AdvMT: Adversarial Motion Transformer for Long-term Human Motion Prediction

Sarmad Idrees, Jongeun Choi and Seokman Sohn

***Abstract*—To achieve seamless collaboration between robots and humans in a shared environment, accurately predicting future human movements is essential. Human motion prediction has traditionally been approached as a sequence prediction problem, leveraging historical human motion data to estimate future poses. Beginning with vanilla recurrent networks, the research community has investigated a variety of methods for learning human motion dynamics, encompassing graph-based and generative approaches. Despite these efforts, achieving accurate long-term predictions continues to be a significant challenge. In this regard, we present the Adversarial Motion Transformer (AdvMT), a novel model that integrates a transformer-based motion encoder and a temporal continuity discriminator. This combination effectively captures spatial and temporal dependencies simultaneously within frames. With adversarial training, our method effectively reduces the unwanted artifacts in predictions, thereby ensuring the learning of more realistic and fluid human motions. The evaluation results indicate that AdvMT greatly enhances the accuracy of long-term predictions while also delivering robust short-term predictions.**

***Index Terms*—Human motion prediction, Deep learning, Adversarial learning, Transformer network**

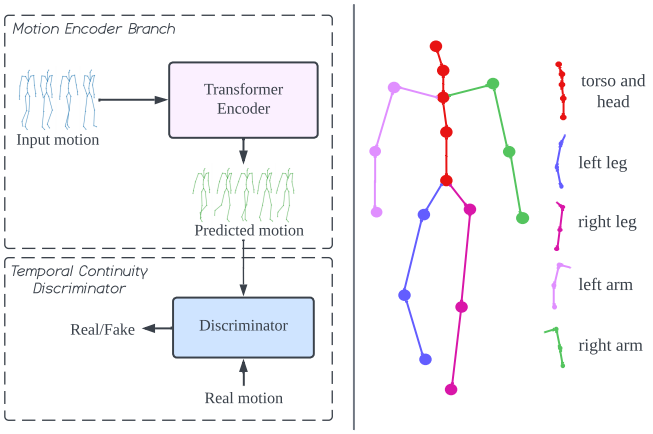

Fig. 1. *Left*: Overview of our proposed AdvMT network to predict future human motion by observing history motion. *Right*: The human body joints link structure consisting of human body parts: the torso and head, left leg, right leg, left arm, and right arm.

I. Introduction

The ability to accurately predict human motion is a crucial aspect in advancing human-robot interactions, an area that has garnered significant attention in both academic and industrial circles. At its core, the problem of human motion prediction aims to understand and forecast the complex and dynamic movements of humans, which is essential for enabling robots to interact safely and effectively with their human counterparts. This technology has broad applications, ranging from assistive robotics in healthcare, where robots aid in patient care, to advanced manufacturing settings, where collaborative robots work alongside human workers. However, the task of predicting human motion is challenging due to a multitude of influencing factors, including individual physical differences, psychological states, and the surrounding environment. These factors collectively complicate the development of realistic and reliable human motion prediction systems, a critical step for ensuring the smooth and harmonious integration of robots in a human-centric environment.

The early stages of human motion prediction research primarily focused on the use of Recurrent Neural Networks (RNNs) and their derivatives, favored for their proficiency in modeling sequential data, as evidenced by several key studies [1]–[4]. As the field evolved, the focus shifted towards convolutional networks, notably Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GCNs). These networks became prominent for their ability to effectively capture human motion representation, particularly by processing the spatial characteristics of human body joints [5]–[8]. Additionally, generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been explored to further enhance the learning of human motion dynamics. The use of adversarial training regimes, in particular, helps to address challenges such as the unrealistic movements and zero-velocity collapse problem, providing more refined and precise motion prediction outcomes [9]–[11].

Recently, the Transformer network [12] has emerged as a potent tool in the domain of human motion prediction, following its impressive performance in both Natural Language Processing (NLP) and computer vision. The inherent attention mechanisms of this network offer enhanced generalization capabilities over human pose datasets. An exemplary application can be found in the work of Cai et al. [13], where they augment a Transformer network with a progressive decoding strategy. This strategy enables the network to sequentially predict movements, cascading from central to peripheral joints in the kinematic chain, thus demonstrating its effectiveness in detailed human motion analysis.

Despite these advancements, accurately estimating long-term predictions remains a challenging aspect in the field of

Sarmad Idrees and Jongeun Choi are with the School of Mechanical Engineering, Yonsei University, Seoul 03722, Korea (e-mail: sarmad@yonsei.ac.kr; jongeunchoi@yonsei.ac.kr).

Seokman Sohn is with the Power Generation Lab, Korea Power Research Institute, 105, Munji-Ro, Yuseong-Gu, Daejeon, 34056, South Korea (email: happysohn@kepco.co.kr).

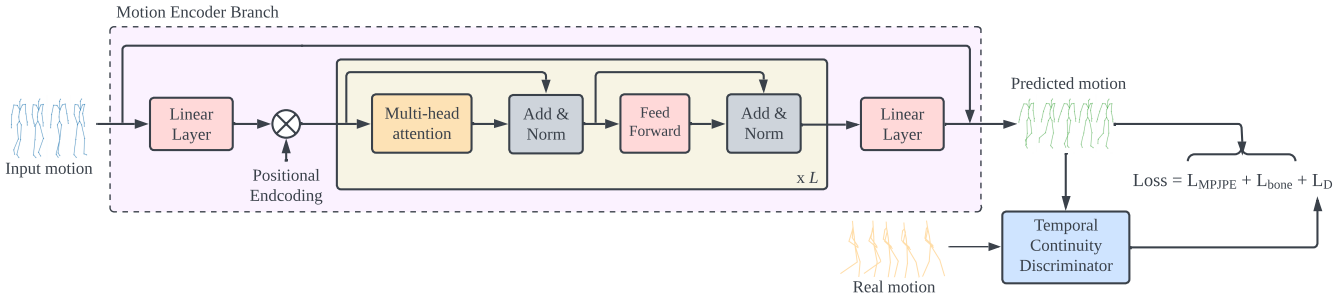

Fig. 2. The architecture of our proposed human motion prediction method primarily comprises of two main branches i.e. motion encoder branch and temporal continuity discriminator. The motion encoder branch, which employs a Transformer encoder layer, is dedicated to learning human motion dynamics. Whereas, the temporal consistency in motion prediction is achieved through our tailored loss function. The bone length error enables the model to maintain consistent bone lengths and adhere to human body constraints over extended periods. Additionally, the discriminator further refines the predicted poses by concentrating on the temporal differences in joint positions. We iteratively use previous predictions as input to forecast future motion, which is particularly effective for long-horizon predictions.

human motion prediction, primarily due to cumulative errors in later frames. The primary contribution of this research is the development of the Adversarial Motion Transformer (AdvMT), a transformer-based auto-regressive approach designed to tackle this challenge. Our contribution lies in the innovative combination of a tailored loss function with our method, which efficiently extracts both temporal and spatial information from motion sequences. By framing the challenge in an adversarial learning context, our model leverages auto-regressive predictions and discriminator feedback to refine its long-horizon forecasts. The results demonstrate that the AdvMT performs favorably compared to existing benchmarks in short-term motion prediction and shows promising results in long-term forecasts, thereby highlighting the versatility and comprehensive efficacy of our methodology.

The organization of the paper is as follows. Section II provides an overview of existing research in human motion prediction. In Section III, we detail our problem formulation and introduce a novel methodology to address challenges in this field. Section IV elaborates on the experimental setup and the results obtained, while Section V discusses the ablation study conducted to substantiate our model selection and the effectiveness of our loss function. The paper concludes with a section that consolidates our key findings and the contributions of our work.

II. Related work

A. Long-term human motion prediction

Human motion prediction has been a cornerstone in the field of computer vision and robotics, witnessing significant evolution over the years. The initial phase of research in this area primarily utilized RNNs with encoder-decoder models [1]–[3], but these models faced challenges in handling complex human motion dependencies. This led to a shift towards CNNs, which provided improved extraction of spatial joint connection information [5], [6], [14]. Later, the exploration of GCNs brought enhanced capabilities in anatomical relationship modeling [7], [8], [15]. However, much of the research has concentrated mainly on short-term predictions, leaving a notable gap in long-term prediction accuracy.

While the field has made advancements in short-term motion prediction, efforts to address the extended horizon prediction challenge have been limited. Tang et al. [16] spearheaded this effort by introducing a Modified Highway Unit (MHU) that effectively removes static joints, thereby focusing on joints in motion to predict reliable long-term motion. To further advance this field, Xu et al. [17] developed an attention-based Error Attenuation Network (EAN) with a focus on three major issues in long-term prediction. Their network aims to reduce error accumulation in future frames, address unbalanced data, and overcome mean pose generalization problems.

Furthermore, to leverage other architectures, Zhao et al. [18] introduced a novel approach by integrating a Transformer network with the capabilities of GAN for long-term prediction challenge. Their method includes a bi-directional Transformer and both frame-level and sequence-level discriminators. While their model achieved enhanced accuracy for long-term predictions, this was accompanied by increased computational demands. In contrast, our approach outperforms this by intelligently combining a curated loss function and a temporal discriminator, resulting in more efficient predictions with a reduced parameter footprint.

B. Adversarial training

Traditional models in human motion prediction often struggled with ensuring motion smoothness and robustness. This challenge prompted researchers to investigate generative architectures and the potential of adversarial training. A significant development in this area was the AGED architecture [3], which utilized a geometry-aware adversarial learning technique. This approach not only improved motion cohesiveness but also introduced a new level of diversity in predictions. Following this, the Q-DCRN model [19] advanced the use of adversarial learning by employing a discriminator to refine motion predictions across various horizons.

The adoption of GANs in human motion prediction rapidly expanded, with studies [10], [20], [21] exploring their capabilities. Despite their potential, GANs encountered challenges, especially in achieving Nash equilibrium. In response, refinement modules were integrated, as seen in [22], enhancing the

TABLE I
COMPARISON FOR SHORT-TERM (<400MS) AND LONG-TERM PREDICTION (>400MS) ON H3.6M DATASET ON FOUR MAIN ACTION CATEGORIES.

	Walking						Eating					
	Short-term		Long-term				Short-term		Long-term			
Time (milliseconds)	160	400	560	720	880	1000	160	400	560	720	880	1000
Res. Sup. [2]	40.9	66.1	71.6	72.5	76.0	79.1	31.5	61.7	74.9	85.9	93.8	98.0
convSeq2Seq [5]	33.5	63.6	72.2	77.2	80.9	82.3	22.4	48.4	61.3	72.8	81.8	87.1
HisRepeat [8]	**19.5**	**39.8**	47.4	52.1	55.5	58.1	**14.0**	36.2	50.0	61.4	70.6	75.7
BiTGAN [18]	-	-	49.8	55.0	58.5	60.5	-	-	48.5	59.2	68.2	73.0
AdvMT (ours)	23.9	39.9	**45.1**	**49.2**	**52.0**	**55.0**	18.3	**36.1**	**44.6**	**51.5**	**56.5**	**59.3**
	Smoking						Discussion					
	Short-term		Long-term				Short-term		Long-term			
Time (milliseconds)	160	400	560	720	880	1000	160	400	560	720	880	1000
Res. Sup. [2]	34.7	65.4	78.1	88.6	96.6	102.1	47.8	91.3	109.5	122.0	128.6	131.8
convSeq2Seq [5]	22.8	48.9	60.0	69.4	77.2	81.7	34.5	77.6	98.1	112.9	123.0	129.3
HisRepeat [8]	**14.9**	**36.4**	**47.6**	**56.6**	**64.4**	**69.5**	**23.4**	**65.4**	86.6	102.2	113.2	119.8
BiTGAN [18]	-	-	48.4	57.5	65.0	70.0	-	-	85.8	101.2	111.6	116.4
AdvMT (ours)	22.8	45.5	56.5	65.5	72.7	77.7	36.0	66.7	**80.2**	**90.9**	**97.8**	**101.0**

accuracy of generated poses and minimizing artifacts. Building on these developments, Lyu et al. [23] took a novel approach by modeling joint motion through stochastic differential equations and utilized GANs for simulating path integrals, further refining the precision of human motion prediction.

C. Transformer network

The advent of the Transformer network marked a significant paradigm shift in sequence modeling [12], offering a solution to the limitations inherent in RNNs, especially when dealing with long sequences. With their attention mechanisms, Transformers excel in focusing on pertinent features, thus efficiently handling long-term dependencies. Beyond NLP, their utility extended to image processing tasks, from classification [24] to object detection [25] and segmentation [26]. In human motion prediction, the works of Aksan et al. [27] and Chen et al. [28] stand out for effectively leveraging Transformers to capture both structural and temporal dependencies. However, we identify potential areas for enhancement in long-term prediction accuracy.

Building on this foundation, we introduce the Adversarial Motion Transformer (AdvMT), a novel approach that seamlessly integrates the strengths of Transformers with the robustness of adversarial training. AdvMT is specifically designed to enhance motion smoothness and achieve superior accuracy over extended prediction horizons, potentially establishing a new benchmark in human motion prediction.

III. ADVERSARIAL MOTION TRANSFORMER (ADVMT)

The overview of the Adversarial Motion Transformer (AdvMT), a system comprised of two main branches, is illustrated in Fig. 2. The motion encoder branch interprets the input motion history and encodes the local and global human joint dependencies. The discriminator branch complements this by refining the predictions from the motion encoder branch to ensure the generation of realistic and consistent human motion.

A. Problem formulation

Human motion prediction fundamentally involves forecasting future movements by interpreting past sequences of human motion data. In a mathematical context, this can be visualized as a function that processes a series of historical human motion data points $X_{1:T} = \{x_1, x_2, \ldots, x_T\}$ and predicts future human poses. Each $x_t = \{j_1, j_2, \ldots, j_N\}$ in this sequence represents a single pose at time t, consisting of N distinct joints. These joints are characterized in a K-dimensional pose representation, where $K = 3$ signifies the 3D position representation in Euclidean space. The model is trained to predict the poses for the forthcoming L time steps, effectively forecasting the sequence $\hat{X}_{T+1:T+L}$ based on the observed historical frames.

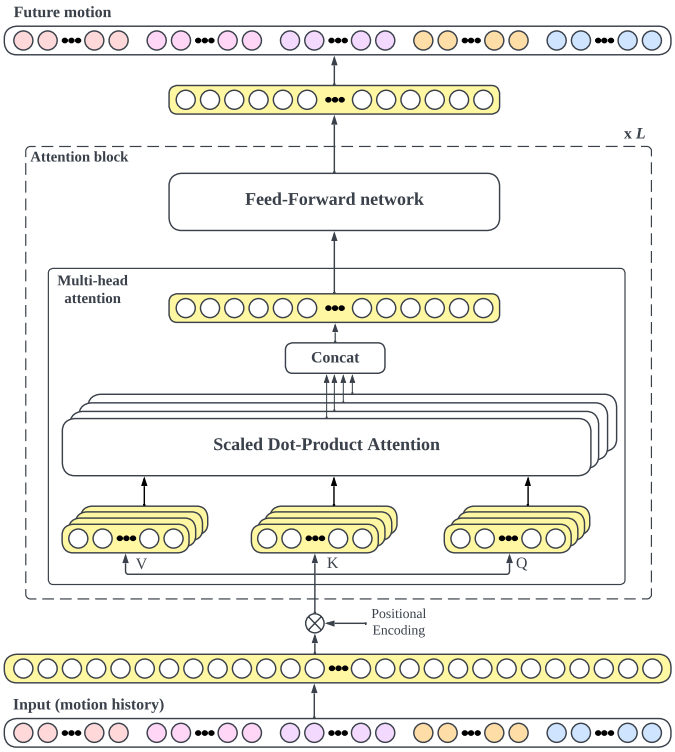

Fig. 3. The detailed architecture of our Tansformer-based motion encoder branch. The local and global dependencies within the human body are extracted through multiple layers of attention blocks. Each block aims to learn different aspects of motion dynamics, enabling a comprehensive understanding of human movement.

TABLE II
COMPARISON FOR LONG-TERM PREDICTION (>400MS) ON H3.6M DATASET ON REMAINING ACTION CATEGORIES.

	Directions				Greeting				Phoning				Posing			
Time (milliseconds)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Res. Sup. [2]	101.1	114.5	124.5	129.1	126.1	138.8	150.3	153.9	94.0	107.7	119.1	126.4	140.3	159.8	173.2	183.2
convSeq2Seq [5]	86.6	99.8	109.8	115.8	116.9	130.7	142.7	147.3	77.1	92.1	105.5	114.0	122.5	148.8	171.8	187.4
HisRepeat [8]	73.8	88.1	100.1	106.4	101.9	118.4	132.7	138.8	67.4	82.9	96.5	105.0	107.5	136.8	161.4	178.2
BiTGAN [18]	**73.3**	**87.9**	99.7	106.3	101.1	117.8	131.4	136.4	**67.3**	82.3	94.9	103.2	**107.1**	134.6	156.7	171.0
AdvMT (ours)	79.2	90.1	**99.4**	**103.5**	**95.1**	**104.5**	**114.1**	**118.5**	68.4	**79.6**	**88.5**	**93.7**	114.1	**128.1**	**138.4**	**145.2**
	Purchases				Sitting				Sitting Down				Taking Photo			
Time (milliseconds)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Res. Sup. [2]	122.1	137.2	148.0	154.0	113.7	130.5	144.4	152.6	138.8	159.0	176.1	187.4	110.6	128.9	143.7	153.9
convSeq2Seq [5]	111.3	129.1	143.1	151.5	82.4	98.8	112.4	120.7	106.5	125.1	139.8	150.3	84.4	102.4	117.7	128.1
HisRepeat [8]	**95.5**	**110.9**	125.0	134.2	76.4	93.1	107.0	116.0	97.0	116.1	132.1	143.5	**72.1**	**90.0**	**105.5**	**115.9**
BiTGAN [18]	99.0	113.7	127.1	135.1	**76.0**	**92.0**	**105.4**	**114.4**	**96.2**	**114.5**	**129.9**	**141.3**	74.2	92.6	107.4	117.7
AdvMT (ours)	99.2	112.3	**121.8**	**127.9**	87.2	101.8	113.1	121.4	100.5	117.9	132.1	142.2	85.4	100.7	113.4	122.0
	Waiting				Walking Dog				Walking Together				Average			
Time (milliseconds)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Res. Sup. [2]	105.4	117.3	128.1	135.4	128.7	141.1	155.3	164.5	80.2	87.3	92.8	98.2	106.3	119.4	130.0	136.6
convSeq2Seq [5]	87.3	100.3	110.7	117.7	122.4	133.8	151.1	162.4	72.0	77.9	82.9	87.4	90.7	104.7	116.7	124.2
HisRepeat [8]	74.5	89.0	100.3	108.2	108.2	120.6	135.9	146.9	**52.7**	**57.8**	**62.0**	**64.9**	**77.3**	91.8	104.1	112.1
BiTGAN [18]	**72.9**	**87.3**	**97.7**	**104.9**	105.4	120.4	136.4	148.3	54.3	59.7	64.2	67.3	**77.3**	91.7	103.6	111.1
AdvMT (ours)	83.3	94.8	103.5	109.5	**102.5**	**115.5**	**127.7**	**136.8**	62.9	71.0	79.0	84.4	80.3	**91.6**	**100.8**	**106.6**

B. Motion encoder branch

The motion encoder branch in our model is based on the Transformer architecture, as described in the foundational work of Vaswani et al. [12]. Our implementation, however, deviates from the conventional Transformer framework by exclusively utilizing the encoder component. This specific design choice is grounded in our findings from extensive ablation studies, which demonstrate that the encoder alone is sufficient for capturing the complexities of human motion prediction. These studies revealed that employing the full Transformer architecture, including both encoder and decoder, tends to introduce unnecessary complexity without proportional benefits in this context. As a result, our focused approach with just the encoder component not only simplifies the model but also effectively enhances its capability to predict human movements over extended timeframes, exceeding the typical one-second prediction horizon seen in prior methods.

A schematic of the motion encoder is depicted in Fig. 3. The process is initiated by transforming the input pose data into joint embeddings through a linear layer. In line with the Transformer framework [12], sinusoidal positional encodings are introduced to these embeddings. This addition is crucial as it allows the model to effectively process sequences of increased length, enabling a better understanding of the temporal relationships and positional dynamics within the motion data.

The architecture of our motion encoder comprises L layers of attention blocks. Each block consists of a multi-head attention mechanism coupled with a position-wise feed-forward network. This configuration allows the model to concurrently learn and integrate various local and global dependencies present in the data. The aggregated representation, forged through these attention layers, is then projected back into the space of human poses through another linear layer. In the final stage of the process, the predicted future human pose is fed into a discriminator. This step is significant as it constrains the motion encoder to focus on learning patterns that result in realistic human motion.

C. Temporal continuity discriminator

Our primary goal in this research is to learn a robust and plausible representation of long-term human motion. Recognizing the inherent uncertainty in human behavior and actions, we observe that some models may overlook human body constraints and fail to predict human-like motion. To incorporate temporal continuity and human body constraints, we include a discriminator branch alongside the motion encoder branch, as in [29]. Our network is specifically trained to focus on the joint positions with an emphasis on maintaining natural body-joint velocities through adversarial learning. The key intuition here is to concentrate on the temporal differences in joint positions rather than their absolute values, ensuring a more continuous and realistic joint movement sequence. The adversarial loss of temporal continuity discriminator D_K is defined as,

$$\mathcal{L}_{D_K} = \sum_{t=T+1}^{T+L} \left(\mathbb{E}_{x_t} \left[\|D_K(\Delta x_t)\|^2 \right] + \mathbb{E}_{\hat{x}_t} \left[\|1 - D_K(\Delta \hat{x}_t)\|^2 \right] \right), \tag{1}$$

where x_t and $\hat{x}_t$ refers to real and predicted motion sequences respectively, and Δx is the temporal change in the motion sequence.

The auto-regressive training regime empowers the discriminator to act as a feedback mechanism for the motion encoder branch, aiding in reducing error accumulation during extended horizon predictions. This approach significantly enhances motion prediction by preventing the tendency to predict zero-velocity motion. The inclusion of the discriminator branch not only enhances the realism of the generated motion but also ensures its smoothness over time.

D. Loss function

Studies have shown that using a vanilla Euclidean loss in human motion prediction often causes models to converge to a mean pose, a phenomenon highlighted in [30]. To address this issue, we propose a modified loss function designed to capture

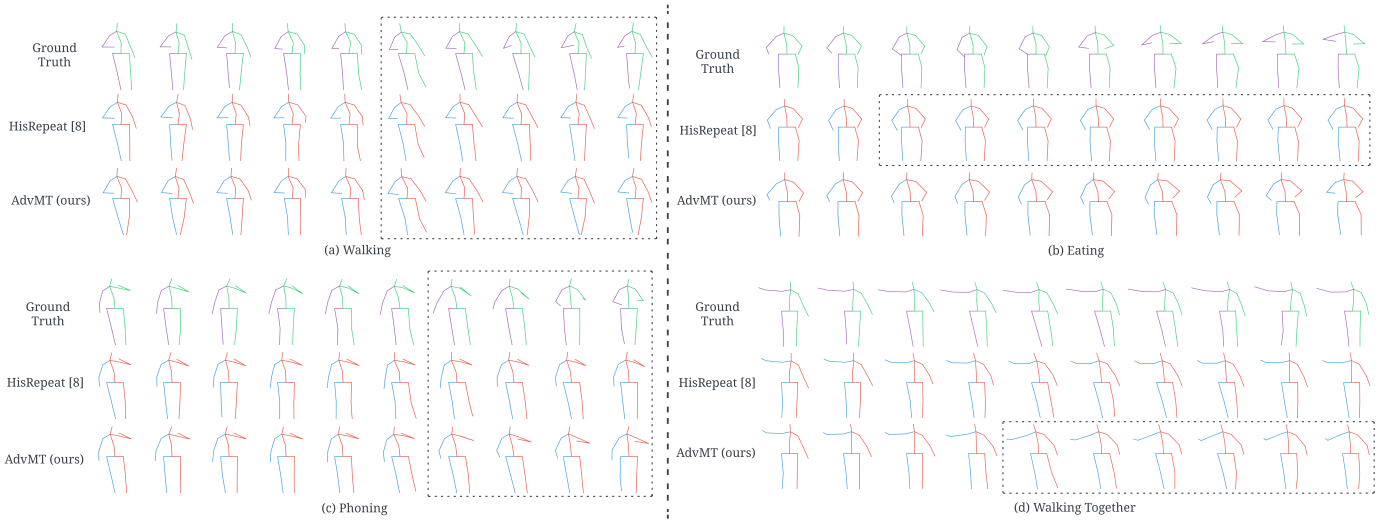

Fig. 4. Qualitative future motion prediction results up to 2 seconds for walking, eating, phoning, and walking together actions from H3.6M dataset. For visualization purposes, the predictions are down-sampled to 5 frames per second. Ground truth poses are drawn in purple and green, whereas the future predictions are marked in blue and red colors. Best visualized in zoomed view.

a better representation of motion data. Additionally, previous research has encountered challenges with zero-velocity collapse in long-term prediction tasks. This means that for predictions extending over 400ms, models often default to predicting static outputs for future frames, failing to capture the dynamic nature of human motion. Our modified loss function aims to mitigate these issues, enhancing the ability of the model to accurately predict longer sequences of human movement.

To effectively learn true human motion representation, our method addresses both spatial and temporal dependencies in the input motion sequence. The spatial relationship between human joints within each frame is captured by the self-attention layer in our motion encoder branch. Simultaneously, we ensure temporal consistency across frames by integrating it into our loss function formulation. This approach is vital for accurate predictions over longer time horizons, embedding a comprehensive understanding of both spatial and temporal dynamics directly into the training regime. Our tailored loss function is defined as

$$\mathcal{L}(X, \hat{X}) = \mathcal{L}_{\text{MPJPE}} + \lambda_B \mathcal{L}_{\text{bone}} + \lambda_D \mathcal{L}_{D_K}, \tag{2}$$

where X and $\hat{X}$ are ground truth and predicted poses. The first term in our loss function corresponds to the Mean Per Joint Position Error (MPJPE) as proposed in [31]. With t and n as the frame and joint number respectively, the $\mathcal{L}_{\text{MPJPE}}$ is defined as

$$\mathcal{L}_{\text{MPJPE}} = \frac{1}{N(T+L)} \sum_{t=T+1}^{T+L} \sum_{n=1}^{N} \|\hat{x}_{t,n} - x_{t,n}\|^2. \tag{3}$$

Moreover, the terms $\mathcal{L}_{\text{bone}}$ and $\mathcal{L}_{D_K}$ represent the bone length error and the adversarial loss, respectively. Each term is weighted by its regularization factor, with λ_B for bone length error and λ_D for adversarial loss, ensuring a balanced contribution of each component to the overall loss function. We refer to these losses as our temporal consistency loss.

TABLE III
LONG-TERM PREDICTION RESULTS FOR H3.6M DATASET.

	Walking		Smoking		Purchases		Sitting	
Time (ms)	1.5k	2k	1.5k	2k	1.5k	2k	1.5k	2k
HisRepeat [8]	64.6	73.4	87.7	101.8	159.1	172.2	156.6	178.9
AdvMT (ours)	62.4	73.1	91.4	99.0	140.4	153.1	151.0	164.2

Our rationale for incorporating bone length error is based on the constant nature of bone lengths over time, which introduces temporal consistency into our motion predictions. Bone lengths are calculated as the Euclidean distance between connected joint positions in predicted and ground truth poses.

However, relying solely on bone length error has a drawback. It may lead the motion encoder to minimize this error without adequately addressing zero-velocity, resulting in static motion predictions. To counteract this, we introduce the temporal continuity discriminator loss $\mathcal{L}_{D_K}$, which penalizes unrealistic human motions. This additional loss encourages the motion encoder to generate more dynamic, realistic, and plausible human motion, thus addressing a critical aspect of motion prediction that bone length error alone cannot resolve.

IV. EXPERIMENTS

Dataset: In our experiments, we conducted our model evaluation using the Human3.6M dataset, widely recognized as a benchmark in the field of human motion prediction. This extensive database contains over 3.6 million 3D poses, recorded with 7 actors performing 15 different types of actions. For training and evaluation, we downsampled the motion data to 25 frames per second. In alignment with the protocol outlined in [2], we used subjects S1, S6, S7, S8, S9, and S11 for training, and S5 is designated for testing. Similar to [8], we report our future motion prediction results on 256 subsequences per action.

Comparison with other methods: In our study, we focused on training our method for 3D joint position prediction. We

compared its performance with other state-of-the-art methods trained for the same motion representation, as detailed in Tables I and II. The primary metric used for evaluation is the MPJPE in millimeters, aligning with the standards used by other methods [8], [18]. We trained the model using an input sequence of 2 seconds (50 frames) to generate predictions for the next 1 second (25 frames). However, our motion encoder branch was trained auto-regressively, enabling it to predict motions extending beyond 1 second (see Table III and Fig. 4). The results for BiTGAN [18] are acquired from the published paper, while the HisRepeat [8] method is evaluated with the provided code and the trained model.

The results reveal that our proposed method consistently surpasses the baseline method [2] in both short-term and long-term predictions. While our performance in short-term prediction is comparable to the current state-of-the-art, it is in long-term prediction where our method particularly excels, outperforming in most of the action tasks evaluated.

The qualitative results further substantiate the efficacy of our model. As demonstrated in Fig. 4, our model excels in accurately predicting joint movements while adhering to the constraints of human body movement. In dynamic sequences like walking, our model significantly improves leg movement accuracy compared to the results in [8]. In addition, for actions like eating and phoning, our model closely approximates actual hand movements, unlike the static outputs predicted in the final frames by [8].

Due to its auto-regressive prediction approach, our method successfully avoids predicting zero-velocity motion for long-term predictions. A prime example is the phoning action, where our model realistically simulates the action of ending a call and putting down the phone, as marked with a dashed box in Fig. 4. It is important to note that while our method excels in dynamic actions, its performance is slightly reduced in static actions such as smoking and waiting.

A limitation we observed in our method is its tendency to focus on specific parts of the human body when predicting future motion. For example, in the walking together action, which requires learning both lower and upper body movements, our method concentrates on lower body movements, resulting in less accurate predictions of hand positions.

V. Ablation Study

A. Architecture

An ablation study was conducted to assess the effectiveness of our proposed AdvMT architecture. We compared the performance of a full Transformer network, comprising both encoder and decoder layers with separate self-attention and multi-head attention mechanisms, against our modified architecture. Originally intended for sequence-to-sequence tasks, the full Transformer architecture was found less effective for human motion prediction compared to our adaptation, which solely utilizes the encoder layer. This suggests that the decoder layer might add unnecessary complexity, hindering the ability to capture human motion dynamics accurately.

Further ablation studies highlighted the critical role of the discriminator branch in enhancing the realism and temporal consistency of generated motion. This branch serves as a feedback mechanism for the motion encoder, guiding it to correct and refine its predictions. Without it, the motion encoder tended to yield less realistic and plausible motions. The discriminator branch functions as a *critic*, ensuring the predicted motions are not only realistic but also align with real-world motion patterns. Its absence leads to the generation of unrealistic or inconsistent motion sequences, particularly in complex human motions where accuracy in joint angles and movements is crucial. By integrating the adversarial training, our model is compelled to generate more lifelike and consistent motion sequences.

TABLE IV
The ablation study results on H3.6M dataset.

Methods	Time (milliseconds)				
	160	400	560	880	1000
Baseline [12]	44.2	79.7	92.2	118.9	126.6
AdvMT ($\mathcal{L}_{\text{MPJPE}}$)	45.8	77.2	88.9	112.9	119.7
AdvMT ($\mathcal{L}_{\text{MPJPE}}$ + $\mathcal{L}_{\text{bone}}$ + $\mathcal{L}_{D_K}$)	33.2	65.3	80.3	100.8	106.6

B. Loss function

Regarding the loss function, we found that using only the bone length error as a loss term with the vanilla MPJPE loss led the model to predict zero-velocity motion for long-term prediction tasks. As the model learns to minimize only the bone length error, leading it to generate static poses. To mitigate this, incorporating the discriminator loss term helped to penalize unrealistic human motion predictions and encouraged the motion encoder to generate more realistic and plausible human motion. We conducted an ablation study to investigate the effectiveness of our modified loss function, which includes a discriminator loss and a bone error loss in addition to the regular MPJPE loss. We compared the performance of our model with the modified loss function against a baseline model that only used the regular MPJPE loss (see Table IV).

The results show that the combination of all three losses achieved the best performance while using only one of the losses resulted in lower performance. The bone error loss contributed the most to the overall performance improvement, followed by the discriminator loss. Our modified loss function, which includes the MPJPE loss, bone length error, and a temporal continuity discriminator loss, is effective in improving the quality of human motion prediction.

VI. Conclusion

In this study, we aim to develop an architecture to improve long-term human motion prediction. The long-time horizon prediction requires modeling the plausibility of the human motion by incorporating the temporal information between frames with the joint-level extraction. We propose a Transformer encoder-based model with a modified loss function to integrate the temporal consistency between the predicted frames. The spatial information is extracted from the transformer encoder branch and a temporal consistency is learned through the loss function. In our auto-regressive training regime, the additional

discriminator branch serves as feedback to the motion encoder, resulting in reducing the error accumulation over the course of time. Our method achieves comparable results in short-term predictions and excels in long-term predictions across most action classes. In future work, the improvement for short-term prediction can be achieved by incorporating the structure-aware model to serve as the motion encoder.

REFERENCES

[1] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," pp. 4346–4354, 12 2015.

[2] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 4674–4683, IEEE Computer Society, jul 2017.

[3] 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)*, pp. 786–803, 2018.

[4] D. Pavllo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," 2018.

[5] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5226–5234, 2018.

[6] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2133–2146, 2020.

[7] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9489–9497, 2019.

[8] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *European Conference on Computer Vision*, pp. 474–489, Springer, 2020.

[9] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1418–1427, 2018.

[10] J. N. Kundu, M. Gor, and R. V. Babu, "Bihmp-gan: Bidirectional 3d human motion prediction gan," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8553–8560, 2019.

[11] B. Chopin, N. Otberdout, M. Daoudi, and A. Bartolo, "Human motion prediction using manifold-aware wasserstein gan," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8, IEEE, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[13] Y. Cai, L. Huang, Y. Wang, T.-J. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, X. Shen, *et al.*, "Learning progressive joint propagation for human motion prediction," in *European Conference on Computer Vision*, pp. 226–242, Springer, 2020.

[14] Y. Li, Z. Wang, X. Yang, M. Wang, S. I. Poiana, E. Chaudhry, and J. Zhang, "Efficient convolutional hierarchical autoencoder for human motion prediction," *The Visual Computer*, vol. 35, pp. 1143–1156, 2019.

[15] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3d human motion prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6519–6527, 2020.

[16] Y. Tang, L. Ma, W. Liu, and W. Zheng, "Long-term human motion prediction by modeling motion context and enhancing motion dynamic," *arXiv preprint arXiv:1805.02513*, 2018.

[17] J. Xu, X. Lan, J. Li, X. Chen, and N. Zheng, "Ean: Error attenuation network for long-term human motion prediction," in *2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI)*, pp. 178–183, IEEE, 2019.

[18] M. Zhao, H. Tang, P. Xie, S. Dai, N. Sebe, and W. Wang, "Bidirectional transformer gan for long-term human motion prediction," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 5, pp. 1–19, 2023.

[19] Q. Men, E. S. Ho, H. P. Shum, and H. Leung, "A quadruple diffusion convolutional recurrent network for human motion prediction," *IEEE transactions on circuits and systems for video technology*, vol. 31, no. 9, pp. 3417–3432, 2020.

[20] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1418–1427, 2018.

[21] A. Hernandez, J. Gall, and F. Moreno-Noguer, "Human motion prediction via spatio-temporal inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

[22] X. Chao, Y. Bin, W. Chu, X. Cao, Y. Ge, C. Wang, J. Li, F. Huang, and H. Leung, "Adversarial refinement network for human motion prediction," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[23] K. Lyu, Z. Liu, S. Wu, H. Chen, X. Zhang, and Y. Yin, "Learning human motion prediction via stochastic differential equations," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4976–4984, 2021.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[26] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2021.

[27] E. Aksan, P. Cao, M. Kaufmann, and O. Hilliges, "Attention, please: A spatio-temporal transformer for 3d human motion prediction," *arXiv preprint arXiv:2004.08692*, vol. 2, no. 3, p. 5, 2020.

[28] L. Chen, R. Liu, X. Yang, D. Zhou, Q. Zhang, and X. Wei, "Sttg-net: a spatio-temporal network for human motion prediction based on transformer and graph convolution network," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, no. 1, p. 19, 2022.

[29] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, and B. Chen, "Motionet: 3d human motion reconstruction from monocular video with skeleton consistency," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 1, pp. 1–15, 2020.

[30] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3d human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022.

[31] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.