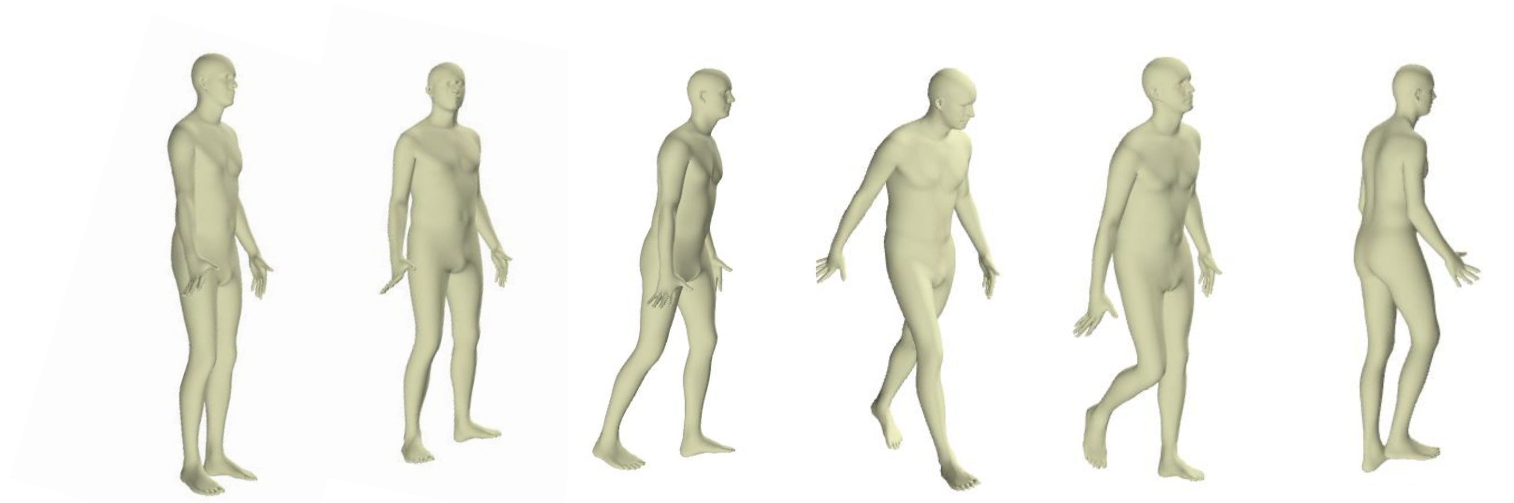
Seamless Human Motion Synthesis: Leveraging Transformer Autoencoders for Interactive Control Inputs

Qianxuan Lin (Supervisor: Hongdong Li)

Australian National University

U7079635@anu.edu.au

Introduction



Human motion synthesis a field within computer vision focused on generating realistic digital human movements for applications like games, simulations, movies, and virtual reality.

Significance: It enhances realism, deepens immersion, supports healthcare, and advances AI research.

State of the Art

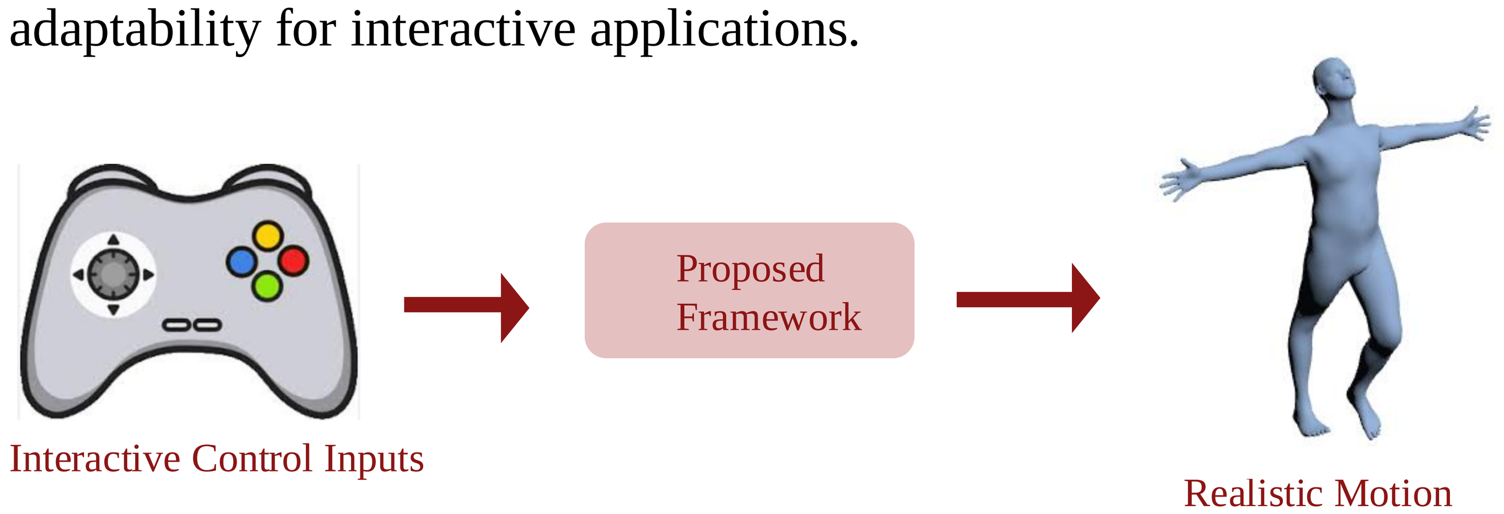
Text-based Motion Synthesis: Generates motion from text inputs using models like CLIP and transformers.

Action-based Motion Synthesis: Produces motion based on actions (e.g., throw, pickup) with models like ACTOR.

Limitations: Text-based synthesis faces issues with ambiguity and data dependency, while action-based synthesis is limited by predefined action categories, affecting adaptability to new inputs.

Problem Definition

Motivated by the above limitations, this thesis tackles the challenge of generating realistic human motion from intuitive controls like joysticks or mobile touch inputs. It aims to create a streamlined process that translates these inputs into natural, responsive movements, improving adaptability for interactive applications.



Methodology

• Model Formulation:

• Transformer Autoencoder:

$$H = \Phi(X) = \text{TransformerEncoder}(X).$$

$$\hat{X} = \Phi^\dagger(H) = \text{TransformerDecoder}(H).$$

$$\text{Cost}(X, \theta) = \|X - \Phi^\dagger(\Phi(X))\|_2^2 + \alpha\|\theta\|_1$$

• Motion Synthesis Pipeline:

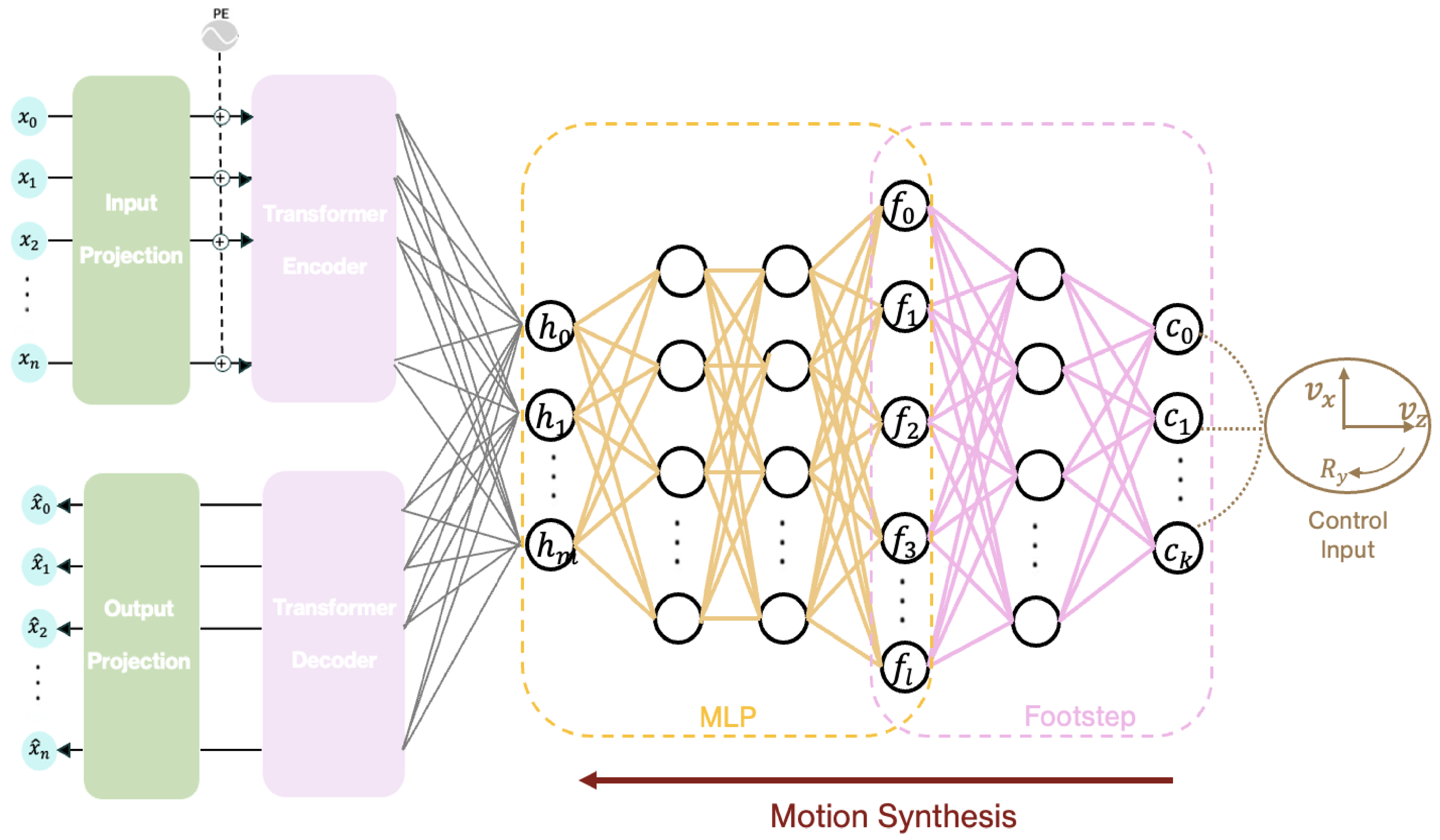
Control Inputs: $T \in \mathbb{R}^{n\times 3}$ representing V_x, V_z, and R_y.

Footstep: $F = \Upsilon(T) = \sigma\left(\sigma\left(TW_f^{(1)} + b_f^{(1)}\right)W_f^{(2)} + b_f^{(2)}\right)$

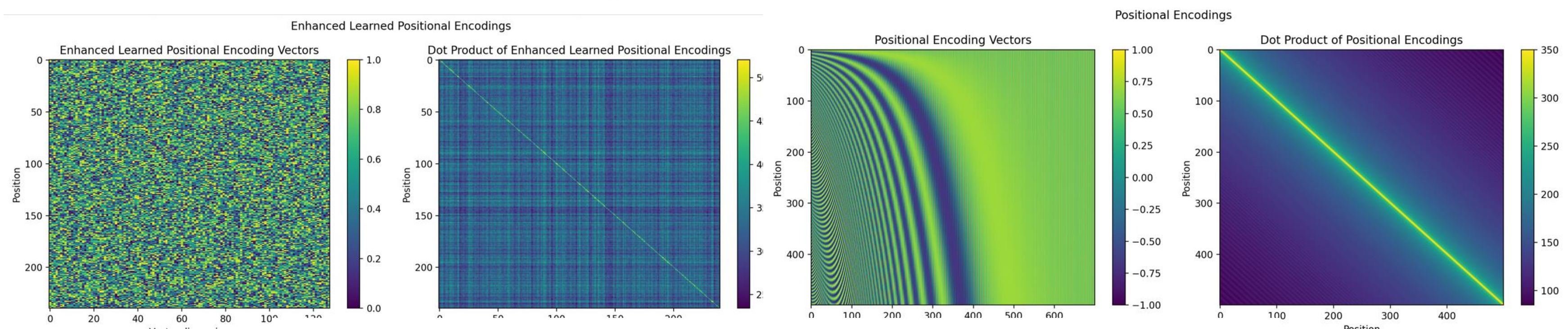
MLP: $H = \Pi([T, F]) = \sigma_3\left(\sigma_2\left(\sigma_1\left([T, F]W_m^{(1)} + b_m^{(1)}\right)W_m^{(2)} + b_m^{(2)}\right)W_m^{(3)} + b_m^{(3)}\right)$

Motion Output: $\hat{X} = \Phi^\dagger(H)$

• Network Architecture:



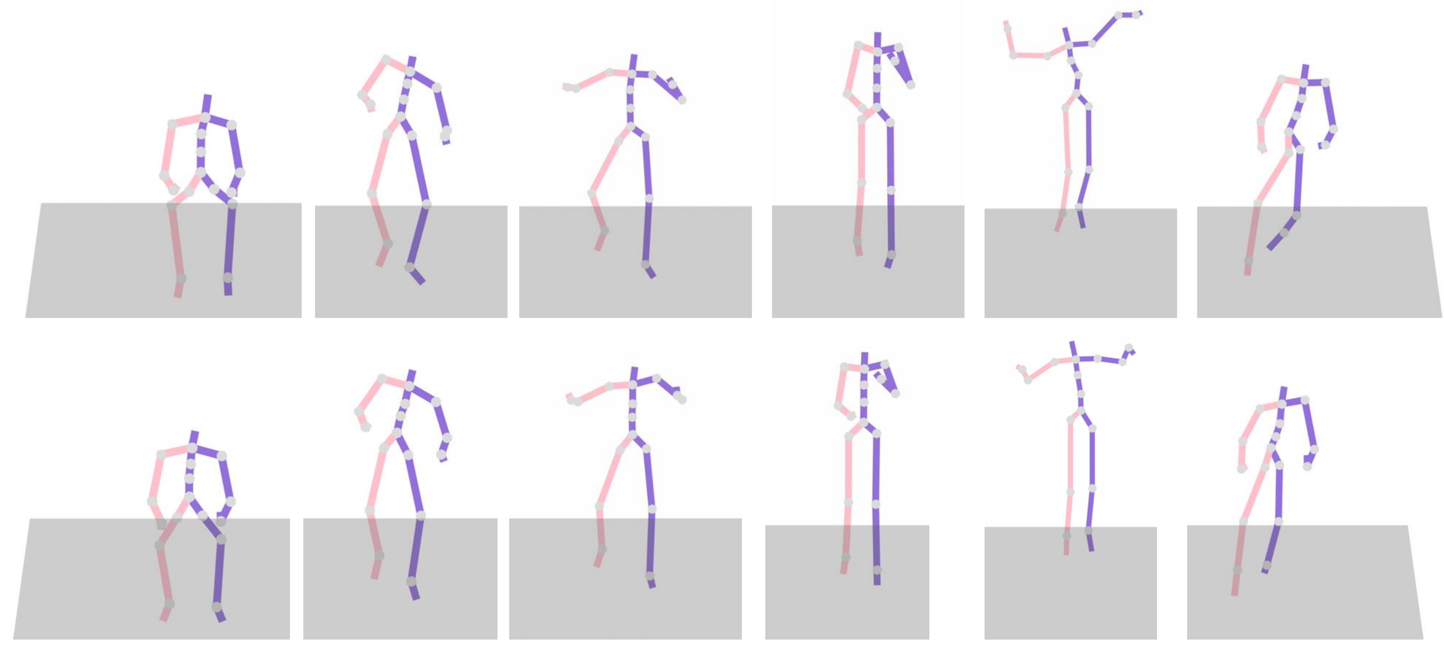
• Positional Encoding Methods:

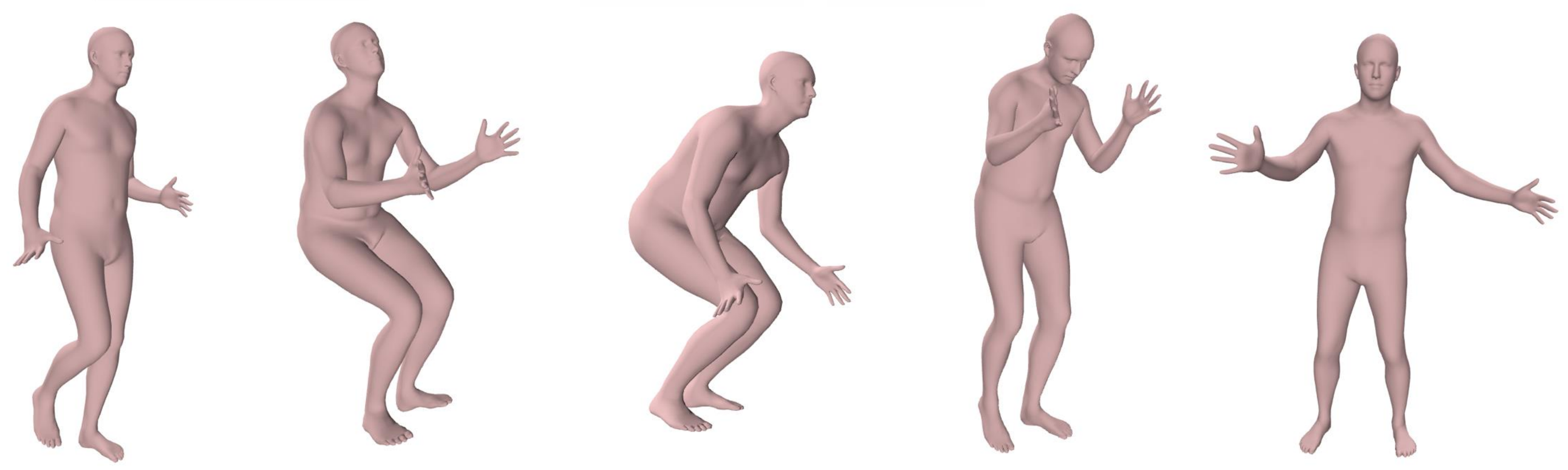


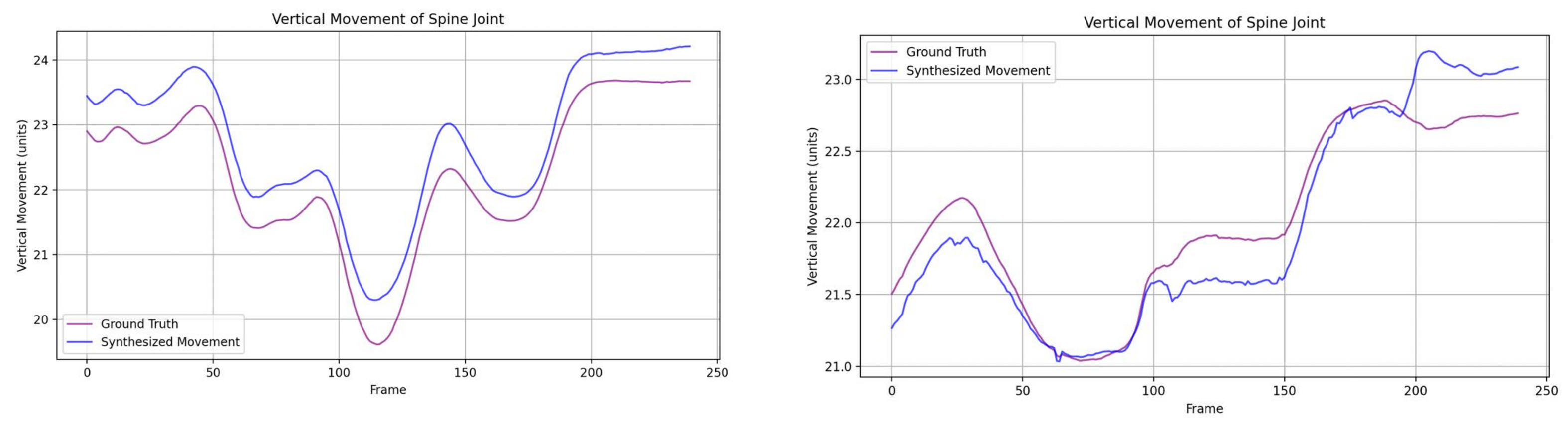
Positional Encoding	Training Loss (Avg)	Test Error (Avg)
Sinusoidal	0.757940	0.313129
Learnable	0.797529	0.120244

Learnable encoding, with its lower test error, captures human motion patterns more effectively than sinusoidal encoding. This choice balances theoretical and practical considerations.

Results







Autoencoder	Dropout	Training Loss	Test Loss	Mean Squared Displacement (MSD)
Convolutional	0.2	1.068	0.925	7.6817
Transformer	N/A	0.798	0.120	**1.4562**

Conclusion

- **Introduces a transformer-based framework** for realistic human motion from intuitive inputs like joysticks or mobile controls.
- **Employs attention mechanisms** to enhance naturalness and responsiveness by capturing temporal and joint dynamics.
- **Offers a streamlined, adaptive process** that addresses current limitations and aligns with user intent.
- **Advances human motion synthesis** with applications in gaming, VR, and other interactive environments.