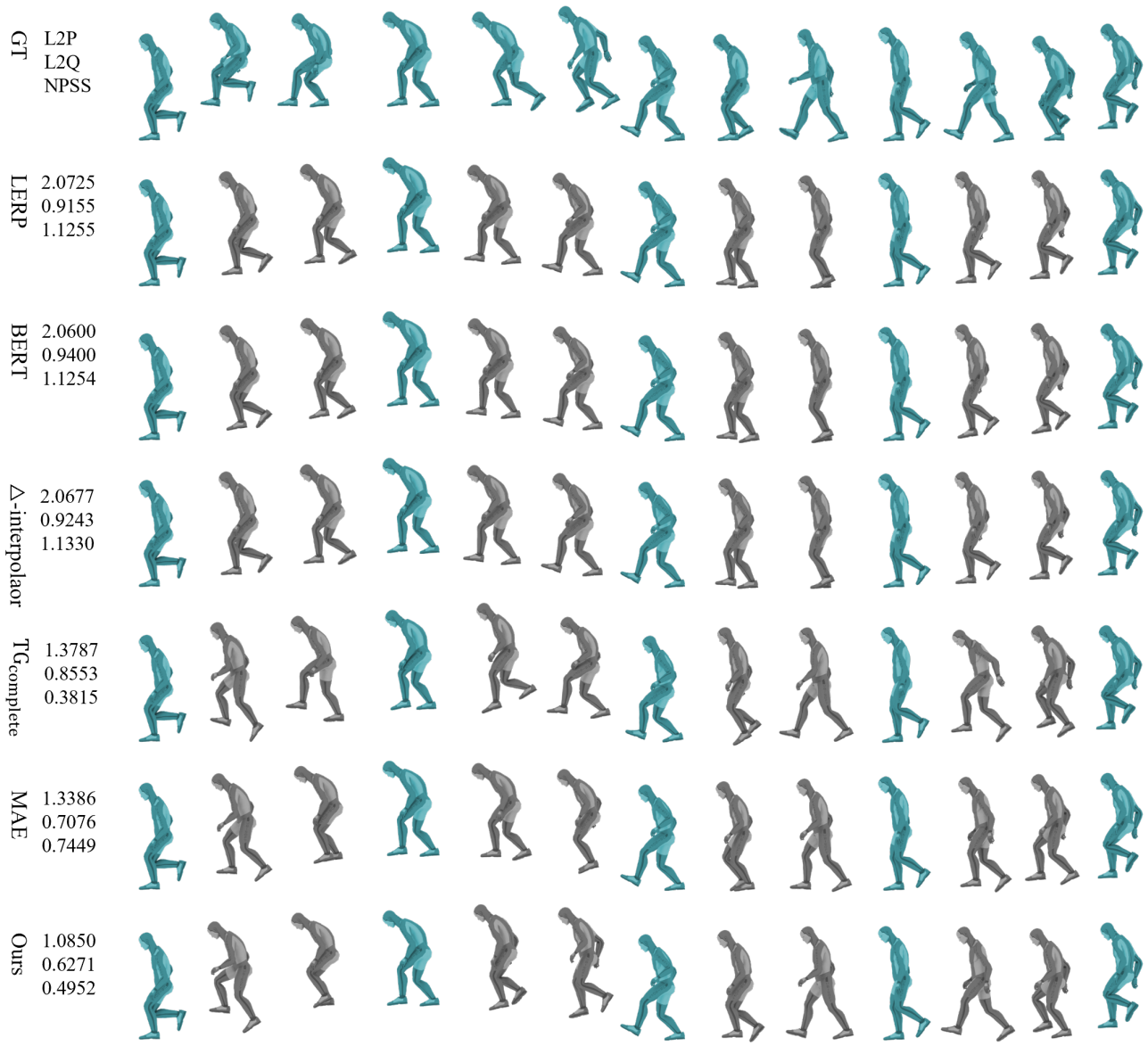

Figure 3. Illustration of the interpolation on a motion sequence of *navigating obstacles*.