Human Motion Prediction in Hallways Using Transformers

Praneel Seth
Computer Science Dept.
University of Texas at Austin
Austin, United States
praneelseth@utexas.edu

Mukund Raman
Computer Science Dept.
University of Texas at Austin
Austin, United States
mkraman@utexas.edu

Akshay Gaitonde
Computer Science Dept.
University of Texas at Austin
Austin, United States
a.g@utexas.edu

Kunal Tiwari
Computer Science Dept.
University of Texas at Austin
Austin, United States
kunaltiwari@utexas.edu

***Abstract*—This paper addresses the challenge of human motion prediction in hallway passing scenarios by introducing a nonautoregressive transformer architecture. Our model leverages the ability of the transformer to capture temporal and spatial dependencies, allowing accurate predictions of human trajectories based on input from markerless motion capture data. Unlike previous approaches focusing on sequential motion continuity, our method generalizes across diverse motion patterns by pre-training the model on various human interactions, and state information provided to the model includes joint position, velocity, and acceleration data. The proposed architecture tokenizes body tracking data and utilizes attention mechanisms to predict motion trajectories, allowing robots to navigate complex environments and respond to human motions effectively. This model is evaluated through the accuracy of its predicted motion trajectory, comparing its predictions with human behavior. This approach aims to improve the accuracy and naturalness of robot navigation in crowded settings, offering a generalized solution for crowded social navigation scenarios.**

***Index Terms*—non-autoregressive transformer, hallway passing, motion prediction, computer vision**

I. Introduction

Motion prediction techniques create more natural movements in entertainment and robotics and can greatly improve real-world interactions in complex environments. For instance, predicting the motion of individuals in crowded spaces could enable robots to navigate seamlessly and respond to human gestures in real-time. Recent advances in machine learning, particularly transformer neural networks, have made it possible to accurately predict a person's next movements based on limited input data, such as body tracking data.

Current technological advancements in sensing and computing have increased robotics applications in human-robot interaction in crowded settings. To illustrate, researchers have specifically investigated the deployment of robots to provide personal mobility services and luggage transport support in complex pedestrian-rich areas [1].

Human motion prediction has been extensively explored through various neural network architectures, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and most recently, transformer-based models. Techniques such as those used by Martinez et al. (2017) [4] and Pavllo et al. (2019) [5] employ RNNs and CNNs to predict motion sequences, focusing on overcoming issues like short-term prediction errors and rotational discontinuities. However, while promising, these models are limited by their focus on sequential frame continuity, which does not effectively handle motion in diverse environments such as crowded hallways.

Initially developed for language modeling, transformer models have been adapted to vision tasks by treating grid segments of images as tokens. This concept can be extended to human motion prediction, where positional encodings are critical in understanding the temporal relationships between key points in body tracking data. Applying transformers to predict motion trajectories involves tokenization and using attention mechanisms to account for temporal and spatial changes.

Even today, robots have difficulty maneuvering smoothly through environments with moving people, such as a crowded hallway. This paper employs a transformer model to predict hallway passing trajectories as a heuristic for how robots can better react in human motion scenarios.

The greater challenge is to create a model that can predict human trajectory motion while also being capable of operating pre-trained or learning in real-time. However, the current approach pre-trains the model to be more generalized across different motion patterns and gestures. This allows the model to handle a variety of human actions without needing to adapt on the fly, providing flexibility while still predicting movements in real-world applications where quick decision-making is essential.

II. Background

Human motion prediction in crowded environments requires technical precision and a deep understanding of social dynamics and temporal-spatial dependencies. Robots navigating such spaces must account for the behaviors and interactions of individuals, adapting to both group movements and isolated gestures.

Ferrer et al. (2013) [2] offers a key framework for understanding how robots can navigate complex environments by incorporating social dynamics into motion planning. This research introduces a socially aware navigation strategy based on a social-force model, which simulates pedestrian interactions such as maintaining personal space and following crowd flow patterns. The model enables robots to move seamlessly

through crowded spaces without disrupting human activity by embedding these social norms into robotic navigation. This study highlights the importance of human-centric navigation strategies in hallway passing scenarios and helped shape the approach of this paper's transformer model.

A. Pose Prediction

Existing human pose prediction techniques can be broadly divided into two main categories: probabilistic and deterministic approaches [7]. Probabilistic methods aim to generate multiple potential scenarios from a sequence of observed frames, mimicking the complexity of human cognitive processing. A notable example in this domain is Diversifying Latent Flows (DLow), which utilizes pre-trained generative models to forecast various 3D human pose hypotheses. While some probabilistic approaches focus solely on trajectory prediction, others explore probability distributions of human kinematic states.

Conversely, deterministic approaches focus on predicting precise pose or trajectory sequences based on observed motion. Traditionally, recurrent neural network (RNN) architectures were employed for motion prediction [8], but these models suffer from cumulative prediction errors and lack computational efficiency due to their sequential processing (autoregressive) nature. The autoregressiveness also renders them non-parallelizable and computationally intensive. Researchers have attempted to address these challenges through various innovative techniques. Some approaches incorporate adversarial losses and geodesic body measurements to improve prediction accuracy and mitigate drift issues [9]. Additionally, spatio-temporal modeling strategies have been developed to better capture intricate joint relationships across individual frames and motion sequences.

Zheng et al. (2021) [3] highlights the potential of transformers in modeling human motion by using attention mechanisms to capture both local joint relationships and global dependencies over time. The authors introduce a spatial transformer module to encode the relationships between individual joints and a temporal transformer module to account for sequence-level dependencies in pose estimation from 2D video data. This architecture demonstrates how transformers handle complex temporal-spatial patterns, making them highly effective for human motion prediction tasks. This work illustrates how attention mechanisms can be applied to human motion tasks to predict future trajectories effectively, and it serves as a major motivator for the transformer-based approach taken in this paper.

Recent efforts in human motion prediction have increasingly focused on leveraging transformer-based architectures for enhanced flexibility and accuracy. The Spatio-Temporal Pose Prediction with Transformers (STPOTR) model, introduced by Petrovich et al. (2022) [6], extends the capabilities of transformers to address both spatial and temporal dimensions in human motion prediction. By modeling human pose sequences as graph structures and utilizing spatio-temporal self-attention, STPOTR achieves highly accurate pose predictions over time. This model highlights the potential of transformers to manage short- and long-term dependencies in body tracking data while improving generalization across diverse environments.

Like STPOTR, this paper's approach employs transformer-based architectures for motion prediction but concentrates on real-time hallway passing scenarios for robotics applications. Both methods utilize the transformer's self-attention mechanism to capture spatial and temporal relationships between body joints. However, the STPOTR model restricts its input space to the positional coordinates of the link segment. In contrast, this work preprocesses the dataset to include continuous state information like the velocity and acceleration along with the positions of all the 3D joints. These time derivatives are encoded explicitly and captured by attention mechanisms in the transformer model. This additional data is hypothesized to improve prediction accuracy farther into the future than STPOTR can currently provide.

III. Methodology

A. System Overview

This research introduces a transformer-based model for human motion prediction, specifically focusing on predicting future movements in 3D space using markerless motion capture data. The system processes motion data from the H3.6M dataset, which includes detailed joint position information across various activities. The input consists of a sequence of 5 frames representing joint positions, while the output predicts 20 subsequent frames. A transformer architecture is utilized to handle the temporal dependencies and spatial relationships inherent in human motion.

Key features of the system include:

- A preprocessing pipeline to normalize and structure data.
- A transformer model with encoder-decoder architecture for sequence-to-sequence prediction.

B. Data Preprocessing

The Human 3.6M dataset is one of the largest and most widely used datasets for human motion prediction and pose estimation. It consists of motion capture data collected from seven professional actors performing a variety of activities, such as walking, sitting, eating, and interacting with objects. Each action is recorded using 4 cameras, which track 3D positions of 32 body joints at a frequency of 50 Hz. The dataset includes a total of 3.6 million frames, with the skeletal data represented as 3D joint coordinates relative to a global origin. Additionally, the dataset provides synchronized RGB video and depth data, allowing researchers to correlate skeletal movements with visual cues. This transformer was trained on the Human 3.6M dataset.

The raw Human 3.6M dataset is processed into structured CSV files containing 3D joint positions for each frame. Each CSV file is a separate recorded scenario such as walking, dancing, giving directions, running, etc. The train-test split was 70% of the scenarios in training directories and 30% of scenarios in testing directories. Preprocessing involves:

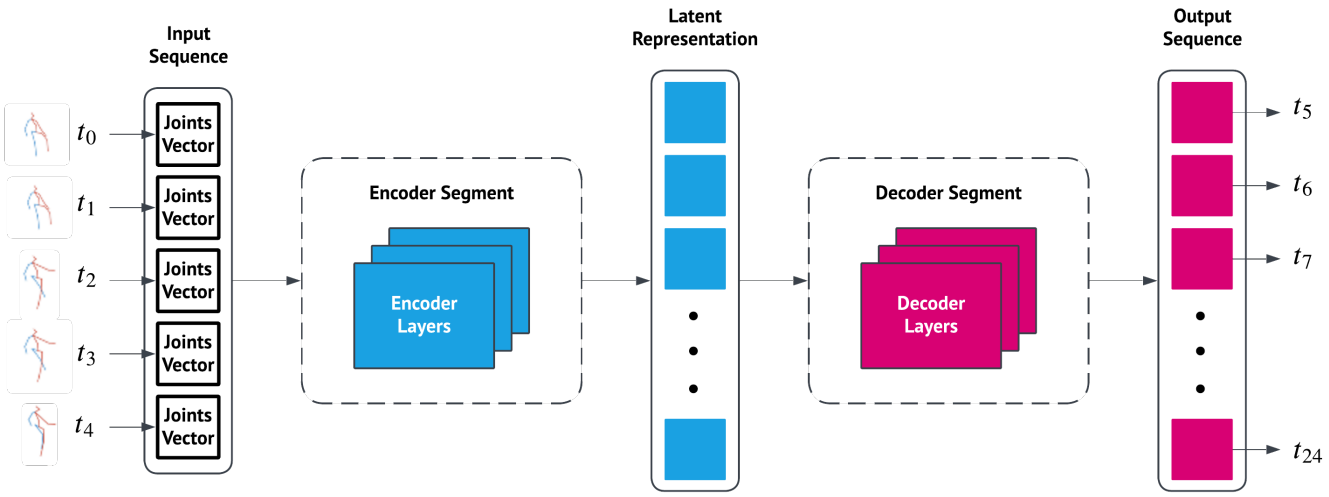

Fig. 1: Architecture Workflow.

1) Data Normalization: Each joint's position is normalized using StandardScaler to ensure numerical stability.
2) Feature Engineering: For specific joints (the pelvis, knees, and spine), the additional features of velocity and acceleration are computed using numerical gradients.
3) Sequence Creation: Input-output pairs are generated by sliding a window over the normalized data:
   - Input: A sequence of 5 frames (t_0 to t_4).
   - Output: A sequence of 20 future frames (t_5 to t_{24}).
4) Batch Processing: The processed data is split into training and testing subsets based on predefined directories, ensuring no overlap between the two sets.

C. Model Architecture

The proposed transformer model follows a standard encoder-decoder structure optimized for sequence-to-sequence tasks. Each input frame is represented as a "joints vector" comprising 120 features. This vector encodes the X, Y, and Z positions of 32 joints (96 features total) along with the directional X, Y, Z velocities and accelerations for the pelvis, left knee, right knee, and spine (24 additional features). These features collectively provide a detailed representation of both the spatial configuration and the motion dynamics of the human body. This input representation ensures the model can capture fine-grained motion details while maintaining a compact feature set suitable for transformer-based processing.

The transformer architecture itself includes both an encoder and a decoder, each consisting of three layers. The encoder focuses on capturing temporal dependencies across the input sequence, leveraging self-attention mechanisms to model long-range interactions between frames. The decoder, in turn, uses masked self-attention to predict future frames without leaking information about subsequent time steps. The decoder relies on both the encoder's outputs and previous predictions to generate accurate motion trajectories. The embedding dimension of the transformer matches the input feature size, ensuring consistency between input and internal representation.

To enhance model capacity and robustness, the transformer employs eight attention heads and a dropout rate of 0.1 to prevent overfitting. Finally, a fully connected output layer maps the transformer's latent representations back to the "joints vector" format, completing the sequence-to-sequence prediction process.

D. Training Procedure

The training process begins with dataset preparation, ensuring a clear separation between training and testing scenarios. The training data comprises scenarios drawn from specific directories (S1, S6, S7, S8, and S9), providing diverse examples to optimize model generalization. Testing data is sourced from entirely separate directories (S5 and S11) to evaluate the model's performance on unseen scenarios, ensuring robust validation of its predictive capabilities. To streamline the training process, data loaders dynamically batch the inputs and outputs, facilitating efficient and scalable processing during training.

The training setup is fine-tuned using carefully selected hyperparameters. The loss function employed is L1 Loss (Mean Absolute Error), chosen for its ability to minimize the absolute differences between the predicted and actual joint positions. The Adam optimizer, with a learning rate of 0.001, is utilized for its efficiency and adaptability in optimizing complex neural network architectures. Due to GPU limitations, the batch size is set at 9,000, which strikes a balance between computational feasibility and model performance. Training is conducted over 60 epochs to ensure that the model converges while capturing the complexities of human motion trajectories.

Each epoch in the training loop involves feeding batches of data into the model, computing gradients through back-propagation, and updating model weights accordingly. To maintain stability during training, special attention is given to

monitoring anomalies, such as NaNs or exploding gradients, which could disrupt the optimization process.

The validation process is integral to the training loop, with the validation loss being monitored after each epoch to assess the model's performance on unseen data. The best-performing model, identified through its validation performance, is saved for testing, ensuring that the final evaluation is conducted on the most accurate version of the model. This rigorous training procedure ensures that the model is well-equipped to predict human motion trajectories in real-world applications, particularly in dynamic and crowded hallway scenarios.

IV. Evaluation

A. Evaluation Metrics

The evaluation of the proposed transformer model relies primarily on the Mean Per Joint Position Error (MPJPE), a widely used metric in 3D human pose estimation. MPJPE calculates the mean Euclidean distance between the predicted 3D joint positions and their corresponding ground truth locations. This metric is particularly effective for assessing pose estimation because it directly measures spatial discrepancies in 3D space, avoiding the ambiguities associated with angle-based metrics like the Mean Angle Error (MAE). In scenarios where multiple joint configurations can produce identical poses, MAE may fail to capture discrepancies effectively, whereas MPJPE provides a clear quantitative comparison of the model's predictions against ground truth data.

To obtain a comprehensive view of the model's performance, MPJPE is averaged across all test scenarios, ensuring that the results are representative of diverse motion patterns and conditions. Furthermore, MPJPE is evaluated separately for each predicted time frame, generating a detailed error profile that captures how prediction accuracy evolves over time. This temporal analysis provides insights into the model's strengths and weaknesses, such as its ability to maintain accuracy for longer-term predictions versus its performance in short-term scenarios. By plotting MPJPE for each joint across different predicted frames, this analysis helps identify which joints are more prone to errors, offering valuable information for refining the model architecture and training process.

In addition to quantifying overall prediction accuracy, MPJPE-based analysis is used to diagnose potential areas for improvement. For instance, higher errors in joints with complex motion dynamics, such as hands or feet, may indicate the need for enhanced spatial representation techniques. Similarly, a rising trend in MPJPE over longer prediction horizons could suggest limitations in the model's temporal dependency encoding, motivating further research into improved attention mechanisms or hybrid architectures. This granular approach to evaluation ensures that the model not only achieves strong average performance but also addresses critical edge cases relevant to real-world applications like hallway navigation.

The focus on MPJPE aligns with the paper's goal of enabling accurate, natural human motion predictions for robotics applications. By identifying both strengths and limitations through this metric, the evaluation framework ensures that the proposed model is robust enough to handle dynamic and crowded environments, paving the way for real-time deployment in human-robot interaction scenarios.

B. Visualization

To better understand the performance of the proposed transformer model, OpenGL was utilized to create visualizations comparing the predicted frames with the ground truth. These visualizations played a crucial role in interpreting the model's predictions and verifying its effectiveness in accurately forecasting human motion. The visualizations were rendered in the form of video sequences, where each frame displayed both the actual human motion (ground truth) and the corresponding predicted motion trajectories generated by the model. This dynamic representation offered an intuitive and clear method for evaluating the model's ability to replicate realistic human movements.

Each frame in the visualization contained the link-segment skeleton of the ground truth frame alongside the skeletons of four predicted frames, specifically the 5th, 10th, 15th, and 20th frames into the future. These specific frames were chosen to provide a comprehensive view of the model's performance at evenly spaced intervals over the prediction horizon. The 5th frame highlighted the model's short-term predictive capabilities, capturing immediate motion trends, while the 20th frame emphasized its long-term accuracy and stability. By including intermediate frames at the 10th and 15th time steps, the visualization provided insights into how prediction errors developed over time, helping to identify potential temporal drift or compounding inaccuracies in motion forecasting.

The link-segment skeletons were displayed in contrasting colors to distinguish between the ground truth and predicted frames. For example, the ground truth was rendered in a distinct color (such as pink), while the predicted frames were shown in a neutral shade (such as white) to avoid visual clutter while maintaining clarity. This clear differentiation made it possible to visually assess the alignment and deviation of the predicted trajectories from the actual motion data. Additionally, the skeletal representation of the joints and links highlighted the relationships between body segments, enabling a detailed examination of errors in specific areas, such as the arms, legs, or torso.

By replaying these sequences in real-time, the visualizations offered an informal yet powerful method of validation, supplementing quantitative metrics like MPJPE. Observers could detect subtle differences between predicted and actual motion, such as unnatural poses, misaligned joint movements, or inconsistent trajectories. This visual feedback was instrumental in diagnosing specific limitations of the model, such as its tendency to overshoot or undershoot certain joint positions in complex movements. In addition, scenarios that involve rapid changes in direction or intricate gestures provided valuable insights into the model's ability to adapt to challenging motion patterns.

These visualizations also served as a tool for identifying patterns in error distribution across joints. For example, joints

with higher prediction errors, as highlighted in the MPJPE analysis, could be visually inspected to determine whether these errors stemmed from systemic inaccuracies in the model or from inherent variability in the dataset. Observing the alignment of critical joints like the pelvis and spine, which anchor overall body motion, was particularly useful in understanding the stability of the model's predictions.

The video visualizations offered a compelling demonstration of the model's ability to handle real-world scenarios, particularly in the context of crowded hallway navigation. They provided an intuitive means of verifying the alignment between predicted and actual motions, ensuring that the transformer's predictions were not only numerically accurate but also visually plausible and natural. By supplementing quantitative metrics with qualitative assessments, the OpenGL-based visualizations contributed to a holistic evaluation of the model's performance, laying the groundwork for further refinements and real-world deployment.

V. Results

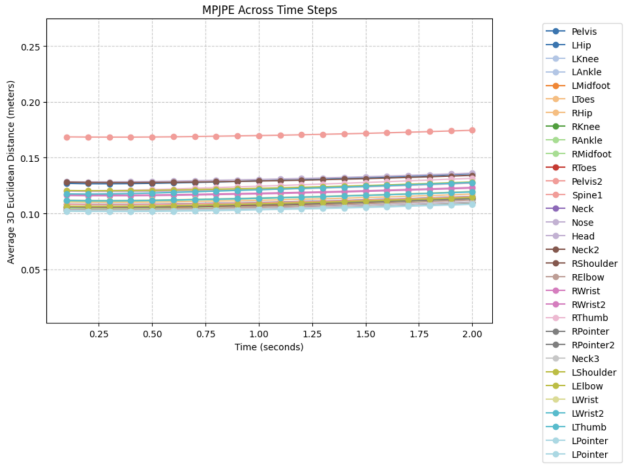

Fig. 2: MPJPE across time steps for this model. This displays data for 32 joints and shows that the MPJPE across all the joints ranges roughly between 0.1 and 0.15 m.

The results of the transformer model, which is run with 100 epochs in the training phase, are shown in Figure 2. For the purposes of benchmarking, the results of the STPOTR are also shown for comparison in Figure 3.

The figures suggest that the proposed transformer is best suited for longer-term predictions as STPOTR has much higher MPJPE error values for all the joints shown. These results are promising in the context of accurately predicting human motion trajectories. The transformer's consistent MPJPE value of 0.15 indicates that the model can make somewhat accurate predictions but only in the longer term and not in the short term, which could be essential for robot following or motion analysis applications. However, it is important to note that the accuracy determined in a simulated environment does not

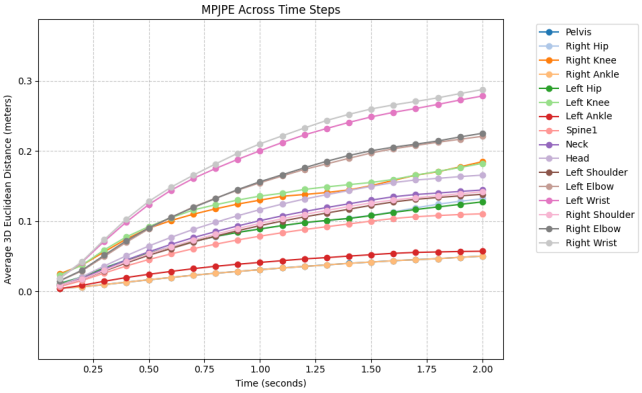

Fig. 3: MPJPE across time steps for STPOTR. This displays data for 16 joints and shows that STPOTR has a compounding increase in error over time from near zero to almost 0.3 m.

often translate to accuracy in real physical environments under real-time constraints.

With our visualization, we informally confirmed the graphed results with a video visual. An example of a frame from this generated visualization is as shown in Figure 4.

VI. Conclusion

The transformer model introduced in this paper effectively integrates positional data alongside velocity and acceleration of 3D human joints to predict human motion trajectories. This innovative approach addresses key challenges in motion prediction by leveraging attention mechanisms to capture spatial and temporal dependencies, providing a significant improvement in computational efficiency compared to existing models. The results indicate that the model performs comparably against the STPOTR architecture over longer prediction periods, demonstrating that it is possible to create an accurate, scalable, and computationally inexpensive framework suitable for real-time human motion prediction.

This work lays the foundation for advancements in human-robot interaction, particularly in dynamic environments such as crowded hallways. The ability to generalize across diverse motion patterns without the need for autoregressive processes highlights its potential for deployment in real-world applications. By focusing on longer-term motion predictions, the model provides a robust solution for tasks requiring anticipatory planning, such as navigation in social settings, collaborative tasks, and human-robot teaming in complex environments.

Key contributions and outcomes of this research include:

- Novel Data Representation: Incorporating velocity and acceleration alongside positional data allows the model to better capture motion dynamics and predict trajectories with higher accuracy.
- Efficiency in Design: The non-autoregressive architecture enables parallelized predictions, making it computationally efficient and suitable for real-time applications.

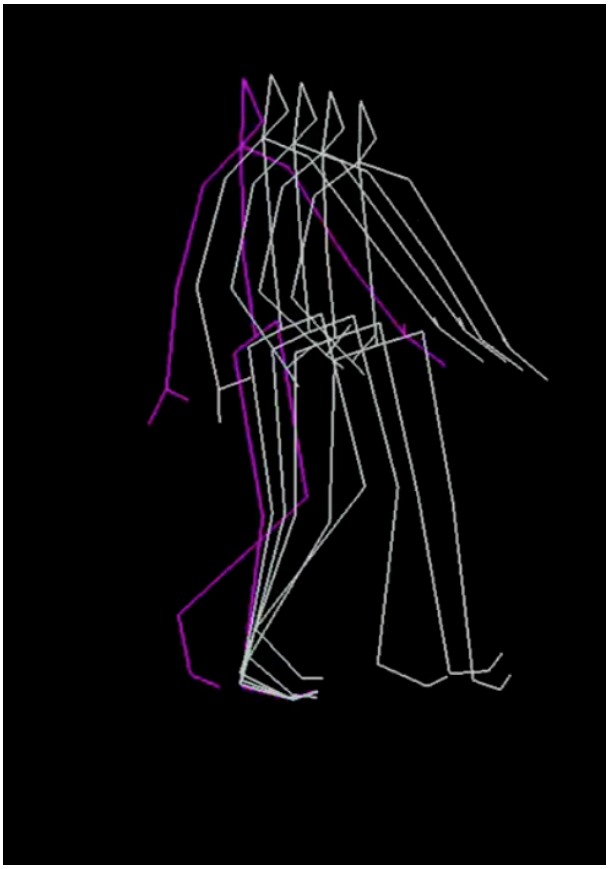

Fig. 4: Visualization of our transformer to verify the efficacy the output. The white figures are the predictions 5, 10, 15, and 20 frames in advance. The pink figure is the ground truth. The visualization is a accurate portrayal of the predictions versus the ground truth in 2D representation of 3D space.

- Visualization Validation: The use of visualization tools confirmed the accuracy of predicted trajectories, reinforcing the practical applicability of the model.
- Potential for Real-World Deployment: The consistent performance of the model across diverse motion scenarios suggests readiness for integration into robotics systems for real-time tasks.

A. Future Directions

While the results are promising, there remain exciting opportunities for further exploration:

- Multi-Person Prediction: Expanding the model to predict the motion of multiple individuals simultaneously. The non-autoregressive design ensures that such predictions can be parallelized, providing significant computational advantages.
- Real-Time Social Navigation: Deploying the model on robotic systems to evaluate its performance in live environments, where dynamic human behavior presents additional complexities.
- Robustness in Diverse Environments: Testing the model in varying physical contexts to enhance its generalizability, especially under real-time constraints.
- Enhanced Input Features: Exploring additional data modalities, such as visual cues or environmental context, to improve prediction accuracy and adaptability.

This study underscores the transformative potential of leveraging transformer architectures in robotics and motion prediction. By bridging the gap between theoretical advances and practical implementations, this work contributes to the broader goal of creating more intelligent, responsive, and socially aware robotic systems. The proposed framework not only demonstrates the feasibility of accurate human motion prediction in computationally constrained settings but also opens pathways for its application in collaborative robotics, autonomous vehicles, and beyond.

In conclusion, the integration of advanced machine learning techniques with real-world applications continues to drive innovation in human-robot interaction. The methods and insights presented here represent a step forward in enabling robots to navigate and adapt to human-centric environments effectively, paving the way for safer, more natural, and more efficient collaborative systems.

VII. Acknowledgments

We acknowledge Dr. Justin Hart from the UT Austin Computer Science Department and Dr. Nick Fey from the Walker Department of Mechanical Engineering for their continued support and advising of this project. We also thank Brandon Nunley from the Department of Biomechanical Engineering, and Jacob Mathew, Ananya Chintalapudi, and Zach Vazhekatt from the UT Austin Living With Robots Laboratory for their suggestions.

References

[1] M. Phillips and M. Likhachev, "SIPP: Safe interval path planning for dynamic environments," in Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA), May 2011, pp. 5628–5635

[2] G. Ferrer, A. Garrell and A. Sanfeliu, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2013, pp. 1688-1694, doi: 10.1109/IROS.2013.6696576.

[3] C. Zheng, S. Zhu, T. Yang, C. Chen, and Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers" (2021)

[4] Martinez, J., Black, M., and Romero, J. On human motion prediction using recurrent neural networks (2017).

[5] Pavllo, D., Feichtenhofer, C., Auli, M., and Grangier, D. Modeling human motion with quaternion-based neural networks (2019).

[6] Mahdavian, Mohammad, et al. "STPOTR: Simultaneous Human Trajectory and Pose Prediction Using a Non-Autoregressive Transformer for Robot Following Ahead." arXiv.org, 27 Sept. 2022, arxiv.org/abs/2209.07600.

[7] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3D Human Motion Prediction: A Survey," Neurocomputing, vol. 489, pp. 345–365, 2022

[8] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent Network Models for Human Dynamics," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4346–4354

[9] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial Geometry-Aware Human Motion Prediction," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 786– 803.