PG20+21
WELLINGTON NZ

Diverse Dance Synthesis via Keyframes with Transformer Controllers

Junjun Pan[1,2], Siyuan Wang[1], Junxuan Bai[1,2], Ju Dai[2]

1 Beihang University   2 Peng Cheng Laboratory

北京航空航天大學
BEIHANG UNIVERSITY

VR 虚拟现实技术与系统国家重点实验室
CHINA'S STATE KEY LABORATORY OF VIRTUAL REALITY TECHNOLOGY AND SYSTEMS

鹏城实验室 PCL Peng Cheng Laboratory

Contents

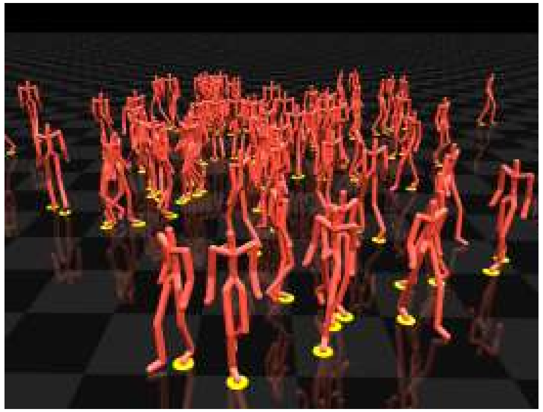
- Introduction
- Method details
- Experimental results
- Limitations and conclusion

Motivation

- Animation production is tedious and time-consuming.
- Little research investigates efficient keyframe-based motion synthesis for an impromptu dance performance with neural networks.
- Transformer can well model long-range dependency by leveraging the query-key correlation to different tokens.

Character in film [*“Wreck-It Ralph”*]

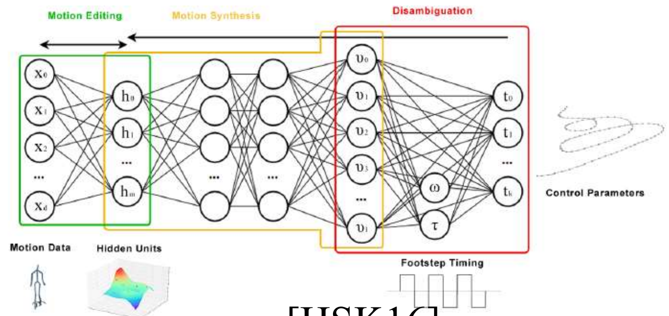
Motion capture



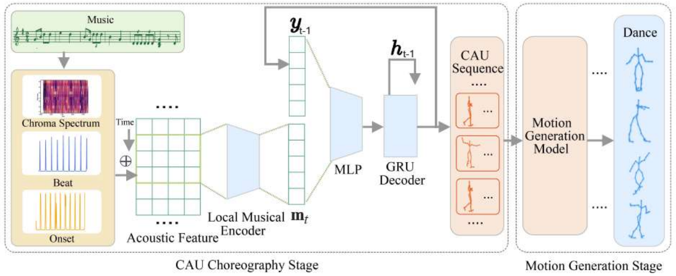
Research on walk animation [WCX21]

Related work

- **Deep learning based motion modeling**
  - Deep learning framework for motion synthesis and editing [HSK16]
  - Spatio-temporal manifold learning [WHSZ21]
- **Dance motion synthesis**
  - Dance with melody [TJM18]
  - Choreonet [YWJ*20]
- **Motion transition generation**
  - Data-driven autocompletion for keyframe animation [ZvdP18]
  - Robust motion in-betweening [HYNP20]



[HSK16]



[YWJ*20]

[HYNP20]

Challenges

- Dance performance is **highly irregular** with complex kinetics, and dance movements are inherently diversified.
- A well-choreographed dance animation requires **collaborative efforts** between animators, dancers, and choreographers, which is an expensive and tedious process.
- We propose a novel **keyframe-based** motion generation network **based on multiple constraints**, which can achieve diverse dance synthesis **via learned knowledge**.

Contributions

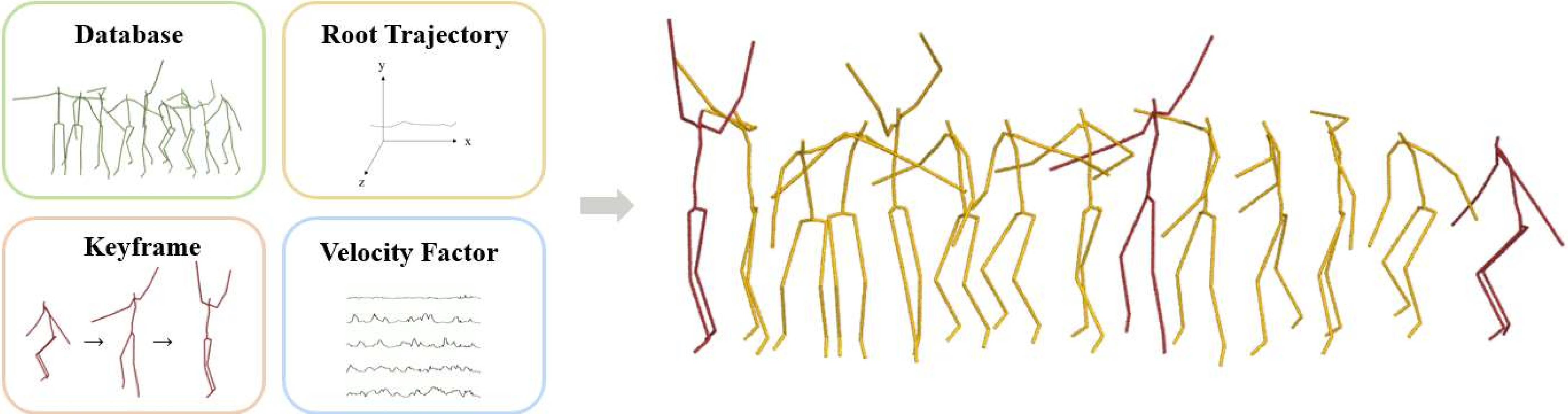
- A novel neural network based on **LSTM** and **Transformer** for complex motion generation via keyframes. The model is **elaborately controlled** under the root trajectory and the velocity factor.
- We design the **velocity factor** constraint for fine-grained dance motion synthesis.
- The data synthesized by our technique on the dance dataset **obtain better accuracy** in terms of various evaluation criteria, and the quality of character animation **is also higher**.

Contents

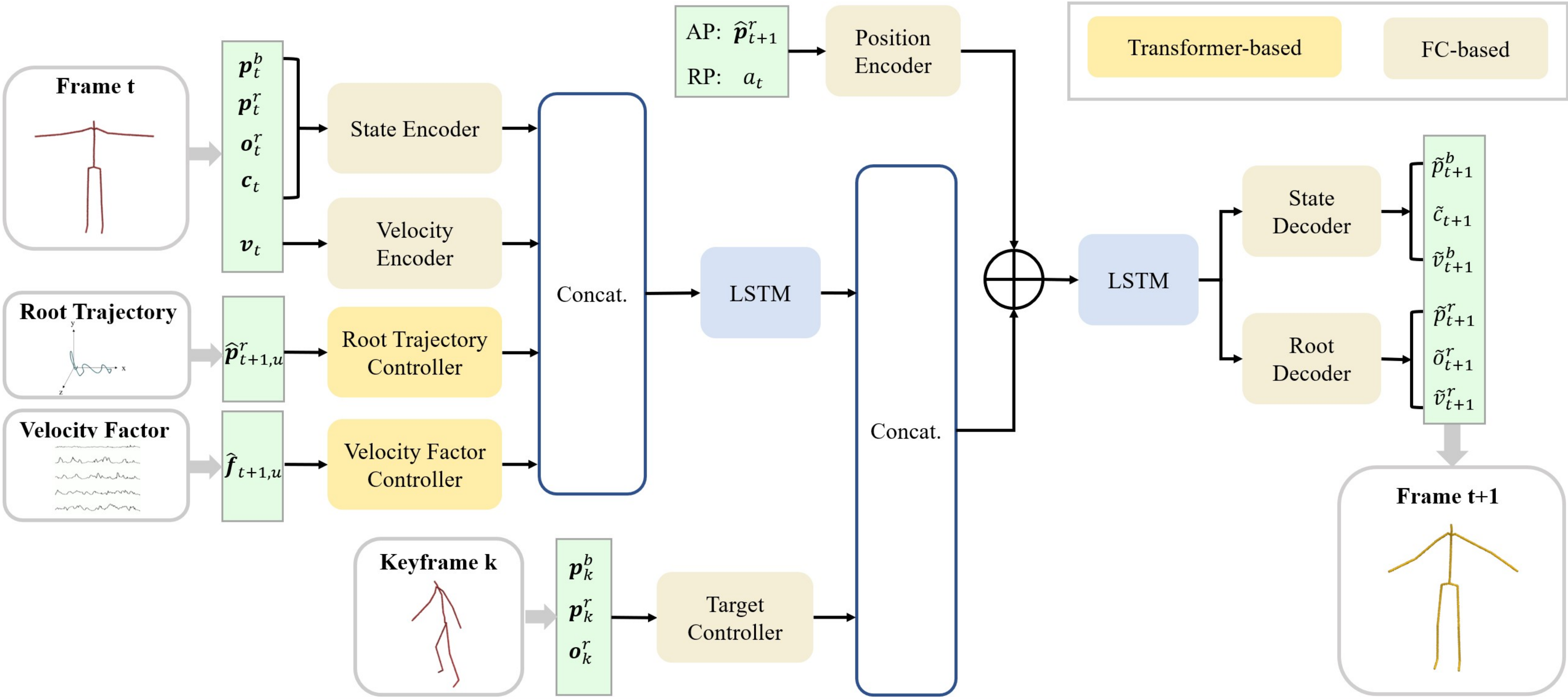
- Introduction
- Method details
- Experimental results
- Limitations and conclusion

Method overview

- A time series modeling and synthesis problem.
- Conditions: keyframes, root trajectory, velocity factor.
- Targets: diverse and natural dance animations.

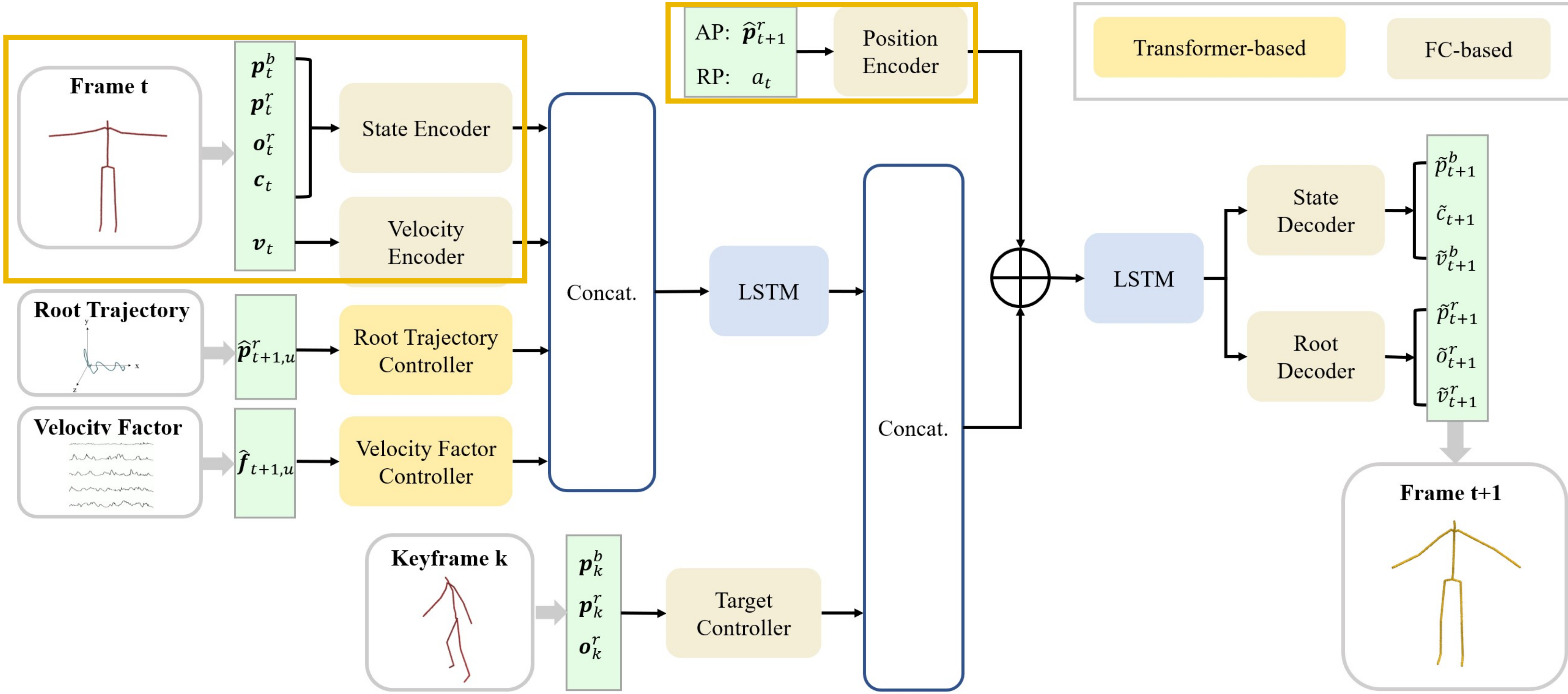


Framework



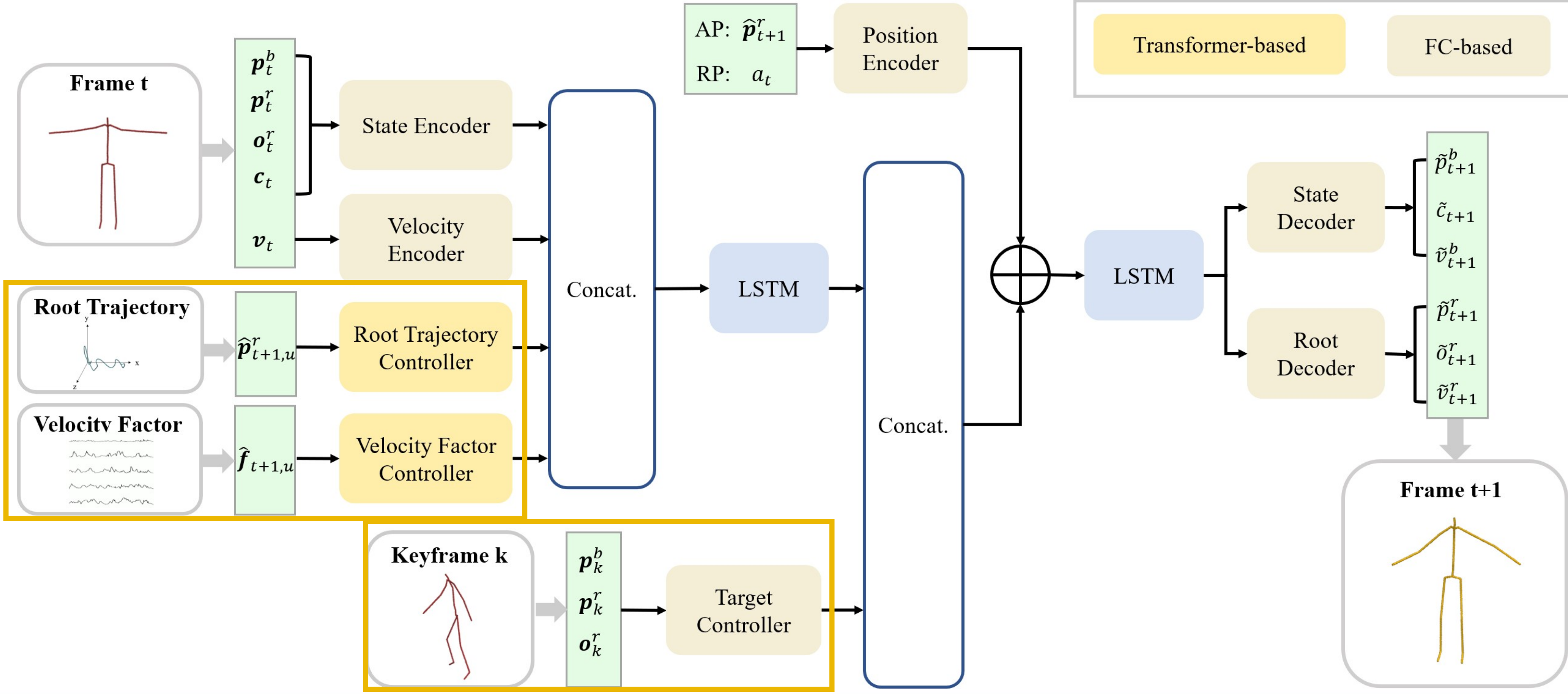
Frame t
p_t^b
p_t^r
o_t^r
c_t
v_t
State Encoder
Velocity Encoder
Root Trajectory
$\hat{p}_{t+1,u}^r$
Root Trajectory Controller
Velocity Factor
$\hat{f}_{t+1,u}$
Velocity Factor Controller
Concat.
LSTM
AP: $\hat{p}_{t+1}^r$
RP: a_t
Position Encoder
Transformer-based
FC-based
Concat.
LSTM
State Decoder
Root Decoder
$\tilde{p}_{t+1}^b$
$\tilde{c}_{t+1}$
$\tilde{v}_{t+1}^b$
$\tilde{p}_{t+1}^r$
$\tilde{o}_{t+1}^r$
$\tilde{v}_{t+1}^r$
Frame t+1
Keyframe k
p_k^b
p_k^r
o_k^r
Target Controller

Framework



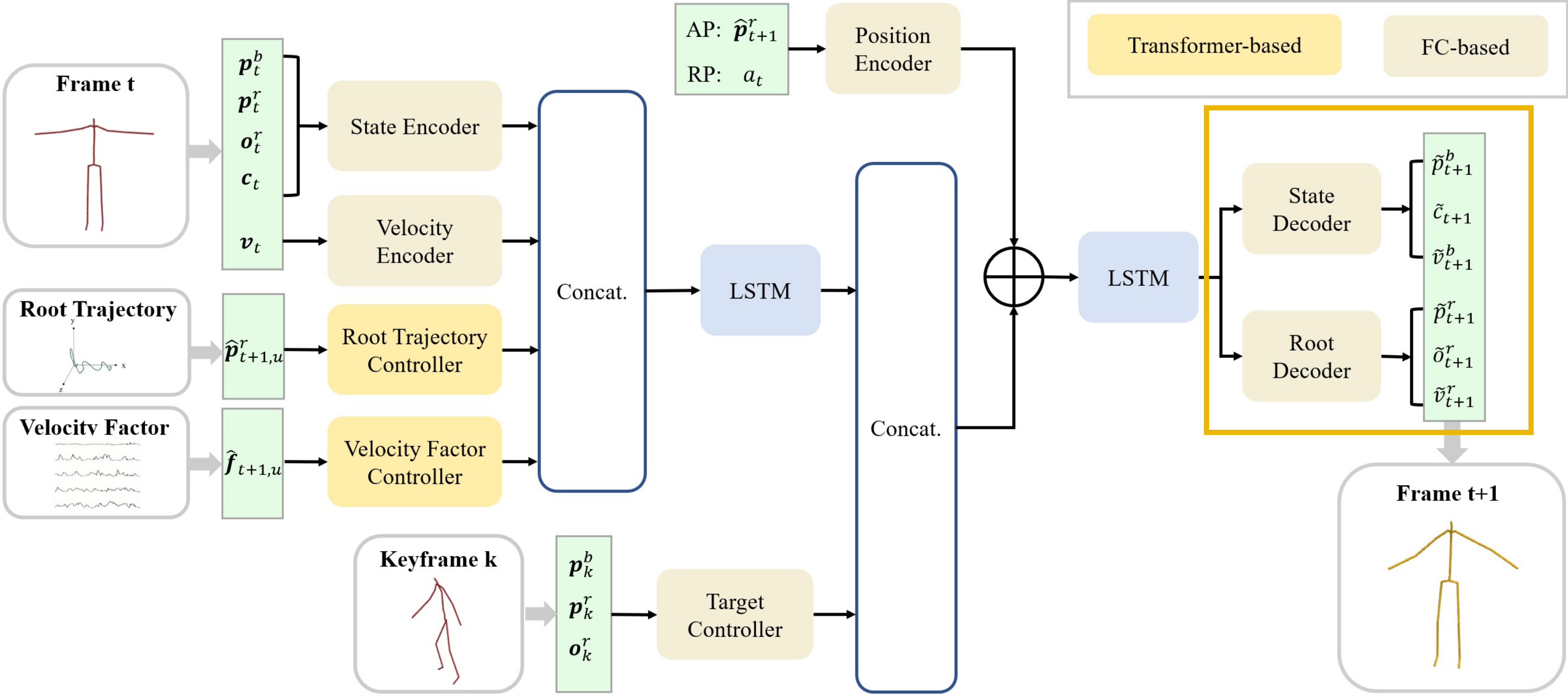
Frame t
p_t^b
p_t^r
o_t^r
c_t
v_t
State Encoder
Velocity Encoder
Root Trajectory
$\hat{p}_{t+1,u}^r$
Root Trajectory Controller
Velocity Factor
$\hat{f}_{t+1,u}$
Velocity Factor Controller
Concat.
LSTM
AP: $\hat{p}_{t+1}^r$
RP: a_t
Position Encoder
Concat.
Keyframe k
p_k^b
p_k^r
o_k^r
Target Controller
Transformer-based
FC-based
LSTM
State Decoder
Root Decoder
$\tilde{p}_{t+1}^b$
$\tilde{c}_{t+1}$
$\tilde{v}_{t+1}^b$
$\tilde{p}_{t+1}^r$
$\tilde{o}_{t+1}^r$
$\tilde{v}_{t+1}^r$
Frame t+1

Framework



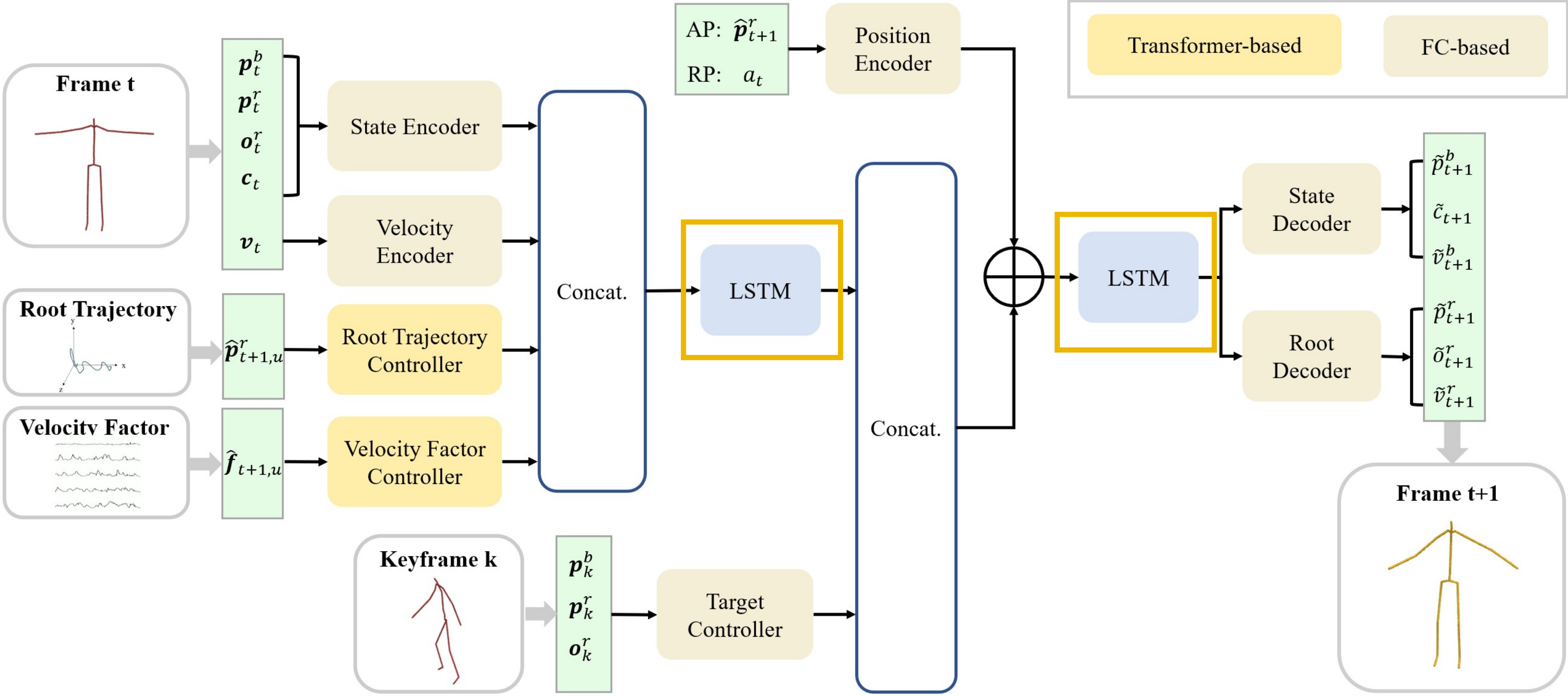
Frame t
p_t^b
p_t^r
o_t^r
c_t
v_t
State Encoder
Velocity Encoder
Concat.
AP: $\hat{p}_{t+1}^r$
RP: a_t
Position Encoder
Transformer-based
FC-based
LSTM
Concat.
LSTM
State Decoder
Root Decoder
$\tilde{p}_{t+1}^b$
$\tilde{c}_{t+1}$
$\tilde{v}_{t+1}^b$
$\tilde{p}_{t+1}^r$
$\tilde{o}_{t+1}^r$
$\tilde{v}_{t+1}^r$
Root Trajectory
$\hat{p}_{t+1,u}^r$
Root Trajectory Controller
Velocity Factor
$\hat{f}_{t+1,u}$
Velocity Factor Controller
Keyframe k
p_k^b
p_k^r
o_k^r
Target Controller
Frame t+1

Framework



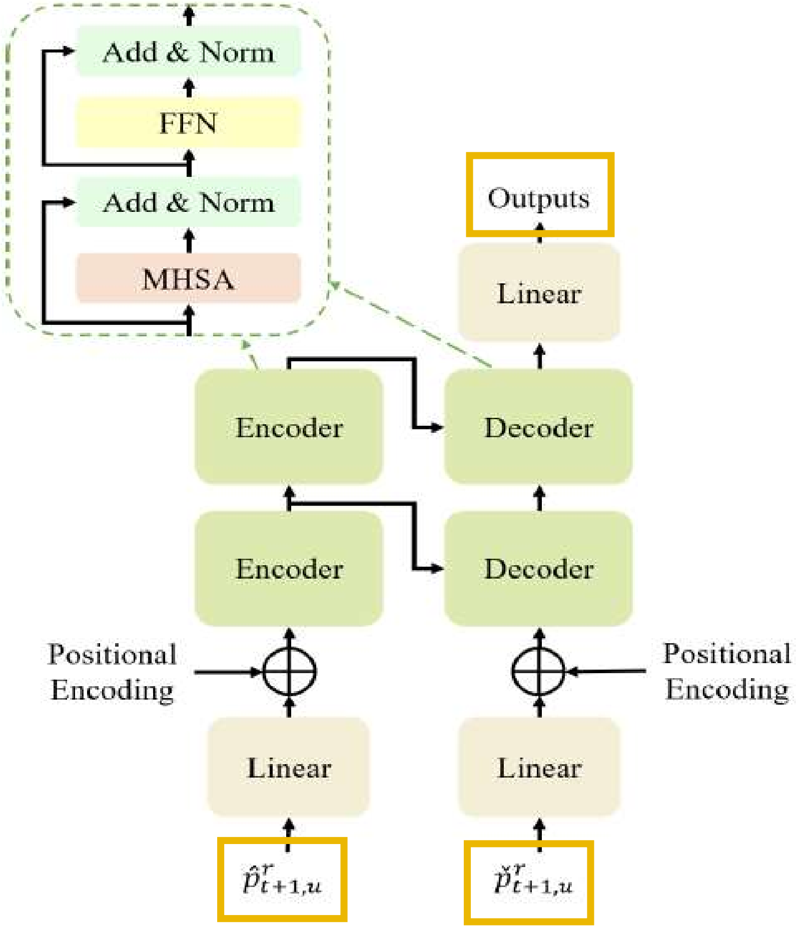
Frame t
p_t^b
p_t^r
o_t^r
c_t
v_t
State Encoder
Velocity Encoder
Root Trajectory
$\hat{p}_{t+1,u}^r$
Root Trajectory Controller
Velocity Factor
$\hat{f}_{t+1,u}$
Velocity Factor Controller
Concat.
LSTM
AP: $\hat{p}_{t+1}^r$
RP: a_t
Position Encoder
Concat.
Keyframe k
p_k^b
p_k^r
o_k^r
Target Controller
LSTM
Transformer-based
FC-based
State Decoder
Root Decoder
$\tilde{p}_{t+1}^b$
$\tilde{c}_{t+1}$
$\tilde{v}_{t+1}^b$
$\tilde{p}_{t+1}^r$
$\tilde{o}_{t+1}^r$
$\tilde{v}_{t+1}^r$
Frame t+1

Framework

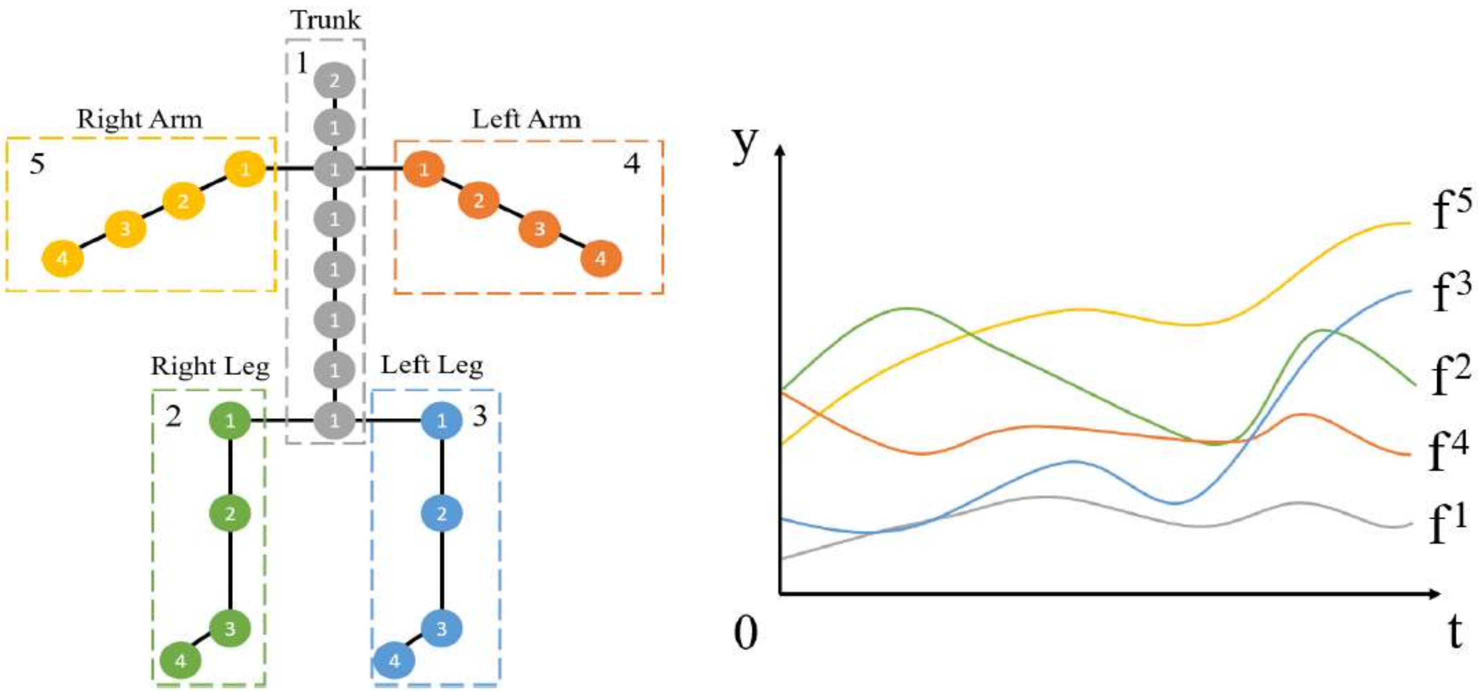

Frame t
p_t^b
p_t^r
o_t^r
c_t
v_t
State Encoder
Velocity Encoder
Root Trajectory
$\hat{p}_{t+1,u}^r$
Root Trajectory Controller
Velocity Factor
$\hat{f}_{t+1,u}$
Velocity Factor Controller
Concat.
LSTM
AP: $\hat{p}_{t+1}^r$
RP: a_t
Position Encoder
Concat.
Keyframe k
p_k^b
p_k^r
o_k^r
Target Controller
LSTM
Transformer-based
FC-based
State Decoder
Root Decoder
$\tilde{p}_{t+1}^b$
$\tilde{c}_{t+1}$
$\tilde{v}_{t+1}^b$
$\tilde{p}_{t+1}^r$
$\tilde{o}_{t+1}^r$
$\tilde{v}_{t+1}^r$
Frame t+1

The root trajectory controller

- Input
  - $\hat{p}^r_{t+1,u}$ refers to the target root trajectory segment
  - $\check{p}^r_{t+1,u}$ is the modification of $\hat{p}^r_{t+1,u}$
- Output
  - The representation of the root trajectory constraint



Velocity factor



(a) Body part division (b) Velocity factor sequences

Loss function

- **The complete loss function**

$$\mathcal{L} = w_{rec}\mathcal{L}_{rec} + w_{con}\mathcal{L}_{con} + w_{root}\mathcal{L}_{root} + w_{key}\mathcal{L}_{key} + w_{vfac}\mathcal{L}_{vfac}$$

- **Reconstruction loss $\mathcal{L}_{rec}$**

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{t=k_1+1}^{k_2} \|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|^2$$

- **Posture consistency loss $\mathcal{L}_{con}$**

$$\mathcal{L}_{con} = \mathcal{L}_{bone} + \mathcal{L}_{contact} + \mathcal{L}_{velocity}$$

$$\mathcal{L}_{bone} = \frac{1}{N}\sum_{t=k_1+1}^{k_2} \Big(\sum_{(i,j)\in\mathcal{B}} \| \|\tilde{\mathbf{p}}_t^i - \tilde{\mathbf{p}}_t^j\|_2 - l_{ij}\|^2 \Big)$$

$$\mathcal{L}_{contact} = \frac{1}{N}\sum_{t=k_1+1}^{k_2} \Big(\sum_{i}^{\mathcal{F}} \tilde{c}_t^i \|\tilde{\mathbf{v}}_t^i\|_2 \Big)$$

$$\mathcal{L}_{velocity} = \frac{1}{N}\sum_{t=k_1+1}^{k_2} \Big(\sum_{i}^{\mathcal{J}} \|\tilde{\mathbf{v}}_t^i - (\tilde{\mathbf{p}}_t^i - \tilde{\mathbf{p}}_{t-1}^i)\|^2 \Big)$$

Loss function

- **Root trajectory smooth loss $\mathcal{L}_{root}$**

$$\mathcal{L}_{root} = \frac{1}{N}\left(\sum_{t=k_1+1}^{k_2} \|\tilde{\mathbf{p}}_t^r - \tilde{\mathbf{p}}_{t-1}^r\|^2 + \sum_{t=k_1+1}^{k_2} \|\tilde{\mathbf{o}}_t^r - \tilde{\mathbf{o}}_{t-1}^r\|^2\right)$$

- **Keyframe consistency loss $\mathcal{L}_{key}$**

$$\mathcal{L}_{key} = \begin{cases} \frac{1}{2m}\left(\sum_{t=k_1+1}^{k_1+m} \|\tilde{\mathbf{p}}_t - \mathbf{p}_{k_1}\|^2 + \sum_{t=k_2-m+1}^{k_2} \|\tilde{\mathbf{p}}_t - \mathbf{p}_{k_2}\|^2\right), & N > 2m \\ \frac{1}{N}\sum_{t=k_1+1}^{k_2}\left(\frac{t-k_1}{N}\|\tilde{\mathbf{p}}_t - \mathbf{p}_{k_1}\|^2 + (1-\frac{t-k_1}{N})\|\tilde{\mathbf{p}}_t - \mathbf{p}_{k_2}\|^2\right), & N \leq 2m \end{cases}$$

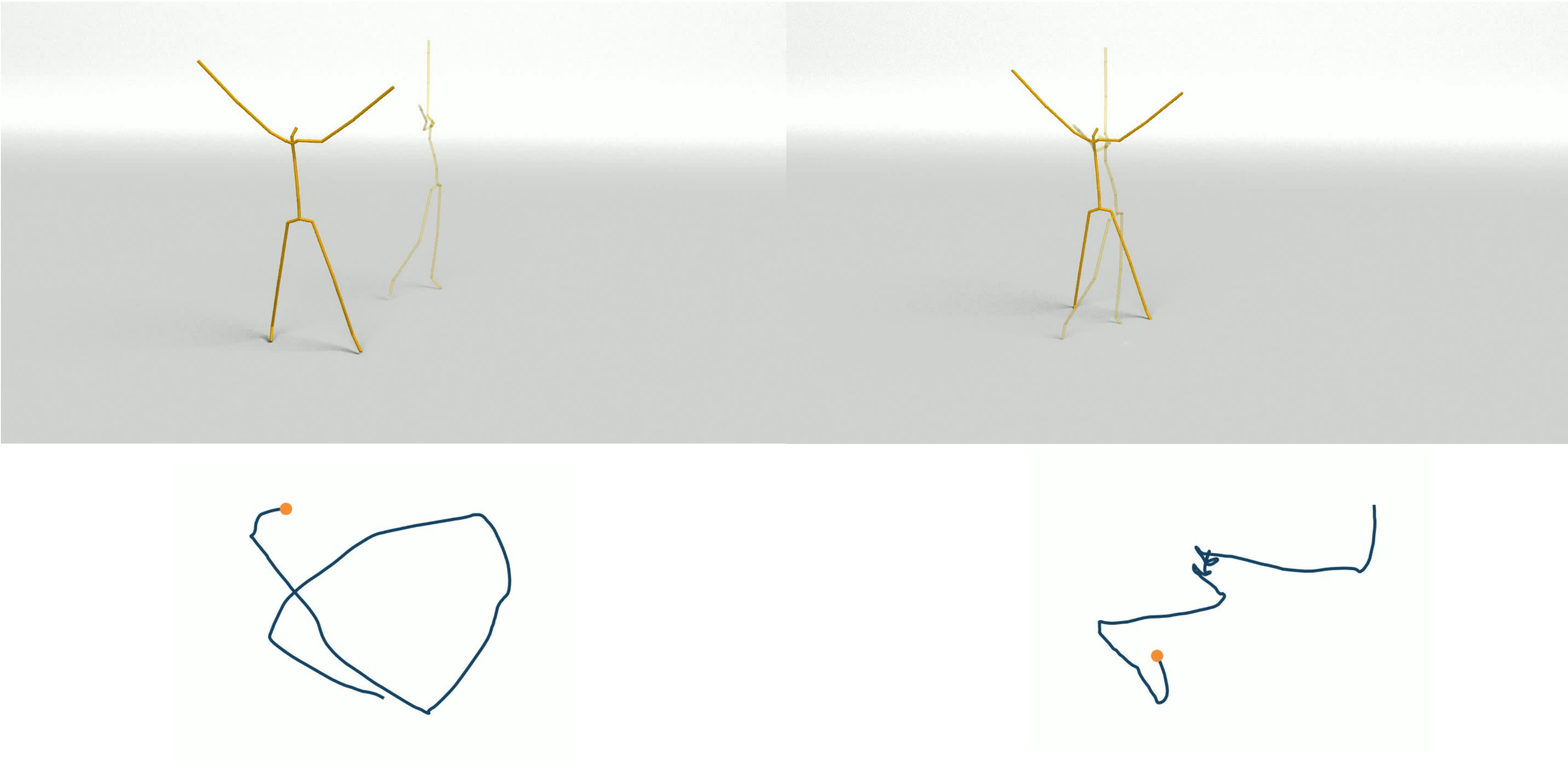
- **Velocity factor consistency loss $\mathcal{L}_{vfac}$x**

$$\mathcal{L}_{vfac} = \frac{1}{N}\sum_{t=k_1+1}^{k_2} \|\tilde{\mathbf{f}}_t - \hat{\mathbf{f}}_t\|^2$$

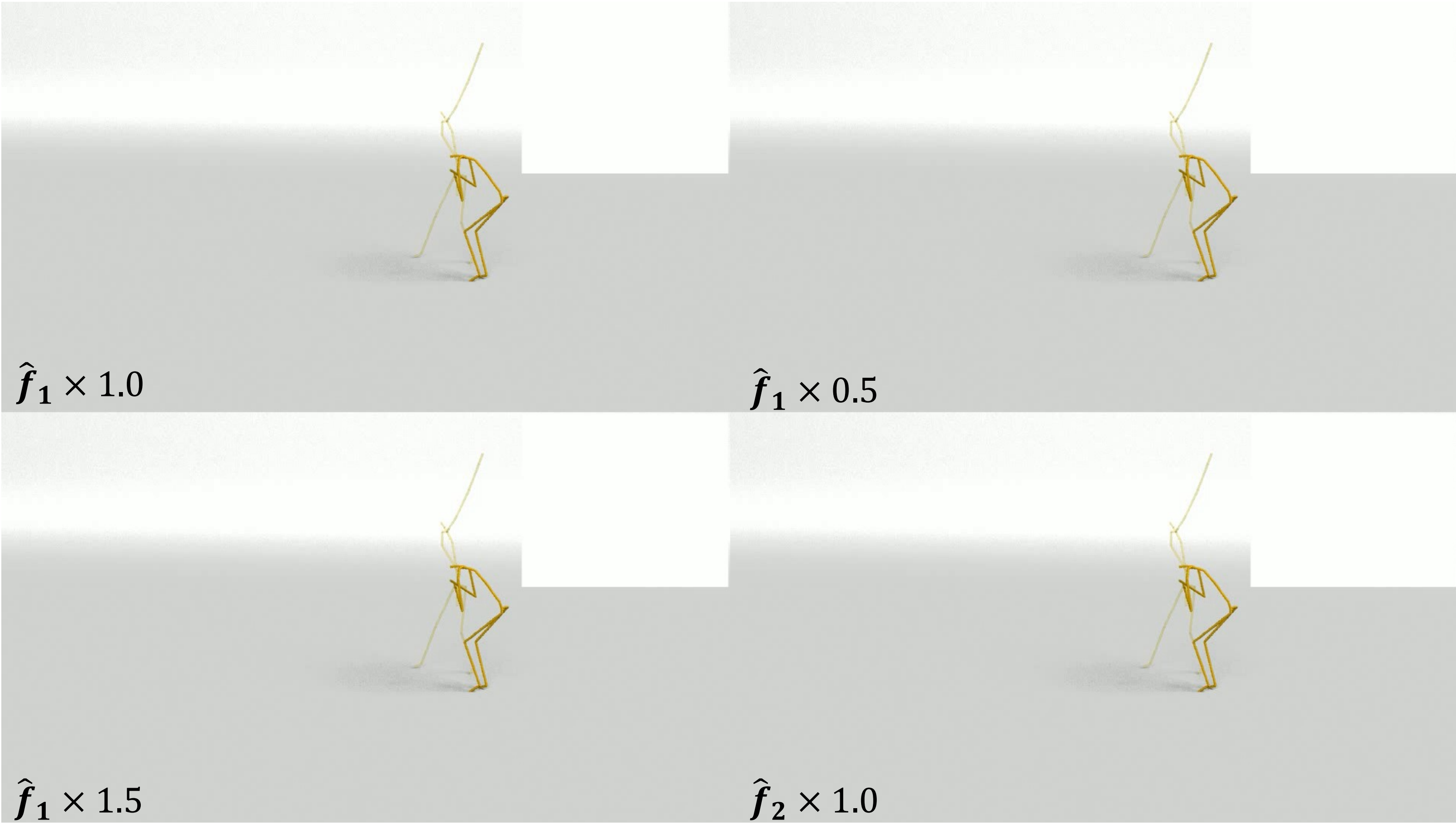
Contents

- Introduction
- Method details
- Experimental results
- Limitations and conclusion

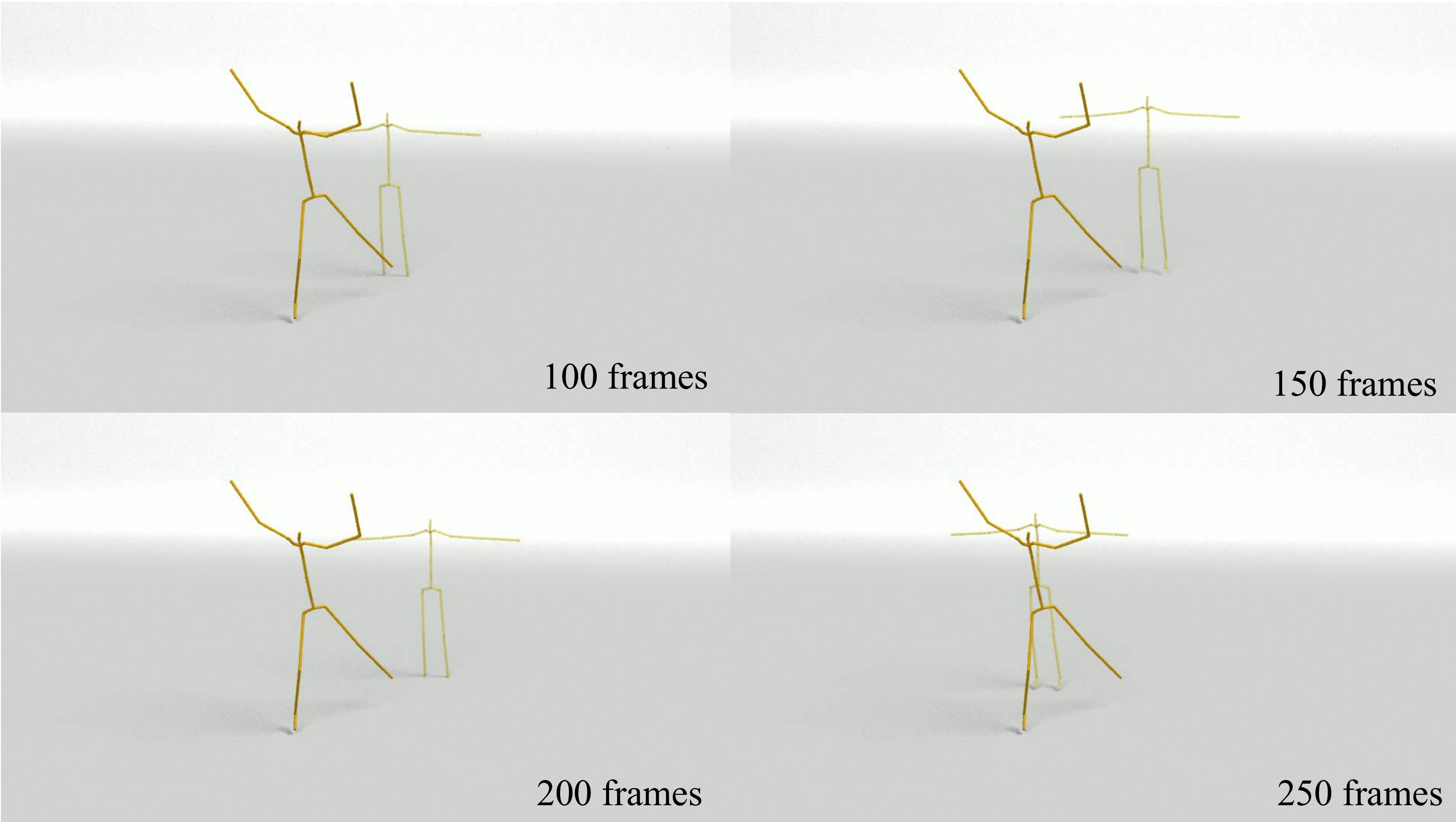
Different root trajectories



Different velocity factors



Motion transitions with different lengths



Quantitative metrics

- L2 distances of root-relative position (LRP)

$$LRP = \frac{1}{|\mathcal{D}|}\frac{1}{N}\sum_{c\in\mathcal{D}}\sum_{t=k_1+1}^{k_2} \|\tilde{\mathbf{p}}_t^b(c) - \mathbf{p}_t^b(c)\|_2$$

- Average accuracy of the velocity factor (AVF)

$$AVF = \frac{1}{|\mathcal{D}|}\sum_{c\in\mathcal{D}}\begin{cases}1, & if \quad g_v(c) < \delta_v\\ 0, & otherwise,\end{cases}$$

$$g_v(c) = \frac{1}{N}\frac{1}{M}\sum_{t=k_1+1}^{k_2}\sum_{i=1}^{M}|\tilde{f}_t^i(c) - \hat{f}_t^i(c)|$$

- Average accuracy of the root trajectory (ART)

$$ART = \frac{1}{|\mathcal{D}|}\sum_{c\in\mathcal{D}}\begin{cases}1, & if \quad g_r(c) < \delta_r\\ 0, & otherwise,\end{cases}$$

$$g_r(c) = \frac{1}{N}\sum_{t=k_1+1}^{k_2}\|\tilde{\mathbf{p}}_t^r(c) - \hat{\mathbf{p}}_t^r(c)\|_2$$

Ablation study

Table 2: *The LRP evaluation for the transition sequence at the length of 10, 50, 100, 150, and the average result (AVG). The best results are shown in bold.*

Models	Frames				AVG
	10	50	100	150	
Interpolation	**13.83**	87.59	113.59	124.47	84.87
Harvey *et al.* [HYNP20]	141.79	116.40	201.43	297.22	189.21
One LSTM	32.81	70.49	88.11	100.81	73.06
Condition-FC	37.30	67.03	85.20	96.39	71.48
One decoder	28.36	51.15	65.04	77.24	55.45
Without Velfac constraint	34.37	62.30	83.81	97.52	69.50
Without $\mathcal{L}_{key}$	41.59	82.95	97.31	105.46	81.83
Whole model	27.07	**50.06**	**63.01**	**74.37**	**53.63**

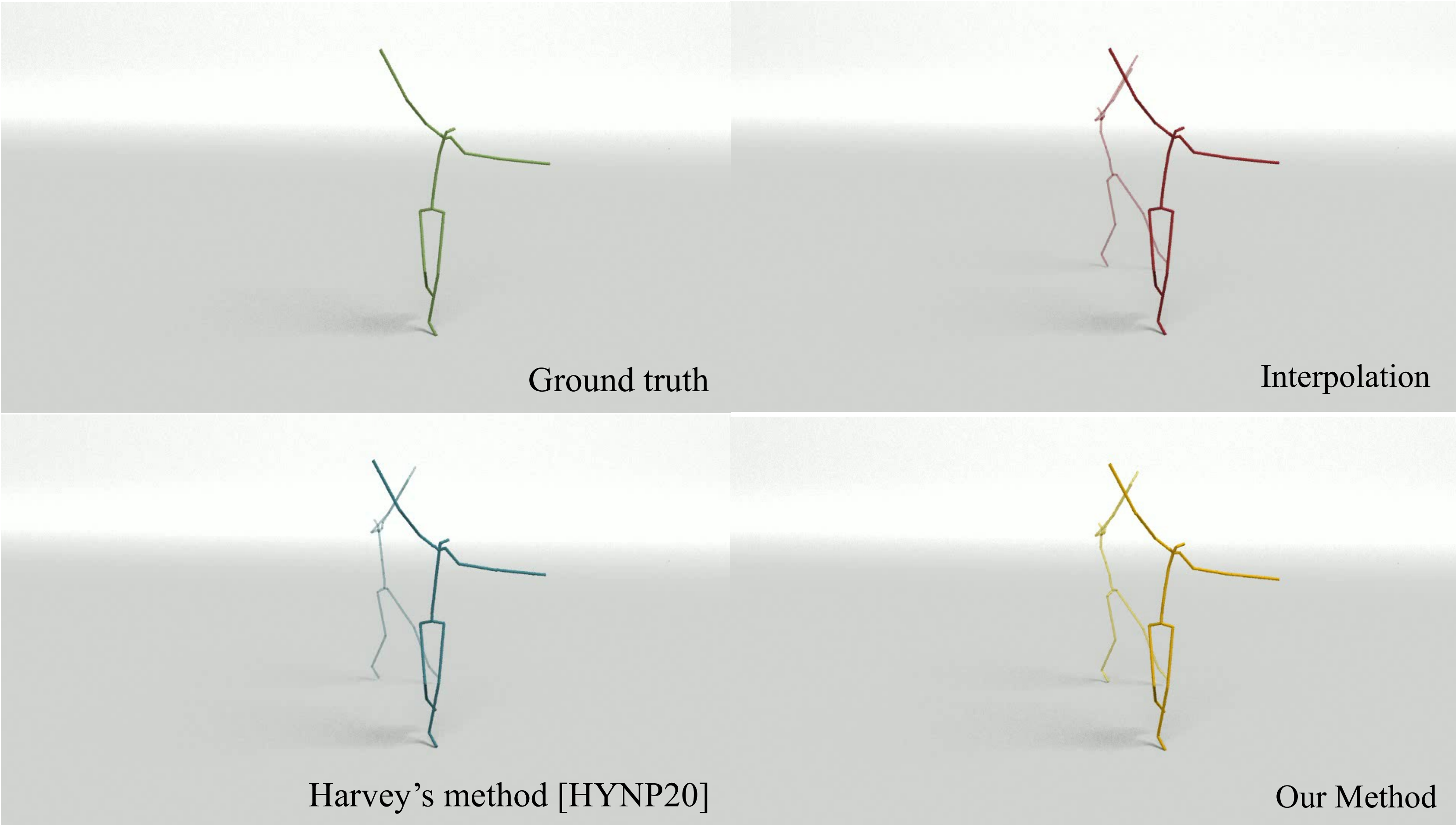
Table 3: *The AVF evaluation for transition sequence at the length of 10, 50, 100, 150, and the average result (AVG). The best results are shown in bold.*

Models	Frames				AVG
	10	50	100	150	
One LSTM	0.68	0.75	0.70	0.67	0.70
Condition-FC	0.68	0.71	0.68	0.65	0.68
One decoder	0.71	0.75	0.73	0.70	0.72
Without Velfac constraint	0.60	0.60	0.53	0.49	0.56
Without $\mathcal{L}_{key}$	0.66	0.68	0.66	0.64	0.66
Whole model	**0.72**	**0.76**	**0.75**	**0.72**	**0.74**

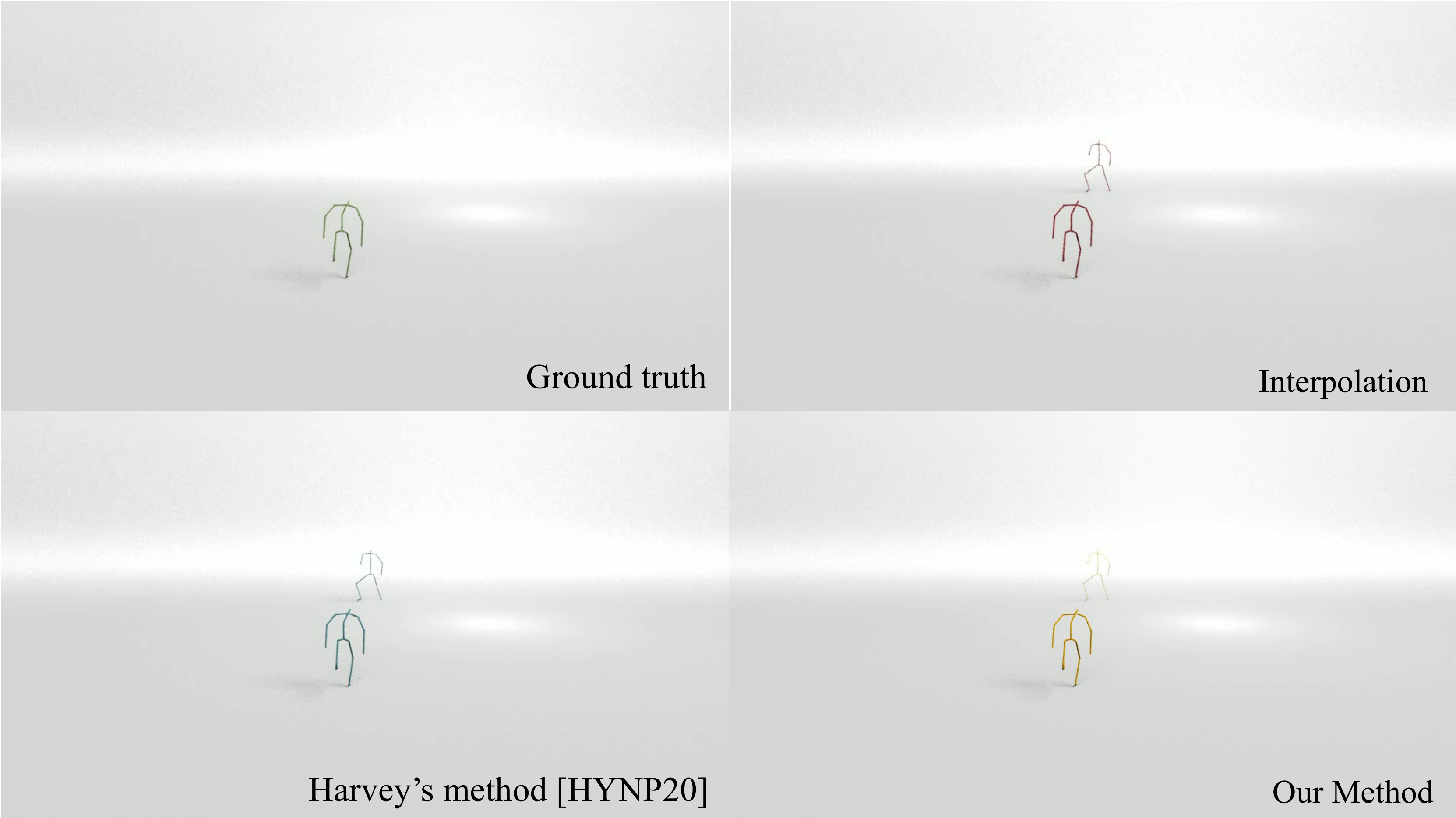
Table 4: *The ART evaluation for transition sequence at the length of 10, 50, 100, 15, and the average result (AVG). The best results are shown in bold.*

Models	Frames				AVG
	10	50	100	150	
One LSTM	0.63	0.74	0.76	**0.74**	0.72
Condition-FC	0.47	0.64	0.67	0.67	0.61
One decoder	0.63	0.55	0.48	0.39	0.51
Without Velfac constraint	0.62	0.59	0.45	0.34	0.50
Without $\mathcal{L}_{key}$	0.62	0.42	0.33	0.27	0.41
Whole model	**0.86**	**0.85**	**0.80**	0.72	**0.81**

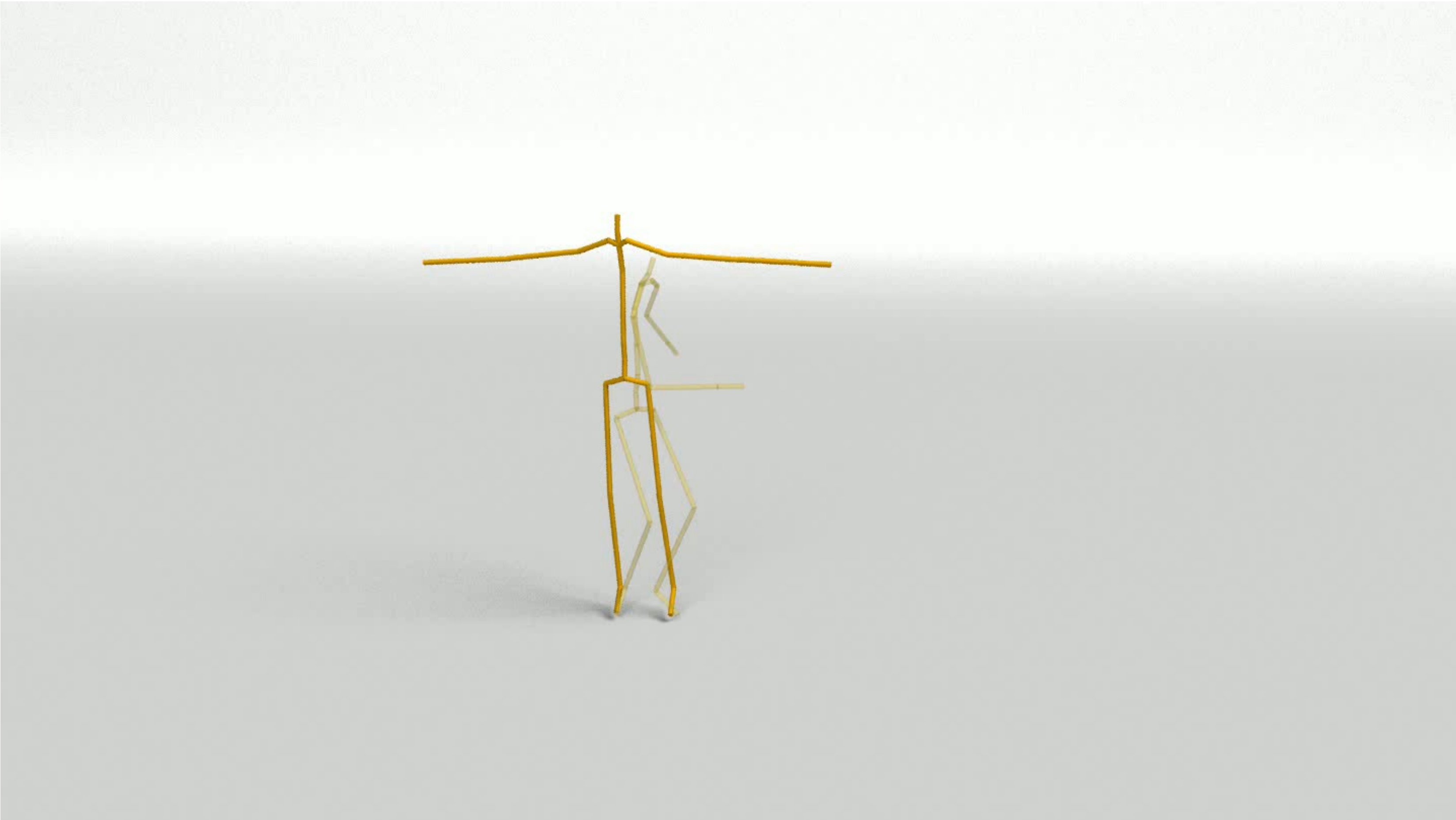
Comparison with other methods



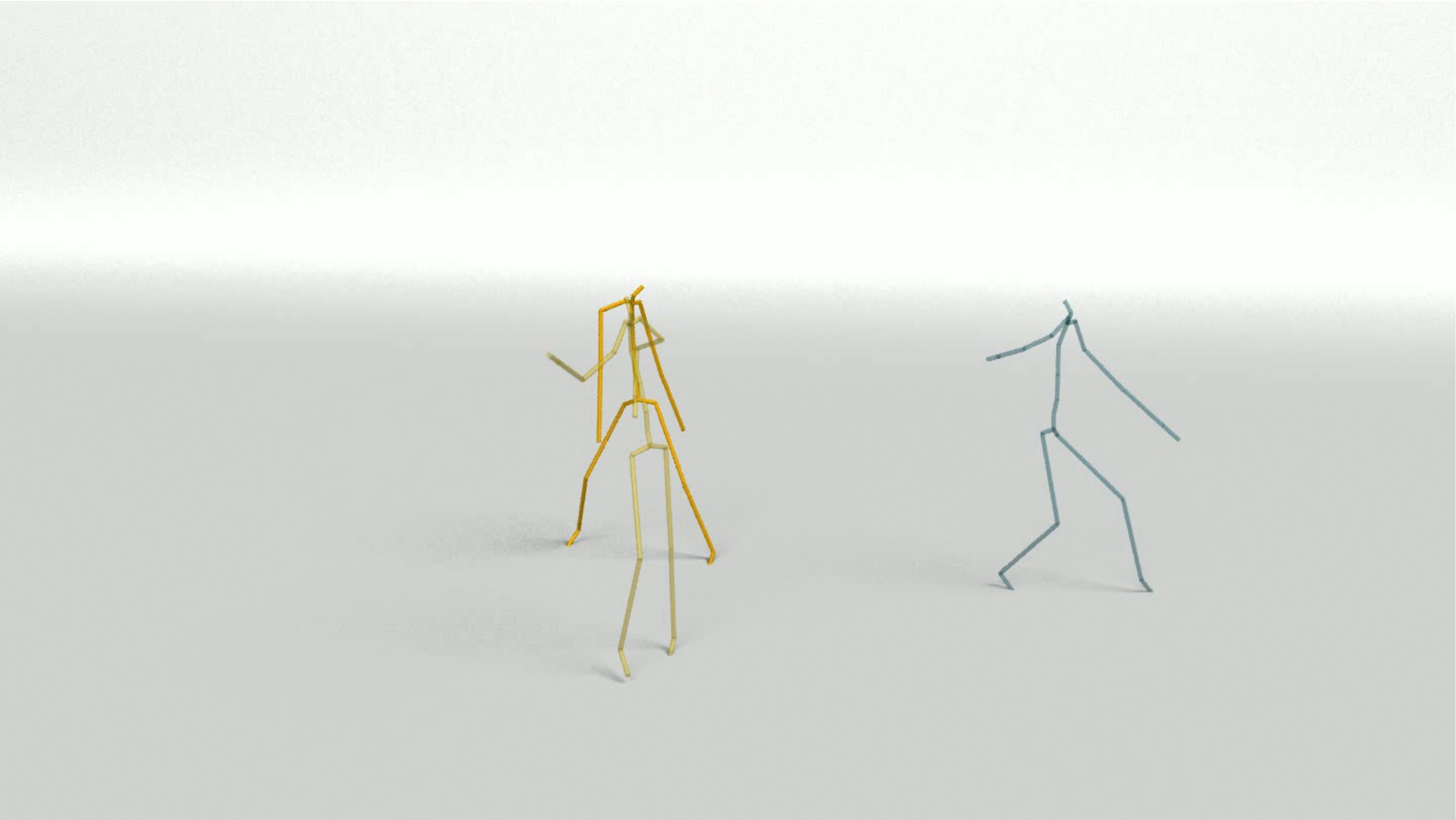
Comparison with other methods



More results



More results



Contents

Limitations

- The generated dance motion have the footstep floating problem.
- When providing control conditions that differ significantly from the training set, our method produces poor results similar to many data-driven tasks.
- The user may give some extremely contradictory control conditions, which will also cause the unreality of the results.

Conclusion

- We propose **a complex motion generation network** from keyframes, which can realize fine-grained control for complex motion through the trajectory sequence of the root joint and the velocity factors of different body parts.

- We have **conducted three categories of quantitative evaluation and ablation experimental analysis**, and have compared our algorithm with the state-of-the-art methods, which proves the superiority of our network.

- In the future, we will focus on the **footstep floating problem** and **multiple constraint contradiction issue** in complex motion synthesis.

Source Code

- https://github.com/godzillalla/Dance-Synthesis-Project

Thank You!

Q&A: Any question can be sent to the authors!

Email:
Ju Dai : daij@pcl.ac.cn