Generative AI for 3D Human Pose Completion Under RFID Sensing Constraints

Ziqi Wang and Shiwen Mao
Deptartment of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
Email: zzw0104@auburn.edu, smao@ieee.org

***Abstract*—Accurate 3D human pose estimation (HPE) from wireless signals is highly challenging due to wireless sensing constraints. A functional platform typically requires a comprehensive setup of transceivers and antennas, such as WiFi devices, FMCW radars, or RFID tags, all of which come with their own limitations. RFID-based 3D HPE often suffers from tag interference and sparse readings, resulting in incomplete skeletal pose estimations, as current RFID systems capture information from only up to 12 joints. To address this challenge, we propose a novel latent diffusion transformer (LDT) framework with a cross-attention conditioning method, termed PoseCompLDT, to accurately complete 3D poses by generating the missing joints, enabling full 25-joint configurations from partial 12-joint inputs. This marks a significant advance, as it is the first approach to achieve over 20 distinct skeletal joints for wireless sensing-based continuous 3D human pose estimation (HPE) using generative AI. Extensive experiments validate the effectiveness of PoseCompLDT in preserving motion fidelity, supported by rigorous qualitative and quantitative studies. The framework offers scalable solutions beyond RFID sensing applications, such as pedestrian tracking and health monitoring in occluded or constrained sensing environments.**

***Index Terms*—Generative AI, latent diffusion transformer, 3D human pose completion, multi-modal conditioning, RF sensing-based 3D human pose estimation.**

I. Introduction

Traditional camera-based human activity recognition systems often face challenges such as occlusion, lighting variability, and privacy concerns. To address these issues, wireless sensing technologies have emerged as a promising alternative, offering non-intrusive and robust solutions for monitoring human activities [1]. Achieving accurate 3D human pose estimations (HPE) without physical contact is particularly valuable in scenarios where privacy concerns, occlusions, or lighting conditions hinder camera-based methods. For instance, Zhao et al. introduced RF-Pose3D [2], a seminal system that can estimate 3D human poses through walls using RF signals.

However, wireless sensing systems often face challenges in achieving the same level of detail and accuracy as camera-based methods. These limitations arise from constrained spatial resolution of wireless signals, interference, and environmental occlusions, which can lead to incomplete pose estimations where key joints are missing or occluded. While deep learning models have advanced the field by leveraging spatiotemporal patterns in wireless signals, they often struggle with overfitting to limited training data and poor generalization across diverse subjects, activities, and environments [3]–[5]. This exacerbates the difficulty of reconstructing anatomically accurate, full-body poses in real-world deployments. In such scenarios, pose completion becomes critical to infer missing joints and ensure reliable 3D pose estimation, especially in occluded or sensor-limited conditions. This capability can extend beyond wireless sensing, highlighting the potential to address critical challenges in other fields such as accurate full-body poses for patient assessment and fall detection, realistic avatar movements to enhance user immersion, and improved safety systems in autonomous driving [6]–[8].

WiFi CSI-based systems [9] and radar-based methods [10], even when utilizing advanced transceiver setups or high-cost equipment, detect significantly fewer joints (typically 17) compared to camera-based systems (34 or 25 joints). This disparity arises from the inherent sensing limitations and model constraints of wireless-based systems. Their lower spatial resolutions compared to optical systems restrict the ability to distinguish closely positioned joints, such as those in the hands or neck region. For accurate creation of a fully rendered metaverse character, 24 major joints—especially detailed neck joints—are essential [11]. Despite advancements, wireless-based 3D HPE still faces considerable challenges in achieving the levels of completeness and reliability required for seamless integration into many emerging applications [12].

Transformers and diffusion models have significantly advanced the possible solutions to the recovery of missing joints in dynamic 3D human pose monitoring. Transformers, as demonstrated in [13], leverage self-attention mechanisms to simultaneously model relationships across all joints, making them effective in capturing complex spatial dependencies. However, their reliance on extensive and diverse training datasets can limit their generalization in out-of-domain (OOD) scenarios. Diffusion models complement these capabilities by modeling the underlying data distribution through a gradual transformation of Gaussian noise into realistic data samples. Their success in 3D HPE lies in their ability to ensure fidelity, diversity, and robust conditioning, enabling broader applications even with less reliance on exhaustive training data [14]. The authors in [15] leveraged a motion diffusion model combined with transformer-based architectures to predict human motions. The model employs a masked motion strategy, where segments of the motion sequence are deliberately masked during training, and the diffusion process progressively denoises these masks to reconstruct plausible trajectories.

In this paper, we focus on enhancing the RFID-based 3D

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

HPE system proposed in our prior work [16], which generates partial poses based on RFID readings from tags attached to 12 joints of the test subject. The problems detailed in [16] include tag interference, reader capacity, very sparse tag readings, and poor wearability in practice, making the existing system difficult to increase the number of deployed tags. We propose an innovative framework, termed PoseCompLDT, that uses conditional LDT to complete 3D human poses, thus overcoming the limitations of RF sensing. By leveraging cross-attention conditioning on partial pose observations, the LDT generates missing joints and produces coherent, full-body pose estimates that are based on RFID sensing, as shown in Fig. 1.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work that proposes a novel latent diffusion transformer architecture that leverages cross-attention conditioning on partial pose latents. This approach enables the model to effectively capture the correlations between partial and full poses in the latent space, facilitating more accurate and contextually relevant pose completion.
- We introduce a two-stage motion-aligned generative framework that combines attention-guided velocity matching with momentum-based refinement. The first stage captures spatial dependencies through cross-attention mechanisms, while the second stage enforces temporal consistency through adaptive velocity alignment, where the alignment weights are dynamically adjusted based on attention scores and enhanced with momentum terms to prevent motion freezing.
- The framework addresses practical limitations of RFID sensing, such as the use of only 12 tags and highly sparse phase data. Completing missing joints significantly enhances the practicality and effectiveness of RFID-based motion capture systems in real-world scenarios.

The remainder of this paper is structured as follows: Section II depicts the proposed system design and the training of the kinematics neural network and latent diffusion transformer models. Section III delineates the data collection and model implementations. Section IV discusses our experimental study. Section V summarizes this paper.

II. System Design

Our proposed framework addresses the challenge of completing 3D human poses in RFID-based sensing, where inherent constraints limit the number of detectable joints. By leveraging LDTs with cross-attention conditioning, we introduce a robust two-stage motion-aligned 3D pose completion system designed to infer missing joints and produce temporally coherent pose trajectories. The system architecture is presented in Fig. 2. We denote 3D pose data as 2D sequential data with frame-wise motion features $x^L_{1:N} = \{x^L_n\}^N_{n=1}$, where N indicates the number of time frames and L specifies the number of RF features.

Figure 1. Joints differences between a complete pose of 25 joints (left) and a partial pose of 12 joints (right).

A. 3D Pose Diffusion in Latent Domains

To overcome the noisy and incomplete nature of RFID-based pose estimation, we adopt an LDT framework that operates in a compressed latent space learned via a transformer-based variational autoencoder (VAE), denoted by $\mathcal{V}$. This design filters noise and models sequential dependencies efficiently. $\mathcal{V}$ consists of two main components: (i) The encoder (ε) encodes motion features $x^L_{1:N}$ into latent variables $z \in \mathbb{R}^{1\times 256}$, where 1 is the latent size and 256 is the latent dimension; (ii) The decoder (ψ) reconstructs the original data $\tilde{x}^L_{1:N}$ from the latent representation z. They both employ a transformer architecture with self-attention mechanisms. Skip connections between the encoder and decoder further enhance the reconstruction quality by preserving low-level information and gradients. The VAE is trained to minimize the following objective:

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL} + \lambda\mathcal{L}_{smooth}, \tag{1}$$

where $\mathcal{L}_{recon}$ is the reconstruction loss to ensure that the output $\tilde{x}^L_{1:N}$ closely matches the input $x^L_{1:N}$, $\mathcal{L}_{KL} = D_{KL}\left(\mathcal{N}(\mu_z, \sigma_z)|\mathcal{N}(0, \mathbf{I})\right)$ is the Kullback-Leibler divergence loss to regularize the latent distribution, and $\mathcal{L}_{smooth}$ is the temporal smoothness loss to preserve the continuity of motion, thus preventing unnatural or abrupt transitions in reconstructed poses, given by

$$\mathcal{L}_{smooth} = \sum_{n=1}^{N-1} \|\Delta x^L_n - \Delta\tilde{x}^L_n\|^2_2, \tag{2}$$

where Δx^L_n and $\tilde{x}^L_n$ denote the ground truth velocity and the reconstructed velocity, respectively.

The diffusion model progressively refines latent representations z using a multimodal-conditional transformer denoiser $\epsilon_\theta(x_t, t, e_\alpha, z_p)$. To guide pose completion, the denoiser incorporates cross-attention, aligning partial pose observations with the full pose. e_α is the activity embeddings mapped from activity labels α, and $z_p \in \mathbb{R}^{1\times 256}$ is the latent representations of the partial pose encoded by $\mathcal{V}$. The forward and reverse diffusion adhere to the design in [17]. We omit the detailed process to focus on our attention-based mechanism.

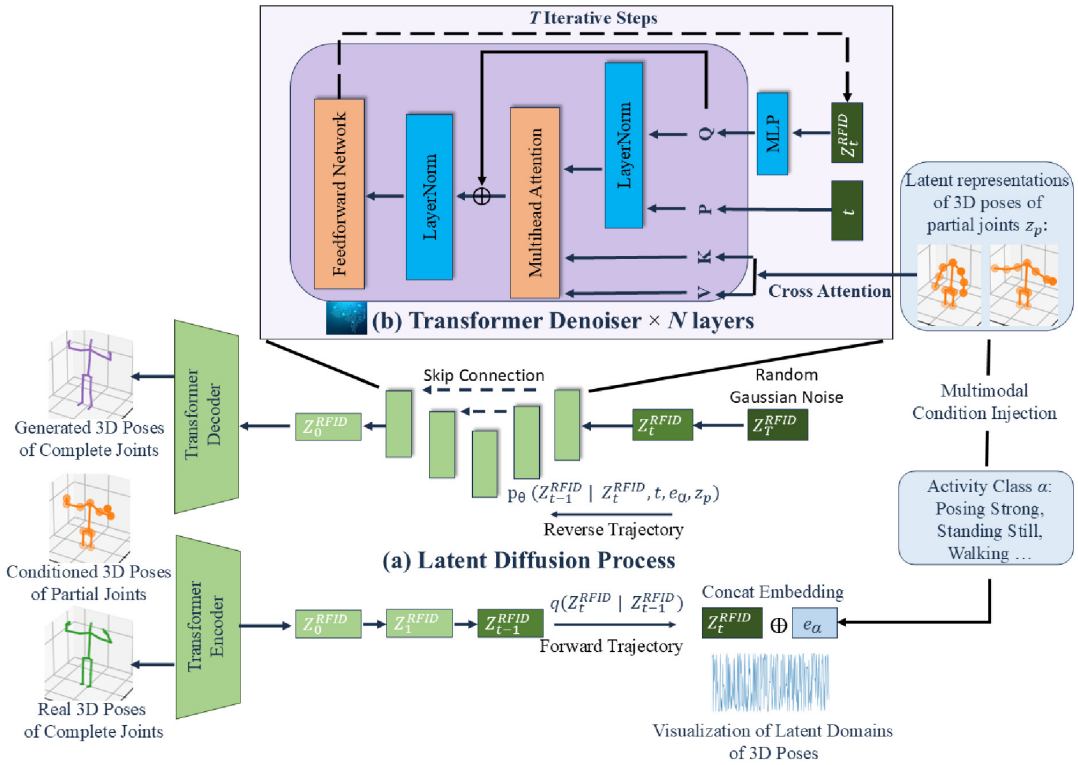

Figure 2. The diagram illustrates the process of conditional 3D pose data generation via LDTs: (a) latent diffusion process after compressing 3D pose data into its respective latent representations; (b) the detailed structure of the transformer denoiser and cross-attention conditioning method. The generated 3D pose of complete joints after its generative latent representations get decoded back to the data domain, as shown on the left, in comparison with the real 3D pose of complete joints.

B. Attention in the Transformer Denoiser

The denoiser ϵ_θ adopts a transformer encoder-only architecture to process noised latent representations $\tilde{z}_t = [z_t; e_t; e_a] \in \mathbb{R}^{3\times256}$, where z_t is the noised latent, e_t is the timestep embedding. Each transformer layer computes self-attention over $\tilde{z}_t$, which allows the model to capture long-range dependencies within the latent domain. For each layer l, the self-attention mechanism computes:

$$\begin{aligned}&\text{SelfAttn}(h_t^{(l-1)}) = \text{softmax}\left(\frac{Q_l(K_l)^T}{\sqrt{d}}\right)V_l,\\ &Q_l = h_t^{(l-1)}W_{Q,l}^{SelfAttn}, K_l = h_t^{(l-1)}W_{K,l}^{SelfAttn},\\ &V_l = h_t^{(l-1)}W_{V,l}^{SelfAttn},\end{aligned} \tag{3}$$

where $W_{Q,l}^{\text{SelfAttn}}$, $W_{K,l}^{\text{SelfAttn}}$, and $W_{V,l}^{\text{SelfAttn}} \in \mathbb{R}^{256\times256}$ are learnable projection matrices, and $h_t^{(l-1)}$ is the input from the previous layer. The self-attention mechanism computes compatibility scores between query and key vectors, normalizes them with a softmax, and applies the scores to value vectors, capturing relationships between latent features. This allows the denoiser to adaptively attend to relevant features in the noised latent and activity embeddings.

To stabilize training, residual connections and layer normalization are applied, followed by a feed-forward network. The output from the final layer predicts the noise $\epsilon_\theta(z_t, t, e_a)$, guiding the denoising process of activity-conditioned 3D pose generation.

To condition pose completion on partial observations, we incorporate cross-attention, which aligns partial pose latents z_p with the noised full pose latent z_t at each timestep. The cross-attention mechanism is defined as:

$$\begin{aligned}&\text{CrossAttn}(z_t, z_p) = \text{softmax}\left(\frac{Q_t K_p^T}{\sqrt{d}}\right)V_p,\\ &Q_t = z_t W_Q^{CrossAttn}, K_p = z_p W_K^{CrossAttn},\\ &V_p = z_p W_V^{CrossAttn},\end{aligned} \tag{4}$$

where $W_Q^{\text{CrossAttn}}$, $W_K^{\text{CrossAttn}}$, and $W_V^{\text{CrossAttn}}$ are learnable projection matrices. The attention weights computed between z_t (query) and z_p (key) emphasize relevant features of the partial pose latent. The resulting value vectors V_p guide the generation of missing joints.

C. Two-Stage Motion-Aligned Completion

A single diffusion pass may lack the ability to fully align reconstructed poses with partial pose trajectories. To address this, we propose a two-stage completion framework. The first stage (initial generation) generates preliminary latent representations of the full pose using cross-attention-guided diffusion, based on the latent representations of the observed partial pose. The second stage (refinement) refines the latent representations

by aligning velocities of initial latents and observed latents, leveraging stored cross-attention maps. The entire process is operated in the latent domain, ensuring structural integrity and alignment with the observed trajectories and with high computation efficiency.

The specific generation guideline is as follows: we utilize distinct attention guidance scales (γ_1, γ_2) and class conditioning scales (λ_1, λ_2) across the two stages. During Stage 1, stronger guidance ($\gamma_1 > \gamma_2$) is required to capture clear action-specific features, while a higher conditioning scale (λ_1) ensures the establishment of a basic pose structure. We typically set $\gamma_1 \approx 10$ and $\lambda_1 \approx 20$ in this stage. Stage 2 focuses on refinement with weaker guidance (γ_2). A lower conditioning scale (λ_2) introduces greater flexibility, facilitating natural motion alignment. Typical values for this stage are $\gamma_2 \approx 7.5$ and $\lambda_2 \approx 12.5$. For unseen scenarios, where the model must rely more heavily on conditional partial poses, λ_1 and λ_2 are increased to approximately 35 and 22.5, respectively, to enhance adaptability and ensure robust completion. A higher guidance scale can cause the movements to be stiff, and lower guidance could confuse the class the model tries to generate. The latent motion alignment process combines base constraints with attention-guided adaptation. We set a minimum partial pose influence using a base weight of 0.3, then scale this weight using cross-attention scores: joints with high attention receive stronger weights (≈ 0.5) for closer alignment with partial poses, while joints with low attention maintain lower weights (≈ 0.3) for more flexible motions. A velocity momentum term is introduced to maintain motion continuity and prevent static poses. For this experiment, we use 80 and 60 denoising steps for Stage 1 and Stage 2, as it only takes 1.63 seconds to generate a full pose with some of the best performance as of several performance metrics.

III. System Implementation

Building upon the foundation established by [16], our system advances RFID-based HPE by transforming RFID phase variations into human joint rotations using a transformer-based autoencoder. This transformation enables the application of forward kinematics to map the estimated joint rotations onto a target skeleton, facilitating dynamic 3D pose estimation. Unlike Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) that process data step by step, transformer architectures can attend to all temporal positions simultaneously. This characteristic makes transformers particularly adept at capturing long-range dependencies in RFID signal sequences. The RNN baseline RFID-Pose requires 36 s and nearly 15 s to estimate one output pose. In contrast, our transformer-based model only takes 0.6 s to initiate and load the model and 0.08 s to estimate one output.

Our RFID sensing setup utilizes an Impinj R420 reader with three S9028PCR polarized antennas and passive ALN-9634 (HIGG-3) tags. Signal processing, model training, and inference are conducted on a Lenovo desktop equipped with an RTX A4000 GPU. As in [16], 12 passive RFID tags are attached to specific joints on a participant to track their movements. We capture the phase variation from the three reader antennas, which reflects changes in the signal phase between consecutive readings, effectively encoding the relative movement of the participant's joints. As Fig. 1 shows, compared to the Kinect captured full pose of 25 joints, the entire joint chains below the knees are missing, alongside small details such as hands and thumbs. The partial pose tracks up to the neck, with information regarding the head entirely missing.

Nine activities—raising the right arm and drinking, waving up and down, jabbing, standing still, body twisting, walking, squatting, kicking, and posing—were performed by seven test subjects within the detection range of the RFID sensing platform and Kinect 2.0 device. Each activity was repeated continuously, resulting in 24 paired RF and 3D pose samples of 72 time frames per class for training and 6 samples for testing. To enhance temporal diversity, these samples were window-slided into 30-frame segments with a 10-frame overlap during training. The dataset included five subjects (four males and one female) with similar body shapes for training and two different subjects with distinct forms for testing.

The transformer-based kinematics neural network employs a 6-layer encoder-decoder architecture with eight attention heads per layer. The feature dimension of each transformer is 256 and expanded to 1,024 in feed-forward layers. Inputs consist of RFID data, while outputs generate quaternion rotations and global offsets. Dropout is set to 0.1, and the model is trained using the Adam optimizer with a learning rate of 1e-4 and batch size 32. Training spans 10,000 epochs, employing local position loss, global offset loss, twist loss, and smoothness loss for optimization.

The VAE encoder and decoder use an 11-layer transformer architecture with eight attention heads and a latent dimension of (1, 256). The transformer-based denoiser features 11 encoder layers with eight attention heads, feed-forward dimensions of 1,024, a dropout rate of 0.2, learned positional embeddings, and cross-attention layers for partial pose conditioning. The diffusion process spans 1,000 timesteps with a linear noise schedule. Classifier-free guidance is implemented with a scale of 7.5 and an unconditional dropout rate of 0.1. Cross-attention enables interaction between noised full poses and partial pose latents for effective completion. Both the VAE and diffusion models are trained using the AdamW optimizer with mixed precision (16-bit) and gradient accumulation (2 batches) for memory efficiency. Training the VAE takes approximately 13,300 epochs, while training the diffusion model takes 13,899 epochs. Validation occurs every 100 epochs. Gradient clipping is applied with a maximum norm of 5 for stability. The entire implementation leverages PyTorch Lightning for distributed training, optimizing GPU resource utilization.

IV. Experimental Study

We align the root position of the conditional partial pose, generated pose, and ground truth full pose before calculating various metrics. The mean per joint position error (MPJPE) is

computed as the discrepancy between the generated full pose and the ground truth full pose, as

$$\text{MPJPE} = \frac{1}{N_f N_j} \sum_{f=1}^{N_f} \sum_{j=1}^{N_j} \|\mathbf{p}_{f,j} - \hat{\mathbf{p}}_{f,j}\|_2, \tag{5}$$

where N_f is the number of frames, N_j is the number of joints, and $\mathbf{p}_{f,j}$ and $\hat{\mathbf{p}}_{f,j}$ are the ground truth and predicted 3D positions of joint j in frame f, respectively. We introduce frame-wise joint position smoothness that measures how smoothly the poses transition between consecutive frames by calculating the Euclidean distance of joint positions between adjacent frames. This metric is sensitive to large movement discrepancies or minimal movements.

$$\text{Smoothness}_{\text{frame}} = \frac{1}{N_f - 1} \sum_{f=1}^{N_f - 1} \|\mathbf{p}_{f+1} - \mathbf{p}_f\|_2, \tag{6}$$

where $\mathbf{p}_f$ is the pose in frame f, and N_f is the number of frames. Furthermore, we incorporate the following metrics to measure the anatomical plausibilities of the generated poses.

$$\text{Bone Length Consistency} = \frac{1}{N_f N_b} \sum_{f=1}^{N_f} \sum_{b=1}^{N_b} |\ell_{f,b} - \hat{\ell}_{f,b}| \tag{7}$$

$$\text{Joint Angle Error} = \frac{1}{N_f N_a} \sum_{f=1}^{N_f} \sum_{a=1}^{N_a} |\theta_{f,a} - \hat{\theta}_{f,a}|, \tag{8}$$

where N_b is the number of bones, $\ell_{f,b}$ and $\hat{\ell}_{f,b}$ are the ground truth and predicted lengths of bone b in frame f, respectively, N_a is the number of angles, $\theta_{f,a}$ and $\hat{\theta}_{f,a}$ are the ground truth and predicted joint angles in degrees for angle a in frame f, respectively. We also use the Frechet Inception Distance (FID) and diversity score to quantitatively measure the motion fidelity and diversity of our generated poses, given by

$$\begin{aligned}\text{FID} = \|\mu_\phi(\mathbf{X}_r) - \mu_\phi(\mathbf{X}_g)\|_2^2 + \\ \text{Tr}(\Sigma_\phi(\mathbf{X}_r) + \Sigma_\phi(\mathbf{X}_g) - 2\sqrt{\Sigma_\phi(\mathbf{X}_r)\Sigma_\phi(\mathbf{X}_g)}),\end{aligned} \tag{9}$$

$$\text{Diversity} = \mathbb{E}_{\mathbf{x}_1, \mathbf{x}_2 \sim \mathbf{X}_g} \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|_2^2, \tag{10}$$

where $\phi(\cdot)$ is the feature extractor, which is a custom pose discriminator we train on the Kinect-captured 3D pose data. $\mathbf{X}_r$ and $\mathbf{X}_g$ are the real and generated pose sets, respectively, $\mu_\phi(\cdot)$ is the mean of extracted features, $\Sigma_\phi(\cdot)$ denotes their covariance matrices, and $\mathbf{x}_1, \mathbf{x}_2$ are randomly sampled pairs of poses (50 samples for a set).

This experiment evaluates the system's ability to generate complete 3D poses while preserving the trajectory of the conditioned partial input. As shown in Fig. 3, the system demonstrates excellence in two critical aspects. The first is trajectory adherence, where the generated poses (purple) closely follow the trajectory of the conditioned partial poses (orange), ensuring alignment with input motion dynamics. Unlike simpler pose completion methods such as joint-based interpolation that risk overfitting to ground truth (green), our model maintains individuality in motion dynamics, even for unseen scenarios; The second is structural accuracy, where the generated poses remain anatomically plausible, completing the skeletal structure without sacrificing bone consistency. We use newly collected poses with the same activities and poses with a different skeleton subject as the conditions to simulate an in-the-wild setting. We can see that in the right column of Fig. 3, raising arms over the shoulder is performed by similar partial skeletons to the ground truth in the left column, but the skeleton is having a different pose. Our generated full pose can follow the partial pose despite the unseen movement patterns. The walking example in the right column demonstrates that with a largely different skeleton (mainly in the structures of foot to hip positions), our generated pose can follow the movement trajectories despite the skeleton difference.

To ensure a comprehensive evaluation of our generated poses, we generate the same number of samples as the ground truth counterparts and calculate their average scores. Table I presents the metrics introduced earlier for both ground truth and unseen partial pose conditioning scenarios. We achieve a considerable lower average joint position error compared to the ground truth full poses. In the context of 3D HPE, where the average limb length is approximately 35 cm, a 10 cm error translates to deviations of roughly $\sim \frac{1}{3}$ for an arm or $\sim \frac{1}{4}$ for a leg. Our error of 11.26 cm is right within this range, particularly as we generate 25 joints and infer 13 missing joints—significantly more than most existing frameworks. This highlights the complexity of our task, which involves motion tracking rather than static pose estimation.

To address the absence of ground truth for unseen scenarios, we introduce two additional metrics: trajectory joint error and trajectory velocity error. These metrics evaluate the average joint position and velocity deviations between the matching 12 joints of our generated full pose with the conditioned partial pose. They offer a quantitative measure of how well the generated poses follow the trajectory of the partial pose beyond mere visual inspection. Even in unseen scenarios, the trajectory error remains within an acceptable range of 8.03 cm, demonstrating the robustness of our approach. Overall, the trajectory errors across both scenarios confirm that our method effectively preserves the natural motion range dictated by the conditioned partial poses, ensuring a balance of structural accuracy and temporal fidelity.

V. Conclusions

In this paper, we introduced a novel framework leveraging LDT for completing 3D human poses from partial RFID-based observations. The system integrates two key components: a transformer-based kinematics predictor and a cross-attention-enabled pose completion model. Our approach tackled challenges inherent in RFID-based 3D pose estimation, including low tag density, sparse readings, and ensuring anatomically plausible pose completion. Through extensive experiments, we validate the proposed system's ability to generate realistic

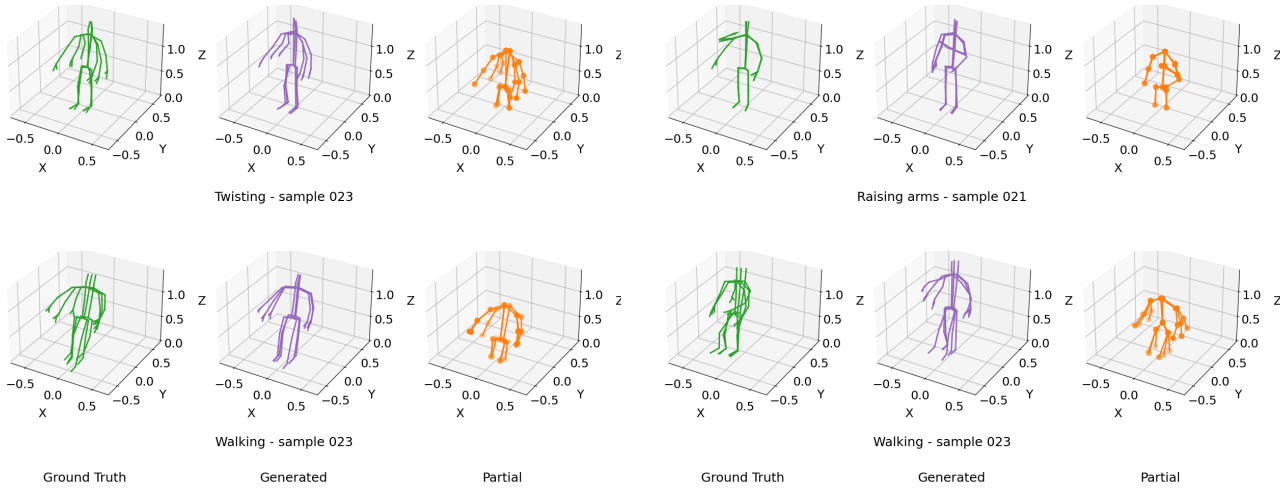

Figure 3. Visual illustration of the trajectories of 3D poses generated by our model (purple) compared to the ground truth complete pose (green) and the conditioned partial pose (orange). Each subplot is plotted every other five frames three times. The left columns demonstrate that our model accurately generates poses identical to the ground truth complete poses when conditioned on ground truth partial poses. In contrast, the right columns illustrate that our generated poses follow the trajectory of conditioned unseen partial poses rather than simply replicating the ground truth complete poses.

Table I
EVALUATION METRICS FOR 3D POSE COMPLETION WITH GROUND TRUTH AND UNSEEN PARTIAL POSE CONDITIONING

Metrics	Ground Truth	Unseen
Avg joint error (cm)	11.26	16.83
Bone consistency (cm)	1.78	2.15
Joint angle error (°)	7.31	10.49
Smoothness (cm/frame)	2.46	1.89
FID (-)	0.81	4.12
Ground truth FID (-)	0.18	0.47
Diversity (-)	22.52	15.51
Ground truth diversity (-)	24.83	30.46
Trajectory joint error (cm) compared with partial pose	7.18	8.03
Trajectory velocity error (cm/frame) compared with partial pose	7.80	8.23

and diverse pose data, achieving low joint errors, high temporal smoothness, and competitive FID and diversity scores. This work advances the integration of wireless sensing with AIGC, providing scalable solutions for pose completion in applications such as smart healthcare, autonomous systems, and immersive metaverse environments.

ACKNOWLEDGMENTS

This work is supported in part by the NSF under Grants CNS-2107190 and CNS-2148382.

REFERENCES

[1] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, Aug 2020.

[2] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3D skeletons," in *Proc. ACM SIGCOMM 2018*, Budapest, Hungary, Aug. 2018, pp. 267–281.

[3] C. Yang, X. Wang, and S. Mao, "Subject-adaptive skeleton tracking with RFID," in *Proc. The 16th IEEE International Conference on Mobility, Sensing and Networking*, Tokyo, Japan, Dec. 2020, pp. 599–606.

[4] C. Yang, L. Wang, X. Wang, and S. Mao, "Environment adaptive RFID based 3D human pose tracking with a meta-learning approach," *IEEE J. Radio Frequency Identification*, vol. 6, no. 1, pp. 413–425, Jan. 2022.

[5] Z. Wang, C. Yang, and S. Mao, "AIGC for RF-based human activity sensing," *IEEE Internet of Things Journal*, 2024, in press. DOI: 10.1109/JIOT.2024.3482256.

[6] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A survey of vision-based human action evaluation methods," *MDPI Sensors*, vol. 19, p. 4129, Sept. 2019.

[7] K. Ahuja, V. Shen, C. M. Fang, N. Riopelle, A. Kong, and C. Harrison, "ControllerPose: Inside-out body capture with VR controller cameras," in *Proc. ACM CHI 2022*, New Orleans, LA, Apr.-May 2022, p. Article No. 108.

[8] A. Zanfir, M. Zanfir, A. Gorban, J. Ji, Y. Zhou, D. Anguelov, and C. Sminchisescu, "HUM3DIL: Semi-supervised multi-modal 3D human-pose estimation for autonomous driving," in *Proc. 6th Annual Conf. Robot Learning*, Auckland, NZ, Dec. 2022, pp. 1–4.

[9] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proc. ACM MobiCom'20*, London, UK, Sept. 2020, p. Art. No. 23.

[10] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10 032–10 044, Sept. 2020.

[11] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF CVPR 2019*, Long Beach, CA, June 2019, pp. 10 975–10 985.

[12] Y. Zhou, H. Huang, S. Yuan, H. Zou, L. Xie, and J. Yang, "MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation," *IEEE Internet of Things J.*, vol. 10, no. 16, pp. 14 128–14 136, Aug. 2023.

[13] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF CVPR 2021*, Virtual, June 2021, pp. 1954–1963.

[14] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, June 2024.

[15] L.-H. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu, "HumanMAC: Masked motion completion for human motion prediction," in *Proc. IEEE/CVF ICCV 2023*, Paris, France, Oct. 2023, pp. 9544–9555.

[16] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3D human pose estimation with RFID," *IEEE Transactions on Reliability*, vol. 70, no. 3, pp. 1218–1231, Sept. 2021.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arxiv:2006.11239*, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2006.11239