

Generative AI-Empowered RFID Sensing for 3-D Human Pose Augmentation and Completion

ZIQI WANG^{ID} (Graduate Student Member, IEEE), AND SHIWEN MAO^{ID} (Fellow, IEEE)

(Selected Best Papers of IEEE International Conference on Communications (ICC'24))

Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA

CORRESPONDING AUTHOR: S. MAO (e-mail: smao@ieee.org)

This work was supported in part by the National Science Foundation (NSF) under Grant CNS-2107190 and Grant CNS-2148382.

ABSTRACT Collecting paired Radio Frequency Identification (RFID) data and the corresponding 3D human pose data is challenging due to practical limitations, such as the discomfort of wearing numerous RFID tags and the inconvenience of synchronization between RFID and camera data. We propose a novel framework that leverages latent diffusion transformers to generate high-quality, diverse RFID sensing data across multiple classes. This synthetic data augments limited datasets by training a transformer-based kinematics predictor to estimate 3D poses with temporal smoothness. More important, we introduce a latent diffusion transformer training stage with cross-attention conditioning and an inference design of two-stage velocity alignment to accurately infer missing joints in skeletal poses, completing full 25-joint configurations from partial 12-joint inputs. This is the first method to detect > 20 distinct skeletal joints using Generative-AI technologies for any wireless sensing-based continuous 3D human pose estimation (HPE) task. The application is particularly useful for RFID-based systems, which typically capture limited joint information due to RFID sensing constraints. Our approach can extend the applicability of wireless-based pose estimation to scenarios where collecting extensive paired datasets is impractical and achieving more fine-grained joint information is infeasible, such as pedestrian and health monitoring in occluded environments.

INDEX TERMS 3D pose augmentation and completion, conditional latent diffusion transformer, generative AI, multi-modal conditioning, RF sensing-based 3D human pose estimation.

I. INTRODUCTION

WIRELESS sensing has revolutionized human activity recognition (HAR) and environmental understanding, offering the advantages of non-intrusive, device-free, privacy-preserving, and low-cost monitoring [1]. Early pioneering works demonstrated the potential of wireless signals to capture human activities without the need for wearable devices. For instance, the seminal work by Adib and Katabi [2] utilized WiFi signals to detect human movements through walls, marking a milestone in device-free wireless sensing. Subsequent developments advance quickly where WiFi Channel State Information (CSI) [3], [4], RFID [5], [6], and FMCW Radar [7], [8] are leveraged for vital sign monitoring and fine-grained HAR, showcasing the utility of wireless signals in inferring spatial and temporal motion patterns. These studies highlighted the robustness

and versatility of wireless sensing across various domains, establishing it as a compelling alternative to traditional camera and sensor-based approaches.

Building on the foundations of wireless sensing, researchers have extended its applications to more complex tasks, such as 3D human pose estimation (HPE). The ability to estimate human poses without contact prevails in scenarios where camera-based methods may fail due to privacy concerns, occlusions, or lighting constraints. Zhao et al. [9] introduced RF-Pose, a groundbreaking system that estimates 2D human poses through walls using RF signals. This work was extended to 3D pose estimation in [10], demonstrating that wireless signals contain rich spatial information capable of reconstructing 3D human poses. Further advancements like [11], [12] proved that mmWave radars and RFID can be effectively used for high-accuracy 3D pose estimation as well.

The development of deep learning has dramatically improved the performance of 3D pose estimation, but obtaining large, labeled datasets for wireless-based HPE remains a significant bottleneck. Collecting paired RF signal and 3D pose data is inherently challenging due to the complexity of capturing synchronized multi-view data and the necessity of specialized hardware setups, which limit data diversity and volume. Güler et al. [13] developed DensePose, a large-scale dataset and model for 3D HPE using RGB data, highlighting the scale required for deep learning-based 3D pose estimation. However, creating such datasets for wireless sensing is even more constrained, as it requires RF setups in addition to expensive motion capture systems. The lack of extensive datasets impedes progress in wireless-based pose estimation, as deep learning models often suffer from overfitting to limited data and struggle with generalization across diverse activities, subjects, environments, and spectra [14], [15], [16].

An additional challenge often ignored in 3D pose estimation, and particularly in wireless-based systems, is *pose completion*—the recovery of occluded or missing parts of the human pose. In real-world scenarios where only a partial view of the human body is available, such as occlusions due to obstacles or limited sensor coverage, recovering the missing parts of a pose becomes essential. Pose completion is critical for applications like healthcare monitoring, where accurate full-body poses are necessary [17]. In VR/AR environments, complete 3D poses enhance user immersion and interaction by enabling realistic avatar movements [18]. Moreover, in autonomous driving and pedestrian monitoring, understanding the full pose of individuals can improve safety systems and human-robot interaction [19].

However, research specifically addressing pose completion in wireless-based pose estimation is still in its infancy, as this domain faces additional challenges from the indirect nature of RF signals and noise sensitivity. We choose to develop an RFID-based 3D pose estimation system [12] as an example due to its resilience to environmental interference as well as its low cost and wide availability. Most existing RFID-based work [12], [14], [15], [16] only deliver partial poses with 12 joints, where a completion scheme can be highly desirable. Our work addresses these challenges by proposing a novel approach that leverages latent diffusion models and transformer-based architectures for RFID-based single-person 3D HPE and pose completion. We develop a multi-purpose latent diffusion transformer system suited for distinct tasks but with the same architecture: one generates synthetic RFID data, which is then transformed into 3D poses using our transformer-based kinematics neural network; the other component distinctively operates on pose data to generate complete 3D poses from the incomplete poses estimated from RFID data, compensating for RF sensing constraints. By conditioning on partial pose observations and activity labels, our model infers missing joints and provides full-body pose estimates. This approach contributes to wireless sensing and 3D pose estimation by enabling more

accurate and complete pose recovery in scenarios where traditional data collection methods are costly or infeasible.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work that introduces a latent diffusion transformer to not only generate high-quality RFID sensing data that can be transferred to rich 3D poses, but also to leverage cross-attention conditioning to infer missing joints of RFID-based 3D pose estimation. By conditioning on partial poses, the model accurately completes full poses, ensuring coherence between the observed and inferred joints.
- We develop a transformer-based kinematics predictor that maps RFID phase variation data to skeletal poses. Trained on a diverse distribution of real data, it can estimate 3D poses even from synthetic RFID data across different skeletons, ensuring temporal smoothness and high accuracy. Poses can be transferred to unseen skeletons to achieve AI-generated content (AIGC) purposes.
- We also employ a U-Net based frame interpolation method to increase the frame rate of RFID-based 3D human pose estimation, from the presently overly downsampled 7.5 Hz (caused by the RFID sampling rate constraints and synchronization constraints during training data collection) to 30 Hz, resulting in smooth, coherent, and anatomically plausible poses.
- Our framework is particularly valuable in situations where wearing additional RFID tags is impractical or where sensing limitations exist. By providing a method to generate extensive paired datasets and infer complete poses, we significantly enhance the applicability of RFID-based motion capture systems in real-world applications.

The remainder of this paper is structured as follows. We first review related work regarding RF-based 3D pose estimation and pose completion in Section II. Then Section III details the challenges for achieving a coherent full pose using RFID-sensing platforms. Section IV depicts the proposed system design and model training, and Section V delineates the data collection and model implementations. Section VI presents our experimental study and Section VII summarizes this paper along with a discussion of future work.

II. RELATED WORKS

A. DIFFERENTIATING HAR AND POSE ESTIMATION HPE

While conventional HAR focuses on classifying discrete human activities, pose estimation involves predicting 3D joint positions, requiring fine-grained spatial and temporal modeling. Furthermore, our pose completion task reconstructs anatomically plausible full-body poses from partial observations, which is beyond the scope of traditional HAR approaches. Generative AI solutions for HAR primarily target activity classification, whereas our framework uniquely addresses pose estimation and completion challenges.

Traditional data augmentation methods such as time stretching, spectrum shifting, and spectrum scaling have been

used for WiFi-based human activity augmentation [20], but the augmented data could not serve as standalone inputs for training models. These methods require real data for effectiveness, which contradicts our goal of alleviating the need for extensive real-world data collection. One of our primary tasks is 3D pose estimation, meaning there is a high regard for the inherent structure of the RFID data. If some properties are changed, it will seriously alter the final estimated 3D human pose, which needs to follow strict anatomical structures.

B. GENERATIVE AI FOR RF SENSING-BASED POSE DATA GENERATION

Generative AI has been applied in wireless sensing to augment data and improve model robustness, but few works have leveraged the generated wireless data for 3D pose estimation. The pivotal work in time-series data generation TimeGAN [21] is primarily designed for generic time-series data and excels in tasks like financial modeling or healthcare monitoring. However, due to the complexities of RF sensing data, TimeGAN is limited in its capacity to model these multidimensional features. The authors in [22] first utilized an Autoencoder-based generative adversarial network (GAN) to synthesize RFID data that can be used to estimate 3D poses. Reference [16] further enhanced the capabilities of the system and extended the application to WiFi and FMCW radar platforms. Moreover, GAN was used in [23] as a signal refiner to make simulated mmWave signals more realistic. Wang and Mao [24] first utilized diffusion models to generate high-quality RFID data, which can be transformed into more natural and smooth 3D poses compared to Autoencoder-GAN-based approaches. The recent work [25] further leveraged latent diffusion models (U-Net as the backbone) to generate hyper-realistic RF data. The RFID-based kinematics predictor trained on these pose data exhibits performance comparable to that trained on real 3D pose data. Another recent work, RF-Diffusion [26], introduced a time-frequency diffusion theory and a hierarchical diffusion transformer design to generate diverse and high-quality RF data. However, this work did not explore if the diffusion-generated RF data can be further utilized for 3D human pose estimation.

C. RF SENSING-BASED 3D POSE ESTIMATION AND COMPLETION

In recent years, the transformer architecture has shown great promise in handling WiFi CSI-based end-to-end 3D pose estimation for its ability to model complex dependencies [27]. However, such a system requires three pairs of antennas at different locations to achieve good pose estimation and focuses on a one-to-one mapping from WiFi signals to a specific human pose representation. This design can benefit from retargetable poses across varying skeletal models to accommodate the diverse needs in the age of AIGC. WiFi CSI-based approaches such as [10] investigated the problem of predicting missing keypoints, where the missing points were inferred based on

the rich spatial-temporal patterns learned from training WiFi signals. In contrast, our system aims to recover missing joints from an incomplete 3D pose, estimated from noisy RFID signals. The process relies solely on the learned priors about human body dynamics and relationships between completed and partial joints.

Pose completion addresses the challenge of recovering missing or occluded joints in 3D HPE. The authors in [28] first introduced semantic modeling for occluded joint recovery, but limited themselves to depth images and simple occlusion patterns. The work in [29] pioneered single image joint recovery using joint distance relationships, but it relied on traditional geometric constraints and could only recover a few joints. Joint recovery was enabled in video-based continuous 3D HPE, where [30] combined 2D pose estimation from images and temporal information across video frames to infer the missing joints. However, the operation of filtering out unreliable joint predictions and the reliance on analyzing movement patterns over frames for joints inference would not be effective for the problem considered in this paper.

Transformers and diffusion models have a great potential to bridge this gap. Reference [31] demonstrated the capabilities of transformers. The self-attention mechanism allows the transformer to simultaneously model relationships between joints of the human skeleton. Unlike traditional convolutional or recurrent neural networks that rely on localized or sequential processing, transformers can capture long-range dependencies between distant joints. The model can infer the positions of absent joints by considering the relationships with the visible ones. However, 3D poses are time-varying motions with outliers and noises. The ability of transformers depends heavily on the knowledge learned from large training datasets, which can make them susceptible to out-of-domain (OOD) situations. Their generalization capability is limited to the diversity and coverage of the training dataset. Diffusion models, on the other hand, are explicitly designed to model the underlying data distribution by progressively transforming Gaussian noise into data samples. This allows them to generate plausible samples even for poses that are underrepresented or absent in the training data, effectively expanding the range of poses that can be inferred. Diffusion models have been implemented with great success in 3D HPE fields with fidelity, diversity, and strong conditioning methods that can enable various applications. For instance, DiffPose [32] employed a denoising diffusion probabilistic model (DDPM) [33] framework, where the reverse diffusion process was conditioned on 2D keypoints through cross-attention mechanisms. This approach allowed the model to focus on the relevant parts of the input, effectively handling partial observations and producing diverse 3D pose outputs.

The key innovations of our proposed system lie in its generative approach, moving beyond deterministic recovery methods. It is the first to address the challenging task of generating complete 3D poses from partial 3D poses, recovering entirely absent joints, by leveraging latent diffusion

models. Furthermore, our system operates in the latent space, enabling end-to-end training without requiring intermediate supervision, thus ensuring both efficiency and adaptability in pose completion tasks. In the latent domain, our cross-attention mechanism works as an intelligent observer that learns to ‘watch’ how different body parts move together. Just as humans can predict how someone’s ankle will move by watching their knee, our model learns such natural movement relationships. The cross-attention mechanism serves two key purposes: First, it acts like a smart connection system that learns which observed joints (like knees) are most relevant for predicting missing joints (like ankles), rather than treating each joint independently. Second, beyond just position, it understands speed and direction of movement - if an observed knee is moving quickly in one direction, the missing ankle likely follows a similar pattern, helping maintain natural, fluid motion in completed poses.

III. CHALLENGES

In the RFID-Pose system [12], 12 passive RFID tags are strategically attached to specific joints on a subject to track their movements. While the subject performs various poses, three reader antennas continuously interrogate these tags, capturing phase variation measurements. This data reflects changes in the phase between consecutive readings, effectively encoding the relative movement of the joints. The phase variation is given by:

$$\Delta\phi = \text{mod}\left\{\frac{4\pi f_{\alpha}}{c}(D_t - D_{t-1}), 2\pi\right\}, \quad (1)$$

where c is the speed of light, D_t and D_{t-1} denote the tag-to-antenna distance for the t th and $(t-1)$ th sampled data on the same channel α , respectively, and $(D_t - D_{t-1})$ denotes the relative shift in distance between the two consecutive samples. RFID phases are sampled at a frequency of 110 Hz.

While this system successfully captures movements of the tagged joints, only 12 joints are monitored, leaving a significant portion of the skeletal structure unaccounted for. As shown in Fig. 1, compared to Kinect captured full pose with 25 joints, the entire joint chains below knees are missing, alongside small details such as hands and thumbs. The partial pose tracks up to the neck position, but the information regarding the head is entirely missing.

This critical issue is due to the constraints of tag interference, reader capacity, sparse readings, and practical wearability: using more tags increases the likelihood of signal interference, especially when tags are closely spaced, leading to degraded signal quality and increased ambiguity in phase readings. Beyond a certain number of tags, collisions in signal acquisition or limitations in the reader’s range and sampling rate can degrade performance, especially for complex movements or occluded body parts. The collision avoidance protocol adopted by the RFID system ensures that only one tag responds to the reader interrogation in each timeslot. As a result, the sampling rate per tag decreases as more tags are deployed, reducing the chances of collecting

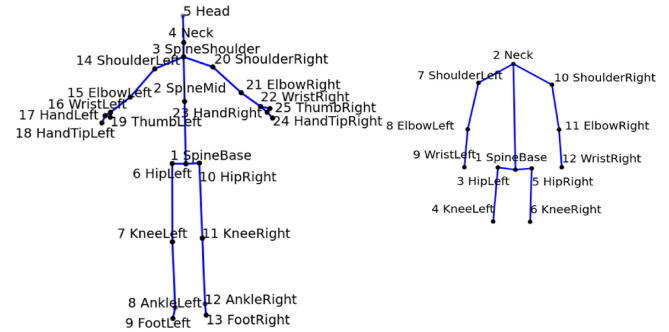


FIGURE 1. Differences between a complete pose consisting of 25 joints (left) and a partial pose of 12 joints (right).

phase data from any single tag. In the current setup with 12 tags, the sampling rate for each tag is already low. For 500 timeslots, only 38 samples (about 7.6% sampling rate) are collected for each tag, leaving more than 90% of data missing. From a usability perspective, affixing more than 12 tags to a human body can become cumbersome and intrusive, limiting the feasibility of deploying such a system in real-world applications.

A follow-up issue to this inherent sparse data problem is that we have to downsample the RFID data to 7.5 Hz to synchronize with the Kinect camera frame rate of 30 Hz. This alignment made it feasible to apply low-rank tensor completion for data imputation. However, this downsampling introduced its own limitation: any pose data estimated from this processed RFID data was restricted to the same 7.5 Hz frame rate. Such a low frame rate resulted in slow and discontinuous pose sequences, which are unsuitable for applications requiring smooth and continuous motions.

WiFi CSI [34] and radar-based [11] systems, although with more comprehensive transceivers, still detect fewer joints (e.g., 17 joints) than their vision counterparts (34 and 25 joints) due to inherent sensing constraints and model limitations. The RF sensing based methods have lower spatial resolution than optical systems, limiting the ability to distinguish closely spaced joints. To create a fully rendered metaverse character, 24 major joints (especially, detailed joints over the neck region) are needed [35]. Wireless-based 3D pose estimation still faces significant challenges in achieving the completeness and reliability required by emerging applications.

Another significant challenge encountered is the limitations of standard diffusion models in generating 3D poses. Training a diffusion model directly on raw 3D pose data poses a severe memory bottleneck, even on hardware such as the RTX A4000 with 16 GB of VRAM. This issue makes it infeasible to process large pose datasets effectively in resource-constrained environments. Generating realistic 3D poses requires more than just minimizing diffusion loss; additional objectives such as temporal smoothness, joint localization loss, and joint rotation loss are essential to ensure anatomical plausibility and motion consistency. Incorporating these pose-specific losses into the diffusion

framework presents further complexity. Specifically, the model needs to sample 1,000 reverse diffusion steps to generate a pose, and this sampled pose is then used to calculate pose-related losses during training, significantly increasing the computational overhead. Switching to a latent diffusion framework resolved the memory issue. However, the latent diffusion model alone cannot generate full poses that faithfully follow the trajectory of the partial poses. A more comprehensive framework that can align the completed pose with the observed partial pose is needed.

IV. SYSTEM DESIGN

In this section, we present the detailed design of our proposed system, which aims to address the challenges of data scarcity and incomplete pose data in RFID-based 3D HPE. As shown in Fig. 2, the framework comprises three main components that operate sequentially: RFID data generation with a latent diffusion transformer, a transformer-based kinematics predictor, and pose completion via cross-attention conditioning. RFID and 3D pose data can both be represented as 2D sequential data with frame-wise motion features $x_{1:N}^L = \{x_n^L\}_{n=1}^N$, where N indicates the number of time frames and L specifies the number of RF features. A detailed problem formulation is presented below to fundamentally clarify the two main tasks of our framework.

A. PROBLEM FORMULATION

This paper addresses two key tasks: (1) *pose augmentation* and (2) *pose completion*, both formulated within the context of latent diffusion models.

1) POSE AUGMENTATION

The objective of pose augmentation is to generate synthetic 3D human pose estimations from RFID data, conditioned on activity labels. This task involves two key stages: (1) generating high-fidelity and temporally smooth synthetic RFID data, and (2) estimating corresponding 3D human poses from this synthetic data using a kinematics predictor. First, the conditional generation of synthetic RFID data is mathematically formulated as modeling the conditional distribution:

$$z_{RFID} \sim p(z | \alpha),$$

where z_{RFID} represents the latent encoding of RFID data, and α is the conditioning information (i.e., activity class labels). The LDT reverses a forward diffusion process, progressively denoising random Gaussian noise into samples from $p(z | c)$, ensuring high-quality and diverse RFID data generation. Next, the generated RFID data z_{RFID} is passed through a kinematics predictor f_{kin} to estimate corresponding 3D poses:

$$[\hat{P}_p = f_{kin}(\psi_{RFID}(z_{RFID})),]$$

where $\hat{P}_p$ represents the estimated 3D human pose of partial joints, and $\psi_{RFID}(\cdot)$ represents the decoder. The fidelity and temporal smoothness of the synthetic RFID data ensure that $\hat{P}_p$ are of high quality, preserving anatomical accuracy and motion dynamics.

2) POSE COMPLETION

Pose completion involves generating a complete 3D pose from partial observations, leveraging the latent representation of partial poses and activity labels. This task is formulated as:

$$z_f \sim p(z | z_p, \alpha), \hat{P}_f = \phi_{Pose}(z_f),$$

where z_p represents the latent encoding of partial pose data, z_f stands for the latent encoding of full pose data, and $\hat{P}_f$ for the decoded full pose data. The cross-attention mechanism in the LDT aligns partial pose trajectories with complete poses.

B. LATENT DIFFUSION TRANSFORMER (LDT)

Diffusion models transform a Gaussian noise distribution into a complex data distribution $p(x)$ through a Markov chain of denoising steps, effectively learning to reverse a forward process $\{x_t\}_{t=1}^T$ that adds noise to the data sample. However, applying diffusion models to RFID or 3D pose data presents significant challenges. They are both characterized by intricate, motion-specific features, and high-frequency outliers, which can hinder the model's ability to accurately learn the underlying data distribution, especially when dealing with various activity classes.

We perform the diffusion process in a representative and low-dimensional latent space learned by our transformer-based variational autoencoder (VAE). This approach reduces computational requirements and improves the quality of the generated data by focusing on essential structural features and filtering out high-frequency noises. Furthermore, we employ a transformer-based denoiser $\epsilon_\theta(x_t, t)$ in the diffusion model. Transformers are well-suited for modeling sequential data and capturing long-range dependencies due to their self-attention mechanisms. The transformer's ability to model relationships across time steps enhances the denoising process in the latent space, leading to more realistic data generation. Our system can effectively generate high-quality RFID data and infer complete 3D poses, while maintaining computational efficiency. This approach builds upon the LDT framework [36], adapting it for RFID data generation and pose completion. We next elaborate on the design of the three sub-modules: Latent Compressor, Conditional Latent Diffusion Process, and Transformer Denoiser.

1) LATENT COMPRESSOR

RF signals, highly responsive to nearby movements, exhibit distinct cyclical patterns that correlate with various human activities. Similarly, 3D human poses have complex temporal dependencies due to the continuity and flow of human motion. These dependencies make it challenging to model them directly, but they are well suited for a transformer architecture, which excels at capturing sequential relationships and learning from time-dependent data. The Variational Autoencoder (VAE) is a pivotal component of our system, designed to learn meaningful latent representations

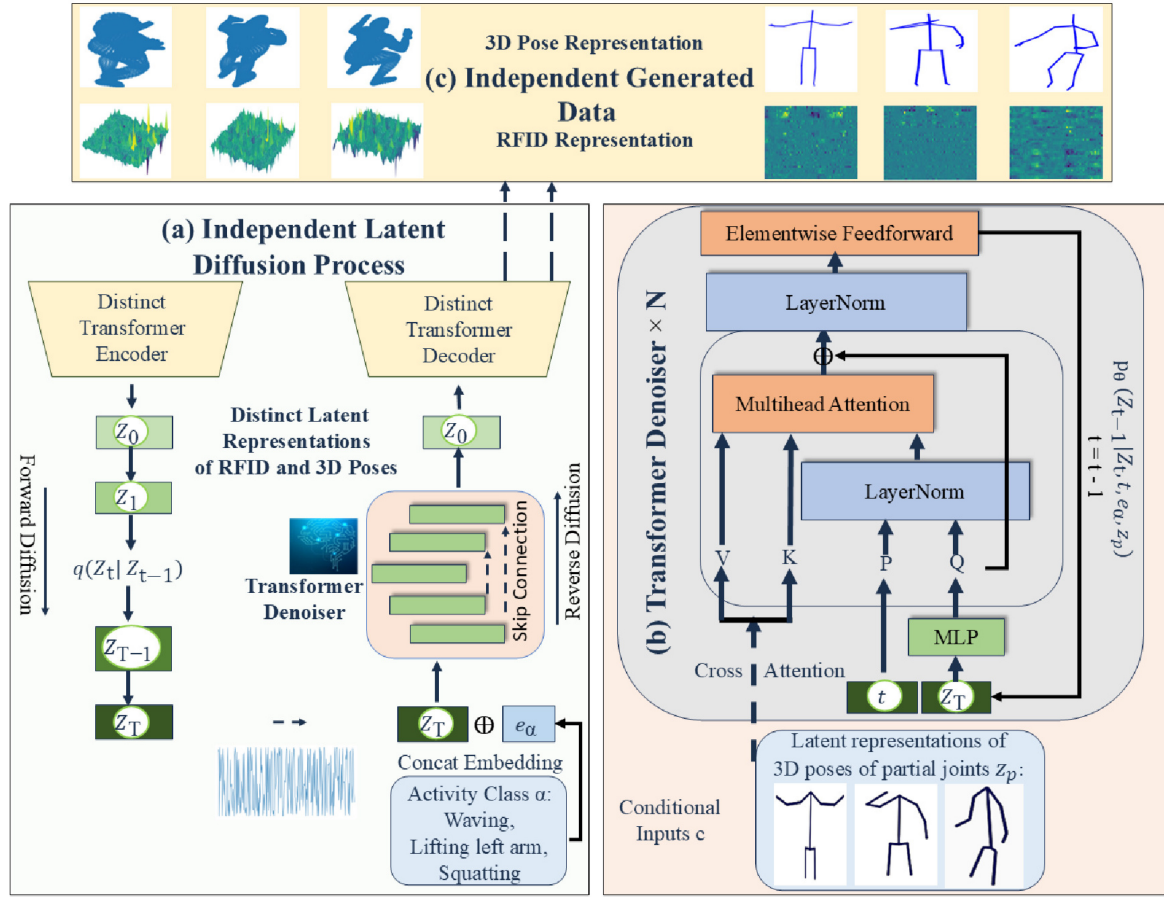


FIGURE 2. The diagram illustrates the process of conditional RFID and 3D pose data generation via latent diffusion transformers: (a) a latent diffusion process after compressing RFID or pose data into its respective latent representations; (b) the detailed structure of the transformer denoiser and cross-attention conditioning method; (c) a demonstration of the generated RFID and completed 3D pose samples after their generative latent representations get decoded back to the data domain. The diagram also depicts the forward and reverse latent diffusion processes.

of RFID or pose data and reconstruct the data from these representations. By treating latent representations as tokens, we enable the transformer to effectively model the sequential and spatial relationships inherent in the data.

We construct our transformer-based VAE, denoted by V , comprising a transformer encoder ϵ and a transformer decoder ψ . The VAE is trained to encode input sequences into a latent $z = \epsilon(x_{1:N}^L)$ and then decode from this latent back to the original data domain. We employ two distinct VAEs, denoted by V_{RFID} and V_{Pose} , each with their own encoder and decoders, to encode RFID and pose data into separate latent spaces. Specifically, the transformer encoder ϵ takes time-series motion features $x_{1:N}^L$ (can be of arbitrary length N), and trainable distribution tokens as input. The encoder consists of a series of transformer layers that employ self-attention mechanisms to capture long-range dependencies within the data. Positional encoding is leveraged to retain temporal order. The outputs corresponding to the latent distribution tokens are extracted and projected to obtain Gaussian distributions μ and σ . We can then use the parameterization trick to sample a latent vector $z \in R^{M \times D}$, where M is the latent size and D is the latent dimension: $z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. This allows gradients to flow

through μ and σ during backpropagation. The decoder ψ employs the cross attention mechanism on top of the similar architecture of the transformer. We initialize N zero vectors as queries and utilize the latent as memory to generate an RFID or pose sequence $\tilde{x}_{1:N}^L$ with their respective decoders. We incorporate skip connections between the encoder and decoder to help preserve low-level information and gradients, improving the reconstruction quality and training stability of the VAE. The training process for the RFID case is typical of a VAE and includes the Mean Squared Error (MSE) loss between $x_{1:N}^L$ and $\tilde{x}_{1:N}^L$ and the Kullback-Leibler (KL) loss. However, for 3D pose data, we additionally employ a temporal smoothness loss to ensure continuous and natural motion patterns. This smoothness loss $\mathcal{L}_{smooth}$ computes the difference between consecutive frame velocities in both reconstructed and ground truth sequences, as

$$\mathcal{L}_{smooth} = \sum_{n=1}^{N-1} \|\Delta x_n^L - \Delta \tilde{x}_n^L\|_2^2, \quad (2)$$

where $\Delta x_n^L = x_{n+1}^L - x_n^L$ is the ground truth velocity, and $\Delta \tilde{x}_n^L = \tilde{x}_{n+1}^L - \tilde{x}_n^L$ is the reconstructed velocity. This velocity-matching objective encourages the VAE to learn a latent

space that preserves activity-specific temporal dynamics, and prevents collapse to mean pose.

2) CONDITIONAL LATENT DIFFUSION WITH CROSS ATTENTION

Distinct from the previous U-Net-based architecture on the RFID and pose data (or their latents), a transformer denoising model with cross attention conditioning is built on the latent representations $z \in \mathbb{R}^{1 \times 256}$. In [37], we reshaped the latent space into a two-dimensional format of 16×16 to accommodate the input requirements of the U-Net, while sacrificing the inherent sequential structure of the latents encoded by the RNN-based VAE.

The LDT operates within the separate latent spaces encoded by $\mathbf{e}_{RFID}$ and $\mathbf{e}_{Pose}$, to generate synthetic RFID data and complete 3D poses. This modular design avoids interference between the two tasks, enabling task-specific fidelity. By utilizing distinct VAEs for RFID and pose data, the system achieves high-quality augmentation and pose completion tailored to the nuances of each domain. The forward diffusion process progressively transforms the clean latent vector $z_0 = x_{1:N}^L$ generated by the latent compressor into a noisy latent z_t over T time steps, effectively converting the data distribution into an isotropic Gaussian. This process is defined by a Markov chain, as

$$q(z_{1:T}|z_0) = \prod_{t=1}^T q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t}z_{t-1}, 1 - \alpha_t\mathbf{I}), \quad (3)$$

where $\alpha_t = \prod_{i=1}^t (1 - \beta_i) \in (0, 1)$, and β_t denotes the variance schedule hyper-parameter. Here z_t can be sampled via $z_t = \sqrt{\alpha_t}z_0 + \sqrt{1 - \alpha_t} \cdot \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. We denote the noising sequence as $\{z_t\}_{t=1}^T$, and the step t denoising as $z_{t-1} = \epsilon_\theta(z_t, t)$.

The reverse diffusion process starts from z_T , and iteratively denoises it to recover z_0 . This is also the generation process of the standard diffusion model. Our RFID-LDT is capable of conditional human activity generation, from simple (e.g., drinking water with one hand) to complex (e.g., boxing). This is done by modeling the conditional distribution $p(z|a)$, where a represents activity labels. Our Pose-LDT implements a multi-modal conditioning mechanism: activity label conditioning through embedded concatenation and partial pose conditioning through cross-attention. This is done by modeling $p(z|a, z_p)$, where z_p represents partial pose latents. The choice of different conditioning methods is motivated by the nature of the data being fused. For activity labels, which are discrete class information, we use embedded concatenation as it is an effective way to incorporate global activity context into the generation process. We employ a condition encoder that maps an activity label α to learnable embeddings e_α . In contrast, for partial pose conditioning, we leverage the VAE-encoded partial pose latents $z_p \in \mathbb{R}^{1 \times 256}$ as the conditioning information for the cross-attention mechanism, which allows our model to dynamically learn which observed joints are most relevant

for completing each missing joint, capturing both spatial and temporal dependencies, in the latent space without the complications of learning intricate structural relationships in the 3D pose space.

The conditional reverse diffusion process can be formulated as

$$p_\theta(z_{t-1}|z_t, e_\alpha, z_p) = \quad (4)$$

$$\mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, e_\alpha, \text{CrossAttn}(z_t, z_p)), \sigma^2 t^2 \mathbf{I}),$$

$$\mu_\theta(z_t, t, e_\alpha, z_p) = \quad (5)$$

$$\frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} (\epsilon - \epsilon_\theta(z_t, t, e_\alpha, \text{CrossAttn}(z_t, z_p))) \right),$$

where μ_θ is the mean of the modeled Gaussian distribution and is key to estimating the denoised latent token at timestep $t - 1$ using both the activity condition e_α and the cross-attention features from partial pose latents z_p . $\tilde{\alpha}_t = \prod_{s=1}^T \alpha_s$ quantifies the total amount of signal retained after t steps in the forward diffusion.

The cross-attention mechanism is defined as:

$$\text{CrossAttn}(z_t, z_p) = \text{softmax} \left(\frac{Q_t K_p^T}{\sqrt{d}} \right) V_p, \quad (6)$$

$$Q_t = z_t W_Q^{\text{CrossAttn}}, K_p = z_p W_K^{\text{CrossAttn}}, \\ V_p = z_p W_V^{\text{CrossAttn}}, \quad (7)$$

where z_t is the noised latent of the full pose at timestep t , and z_p is the partial pose latent. The cross-attention operation projects these latents into query (Q_t), key (K_p), and value (V_p) spaces through learnable projection matrices $W_Q^{\text{CrossAttn}}$, $W_K^{\text{CrossAttn}}$, and $W_V^{\text{CrossAttn}} \in \mathbb{R}^{256 \times 256}$. The dimension of the latent space $d = 256$ is the scaling factor. The attention mechanism computes compatibility scores between the query (derived from noised latents of the full pose) and keys (derived from latents of the partial pose), normalized through softmax. These scores then weigh the values to produce context-aware features. This allows the denoiser to selectively attend to relevant information in the partial pose latent while generating the full pose.

The conditional training objective is as follows:

$$\mathcal{L}_{\text{Pose-LDT}}(\theta) \\ = \mathbb{E}_{\epsilon, t, e_\alpha, z_p} \left\| (\epsilon - \epsilon_\theta(z_t, t, e_\alpha, \text{CrossAttn}(z_t, z_p))) \right\|_2^2. \quad (8)$$

This training objective minimizes the mean squared error (MSE) between the predicted noise and the true noise added during the forward diffusion process. For pose augmentation, minimizing this loss ensures that the generated z_{RFID} samples follow the distribution $p(z|e_\alpha)$, enabling diverse and realistic synthetic RFID data generation, which further allows accurate downstream pose estimation. For pose completion, minimizing the noise prediction error allows the model to infer missing pose information while maintaining anatomical and temporal consistency. This unified formulation highlights the adaptability of the latent diffusion framework to address both tasks effectively.

To enhance the class conditioned generation, we also equip our denoiser ϵ_θ with classifier-free guidance [38] using a guidance scale s . It learns both the conditioned and unconditioned distribution with a 10% dropout of the samples, and a linear combination is operated.

$$\epsilon_\theta^{\text{guided}}(z_t, t, e_\alpha) = \epsilon(z_t, t) + s(\epsilon(z_t, t, e_\alpha) - \epsilon(z_t, t)), \quad (9)$$

where $s > 1$ can strengthen the effect of guidance. We use the pretrained and frozen latent compressor V to transform the RF or pose data into latent tokens across various activity classes before initiating the diffusion process, which can offload some computational burdens. After conditional reverse diffusion, the decoder ϕ efficiently reconstructs the activity data from the predicted $\tilde{z}_0$.

3) TRANSFORMER-BASED DENOISER

To construct a conditional denoising network $\epsilon_\theta(z_t, t, e_\alpha, z_p)$ that can effectively model long-range dependencies in the latent domain, our denoising process employs a transformer encoder-only architecture with long skip connections, comprising 11 layers. Diverging from traditional encoder-decoder structures, our model processes a concatenated sequence of the noised latent token and condition embeddings directly through self-attention mechanisms to extract the final denoised latent for noise prediction. The input to the transformer encoder is formulated as $\tilde{z}_t = [z_t; e_t; e_a \in \mathbb{R}^{3 \times 256}]$, where e_t is the time step embedding and e_a is the activity embedding. This way of input processing is more computationally efficient than the encoder-decoder structure. The self-attention mechanism allows for direct interaction among the noised latent features, temporal information, and conditional embeddings within the same layer.

At each layer l of the encoder, the self-attention mechanism operates on $h^{(l-1)}$, i.e., the output from the previous layer (with $h^0 = \tilde{z}_t$), as

$$\text{SelfAttn}(h_t^{(l-1)}) = \text{softmax}\left(\frac{Q_l(K_l)^T}{\sqrt{d}}\right)V_l, \quad (10)$$

$$Q_l = h_t^{(l-1)}W_{Q,l}^{\text{SelfAttn}}, K_l = h_t^{(l-1)}W_{K,l}^{\text{SelfAttn}}, \\ V_l = h_t^{(l-1)}W_{V,l}^{\text{SelfAttn}}, \quad (11)$$

where W_Q^{SelfAttn} , W_K^{SelfAttn} , and $W_V^{\text{SelfAttn}} \in \mathbb{R}^{256 \times 256}$ are learned projection matrices of layer l . To facilitate stable training and effective gradient flow, each encoder layer incorporates residual connections and layer normalization $h_t^{(l)} = \text{LayerNorm}(h_t^{(l-1)} + \text{SelfAttn}(h_t^{(l-1)}))$. This long skip connection aids in mitigating vanishing gradient issues and allows the model to learn identity mappings, ensuring that the original latent information is preserved across layers. In the end, an element-wise feed-forward network (FFN) is included, as $h_t^{(l)} = \text{LayerNorm}(h_t^{(l)} + \text{FFN}(h_t^{(l)}))$, which consists of two linear transformations with a ReLU activation in between. After processing through N encoder layers, the

added noise $\epsilon_\theta(z_t, t, e_\alpha, z_p) = h_t^{(N)}$ is predicted. An MSE-based training minimizes the difference between predicted and actual noise. The implementation of our transformer denoiser is shown in Fig. 2. The above formulas are inspired by [39].

C. TRANSFORMER-BASED KINEMATICS PREDICTOR (RF-FORMER)

Building upon the foundation established in [12], 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. The detailed signal processing, including obtaining RFID phase variation, synchronizing RFID and Kinect camera data, low-rank tensor completion for data imputation, and forward kinematics remain as the original setup in [12]. Unlike traditional sequential models such as Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) that process data step by step, transformers can attend to all temporal positions simultaneously. This characteristic makes transformers particularly adept at capturing long-range dependencies in RFID signal sequences. Our transformer-based approach excels in learning complex spatiotemporal correlations between RF signal patterns and joint movements. By processing parallel information from multiple RFID tags concurrently, the model effectively understands the hierarchical relationships inherent in joint kinematics. These capabilities collectively contribute to a more accurate and robust RFID-based HPE system.

As shown in Fig. 3, our system consists of a transformer encoder-decoder structure. The input training data includes the preprocessed RFID phase variation sequence, the 3D pose data sequence, and a target skeleton. The encoder stack processes RFID signals through multi-head self-attention to learn temporal correlations in RF signals and the relationships between different tag positions. A regular sinusoidal-based positional encoding is used to inject sequence order information into the input embedding. The decoder stack, consisting of six layers, transforms the encoded RF features into joint rotations. It utilizes masked multi-head self-attention to maintain temporal consistency established in the output sequences. The multi-head cross-attention mechanism is pivotal, mapping RF features to the joint rotation space. Encoded RF features serve as keys and values, which allows the decoder to focus on relevant parts of the encoded signals, and establish direct mappings between phase variations and specific joint rotations. Our system significantly enhances the capability to interpret RFID signals for accurate 3D HPE. This approach addresses the limitations of previous sequential models and paves the way for more advanced applications in RFID-based sensing.

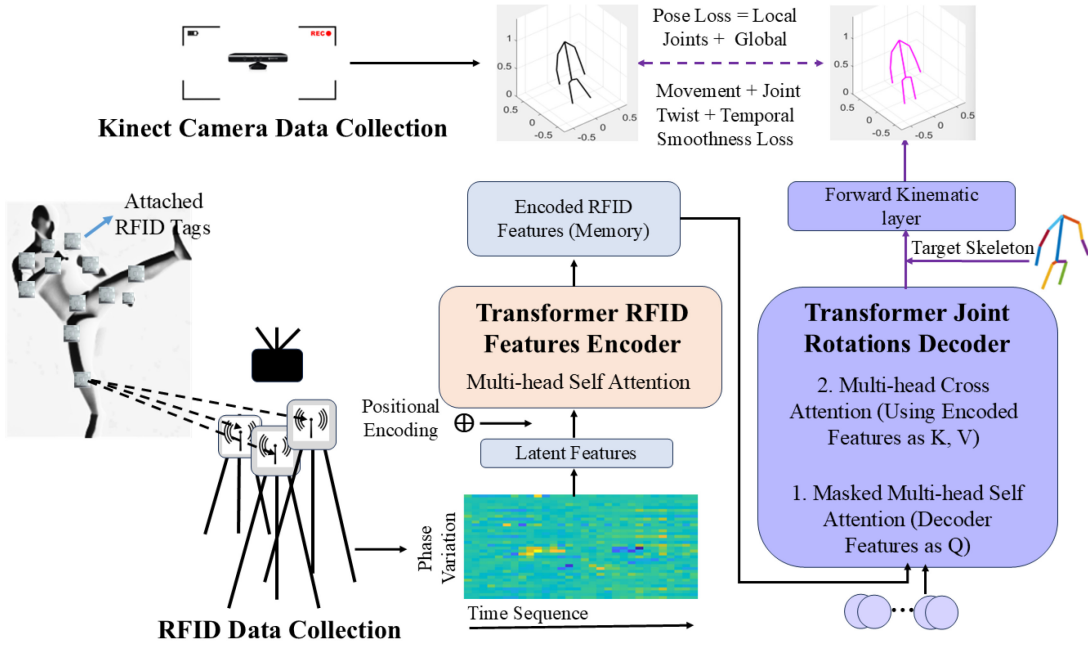


FIGURE 3. Illustrate the design of the transformer-based kinematics neural network, termed RF-Former.

D. MOTION-ALIGNED TWO-STAGE 3D POSE COMPLETION

Even with the implementation of cross-attention during training, a single reverse diffusion process does not fully guarantee 3D pose completion that accurately follows the trajectories of the conditioned partial pose, despite demonstrating satisfactory performance with regard to ground truth (GT) partial and full poses. Our application necessitates the ability to generate complete full poses based on observed partial poses, ensuring reliable performance even in occluded sensing scenarios. Hence we propose a two-stage generation process that leverages attention maps for temporal alignment without the need for model retraining.

Drawing inspiration from [40], we first decide to utilize cross-attention mechanisms to inject more sophisticated conditions beyond activity class labels. The authors in [41] implemented their cross-attention mechanism for different motion representations and regions of the human body, but mainly based on text-motion attention. Instead, we capture cross-attention patterns during initial generation, storing attention maps $\alpha_t = \text{softmax}(\frac{Q_t K_p^T}{\sqrt{d}})$ that encode the relationships between partial and full pose representations in the latent space. Each element $\alpha_t[i, j]$ represents the attention weight between the i th token of the full pose and j th token of the partial pose. Higher attention weights indicate stronger relationships between corresponding joints. These attention maps then guide a motion-aligned refinement process. The refinement stage employs velocity-based alignment, where velocities from partial poses guide the completion through adaptive attention-weighted interpolation: $v_{\text{aligned}} = (1 - w_t)v_t + w_tv_p$, with w_t derived from stored attention maps. This approach builds upon observation-conditioned motion

prediction [42] and velocity-based motion synthesis [43]. The latent velocity weighting scheme works as follows: a base weight (0.3) is selected to ensure minimum partial pose influence, and then an adaptive weighting based on attention strength ensures natural transitions with higher weights (≈ 0.5) for joints strongly correlated with partial pose, and lower weights (≈ 0.3) for joints that need more freedom.

Our two-stage framework is further validated by the success of progressive refinement in conditional generation [44], [45]. Using different attention guidance scales (γ_1, γ_2) and class conditioning scales (λ_1, λ_2) for the two stages is beneficial because Stage 1 of initial generation process needs stronger guidance ($\gamma_1 > \gamma_2$) to establish clear action-specific features and a higher conditioning scale (λ_1) helps establish the basic pose structure. Typically we set $\gamma_1 \approx 10$ and $\lambda_1 \approx 20$. After that, Stage 2 of refinement process can use weaker guidance (γ_2) as basic activity features are established. A lower conditioning scale (λ_2) allows more flexibility for motion alignment. Typically we set $\gamma_2 \approx 7.5$ and $\lambda_2 \approx 12.5$. For unseen scenarios, we increase λ_1 to 35 and λ_2 to 22.5 to allow stronger reliance on conditional partial pose. A higher guidance scale can cause the movements to be stiff, and lower guidance could confuse the class the model tries to generate. Similarly, a high conditioning scale can result in unnatural motion. It takes careful tuning to reach an optimal performance for different scenarios. Notably, our method operates purely in the latent space of independently trained VAEs, differing from conventional latent diffusion models that typically require joint training of the VAE and diffusion model. This design choice enables more flexible and modular pose completion while maintaining temporal coherence through

Algorithm 1 Two-Stage Motion-Aligned Latent Generation

Input: Action embedding e_α , partial pose latents z_p , guidance scales γ_1, γ_2 , conditioning scales λ_1, λ_2 , Stage-1 steps N_1 , Stage-2 steps N_2 ;

```

1: Stage 1: Initial Generation with Attention Capture
2:  $z_T \sim \mathcal{N}(0, \mathbf{I})$ ;  $\triangleright$  Initialize noise
3: for  $t = N_1, \dots, 1$  do
4:    $Q_t = z_t W_Q^{CrossAttn}$ ;  $\triangleright$  Cross-attention computation
5:    $K_p = z_p W_K^{CrossAttn}$ ;
6:    $\alpha_t = \text{softmax}(Q_t K_p^T / \sqrt{d})$ ;  $\triangleright$  Store attention maps
7:   if do_classifier_free_guidance then
8:      $\epsilon_{\text{uncond}} = \epsilon_\theta(z_t, t, \emptyset)$ ;  $\triangleright$  Only action-conditional
9:      $\epsilon_{\text{cond}} = \epsilon_\theta(z_t, t, e_\alpha)$ ;
10:     $\epsilon_{\text{pred}} = \epsilon_{\text{uncond}} + \gamma_1(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$ ;
11:   end if
12:    $z_{t-1} = \mu_\theta(z_t, t, \epsilon_{\text{pred}}, \lambda_1) + \sigma_t \epsilon$ ;  $\triangleright$  Denoising step
13: end for
14:  $z_{\text{init}} = z_0$ ;  $\triangleright$  Initial completion
15: Stage 2: Motion-Aligned Refinement
16:  $z = z_{\text{init}}$ ;
17:  $v_p = z_p[:, 1:] - z_p[:, : - 1]$ ;  $\triangleright$  Compute velocities:
   frame-to-frame differences
18: for  $t = N_2, \dots, 1$  do
19:    $v_t = z[:, 1:] - z[:, : - 1]$ ;  $\triangleright$  Current velocities
20:    $w_t = \beta_1 + \beta_2 * \text{mean}(\text{attention\_maps}[t])$ ;  $\triangleright$ 
    $\beta_1 = 0.3, \beta_2 = 0.2$ , selected based on empirical results,  $\beta_2$ 
   dictates more the velocity matching
21:    $v_{\text{aligned}} = (1 - w_t)v_t + w_t * v_p$ ;  $\triangleright$  Latent velocity alignment
22:    $z_{\text{noisy}} = \text{add\_noise}(z, t)$ ;  $\triangleright$  Add noise for refinement
23:    $\epsilon_{\text{pred}} = \epsilon_\theta(z_{\text{noisy}}, t, e_\alpha)$ ;  $\triangleright$  Predict noise
24:   if do_classifier_free_guidance then
25:      $\epsilon_{\text{uncond}}, \epsilon_{\text{cond}} = \epsilon_{\text{pred}}.\text{chunk}(2)$ ;
26:      $\epsilon_{\text{pred}} = \epsilon_{\text{uncond}} + \gamma_2(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$ ;
27:   end if
28:    $z = \mu_\theta(z_{\text{noisy}}, t, \epsilon_{\text{pred}}, \lambda_2) + \sigma_t \epsilon$ ;  $\triangleright$  Denoising step
29:    $z[:, 1:] = z[:, : - 1] + v_{\text{aligned}}$ ;  $\triangleright$  Update positions: Apply
   aligned velocities
30: end for
31: return  $z$ ;  $\triangleright$  Final motion-aligned latents

```

attention-guided motion alignment. Algorithm 1 explains the details for our 3D pose completion process.

V. SYSTEM IMPLEMENTATION

A. PROTOTYPE SYSTEM AND DATASET COLLECTION

Our RFID sensing setup includes (i) an off-the-shelf ImpinJ R420 reader equipped with three S9028PCR polarized antennas; (ii) passive ALN-9634 (HIGG-3) RFID tags; (iii) a ROG laptop equipped with an RTX 3070 GPU (8GB VRAM), and (iv) a Lenovo desktop with an RTX A4000 used for signal processing, model training, and inference.

Nine distinct activities are performed by seven different test subjects positioned within the detection range of the RFID sensing platform and a Kinect 2.0 device. The activities include: raising the right arm and drinking, waving up and down, jabbing, standing still, body twisting, walking, squatting, kicking, and posing. The full range of these activities is repeated continuously to facilitate consistent data sampling. There are 24 paired RF and 3D pose datasets of 72 time frames per class collected to facilitate the training

of the RF-Former system and 6 for testing purposes. The samples are transformed into 30 frames using a sliding window with a 10-frame size. This helps conveniently create more training samples with temporal diversity. Seven test subjects are involved. The first five subjects include 4 males of similar body shape and 1 female. The last two subjects are relatively taller, bulkier, or shorter, thinner than the first 4 males, respectively.

B. MODEL IMPLEMENTATION

Our transformer-based kinematics neural network employs a standard encoder-decoder architecture with 6 encoder and 6 decoder layers, each utilizing 8 attention heads. The feature dimension is set to 256, with the feed-forward network in each transformer layer expanding to 1,024 dimensions. The input dimension is configured to handle RFID signals ($3 \times \text{number of joints} + 4$ features, accounting for local positions and global offset), while the output dimension is set to $4 \times (\text{number of joints} + 1)$ to generate quaternion rotations and global offset [12]. We maintain a dropout rate of 0.1 throughout the network for regularization. The model is trained using the Adam optimizer with a learning rate of $1e-4$ and a batch size of 32. The loss function combines local position loss, global offset loss, twist loss, and smoothness loss. Training is conducted for 10,000 epochs with early stopping to prevent overfitting. Along with 3D pose generation and completion, the VAE encoder and decoder are implemented as transformers with 8 attention heads and 11 layers, with a latent dimension of (1, 256). We use learned positional encoding and GELU activation functions throughout the network. The training takes around 13,300 epochs with a batch size of 8.

For the diffusion process, we employ a transformer-based denoiser with cross-attention-enabled encoder layers. The key parameters are shown in Table 1. The implementation leverages PyTorch Lightning for distributed training. The hyperparameters for the demo stage were determined through extensive ablation studies (as shown in Table 4), where we evaluated different inference steps (1000/800 vs 80/60 vs 50/30) and their impact on generation quality and efficiency. The two-stage approach with 80/60 steps achieved the best balance between quality (11.74cm joint error) and speed (1.59s). Other key parameters like the latent dimension (256), number of attention heads (6), and transformer layers (10) were chosen based on both computational constraints and model performance.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

Latent diffusion models using U-Net as denoiser have been systematically analyzed and evaluated for generating RFID data, using metrics such as SSIM, FID, diversity score, and class recognition accuracy [25]. In this section, we employ a fitting model for estimating 3D poses and we measure its performance comprehensively across key metrics in the field of 3D HPE. This includes metrics for structural accuracy, temporal consistency, and applicability to AIGC-empowered

TABLE 1. Diffusion implementation details and hyperparameters.

Model Architecture Parameters	
Latent Dimension	256
Attention Heads	8
Transformer Encoder Layers	11
Feed-Forward Dimension	1024
Dropout Rate	0.2
Diffusion Process	
Noise Schedule	Linear β from 1×10^{-4} to 0.02
Number of Diffusion Steps	1000 (training), 80/60 (inference)
Guidance Scale	7.5
Unconditional Dropout (for Classifier-Free Guidance)	0.1
Training Parameters	
Batch Size	8
Learning Rate	1×10^{-4}
Optimizer	AdamW
Weight Decay	0.01
Gradient Clipping	maximum norm of 5
Training Epochs	13899
Mixed Precision Training	16-bit
Batch Gradient Accumulation	2 steps

downstream tasks, aiming to unlock the full potential of latent diffusion-generated RFID data.

A. TRANSFORMER-BASED 3D HPE FROM REAL DATA

Before evaluating the advantages of 3D pose estimations from LDT generated RFID data, we first measure the effectiveness of our transformer-based kinematics neural network. For comparative analysis, the RNN-based RFID-Pose method [12] and its subject-adaptive enhancement Cycle-pose [14] were utilized as baselines, which train their networks using paired RFID and vision data. Our network and the baselines were trained and tested on identical datasets to ensure a fair comparison. The MPJPE (mean per joint position error) is computed as the discrepancy between the estimated 3D pose data and the ground truth vision data, given by

$$\text{MPJPE} = \frac{1}{N_f N_j} \sum_{f=1}^{N_f} \sum_{j=1}^{N_j} \|\mathbf{p}_{f,j} - \hat{\mathbf{p}}_{f,j}\|_2, \quad (12)$$

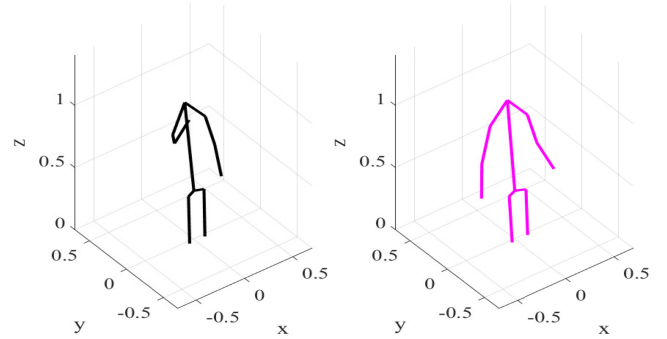
where N_f is the number of frames, N_j is the number of joints, $\mathbf{p}_{f,j}$ and $\hat{\mathbf{p}}_{f,j}$ are the ground truth and predicted 3D positions of joint j in frame f , respectively. Since our goal is to estimate multiple 3D poses over various skeletons for pose augmentation, we conduct experiments with the first five subjects to examine the impact of different target skeletons. The MPJPE results are presented in Table 2.

We find that not only does our network achieve better mean estimation errors without the instability issues

TABLE 2. Estimation errors of different subjects under different kinematics neural networks¹.

Subject Index	Estimation Error RFID-Pose (cm)	Estimation Error Cycle-Pose (cm)	Estimation Error Proposed (cm)
Subject 1	3.75	4.12	3.34
Subject 2	4.55	4.43	3.47
Subject 3	3.58	3.79	3.05
Subject 4	5.32	4.51	4.91
Subject 5	8.17	4.97	5.65

¹ Note: Subjects 1-3 are trained; Subjects 4-5 are untrained.

**FIGURE 4.** 3D HPE comparison results between ground truth (left) and RNN baseline RFID-Pose (right) when the subject is raising arms over the shoulder.

commonly associated with training a Cycle GAN, but it also delivers significantly improved outlier-free estimation. Notably, the maximum joint error is less than 50 cm, highlighting its robustness. Use the results from untrained Subject 5 as an example. As seen in Fig. 4, the estimated pose by the RNN baseline RFID-Pose tends to struggle with raising arms over shoulder, causing large deviations of arm movements. This happens in trained subject scenarios as well, possibly due to the nature of drastic vertical movements and RNN limitations, which tend to regress toward mean poses.

Fig. 5 demonstrates the 3D pose estimation performance more comprehensively, while mean estimation errors alone cannot measure the temporal-wise joint changes over time. For the example activity, the subject mostly moves his/her left arm upward and downward. The right figure in Fig. 5 shows that the RNN baseline method introduces many discrepancies in the left arm region, with drastically changing positions (60 cm offsets) over time. It also has relatively large deviations in the right wrist. However, the left figure in Fig. 5 shows that our transformer counterpart has much smaller discrepancies over time, and the discontinuity only exists briefly.

We hence introduce a new metric called *frame-wise joint position smoothness*. It measures how smoothly the poses transition between consecutive frames by calculating the Euclidean distance of joint positions between adjacent

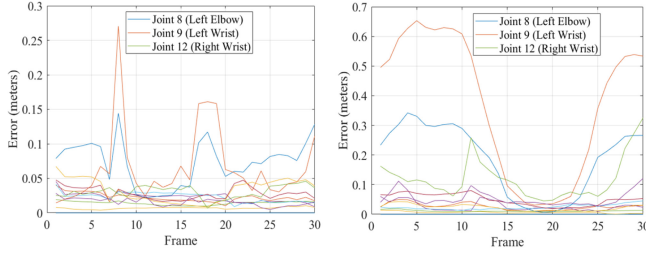


FIGURE 5. A comparison between our transformer-based kinematics neural network (left) and RNN baseline RFID-Pose (right) regarding joint errors over time frames.

frames. This metric is sensitive to large movement discrepancies.

$$\text{Smoothness}_{\text{frame}} = \frac{1}{N_f - 1} \sum_{f=1}^{N_f-1} \|\mathbf{p}_{f+1} - \mathbf{p}_f\|_2, \quad (13)$$

where $\mathbf{p}_f$ is the pose in frame f and N_f is the number of frames. The left example in Fig. 5 has a smoothness index of 1.42 cm/frame, while the right example in Fig. 5 has a smoothness index of 1.9cm/frame. For the overall Subject 5 testing dataset, our method achieves a smoothness index of 2.59 cm/frame, the baseline has a smoothness index of 2.74 cm/frame, while the ground truth has one of 2.49cm/frame.

We also measure the time it takes to initiate models and estimate one complete 3D pose. The RNN baseline RFID-Pose requires 36 seconds and nearly 15 seconds just to estimate one output. However, our transformer-based model only takes 0.6 second to initiate and load the model, and 0.08 second to estimate one output. This validates the computational efficiency and real-time deployment ability of our model.

B. TRANSFORMER-BASED 3D HPE FROM GENERATED DATA

We next assess the estimated 3D poses from our LDT-generated RFID data with a comprehensive series of metrics to ensure the quality of the synthetic dataset for pose augmentation and AIGC tasks [46]. In addition to MPJPE and smoothness score, we incorporate the following metrics:

$$\text{Bone Length Consistency} = \frac{1}{N_f N_b} \sum_{f=1}^{N_f} \sum_{b=1}^{N_b} |\ell_{f,b} - \hat{\ell}_{f,b}|, \quad (14)$$

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.

$$\text{Joint Angle Error} = \frac{1}{N_f N_a} \sum_{f=1}^{N_f} \sum_{a=1}^{N_a} |\theta_{f,a} - \hat{\theta}_{f,a}|, \quad (15)$$

where 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

TABLE 3. Evaluation metrics for estimated 3D human poses using LDT generated RFID data.

Metrics	LDT	Metrics	LDT
Average joint error (cm)	8.99	FID	1.42
Bone consistency (cm)	2.25	GT FID	0.73
Joint angle error (°)	6.91	Diversity	10.98
Smoothness (cm/frame)	1.51	GT Diversity	10.35
GT Smoothness (cm/frame)	1.40		

frame f , respectively. The Frechet Inception Distance (FID) is used to assess how close the distributions of generated and actual pose data are, given by

$$\text{FID} = \|\mu_\phi(\mathbf{X}_r) - \mu_\phi(\mathbf{X}_g)\|_2^2 + \text{Tr}\left(\Sigma_\phi(\mathbf{X}_r) + \Sigma_\phi(\mathbf{X}_g) - 2\sqrt{\Sigma_\phi(\mathbf{X}_r)\Sigma_\phi(\mathbf{X}_g)}\right), \quad (16)$$

where $\phi(\cdot)$ is the feature extractor, which is a custom pose discriminator we train on the Kinect-based 3D pose data, $\mathbf{X}_r$ and $\mathbf{X}_g$ are the real and generated pose sets respectively, $\mu_\phi(\cdot)$ represents the mean of extracted features, and $\Sigma_\phi(\cdot)$ denotes their covariance matrices.

$$\text{Diversity} = \mathbb{E}_{\mathbf{x}_1, \mathbf{x}_2 \sim \mathbf{X}_g} \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|_2^2, \quad (17)$$

where $\phi(\cdot)$ is the same feature extractor as used in FID, and $\mathbf{x}_1, \mathbf{x}_2$ are randomly sampled pairs of poses.

We refer to estimated pose data from LDT generated RFID data as synthetic poses. We measure the scores on synthetic poses by repeatedly estimating 3D poses of various subjects from rounds of generated RFID data and comparing them to ground truth 3D poses of these same subjects. The results are reported in Table 3. The low joint angle error, temporal smoothness, and bone length consistency clearly demonstrate that our synthetic poses are anatomically realistic motions like the ground truth. The average joint error of 8.99 cm, combined with a smoothness score close to ground truth, is highly competitive. This shows that the natural motion features are not random but resemble the ground truth poses without deviation from the original trajectory too much. The scores directly demonstrate that our LDT generated RFID data learns fine-grained movement features embedded in RFID signals, which are then seamlessly transformed to 3D human poses. Some of these skeletons are largely different unseen skeletons during our RF-Former's training. This shows the generalization abilities of the generated RFID data for RF-Former and proves that our estimated 3D poses can be extended in various ways for AIGC-empowered wireless sensing purposes.

To test the generalization of our proposed solutions to different datasets, for the same nine activities, we collect 30 subcarrier-level WiFi CSI (Channel State Information) phase information from each antenna to obtain 90 phase difference samples for one time frame of the captured activity. A mean absolute deviation-based selection method is chosen to filter out the top most reliable subcarriers for activity sensing. A standard Intel 5300 network interface

TABLE 4. Evaluation metrics for estimated 3D human poses using LDT generated WiFi CSI data.

Metrics	LDT	Metrics	LDT
Average joint error (cm)	9.33	FID	4.46
Bone consistency (cm)	2.33	GT FID	0.85
Joint angle error (°)	7.52	Diversity	11.53
Smoothness (cm/frame)	1.03	GT Diversity	11.75
GT Smoothness (cm/frame)	1.38		

card (NIC) operating at 5.3 GHz is used as the WiFi CSI platform. We use the same Kinect camera setup to capture the ground truth 3D pose data. After training the LDT model with the WiFi CSI data, we use the same transformer-based kinematics predictor to estimate synthetic 3D poses from generated WiFi CSI data. We report the same metrics evaluated on RFID data in Table 4. Due to the limited setup and sensing coverage (1 WiFi receiver and 1 WiFi transmitter with 3 antennas), the kinematics predictor displays slight disadvantages compared to the performance of RFID data, especially the worse temporal smoothness (e.g., 1.03 cm/frame compared to 1.38 cm/frame of the ground truth or 1.51 cm/frame of the synthetic RFID-based pose). While RFID data demonstrates slightly better joint accuracy and fidelity (e.g., 8.99 cm average joint error, FID 1.42), the WiFi CSI data maintains robust performance (e.g., 9.33 cm average joint error, FID 4.46). Both synthetic datasets display similar diversity levels, affirming the generalization capability of the LDT framework across RF modalities.

Leveraging the high-quality metrics achieved on 3D poses estimated from LDT-generated RFID data, we extend our framework to perform pose augmentation. By generating diverse synthetic RFID data and mapping them to 3D poses, we significantly expand the dataset size and diversity. Specifically, we generate 80 pairs of RFID and corresponding pose data per class (48 minutes in total), compared to only 18 pairs per class available in the original real dataset (increasing diversity score from 6.64 to 10.98). During the pose estimation process, we utilize a variety of subject skeletons, introducing substantial diversity in skeletal structures and motion patterns, and enriching the existing dataset.

A “Train on Synthesis, Test on Realistic (TSTR)” evaluation is conducted next. We train purely synthetic models (i.e., our LDT, latent diffusion with U-Nets, and diffusion with U-Nets) on the generated RFID-pose pairs. The trained models are then evaluated on real-world testing data (from the last two untrained subjects: Subject 6 and Subject 7). Remarkably, in Fig. 6, the LDT (blue curve) achieves the best performance, with a median error of 4.61 cm and a maximum error capped at 30 cm. Its steeper CDF curve compared to the real-data model indicates that it delivers more consistent estimations, with a majority of joints exhibiting low error rates. This result is a testament to the effectiveness of LDT-based synthetic data for model training

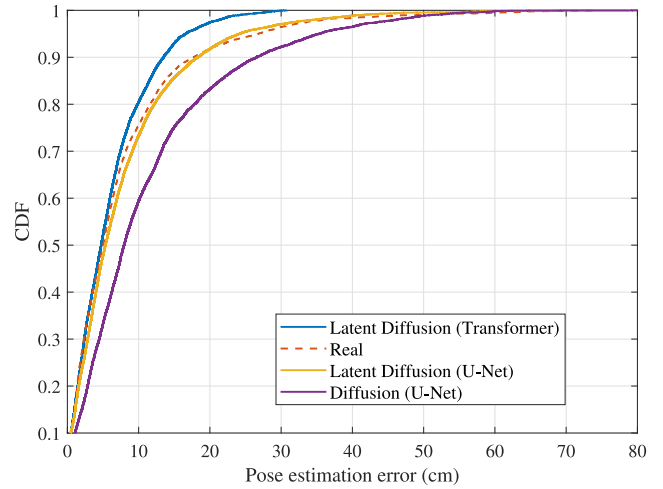


FIGURE 6. Overall pose estimation performance when tested on untrained subjects (Subject 6 and Subject 7) in the form of Cumulative distribution function (CDF) of estimation errors.

and generalization. The Latent Diffusion U-Net (yellow curve) demonstrates decent structural accuracy due to U-Net’s inherent ability to preserve spatial relationships, but its temporal accuracy and smoothness fall short of real-data models. The disadvantages of diffusion models are observed where their estimated poses lack temporal accuracy. Their structural accuracies are maintained due to the benefits of U-Net models. Still, there exists a large number of outliers because diffusion models tend to struggle to effectively learn features of temporal, high-dimensional data in raw data space.

Self-supervised learning (SSL) models typically map RFID signals to 3D poses using paired datasets. However, these models face limitations in data scarcity, sensitivity to noise, and generalization to unseen scenarios. Our LDT framework addresses these challenges by generating high-fidelity synthetic RFID data paired with 3D poses, providing a robust alternative. Results in Table 3 and Fig. 6 are the proof that synthetic data generated by LDT enhances model performance, achieving lower MPJPE and higher temporal smoothness compared to SSL models trained exclusively on real data.

C. 3D POSE COMPLETION

With cross attention conditioning implemented during training, and a novel two-stage motion-aligned generation process, our model generates anatomically consistent and temporally aligned full-body 3D poses from incomplete 12-joint inputs. This experiment evaluates the system’s ability to generate complete poses while preserving the conditioned trajectory of the partial input. Visual demonstrations of the generated 3D poses from Fig. 7 reveal that our system excels in two critical aspects. The first is trajectory adherence where the generated poses (blue) closely follow the trajectory of the conditioned partial poses (green), ensuring that the output motion dynamics are aligned well with the

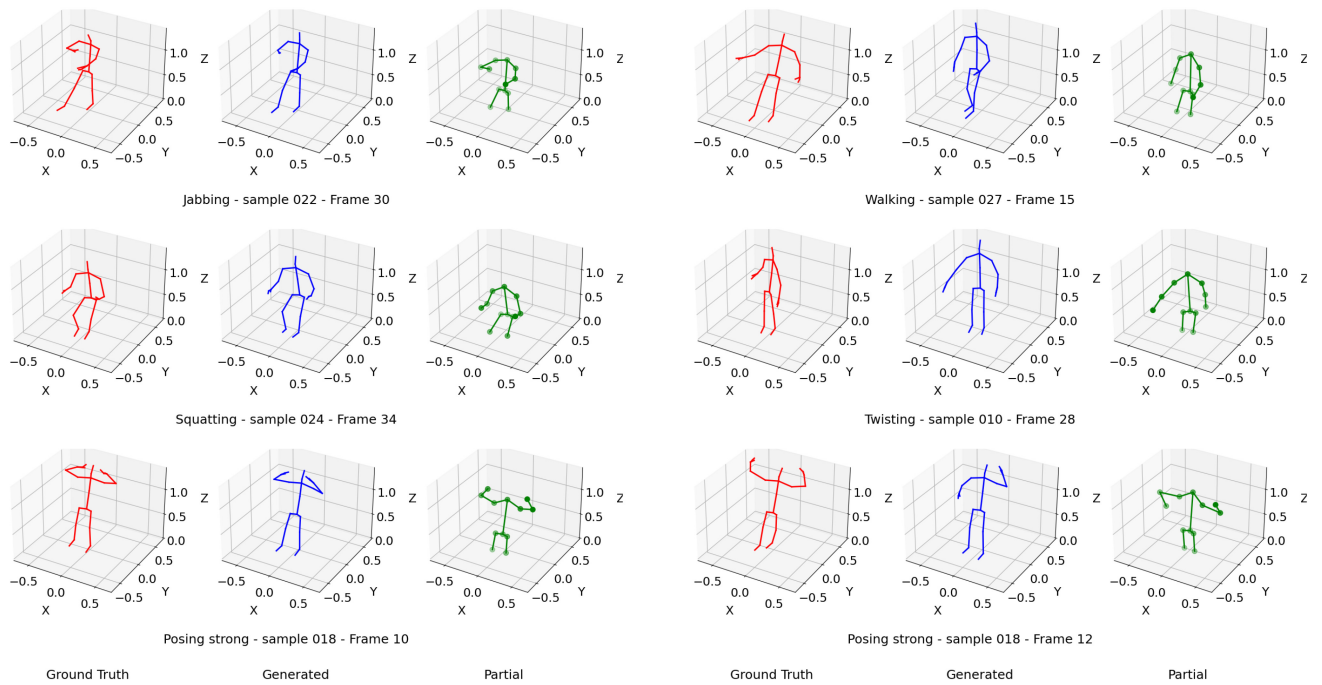


FIGURE 7. Visual demonstration of 3D poses generated by our model (blue) compared to the ground truth complete pose (red), and the conditioned partial pose (green). The left three columns show that our model can generate identical poses conditioned on ground truth partial poses when compared to ground truth complete poses. The right three columns show that our generated pose can follow the trajectory of conditioned unseen partial poses, instead of just copying ground truth complete poses.

input. Unlike simpler pose completion methods, which risk overfitting to ground truth data (red), our model maintains individuality in motion dynamics by conditioning directly on the observed input, even for unseen partial poses. The second is structural accuracy, where the generated poses remain anatomically plausible, accurately reconstructing missing joints to complete the skeletal structure without sacrificing bone consistency.

For testing purposes, we employ a set of most cost-effective metrics. The average joint errors compared to ground truth full poses can be as small as 8.52 cm for simpler activities such as raising an arm (without whole-body movements). The same pose also achieves a bone consistency error of 1.25 cm, joint angle error of 3.5814° , and temporal smoothness of 2.08 cm/frame. These detailed human body-specific metrics explicitly demonstrate the framework's strong capability of structural accuracy. Regarding the unseen scenario, we use LDT-generated partial poses and newly collected poses of the same activities as conditions to simulate an in-the-wild setting. The average joint error increases to 19.23 cm in this case, and the joint angle error increases to 11.83° , highlighting that the generated full pose is not simply copying the trajectories. The bone consistency error of 2.12 cm and temporal smoothness of 1.90 cm/frame show that realistic human motions are maintained.

To ensure fair comparison, we generate as many samples as ground truth counterparts and measure their average scores. Table 5 provides the comprehensive metrics for both ground truth and unseen partial pose conditioning scenarios. The FID and diversity scores for real data distributions used

TABLE 5. Evaluation metrics for 3D pose completion with ground truth and unseen partial pose conditioning.

Metrics	Ground Truth	Unseen
Avg joint error (cm)	11.74	19.23
bone consistency (cm)	1.77	2.12
Joint angle error ($^\circ$)	6.65	11.13
Smoothness (cm/frame)	2.46	1.90
FID (-)	0.87	4.67
Diversity (-)	26.59	13.71
Trajectory joint error (cm)		
compared with partial pose	7.24	8.11
Trajectory velocity error (cm/frame)		
compared with partial pose	7.56	7.80

here are 0.15 and 27.12. The diversity of our generated data mainly depends on the ground truth data, and the scores here and in Table 3 show that our LDT generated data possess as much diversity as real poses. Smoothness scores of generated poses also need to adhere to that of real poses. Anything too low or too high compromises the integrity of Kinect captured ground truth. For unseen scenarios, the smoothness score is only 1.9 cm/frame, considerably lower than the 2.51 cm/frame of ground truth. This means for some generated poses, some joints lack sufficient movements. Due to the absence of ground truth samples, we introduce two metrics termed trajectory joint error and trajectory velocity error. They compare the average joint position and joint velocity errors using the matching 12 joints of our

generated full pose and that of the conditioned partial pose. These help us quantify exactly how much the generated full pose follows the trajectory of partial pose besides visual illustration. For unseen scenarios, the trajectory error is still adequate. One thing to keep in mind is that in 3D HPE, the average length of the limbs is roughly 35cm. A 10cm error means $\sim \frac{1}{3}$ of an arm length deviation or $\sim \frac{1}{4}$ of a leg length deviation. The 11.74cm error of the proposed scheme is reasonable considering we generate more joints (i.e., 25 vs. 17) and inferring more missing joints (11) than most existing works. More important, we are handling the challenging task of motion generation, not just static pose estimation. Fig. 7 shows that the foot, thumb, handtip, and head positions display intricate movements. It is extremely challenging to maintain that kind of precision for Kinect cameras, let alone a generative model. Our trajectory errors in both scenarios indicate that we have achieved a natural motion range preservation from the conditioned partial pose.

We implement and thoroughly evaluate two strong SSL baselines against our approach: Our temporal autoencoder baseline incorporates bidirectional LSTM layers for temporal encoding and decoding, with a spatial encoder to capture joint relationships. This architecture was specifically designed to learn temporal dependencies in pose sequences, processing partial poses (12 joints) to reconstruct full poses (25 joints) while maintaining temporal coherence across time sequences. The KNN baseline was implemented to process poses frame-by-frame, using a 5-nearest neighbor approach to map from partial to full pose configurations.

The autoencoder, despite its sophisticated temporal-aware architecture, shows severe overfitting to the training data. While achieving satisfactory joint errors, it exhibits high trajectory errors and produces unnatural joint twisting (average joint angle error can be 16°) and low-level temporal smoothness (0.3 cm/frame), indicating it's merely learning to minimize reconstruction loss rather than understanding the underlying pose structure. The KNN approach, while maintaining better trajectories, shows limited creativity in pose completion, essentially copying and stitching together existing poses from the training data. This is why in unseen scenarios, the autoencoder approach can achieve slightly better performance. Both methods achieve deceptively good FID scores (lower than 5), but our analysis reveals this is primarily due to statistical similarity to the training data rather than true understanding of pose dynamics and physical constraints. Unlike these baselines, which either average/interpolate between seen poses (autoencoder) or directly copy existing poses (KNN), our LDT framework learns the underlying manifold of human motion, enabling it to generate diverse yet physically plausible pose completions. This is particularly evident in the quality of generated motions, which maintain both local joint consistency and global motion coherence without the artifacts observed in the baseline methods.

D. U-NET-BASED POSE INTERPOLATION

In our framework, the sparse pose frame rate of 7.5 Hz due to the RFID system constraints results in temporally incoherent pose sequences. This low temporal resolution may be insufficient for applications requiring smooth and realistic motions, such as virtual reality, biomechanics, or human activity analysis. In this study, we employ a 2D U-Net-based frame interpolation method to upsample the estimated poses to 30 Hz. Our architecture addresses temporal dependencies using a combination of self-attention mechanisms and temporal feature fusion layers. Self-attention captures motion trajectories, joint dependencies, and frame-to-frame correspondences by modeling long-range relationships across frames. Additionally, temporal convolutional layers in the decoder aggregate information across multiple frames, ensuring smooth transitions and coherent motion sequences.

The encoder consists of four convolutional blocks, each comprising 2D convolutional layers with a kernel size of 3×3 , batch normalization, and ReLU activation. To reduce computational complexity, spatial downsampling with a stride of 2 is applied in selected layers. Skip connections bridge the encoder and decoder, fusing low-level joint details with high-level global motion patterns to ensure precise reconstruction. Self-attention is integrated after each convolutional layer to emphasize motion boundaries and model dependencies between joints over time. The decoder mirrors the encoder, progressively upsampling the features with strides of 2, followed by temporal feature fusion through additional convolutional layers to ensure frame-to-frame smoothness. The architecture also incorporates a combination of loss functions, including mean squared error for spatial reconstruction, temporal smoothness loss to minimize abrupt transitions, and bone-length consistency loss to preserve anatomical plausibility.

During interpolation, the model takes several frames before and after the target interval as input to predict the intermediate pose frames. This approach ensures that the reconstructed data is temporally consistent with the original Kinect video, enabling more precise downstream analysis in motion generation tasks. The frame-to-frame continuity of the 3D poses demonstrates the superiority of our proposed method over traditional interpolation techniques such as linear and cubic interpolation. For instance, in highly dynamic activities like jabbing, kicking, and twisting, the 2D U-Net achieves temporal smoothness errors of 4.76, 1.64, and 3.14 cm, respectively, compared to higher errors from linear interpolation (6.47, 2.86, 5.22 cm) and cubic interpolation (6.51, 2.89, 5.31 cm). This trend is also observed in more static activities such as standing still and squatting, where the 2D U-Net demonstrates approximately a 30% reduction in smoothness error. The performance of our interpolation is illustrated in Fig. 8, where three interpolated frames (green) are generated between every two original frames (blue).

In the end, this higher frame rate preserves the natural dynamics of human motion, aligns well with the standard resolution of motion capture systems, and significantly

TABLE 6. Ablation study on different generation strategies and inference steps².

Generation Strategy	Inference Steps	FID ↓	Avg. Joint Error(cm) ↓	Bone Consistency(cm) ↓	Joint Angle Error(°) ↓	Time(s) ↓
Two-stage	1000, 800	0.71	11.61	1.76	6.60	22.36
	80, 60	0.87	11.74	1.77	6.65	1.59
	50, 30	1.05	13.06	1.82	7.71	1.52
Single-stage	80	1.87	20.18	2.28	11.93	0.67
Single-stage + Velocity Alignment	80	1.27	18.36	2.12	10.25	1.20

² Note: Lower is better for all metrics (↓). Two-stage refers to our proposed approach with initial generation and refinement. Generation time is measured on a single NVIDIA A4000 GPU.

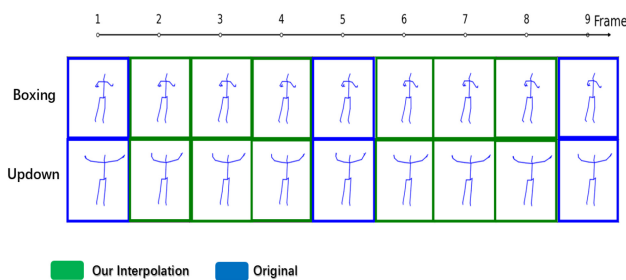


FIGURE 8. Demonstrate of the interpolation process of generating additional frames between original frames.

enhances the realism and usability of the generated pose data for downstream tasks.

E. ABLATION STUDIES

To better understand the relationships between training, generation and hardware setup, and generation performance of the framework, in-depth ablation studies are conducted. The ablation study examined various configurations, including varying noise steps for generation (50, 80, and 1,000), and different training and generation strategies. Additionally, we measure training and inference time to understand how each setup impacts training and generation speed.

Table 6 shows the detailed generation performance between different generation strategies. We find that 80 and 60 inference steps for the first and second stages are the most efficient setup to achieve optimal performance and cost very little time. 1,000 and 800 inference steps take too much time (up to 16s) with diminishing returns on generation performance, while 50 and 30 steps achieve inferior performance while not saving enough time. To prove the effectiveness of two-stage generation and refinement, we further explore a single-stage strategy with only initial generation. Without latent velocity alignment, the generated pose struggles to follow the trajectory of partial poses. Last, we explore only ditching the second stage of refinement but keeping the velocity alignment. Considering the nominal time saved, our two-stage strategy with 80 and 60 steps becomes the default setup.

We also present our findings during the challenging process of training different models leading to a successful one, as seen in Table 7 when directly training on raw data space of 3D poses and only using diffusion's noise prediction loss, the trained model lacks the basic understanding of sequential 3D human poses. The generated pose is a smaller figure than real pose and is moving diagonally across the frames. Hence we incorporate pose-related loss including position error into the training scheme, which greatly increases the computation load to up to 15.14 GB VRAM. It takes more than three days to finish training 799 epochs for RTX 3070, with the performance lacking fidelity as well. The position of the generated pose is fixed across frames, but its limbs are jittering. This method can be further optimized with a better-devised training scheme and GPU, but we intend to explore a cost-effective alternative. Leveraging latent diffusion, and a well-trained VAE for 3D poses, we can largely decrease computation load to around 1 GB of VRAM and less than 5 hours of training, while achieving moderate success on the generation quality. This then becomes the core of our RFID generation and 3D pose completion framework. When combined with cross attention conditioning on 3D pose completion, the computation load is moderately increased, and the generation quality for the full pose reaches a new high (FID of 0.71 and recognition accuracy of 83.22%). This recognition accuracy means our generated poses maintain action-specific features, and the motions are clear enough to be classified correctly. Besides FID score, the recognition accuracy can indicate that the generated poses are similar to real poses in a meaningful way. When trained and tested on real pose data, the FID score is 0.15, and recognition accuracy is 96.8%.

We also conduct extensive experiments to determine the optimal latent dimension, considering both theoretical and empirical factors. Our choice of 256 dimensions aligns with successful practices in related work: VIBE [47] uses 256 dimensions for human pose and motion representation, while Action2Motion [48] employs a 256-dimensional latent space for motion synthesis. While a larger latent space theoretically offers more capacity for pose representation, it must balance against computational efficiency and the risk

TABLE 7. Ablation study on different training frameworks³.

Model	Time (hour)	Train Epoch	GPU Usage (GB)	Generation Time (s)	FID	Recognition Accuracy (%)
RFID-LatentDiffusion	1.63	16099	1.02	10.34	1.42 (from estimated poses)	80.04 (from estimated poses)
Pose12Joints-VAE	4.54	15099	0.65	-	-	-
Pose12Joints-LatentDiffusion	2.91	9899	0.69	1.19	3.66	52.50
Pose25Joints-VAE	4.54	13899	0.67	-	-	-
Pose25Joints-LatentDiffusion	3.70	14499	0.75	1.23	4.12	48.42
Pose25Joints-LatentDiffusion With Cross Attention	12.02	13899	1.97	1.23	0.71	83.22
Pose12Joints-Diffusion	3.73×24	799	15.14	10.34	45.88	10.53
Pose12Joints-Diffusion-only	7.74	12299	0.96	1.32	29.67	20.10

³ Note: Recognition accuracy is designed as the activity class-based 3D pose classification accuracy. We train a pose discriminator that shares the same backbone architecture as our FID feature extractor to classify poses into activity classes.

of overfitting. For training the LDT framework with cross-attention, our experiments with 128 dimensions showed limited capacity to capture fine-grained pose details despite faster training (7 hours, 1.33GB VRAM). Increasing to 512 dimensions demonstrated diminishing returns, with only marginal improvement in reconstruction ($< 2\%$) at significantly higher computational cost (17 hours, 6.21GB VRAM). The chosen 256 dimensions achieved an optimal balance, showing both stable training dynamics and strong reconstruction quality while maintaining reasonable computational requirements (12 hours, 1.97GB VRAM).

VII. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel framework leveraging latent diffusion transformers (LDT) for generating RFID sensing data and completing 3D human poses from partial observations. Our system has comprised three integrated components: RFID data generation, a transformer-based kinematics predictor, and a cross-attention-based two-stage velocity alignment pose completion model. By addressing key challenges, including sparse RFID data, memory constraints, and anatomical plausibility in pose generation, we have demonstrated the effectiveness of LDT in producing realistic and diverse data, as well as completing partial poses. Extensive experiments have validated the quality of our estimated poses, achieving low joint errors, high temporal smoothness, and competitive FID and diversity scores, highlighting the system's suitability for wireless sensing applications. This work has bridged the gap between wireless sensing and artificial intelligence-generated content (AIGC), offering effective solutions for RFID-based pose augmentation and completion in various application domains.

Although LDT has enabled scalable synthetic data generation, the diversity of generated RFID data and corresponding pose data has remained constrained by the training dataset used to learn the diffusion process. This limitation may hinder the system's ability to handle highly diverse real-world scenarios, necessitating the collection of more comprehensive datasets. We plan to investigate new RFID sensing

technologies or configurations to increase the number of tags while mitigating interference, enabling finer-grained partial pose data. Additionally, we plan to incorporate other RF modalities with the LDT approach proposed in this paper. A multi-person capable 3D pose estimation and completion framework can also largely increase the deployment ability in real-world settings.

REFERENCES

- [1] J. Wang, Q. Gao, M. Pan, and Y. Fang, "Device-free wireless sensing: Challenges, opportunities, and applications," *IEEE Netw.*, vol. 32, no. 2, pp. 132–137, Mar./Apr. 2018.
- [2] F. Adib and D. Katabi, "See through walls with WiFi!" in *Proc. ACM SIGCOMM*, Hong Kong, Aug. 2013, pp. 75–86.
- [3] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-Eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures," in *Proc. ACM Mobicom*, Sep. 2014, pp. 617–628.
- [4] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surveys*, vol. 46, no. 2, p. 32, Dec. 2013.
- [5] H. Li, C. Ye, and A. P. Sample, "IDSense: A human object interaction detection system based on passive UHF RFID," in *Proc. ACM CHI*, Seoul, South Korea, Apr. 2015, pp. 2555–2564.
- [6] C. Yang, X. Wang, and S. Mao, "Unsupervised detection of APNEA using commodity RFID tags with a recurrent variational autoencoder," *IEEE Access*, vol. 7, pp. 67526–67538, Jun. 2019.
- [7] C. Ding et al., "Continuous human motion recognition with a dynamic range-doppler trajectory method based on FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6821–6831, Sep. 2019.
- [8] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range FMCW monopulse radar for hand-gesture sensing," in *Proc. IEEE Radar Conf.*, Arlington, VA, USA, May 2015, pp. 1491–1496.
- [9] M. Zhao et al., "Through-wall human pose estimation using radio signals," in *Proc. IEEE/CVF CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 7356–7365.
- [10] M. Zhao et al., "RF-based 3-D skeletons," in *Proc. ACM SIGCOMM*, Budapest, Hungary, Aug. 2018, pp. 267–281.
- [11] 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. 10032–10044, Sep. 2020.
- [12] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3-D human pose estimation with RFID," *IEEE Trans. Rel.*, vol. 70, no. 3, pp. 1218–1231, Sep. 2021.
- [13] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 7297–7306.
- [14] C. Yang, X. Wang, and S. Mao, "RFID based 3-D human pose tracking: A subject generalization approach," *KeAi Digit. Commun. Netw.*, vol. 8, no. 3, pp. 178–288, Aug. 2022.

- [15] C. Yang, L. Wang, X. Wang, and S. Mao, "Environment adaptive RFID based 3-D human pose tracking with a meta-learning approach," *IEEE J. Radio Freq. Identification*, vol. 6, no. 1, pp. 413–425, Jan. 2022.
- [16] Z. Wang, C. Yang, and S. Mao, "AIGC for RF-based human activity sensing," *IEEE Internet Things J.*, vol. 12, no. 4, pp. 3991–4005, Feb. 2025.
- [17] 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, Sep. 2019.
- [18] 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*, New Orleans, LA, USA, Apr./May 2022, p. 108.
- [19] A. Zanfir et al., "HUM3DIL: Semi-supervised multi-modal 3-D humanpose estimation for autonomous driving," in *Proc. 6th Annu. Conf. Robot Learn.*, Auckland, NZ, USA, Dec. 2022, pp. 1–4.
- [20] J. Zhang et al., "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.
- [21] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proc. NeurIPS*, Dec. 2019, p. 494.
- [22] Z. Wang, C. Yang, and S. Mao, "Data augmentation for RFID-based 3-D human pose tracking," in *Proc. IEEE VTC-Fall*, London, U.K., Sep. 2022, pp. 1–2.
- [23] H. Xue et al., "Towards generalized mmWave-based human pose estimation through signal augmentation," in *Proc. ACM MobiCom*, Madrid, Spain, Oct. 2023, p. 88.
- [24] Z. Wang and S. Mao, "AIGC for wireless data: The case of RFID-based human activity recognition," in *Proc. IEEE ICC*, Denver, CO, USA, Jun. 2024, pp. 4060–4065.
- [25] Z. Wang and S. Mao, "AIGC for wireless sensing: Diffusion-empowered human activity recognition," *IEEE Trans. Cogn. Commun. Netw.*, early access, Jan. 3, 2025, doi: [10.1109/TCCN.2025.3525588](https://doi.org/10.1109/TCCN.2025.3525588).
- [26] G. Chi et al., "RF-Diffusion: Radio signal generation via time-frequency diffusion," in *Proc. ACM MobiCom*, Washington, DC, USA, Nov. 2024, pp. 77–92.
- [27] 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 Things J.*, vol. 10, no. 16, pp. 14128–14136, Aug. 2023.
- [28] U. Rafi, J. Gall, and B. Leibe, "A semantic occlusion model for human pose estimation from a single depth image," in *Proc. IEEE/CVF CVPR Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 67–74.
- [29] X. Guo and Y. Dai, "Occluded joints recovery in 3-D human pose estimation based on distance matrix," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 1325–1330.
- [30] Y. Cheng, B. Yang, B. Wang, Y. Wending, and R. Tan, "Occlusion-aware networks for 3-D human pose estimation in video," in *Proc. IEEE/CVF ICCV*, Seoul, South Korea, Oct./Nov. 2019, pp. 723–732.
- [31] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 1954–1963.
- [32] K. Holmquist and B. Wandt, "DiffPose: Multi-hypothesis human pose estimation using diffusion models," in *Proc. IEEE/CVF ICCV*, Paris, France, Oct. 2023, pp. 15931–15941.
- [33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Dec. 2020, *arxiv:2006.11239*.
- [34] W. Jiang et al., "Towards 3-D human pose construction using WiFi," in *Proc. ACM MobiCom*, London, U.K., Sep. 2020, p. 23.
- [35] G. Pavlakos et al., "Expressive body capture: 3-D hands, face, and body from a single image," in *Proc. IEEE/CVF CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 10975–10985.
- [36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR*, New Orleans, LA, USA, Jun. 2022, pp. 10674–10685.
- [37] Z. Wang and S. Mao, "AIGC for RF sensing: The case of RFID-based human activity recognition," in *Proc. ICNC*, Big Island, HI, USA, Feb. 2024, pp. 1092–1097.
- [38] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS Workshops*, Dec. 2021, pp. 1–8.
- [39] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.
- [40] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *Proc. ICLR*, May 2023, pp. 1–16.
- [41] M. Zhang et al., "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, Jun. 2024.
- [42] L.-H. Chen et al., "HumanMAC: Masked motion completion for human motion prediction," in *Proc. IEEE/CVF ICCV*, Paris, France, Oct. 2023, pp. 9544–9555.
- [43] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Trans. Graph.*, vol. 35, no. 4, p. 138, Jul. 2016.
- [44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF ICCV*, Paris, France, Oct. 2023, pp. 3836–3847.
- [45] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. ICLR*, Apr. 2022, pp. 1–21.
- [46] J. Wang et al., "Guiding AI-generated digital content with wireless perception," *IEEE Wireless Commun.*, vol. 31, no. 4, pp. 147–154, Aug. 2024.
- [47] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF CVPR*, Seattle, WA, USA, Jun. 2020, pp. 5252–5262.
- [48] C. Guo et al., "Action2Motion: Conditioned generation of 3-D human motions," in *Proc. ACM Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2021–2029.



ZIQI WANG (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Auburn University, Auburn, AL, USA, in 2022, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His current research focuses on artificial intelligence of things and wireless sensing. He is a recipient of the IEEE ICC 2024 NSF Student Travel Grant, and a co-recipient of the Best Demo Award of IEEE INFOCOM 2024.



SHIWEN MAO (Fellow, IEEE) is a Professor and an Earle C. Williams Eminent Scholar, and the Director of the Wireless Engineering Research and Education Center, Auburn University. His research interest includes wireless networks, multimedia communications, and smart grid. He received the IEEE ComSoc MMTC Outstanding Researcher Award in 2023, the SEC 2023 Faculty Achievement Award for Auburn, the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019, the Auburn University Creative Research and Scholarship Award in 2018, the NSF CAREER Award in 2010, and several service awards from IEEE ComSoc. He is a co-recipient of the 2022 Best Journal Paper Award of IEEE ComSoc eHealth Technical Committee, the 2021 Best Paper Award of *Digital Communications and Networks* (Elsevier/KeAi), the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the 2018 Best Journal Paper Award, the 2017 Best Conference Paper Award from IEEE ComSoc MMTC, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a co-recipient of the best paper/demo awards of 12 conferences. He is the Editor-in-Chief of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, a member-at-large on the Board of Governors of IEEE Communications Society, and the Vice President of Technical Activities of IEEE Council on Radio Frequency Identification.