修士論文概要書

Master's Thesis Summary

Date of submission: 07/21/2025 (MM/DD/YYYY)

専攻名（専門分野） Department	Computer Science and Communications Engineering	氏 名 Name	Xinyi Liu	指 導 教 員 Advisor	Hiroshi Watanabe 印 Seal
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	CD 5123FG46-6		
研究題目 Title	A Conditional Diffusion Framework for 3D Human Pose Refinement: Leveraging Temporal Consistency				

1. Introduction

3D human pose estimation from single images has made good progress in recent years. But temporal inconsistencies in sequential predictions are still a big challenge. Current methods usually process video frames one by one. This leads to jittery and unnatural motion in pose sequences. Single-frame approaches like ZeDO [1] show excellent performance. But they cannot ensure temporal consistency across sequences. Temporal modeling approaches like SmoothNet [2] focus on post-processing. However, they lack the powerful pose priors that diffusion models provide.

This research proposes a new conditional diffusion framework to solve this problem. The framework uses Transformer [3] architecture for temporal 3D human pose refinement. The approach treats pose sequence correction as a denoising process. Noisy pose predictions are refined step by step to achieve temporal consistency. The framework uses a Transformer denoising model that processes sequential pose data. This allows effective modeling of temporal dependencies across joint positions. The method conditions the diffusion process on initial pose predictions. It learns to correct inconsistencies while preserving underlying motion patterns.

2. Related Works

In 3D human pose estimation, important works can be divided into single-frame and temporal modeling approaches. Optimization methods estimate poses by reducing reprojection errors. Recent breakthrough methods include ZeDO, which uses diffusion models as optimization tools. This achieves top performance without needing 2D-3D paired training data.

For temporal consistency solutions, traditional filtering methods are widely used for smoothing pose sequences. These include Gaussian filtering and Savitzky-Golay filtering. SmoothNet represents a major breakthrough in data-driven temporal refinement methods. It proposes a plug-and-play temporal refinement network. This network is designed to reduce jitters in outputs from existing pose estimators.

For diffusion models, Ho et al. introduced Denoising Diffusion Probabilistic Models (DDPM) [4]. These models learn to create data by reversing a diffusion process. Song et al. proposed DDIM [5] for faster sampling. This addresses the computational bottleneck of standard diffusion sampling. Recent applications to 3D pose estimation include DiffPose [6]. It treats pose estimation as a reverse diffusion process.

Compared to these existing works, this research focuses on extending diffusion models to temporal pose sequences. The method conditions the denoising process on predicted poses. The Transformer-based architecture enables effective modeling of both spatial joint relationships and temporal motion patterns within a unified framework.

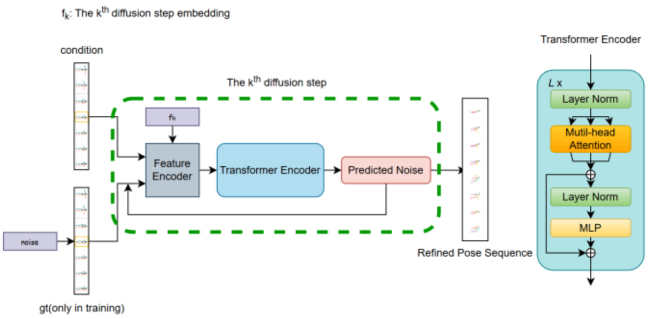


Fig. 1 Overview of the proposed method.

3. Proposed Method

As shown in Fig. 1, the key innovations proposed in the methodology are summarized as follows:

Temporal Pose Diffusion Framework: The approach treats temporal pose refinement as a conditional sequence-to-sequence denoising task. Unlike unconditional diffusion models that start from pure noise, this method conditions the entire denoising process on predicted poses. This serves two purposes. First, it limits the solution space to anatomically reasonable poses. Second, it preserves general motion patterns while correcting errors.

Forward Process for Pose Sequences: Following the standard DDPM framework, the forward process systematically corrupts ground truth pose sequences. It does this by gradually adding Gaussian noise. For a ground truth pose sequence, noisy versions can be directly sampled at any timestep t using: $x_t = \sqrt{\overline{\alpha}_t}\, x_0 + \sqrt{1-\overline{\alpha}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$.

Conditional Reverse Process: The reverse process conditions every denoising step on the predicted pose sequence, modeled as: $p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2 I)$, where c represents the conditioning information.

Temporal Denoising Architecture: The network integrates three information sources. These are current noisy sequence state, diffusion timestep, and condition poses. A Transformer encoder processes flattened joint-time tokens. The encoder has 6 layers, 8 attention heads, and dimension 256. This enables global modeling of spatial-temporal relationships.

4. Data

The experiments are conducted on the Human3.6M [7], [8] dataset, a standard dataset for 3D human pose estimation containing 3.6 million video frames with accurate 3D pose annotations.

Pose predictions generated by ZeDO serve as input to the temporal refinement approach. ZeDO is a state-of-the-art single-frame 3D pose estimation method. Temporal sequences are extracted using sliding windows of 16 frames. The stride is 8 during training and 1 during evaluation. All poses are normalized by subtracting the root joint position. This ensures translation invariance.

5. Experiment

Evaluation Metrics: Three standard metrics are used. MPJPE (Mean Per Joint Position Error) measures absolute pose accuracy. P-MPJPE (Procrustes-aligned MPJPE) evaluates pose structure. MPJAE (Mean Per Joint Acceleration Error) measures temporal smoothness.

Table 1. Quantitative comparison on Human3.6M. Mm for MPJPE/P-MPJPE and mm/frames² for MPJAE.

	MPJPE	P-MPJPE	MPJAE
ZeDO	54.77	37.48	2.52
SmoothNet	53.91	37.45	**0.98**
Ours	**38.91**	**27.18**	2.58

Results: The method achieves significant improvements over baseline methods. Compared to ZeDO, the method reduces MPJPE by 15.86mm. This goes from 54.77mm to 38.91mm. It also reduces P-MPJPE by 10.30mm, from 37.48mm to 27.18mm. The approach also outperforms SmoothNet. It demonstrates superior pose accuracy while maintaining temporal consistency.

Ablation Studies: Experiments reveal that longer sequence lengths provide consistent accuracy improvements. Using 32 frames instead of 16 frames shows better results. DDIM sampling offers significant speed advantages with a 22× speedup. But it introduces temporal quality trade-offs. These trade-offs can be reduced through longer temporal context.

Table 2. Ablation study results on Human3.6M.

	MPJPE	P-MPJPE	MPJAE	Inference FPS
16 frames + DDPM	38.91	27.18	2.58	13.9
16 frames + DDIM	39.09	27.23	10.45	306.2
32 frames + DDPM	38.21	27.02	2.38	7.0
32 frames + DDIM	37.77	26.86	7.81	151.5

6. Conclusion

This research presents a new temporal pose diffusion framework that successfully addresses temporal inconsistencies in 3D human pose estimation. The method conditions diffusion models on predicted poses and uses Transformer-based denoising architecture. The approach achieves significant improvements in pose accuracy and maintains temporal consistency.

Experimental results on Human3.6M show substantial improvements over baseline methods. The method reduces MPJPE by 15.86mm compared to ZeDO. It also maintains plug-and-play compatibility for practical deployment. The framework opens new possibilities for applying diffusion models to temporal motion analysis.

References

[1] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, "Back to optimization: Diffusion-based zero-shot 3d human pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6142–6152, 2024.

[2] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "Smoothnet: A plug-and-play network for refining human poses in videos," in *Proceedings of European Conference on Computer Vision*, pp. 625–642, 2022.

[3] A. Vaswani *et al.*, "Attention is all you need," *Neural Information Processing Systems*, vol. 30, 2017.

[4] J. Ho, A. Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[5] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[6] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Diffpose: Toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051, 2023.

[7] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[8] C. S. Catalin Ionescu Fuxin Li, "Latent Structured Models for Human Pose Estimation," in Proceedings of the *International Conference on Computer Vision*, 2011.

A Conditional Diffusion Framework for 3D Human Pose Refinement: Leveraging Temporal Consistency

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: July 21st, 2025

Xinyi Liu
(5123FG46-6)
Advisor: Prof. Hiroshi Watanabe
Research guidance: Research on Audiovisual Information Processing

Acknowledgements

I want to thank my supervisor, Professor Hiroshi Watanabe, for his helpful guidance, patience, and support during my research. His knowledge and feedback have been very important for this work. His help has also been key to my growth as a researcher. He always gave me useful advice when I had problems.

I also want to thank all members of the Advanced Multimedia Systems Laboratory. Their discussions and teamwork have created a good research environment. This helped my learning a lot. The knowledge sharing and support in our laboratory have been very helpful for my development. My lab members always helped me when I needed it. Their friendship made the research process more enjoyable.

I am also grateful to my family and friends for their love, understanding, and support during this time. They have been my strength throughout the way. Their belief in me and encouragement have helped me overcome difficulties. Their support made this challenging process more meaningful. Without them, this work would not have been possible.

Abstract

3D human pose estimation from monocular images has achieved significant progress, yet temporal inconsistencies in sequential predictions remain a critical challenge. To address this problem, we propose a novel conditional diffusion framework that leverages Transformer architecture for temporal 3D human pose refinement.

Our approach treats pose sequence correction as a denoising process. In this process, noisy pose predictions are step by step refined to achieve temporal consistency. The framework uses a Transformer-based denoising model that processes sequential pose data. This allows effective modeling of temporal dependencies across joint positions. Our method conditions the diffusion process on initial pose predictions. This way, our method learns to correct inconsistencies while keeping underlying motion patterns.

Extensive experiments on the Human3.6M dataset show that our approach greatly improves temporal consistency in pose sequences while maintaining spatial accuracy. The proposed framework offers a solution for post-processing pose estimation results and can be easily integrated into existing pipelines. This opens new possibilities for applying diffusion models to temporal motion analysis.

Keywords: 3D Human Pose Estimation, Conditional Diffusion Model, Pose Refinement, Temporal Consistency

Contents

List of Figures

List of Tables

1 Introduction

1.1 Research Background

3D human pose estimation is an important computer vision task with applications in motion capture, human-computer interaction, and virtual reality [1], [2], [3]. The goal is to predict 3D joint positions from input images or videos. While single-frame methods have achieved good performance, video-based pose estimation faces additional temporal challenges.

Most current methods process video frames independently, which leads to temporal inconsistencies in pose sequences [4], [5]. These inconsistencies lead to jittery and unnatural motion, which affects the quality of reconstructed human poses. Recent optimization methods like ZeDO [6] and learning-based diffusion approaches like DiffPose [7], [8] have shown great results, but temporal consistency is still a challenge. The lack of temporal information limits the applications of these systems.

To solve temporal inconsistencies, researchers have proposed lots of smoothing and refinement approaches [9], [10]. Some methods focus on post-processing temporal smoothing [11]. Others use spatial-temporal information directly in the estimation process [12]. However, these approaches still face difficulties in handling complex motion patterns and long-range temporal sequences.

Diffusion models have been proved as powerful generative tools that learn to reverse noise corruption processes [13], [14]. They have shown great success in image generation and are being used for various computer vision tasks. The iterative denoising nature of diffusion models makes them well-suited for refinement problems that require gradual improvement.

1.2 Research Objectives

In this work, we aim to develop a conditional diffusion framework for improving temporal consistency in 3D HPE from video sequences. The objectives of this study are listed as follows:

1. We develop a Transformer-enhanced [15] conditional diffusion model that uses temporal and spatial positional encodings to model sequential pose data. The framework treats pose sequence refinement as a denoising process, where noisy pose predictions are gradually refined using learned temporal dependencies and joint relationships to achieve smooth and consistent motion.

2. We evaluate the proposed framework on standard 3D human pose estimation benchmarks, using Human3.6M dataset[16], [17]. Experiment results shows that our approach significantly improves temporal consistency metrics while maintaining competitive pose accuracy compared to existing video-based pose estimation methods and single-frame approaches.

3. We conduct ablation studies to analyze the key components and designs in the proposed architecture. This includes looking at different diffusion sampling strategies, optimal temporal window sizes, and the impact of various architectural components on pose refinement performance. The results provide insights into how effective diffusion-based approaches are for temporal pose sequence modeling.

1.3 Thesis Outline

The structure of this thesis is as follows:

Chapter 1: This chapter provides an overview of 3D human pose estimation and the temporal consistency problem. We introduce the motivation for using conditional diffusion models with Transformer architectures and present the research objectives and main contributions.

Chapter 2: This chapter reviews existing literature in four main areas. These areas include 3D human pose estimation methods with single-frame and video-based approaches, diffusion models covering DDPM/DDIM sampling strategies and their computer vision applications, Transformer-based architectures for pose estimation with attention mechanisms, and temporal smoothing methods including SmoothNet and other pose refinement techniques.

Chapter 3: This chapter presents our conditional diffusion framework for temporal pose refinement. We describe the Transformer-based denoising model, conditioning mechanism, diffusion processes, training objectives, and DDPM/DDIM sampling methods for inference.

Chapter 4: This chapter presents experimental evaluation on Human3.6M dataset. This includes comparisons with baseline methods, ablation studies and qualitative results showing improved temporal consistency.

Chapter 5: This chapter summarizes the main findings and contributions. We discuss current limitations and propose future directions including multi-person scenarios, real-time applications, and other temporal sequence modeling tasks.

2 Related Works

2.1 3D Human Pose Estimation

2.1.1 Overview of 3D Human Pose Estimation

The goal of 3D Human Pose Estimation (3D-HPE) is predicting the 3D coordinates of human skeleton joints from images or videos [2]. Methods can be categorized based on camera setup (monocular vs. multi-view), number of subjects, and input type (image vs. video).

Monocular approaches face the challenge of depth ambiguity, where multiple 3D poses can be paired to identical 2D appearances [18]. Two-stage approaches that first estimate 2D poses and then uplift them to 3D have become popular [19], [20]. Recent work has explored Graph Convolutional Networks (GCNs) to model skeletal relationships [21] and transformer-based methods that leverage attention mechanisms [22], [23].

A critical limitation of frame-based methods is their inability to leverage temporal information, which is particularly problematic when dealing with sequences that have limited or missing temporal context. Recent diffusion-based methods have shown promise in addressing this uncertainty. DiffPose treats 3D pose estimation as a reverse diffusion process to deal with inherent indeterminacy [8]. ZeDO uses optimization and learning at the same time, achieving great performance without training on 2D-3D pairs by using diffusion models for multi-hypothesis pose generation [6].

Video-based use temporal information by temporal convolutional networks [24] and transformer architectures like MixSTE, which processes spatial and temporal information jointly [25]. However, when temporal information is sparse or absent, these methods will face challenges, as motion context that could help resolve ambiguities and infer occluded joints becomes unavailable.

The field faces persistent challenges including data scarcity, occlusion handling, and cross-domain generalization. The absence of temporal information in single-frame scenarios further exacerbates these challenges, requiring specialized techniques for isolated frames or sparsely sampled sequences.

2.1.2 Single-Frame Approach

In the broader landscape of 3D human pose estimation, single-frame approaches focus on extracting pose information from individual images without temporal context. These methods are essential when dealing with isolated frames or sparsely sampled sequences.

Traditional Single-Frame Methods include learning-based and optimization-based approaches. Learning-based methods such as the 2D-to-3D lifting baseline by Martinez et al. [26] and end-to-end methods that directly estimate 3D poses from images have shown promising results. Optimization-based methods like SMPLify [27] estimate poses by minimizing reprojection errors while satisfying anatomical constraints. However, these methods often suffer from performance degradation in cross-domain scenarios—learning methods are limited by training data distributions, while optimization methods typically underperform compared to learning approaches.

The ZeDO method [6] represents a breakthrough in single-frame 3D pose estimation. This method innovatively employs diffusion models as optimization tools instead of direct generators, combining geometric constraints with learned priors within an iterative optimization framework.

ZeDO's core idea involves starting with an initial pose hypothesis, applying geometric optimization to minimize 2D reprojection errors, using a pre-trained diffusion model for denoising to ensure pose plausibility, and iterating this process until convergence. This design cleverly avoids the limitations of traditional methods—maintaining the cross-domain adaptability of optimization approaches while taking advantage of the powerful pose priors of diffusion models.

Experimental results show that ZeDO achieves state-of-the-art performance without requiring any 2D-3D paired training data: 51.4mm MPJPE on Human3.6M and 40.3mm PA-MPJPE on 3DPW datasets. However, ZeDO still processes frames independently without considering temporal consistency, which provides opportunities for subsequent temporal optimization research.

2.1.3 Temporal Modeling Method

While traditional methods rely heavily on temporal information for accurate 3D pose estimation, recent work has explored new approaches to handle uncertainty when temporal context is limited or unavailable.

DiffPose [8] represents a major advancement in handling pose uncertainty through diffusion models. Unlike conventional methods that directly regress 3D poses, DiffPose treats the estimation process as a reverse diffusion procedure. It starts with a highly uncertain 3D pose distribution and step by step reduces uncertainty through multiple denoising steps.

DiffPose supports two working modes: video-based estimation and frame-based estimation. In video mode, the method leverages temporal information for pose estimation, while in frame mode, it extracts context information only from single images.

2.1.4 Temporal Consistency Solutions

Apart from methods that directly handle pose uncertainty, there is another important research direction that focuses on improving temporal consistency and smoothness of existing pose estimation results. These methods typically serve as post-processing steps to perform temporal modeling and optimization on existing pose estimation results.

Traditional filtering methods such as Gaussian filtering [28], Savitzky-Golay filtering [29], and One-Euro filtering [30] are widely used for smoothing pose sequences. These methods perform low-pass filtering on pose sequences through fixed temporal windows. This effectively reduces high-frequency jitters. However, these methods face trade-offs between smoothness and lag and perform poorly in handling long-term jitters.

SmoothNet [10] represents a major breakthrough in data-driven temporal refinement methods. This approach proposes a plug-and-play temporal refinement network specifically designed to reduce jitters in outputs from existing pose estimators. The core innovation of SmoothNet lies in adopting a purely temporal modeling strategy, avoiding the bottleneck of joint spatial-temporal optimization.

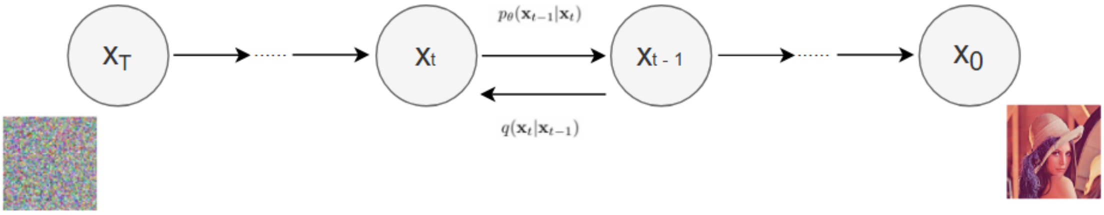
The method uses motion-aware fully connected networks to learn long-term temporal relationships for each joint without considering noisy correlations among joints. SmoothNet supports multiple motion modalities (2D/3D positions, 6D rotation matrices) and shows good generalization across different backbone networks, modalities, and datasets. Experimental results show significant performance improvements on standard datasets such as Human3.6M [16] and MPI-INF-3DHP [31].

2.2 Diffusion Models

2.2.1 Overview of Diffusion Models

Diffusion model is a generative model that learns to create data by reversing a diffusion process [13]. These models have emerged as a powerful alternative to GANs [32] and VAEs [32], achieving SOTA results in a lot of tasks. The key insight is that complex data distributions can be learned by modeling a simple denoising process.

The training process involves two main components: a forward diffusion process that gradually corrupts data with noise, and a reverse process that learns to undo this corruption. This approach has several advantages including training stability, mode coverage, and high-quality sample generation. Recently, diffusion models have been successfully applied to 3D pose estimation tasks [8], showing promise for handling the temporal smoothing challenges in human pose sequences.



X_T
X_t
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
X_{t-1}
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
X_0

Fig. 1 The denoising diffusion probabilistic model.

2.2.2 Forward Process

The forward process, or the diffusion process, gradually adds Gaussian noise to the data with T timesteps. This process is simple and tractable, following a Markov chain where each step depends only on the previous one. At each step t, noise is added as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I\right), \tag{2.1}$$

where β_t is the variance schedule that controls the noise level at step t. It is typically chosen to increase from small values (e.g., 10^-4) to larger values (e.g., 0.02) over the T steps. This ensures that the data gradually becomes pure noise.

A key property of this process is that it can be computed in closed form for any timestep t, without requiring all intermediate steps:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} \cdot x_0, (1-\bar{\alpha}_t) \cdot I\right), \tag{2.2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. This allows efficient sampling of noises of the data at any timestep during training.

2.2.3 Reverse Process

The reverse process learns to remove noise and reconstruct the original data distribution. Since the true reverse process is intractable, it is approximated using a neural network. The reverse process is modeled as:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right). \tag{2.3}$$

In practice, the covariance $\Sigma_\theta(x_t, t)$ is often fixed to $\sigma_t^2 \mathrm{I}$, where $\sigma_t^2 = \beta_t$ or $\widetilde{\beta_t} = \frac{1-\widetilde{\alpha_{t-1}}}{1-\widetilde{\alpha_t}} \cdot \beta_t$. The neural network aims to predict the noise $\epsilon_\theta(x_t, t)$ that was added at each step.

The mean of the reverse process can be computed using the predicted noise:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_\theta(x_t, t)\right). \tag{2.4}$$

This formulation allows the model to learn the reverse process by simply predicting the noise that was added during the forward process, which has been shown to be more stable than directly predicting the mean.

2.2.4 Loss Function

The training objective comes from maximizing the variational lower bound of the log-likelihood. The full objective involves multiple terms, but in practice, it becomes much simpler. The simplified loss function focuses on the noise prediction task:

$$\mathrm{L}_{\mathrm{simple}} = \mathrm{E}_{\mathrm{t,x_0,\varepsilon}}[|\varepsilon - \varepsilon_\theta(\mathrm{x_t}, \mathrm{t})|^2], \tag{2.5}$$

where $\epsilon_\theta(\mathrm{x_t}, \mathrm{t})$ is the noise added at step t, and ϵ_θ is the predicted noise from the neural network. This loss function is computationally efficient and has been shown through experiments to work better than the full variational objective.

The timestep t is typically sampled uniformly from $\{1, 2, \ldots, T\}$, and the loss is computed using mean squared error. This compares the true noise with the predicted noise values. This approach transforms the complex generative modeling problem into a relatively simple denoising task, making training stable and efficient.

2.2.5 Sampling Methods

Standard sampling from diffusion models follows the reverse process by iteratively denoising from pure Gaussian noise. The process starts with $\mathrm{x_T} \sim \mathcal{N}(0, \mathrm{I})$ and gradually denoises over T steps. At each step, we sample:

$$\mathrm{x_{t-1}} = \mu_\theta(\mathrm{x_t}, \mathrm{t}) + \sigma_\mathrm{t} \cdot \mathrm{z}, \quad \mathrm{z} \sim \mathcal{N}(0, \mathrm{I}), \tag{2.6}$$

where $\mathrm{z} \sim \mathcal{N}(0, \mathrm{I})$ is random noise and σ_t is the noise variance at step t. The random nature of this sampling process means that multiple diverse samples can be generated from the same starting noise.

However, this ancestral sampling approach has a major drawback: it requires T forward passes through the neural network, where T is typically 1000 or more. This makes sampling computationally expensive and slow. This limits the practical applications of diffusion models, especially in real-time scenarios.

2.2.6 DDIM: Accelerated Sampling

DDIM (Denoising Diffusion Implicit Models) addresses the computational bottleneck of standard diffusion sampling by proposing a deterministic sampling process [14]. Unlike the stochastic DDPM sampling, DDIM provides a family of non-Markovian processes that can generate high-quality samples with fewer steps.

The key insight of DDIM is that the forward process can be extended to a broader class of processes while keeping the same training objective. The DDIM update rule is:

$$\mathrm{x}_{\mathrm{t}-1} = \sqrt{\overline{\alpha_{\mathrm{t}-1}}} \cdot \frac{\mathrm{x}_{\mathrm{t}} - \sqrt{1 - \overline{\alpha_{\mathrm{t}}}} \cdot \varepsilon_{\theta}(\mathrm{x}_{\mathrm{t}}, \mathrm{t})}{\sqrt{\overline{\alpha_{\mathrm{t}}}}} + \sqrt{1 - \overline{\alpha_{\mathrm{t}-1}}} \cdot \varepsilon_{\theta}(\mathrm{x}_{\mathrm{t}}, \mathrm{t}). \tag{2.7}$$

This deterministic formulation allows skipping timesteps during sampling. Instead of using all T timesteps, DDIM can use a subset of timesteps (e.g., 50 or 100 steps) while keeping comparable sample quality. This acceleration makes diffusion models more practical for applications requiring fast inference, such as real-time 3D pose estimation and smoothing tasks.

2.3 Transformer

2.3.1 Transformer Architecture and Self-Attention

The Transformer architecture, introduced by Vaswani et al. [15], has become a cornerstone of modern deep learning. Transformers represent a basic shift in neural network design. While traditional RNNs process information through recurrent operations and CNNs rely on convolution, Transformers remove both approaches entirely. They use only self-attention mechanisms to handle information processing.

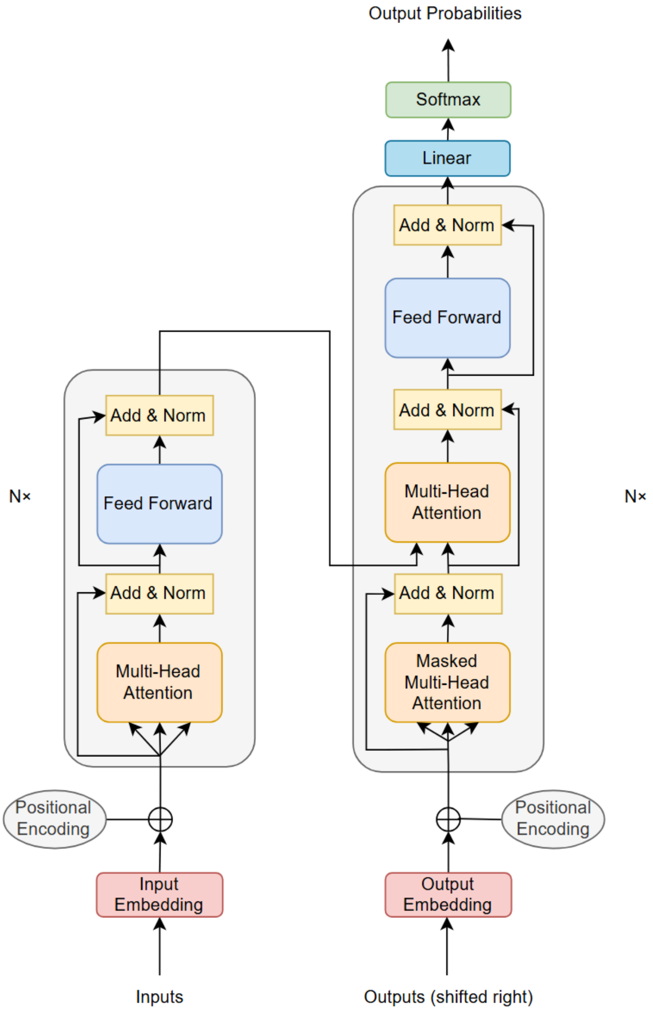


Fig. 2 The Transformer architecture.

The core of the self-attention mechanism lies in computing relationships between different positions in a sequence through Query (Q), Key (K), and Value (V) vectors. The attention is computed as:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}, \tag{2.8}$$

where d_k is the dimension of the key vectors. Multi-head attention builds on the basic attention mechanism. It allows the model to focus on information from multiple representation subspaces at the same time.

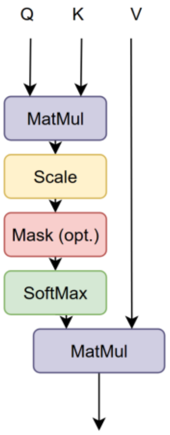


Fig. 3 Transformer’s attention computation.

The self-attention mechanism offers several advantages over sequential models. It enables parallel processing of all sequence elements, greatly improves training efficiency. Also, it can directly model long-range dependencies without the vanishing gradient problems that affect RNNs. These properties make Transformers particularly suitable for capturing complex spatial and temporal relationships in human pose sequences.

2.3.2 Positional Encoding Strategies

Since Transformers process sequences in parallel without built-in positional awareness, positional encoding becomes crucial for adding sequence order information. The original Transformer paper proposed sinusoidal positional encoding, which uses different functions of different frequencies:

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin\left(\frac{\mathrm{pos}}{10000^{2\mathrm{i}/\mathrm{d}_{\mathrm{model}}}}\right), \tag{2.9}$$

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i}+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{2\mathrm{i}/\mathrm{d}_{\mathrm{model}}}}\right), \tag{2.10}$$

where pos is the position, i is the index of dimension, and d_{model} is the model dimension.

The sinusoidal design has several good properties. It allows the model to learn relative positional relationships, as PE_{pos+k} can be represented as a linear function of PE_{pos} for any fixed offset k. This property allows the model to handle longer sequences than those used in training. The model can work beyond its original training data.

Other positional encoding strategies have appeared in recent work. Learned positional encoding treats position embeddings as trainable parameters [33]. Relative positional encoding methods, such as those proposed by Shaw et al. [34], directly encode relative distances between positions rather than absolute positions. More recent approaches like RoPE (Rotary Position Embedding) [35] and ALiBi (Attention with Linear Biases) [36] have shown improved extrapolation capabilities for longer sequences.

2.3.3 Transformers in Human Pose Estimation

The application of Transformers to human pose estimation has gained significant momentum in recent years. These models have shown superior performance in capturing spatiotemporal correlations compared to traditional LSTM-based and convolution-based approaches.

PoseFormer [22] represents a pioneering work that applies pure Transformer architecture to 3D human pose estimation. The model uses spatial-temporal transformers for comprehensive modeling. These transformers capture human joint relations within individual frames. They also model temporal correlations that occur across different frames. This approach removes the need for convolutional architectures entirely, and shows the effectiveness of attention mechanisms for pose estimation tasks.

For temporal smoothing and post-processing applications, Transformers offer particular advantages in modeling sequential dependencies without the computational bottlenecks of recurrent architectures. The parallel processing capability enables efficient handling of long pose sequences, making them well-suited for post-processing tasks that require global temporal context.

3 Proposed Method

3.1 Temporal Pose Diffusion Framework

3.1.1 Framework Overview

Our method extends diffusion models to temporal pose sequences, addressing the limitation of existing approaches that process poses independently. While methods like ZeDO [6] achieve excellent single-frame refinement, they cannot ensure temporal consistency across sequences. Similarly, temporal modeling approaches like SmoothNet [10] focus on post-processing but lack the powerful pose priors that diffusion models provide.

We formulate temporal pose refinement as a conditional sequence-to-sequence denoising task. The key insight is that predicted pose sequences, despite containing errors, provide valuable structural information that can guide the diffusion process toward anatomically plausible and temporally consistent solutions.

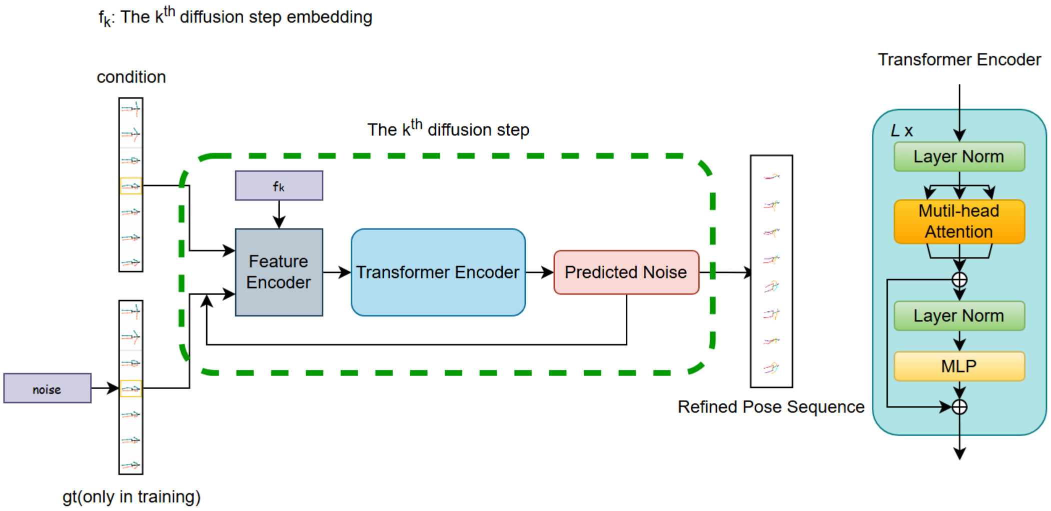


Fig. 4 The structure of proposed method.

Input: A predicted pose sequence $X^{pred} = \{x_1, x_2, \dots, x_T\} \in R^{T \times J \times 3}$ from any pose estimator, where T is sequence length and $J = 17$ is the number of joints following the Human3.6M skeleton structure.

Output: A refined pose sequence $X^{refined} \in R^{T \times J \times 3}$ with improved temporal consistency and accuracy.

Unlike unconditional diffusion models that start from pure noise, our approach conditions the entire denoising process on predicted poses. This conditioning mechanism serves two purposes. It limits the solution space to anatomically reasonable poses, and keeps the general motion pattern while correcting errors.

The framework operates in two phases. During training, we learn to predict noise in ground truth pose sequences while conditioning on corresponding predicted poses. During inference, we start from random noise and gradually denoise while being guided by the predicted sequence, ultimately producing a refined version that maintains the original motion semantics while improving accuracy and temporal smoothness. The overall framework is illustrated in Fig. 4, which shows how predicted poses serve as conditioning information throughout the iterative denoising process.

3.1.2 Forward Process for Pose Sequences

Following the standard DDPM framework [13], our forward process systematically corrupts ground truth pose sequences by gradually adding Gaussian noise. However, we extend this process to handle temporal pose sequences while preserving their structural properties.

For a ground truth pose sequence $x_0 \in R^{T \times J \times 3}$, we can directly sample noisy versions at any timestep t using the closed-form expression:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon, \tag{3.1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is random noise with the same shape as the pose sequence, and $\overline{\alpha_t}$ follows the standard noise schedule.

An important consideration in our design is that the noise addition process preserves the temporal structure of the sequence. Each pose frame is corrupted independently, but the overall sequence dimensionality remains consistent, allowing our model to learn both spatial (joint relationships) and temporal (motion) patterns during the reverse process.

3.1.3 Conditional Reverse Process

The reverse process is where our key innovation lies. Instead of learning an unconditional denoising process, we condition every denoising step on the predicted pose sequence. This conditioning provides important guidance that helps the model generate refined poses that are both accurate and consistent with the original motion.

We model the conditional reverse process as:

$$p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2 I), \tag{3.2}$$

where $c = X^{pred}$ represents the conditioning information (predicted poses). The key difference from standard diffusion models is the explicit conditioning term c that appears in every reverse step.

The conditioning mechanism works by combining predicted pose information at multiple levels within our denoising network. Rather than simply concatenating the predicted poses, we use dedicated encoding pathways that allow the model to selectively attend to relevant conditioning information while maintaining the flexibility to correct errors.

During training, we learn to predict the noise that was added to ground truth sequences while conditioning on corresponding predicted poses:

$$L = E_{t,x_0,\epsilon}[|\epsilon - \epsilon_\theta(x_t, t, c)|^2], \tag{3.3}$$

where $\epsilon_\theta(x_t, t, c)$ is our temporal denoising network that predicts noise conditioned on both the current noisy sequence and the predicted poses.

During inference, we start from random noise $x_T \sim \mathcal{N}(0, I)$ and iteratively apply the conditional reverse process for T steps to generate the final refined pose sequence. For faster inference, we support DDIM sampling [14] which maintains quality while reducing the number of required denoising steps from 1000 to typically 50-100 steps. A complete overview of our temporal pose diffusion pipeline is shown in Fig. 4, highlighting the conditional reverse process and the Transformer-based denoising architecture.

The conditioning ensures that the refined sequence maintains the overall motion characteristics of the input while correcting pose errors and improving temporal consistency. This approach is particularly effective because it leverages the structural information already present in the predicted poses rather than generating poses from scratch.

3.2 Temporal Denoising Architecture

3.2.1 Multi-Modal Feature Integration

Building upon the conditional diffusion framework described in Section 3.1, we now detail the architecture of our temporal denoising network $\epsilon_\theta(x_t, t, c)$. The network must effectively integrate three distinct sources of information: the current noisy sequence state, the diffusion timestep, and the condition poses. Each requires specialized encoding to maximize their contribution to the denoising process.

The diffusion timestep t indicates the current noise level and guides the denoising intensity. We employ sinusoidal positional embedding to encode the timestep, followed by a multilayer perceptron projection to the model dimension $d_{model} = 256$. This time embedding is broadcast across all sequence positions and joints, ensuring uniform temporal awareness throughout the network.

Both the noisy sequence x_t and conditioning poses c are projected from their original 3D coordinates to the model dimension through separate linear layers. This parallel encoding strategy allows the network to process current state and guidance information in the same feature space while maintaining their distinct semantic meanings.

To capture structural relationships, we incorporate two types of learnable positional embeddings. Temporal embeddings encode the sequential position of each frame, while joint embeddings encode the anatomical hierarchy of the human skeleton. These embeddings provide essential inductive biases for modeling human motion patterns.

All encoded features are integrated through element-wise addition. This additive fusion preserves the contribution of each modality while enabling the Transformer to learn optimal feature combinations through attention mechanisms.

3.2.2 Transformer-Based Sequence Modeling

The fused features undergo spatial-temporal modeling through a Transformer encoder. We reshape the feature tensor from $R^{T\times J\times 256}$ to a flattened token sequence $h_{tokens} \in R^{(T\cdot J)\times 256}$. Each token represents a specific joint at a specific time frame, creating $T \times 17$ tokens per sequence. This flattening enables the self-attention mechanism to capture both intra-frame joint dependencies and inter-frame temporal relationships within a unified framework.

Our Transformer encoder consists of 6 layers, each containing multi-head self-attention (8 heads) and feed-forward networks (dimension 1024). The self-attention mechanism computes relationships between all joint-time pairs, allowing the model to learn complex motion patterns such as kinematic chains and temporal smoothness constraints. Unlike convolutional or recurrent approaches that process information locally, the Transformer's global receptive field enables each joint to directly attend to any other joint at any time frame. This is particularly beneficial for capturing long-range dependencies in human motion, such as coordination between upper and lower body movements.

3.2.3 Noise Prediction and Output

The Transformer output undergoes final processing to generate noise predictions in the original sequence format. The processed tokens are projected back to 3D coordinates through a linear layer, then reshaped to reconstruct the sequence structure:

$$\epsilon_\theta(x_t, t, c) \in R^{T\times J\times 3}. \tag{3.4}$$

The predicted noise directly integrates with the loss function defined in Equation 3.3, enabling end-to-end training of the entire architecture. The network learns to predict the specific noise pattern that transforms the current noisy state toward the ground truth, guided by the conditioning information.

Inference Efficiency: During sampling, the network processes entire sequences in parallel rather than frame-by-frame, significantly improving computational efficiency compared to autoregressive approaches. The Transformer's parallel processing capability makes real-time pose refinement feasible for practical applications.

4 Experiments

4.1 Implementation Details

We test our model in PyTorch with mixed precision training using Adam optimizer (lr=4e-4) for 1000 epochs with batch size 512. The model uses early stopping based on validation loss with automatic best checkpoint saving.

Temporal sequences are extracted using sliding windows of 16 frames with stride 8 during training and stride 1 during evaluation. All poses are normalized by subtracting the root joint position for translation invariance.

The temporal denoising network employs a 6-layer Transformer encoder with 8 attention heads and dimension 256. We use 1000 diffusion timesteps with linear noise schedule (β_1=1e-4, β_t=0.02). For faster inference, DDIM sampling with 50 steps is supported.

During evaluation, we apply sliding window processing with weighted averaging for long sequences. The model processes sequences in parallel for computational efficiency.

4.2 Dataset and Metrics

We conduct experiments on the Human3.6M dataset [16], a 3D human dataset for 3D human pose estimation containing 3.6 million video frames with accurate 3D pose annotations. Following standard protocols in prior work, we use subjects S1, S5, S6, S7, S8 for training and subjects S9, S11 for testing.

We use pose predictions generated by ZeDO [6], a state-of-the-art single-frame 3D pose estimation method, as input to our temporal refinement approach.

ZeDO represents current best practices in cross-domain pose estimation without requiring 2D-3D paired training data, making it an ideal baseline for evaluating temporal consistency improvements.

We evaluate our method using three standard metrics for 3D pose estimation. MPJPE (Mean Per Joint Position Error) is the Euclidean distance between estimated and ground truth positions after root joint alignment, capturing absolute pose accuracy. P-MPJPE (Procrustes-aligned MPJPE) evaluates pose accuracy after optimal rigid alignment between estimated and ground truth positions, focusing on pose structure by removing global transformations.

MPJAE (Mean Per Joint Acceleration Error) measures temporal smoothness by computing acceleration differences between predicted and ground truth sequences using second-order finite differences. This metric is particularly important for evaluating the temporal consistency improvements achieved by our method.

All metrics are reported in millimeters for position errors and mm/frames^2 for acceleration errors, following standard conventions in the pose estimation literature.

4.3 Experimental Result

We compare our temporal pose diffusion method against the baseline ZeDO predictions and SmoothNet [10], a temporal refinement method, on the Human3.6M test set. Table 1 summarizes the quantitative results across all evaluation metrics. The best results are highlighted in bold font.

Table 1 Quantitative comparison on Human3.6M test set. All values in mm for MPJPE/P-MPJPE and mm/frames² for MPJAE.

	MPJPE	**P-MPJPE**	**MPJAE**
ZeDO (Baseline)	54.77	37.48	2.52
ZeDO + SmoothNet	53.91	37.45	**0.98**
Ours	**38.91**	**27.18**	2.58

Our method achieves significant improvements over both the baseline ZeDO predictions and the SmoothNet refinement approach. Compared to ZeDO, our method reduces MPJPE by 15.86 mm, and P-MPJPE by 10.30 mm. The comparison with SmoothNet demonstrates that our diffusion-based approach provides superior pose accuracy.

Table 2 provides detailed results for individual actions to analyze where our method provides the most benefit.

Table 2 Per-action results on Human3.6M test set. MPJPE values in mm.

	ZeDO	**Ours**	**Improvement(%)**
Directions	47.75	40.70	14.76
Discussion	47.47	41.25	13.10
Eating	55.33	34.57	37.52
Greeting	58.88	38.31	34.95
Phoning	53.72	39.46	26.55
Photo	67.01	43.59	34.95
Posing	50.91	43.53	14.49
Purchases	45.95	35.99	21.68
Sitting	68.02	39.79	41.5

SittingDown	77.13	40.78	47.13
Smoking	51.83	38.97	24.82
Waiting	57.01	38.90	31.76
WalkDog	57.06	37.72	33.90
WalkTogether	37.55	37.51	0.12
Walking	46.01	32.62	29.09
Average	54.77	38.91	28.96

The results show consistent improvements across all actions, with particularly notable gains in dynamic actions where temporal modeling is most beneficial.

4.4 Ablation Studies

We conduct ablation studies to test the effectiveness of components in our method.

Table 3 Ablation study results on Human3.6M.

	MPJPE	**P-MPJPE**	**MPJAE**	**Inference FPS**
16 frames + DDPM	38.91	27.18	2.58	13.9
16 frames + DDIM	39.09	27.23	10.45	306.2
32 frames + DDPM	38.21	27.02	2.38	7.0
32 frames + DDIM	37.77	26.86	7.81	151.5

Increasing sequence length from 16 to 32 frames provides consistent improvements in pose accuracy across both sampling methods. For DDPM sampling, MPJPE improves from 38.91mm to 38.21mm, while for DDIM sampling, the improvement is more significant, reducing from 39.09mm to 37.77mm. This indicates that longer temporal context enables better pose refinement, with the effect being more pronounced for deterministic DDIM sampling.

DDIM sampling demonstrates a clear speed-accuracy trade-off. The 16-frame DDIM configuration achieves a dramatic 22× speedup (306.2 vs 13.9 FPS) compared to DDPM, but suffers from significant temporal smoothness degradation, with MPJAE increasing from 2.58 to 10.45 mm/frames^2. This temporal quality loss manifests as visible jittering in the output sequences.

The 32-frame configurations reveal an interesting pattern. While 32-frame DDPM achieves the best temporal smoothness (2.38 mm/frames^2) with reduced inference speed (7.0 FPS), the 32-frame DDIM provides the best pose accuracy (37.77mm MPJPE) while maintaining moderate temporal quality (7.81 mm/frames^2) and reasonable speed (151.5 FPS).

For practical deployment, the choice depends on application requirements. The 16-frame DDPM configuration offers the best balance for scenarios prioritizing temporal consistency, while 32-frame DDIM provides superior pose accuracy for applications where slight temporal artifacts are acceptable in exchange for better overall quality and faster inference.

4.5 Qualitative Results

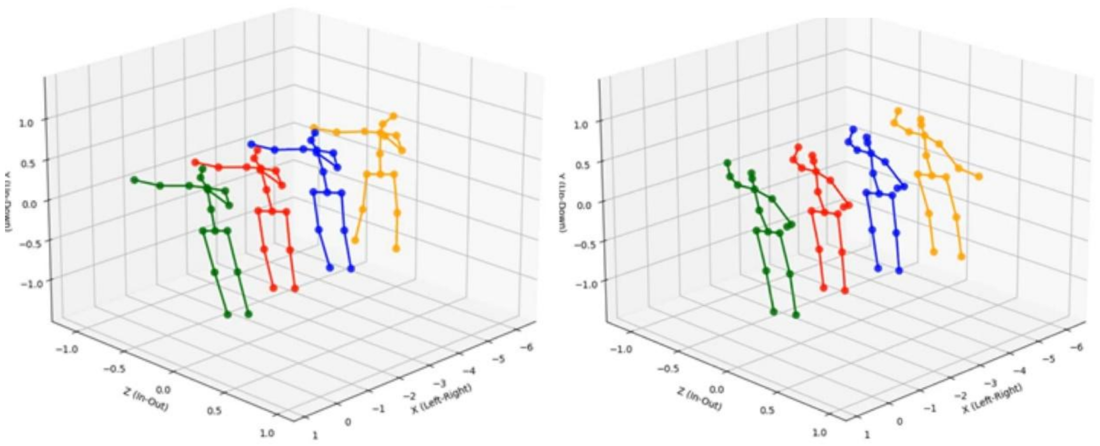


Fig. 5 Cases of refinement results' visualization.

Visualization examples of pose sequence refinement is presented in Fig. 5. In the figure, the green skeleton represents the ground truth poses, the yellow skeleton shows ZeDO's estimated poses, and the red and blue skeletons correspond to the refined results from our model using DDPM and DDIM sampling, respectively. As can be observed, although ZeDO's estimated poses exhibit significant errors compared to the ground truth like right hand or left foot, our proposed method is able to effectively correct these errors and substantially improve pose estimation accuracy. Both DDPM and DDIM variants demonstrate superior performance in maintaining anatomical plausibility while preserving the overall motion characteristics of the sequence.

5 Conclusion and Future Work

5.1 Conclusion

We present a novel temporal pose diffusion framework that addresses the critical limitation of existing single-frame pose refinement methods. By extending diffusion models to temporal sequences and conditioning the denoising process on predicted poses, our approach achieves significant improvements in both pose accuracy and temporal consistency.

Our method formulates temporal pose refinement as a conditional sequence-to-sequence denoising task, where predicted pose sequences guide the diffusion process toward anatomically plausible and temporally consistent solutions. The key innovation lies in the multi-modal feature integration strategy that effectively combines time embeddings, pose representations, condition information, and positional encodings within a Transformer-based architecture.

Experimental results on Human3.6M demonstrate substantial improvements over baseline methods. Compared to ZeDO predictions, our approach reduces both MPJPE and P-MPJPE. The method also outperforms SmoothNet, a state-of-the-art temporal refinement approach, across all evaluation metrics.

Our ablation studies reveal important design insights. Longer sequence lengths provide consistent accuracy improvements, particularly when combined with DDIM sampling. While DDIM sampling offers significant speed advantages, it introduces temporal quality trade-offs that can be mitigated through longer temporal context.

The proposed framework maintains the plug-and-play nature essential for practical deployment, requiring no retraining when applied to different pose estimators. This cross-domain adaptability, combined with the substantial performance improvements, makes our method valuable for both research and real-world applications requiring high-quality temporal pose sequences.

5.2 Future Work

Several promising directions could further advance temporal pose refinement research.

Adaptive sequence lengths that automatically adjust based on motion complexity could optimize the trade-off between accuracy and computational efficiency. Different actions may benefit from varying temporal contexts, and a dynamic approach could better utilize computational resources while maintaining quality.

More sophisticated attention-based conditioning strategies could enhance the guidance provided by predicted poses. Cross-attention mechanisms between noisy sequences and conditioning information might enable more selective error correction compared to the current feature fusion approach.

The framework could be extended to handle variable-length inputs, eliminating the need for fixed-window processing and enabling end-to-end refinement of complete video sequences. Additionally, incorporating multi-view information when available could enhance robustness in challenging scenarios.

The temporal diffusion approach shows potential for other sequential pose-related tasks such as motion prediction, pose completion for missing frames, or joint optimization of pose estimation and tracking. The general framework of conditioning diffusion processes on imperfect predictions may prove valuable across various computer vision domains.

Finally, integrating physics-based constraints within the diffusion process could further improve the plausibility of refined poses, particularly for complex motions involving contact or interaction with the environment.

List of Publication

[1] Xinyi Liu, and Hiroshi Watanabe, “A Study on 3D Human Pose Estimation via Diffusion Models,” ITE Annual Convention, 2025 (to appear).

Bibliography

[1] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, “A review of 3D human pose estimation algorithms for markerless motion capture,” *Computer Vision and Image Understanding*, vol. 212, p. 103275, 2021.

[2] R. B. Neupane, K. Li, and T. F. Boka, “A survey on deep 3D human pose estimation,” *Artificial Intelligence Review*, vol. 58, no. 1, p. 24, 2024.

[3] J. Wang *et al.*, “Deep 3D human pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.

[4] M. Rayat Imtiaz Hossain and J. J. Little, “Exploiting temporal information for 3D pose estimation,” *arXiv preprint arXiv:1711.08585*, 2017.

[5] Y. Li, K. Li, X. Wang, and R. Y. Da Xu, “Exploring temporal consistency for human pose estimation in videos,” *Pattern Recognition*, vol. 103, p. 107258, 2020.

[6] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, “Back to optimization: Diffusion-based zero-shot 3d human pose estimation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6142–6152, 2024.

[7] K. Holmquist and B. Wandt, “Diffpose: Multi-hypothesis human pose estimation using diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15977–15987, 2023.

[8] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, “Diffpose: Toward more reliable 3d pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051, 2023.

[9] Z. Luo, S. A. Golestaneh, and K. M. Kitani, “3d human motion estimation via motion compression and refinement,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[10] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, “Smoothnet: A plug-and-play network for refining human poses in videos,” in *Proceedings of the European Conference on Computer Vision*, pp. 625–642, 2022.

[11] M. Véges and A. Lőrincz, “Temporal smoothing for 3D human pose estimation and localization for occluded people,” in *Proceedings of the Neural Information Processing: 27th International Conference*, pp. 557–568, 2020.

[12] Z. Niu, K. Lu, J. Xue, H. Ma, and R. Wei, “Multi-view 3D smooth human pose estimation based on heatmap filtering and spatio-temporal information,” in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 442–450, 2021.

[13] J. Ho, A. Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” *Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[14] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.

[15] A. Vaswani *et al.*, “Attention is all you need,” *Neural Information Processing Systems*, vol. 30, 2017.

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[17] C. Ionescu and F. Li, "Latent Structured Models for Human Pose Estimation," in *Proceedings of the International Conference on Computer Vision,* 2011.

[18] H. Ci, X. Ma, C. Wang, and Y. Wang, "Locally connected network for monocular 3D human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 3, pp. 1429–1442, 2020.

[19] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 7753–7762, 2019.

[20] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 7103–7112, 2018.

[21] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 3425–3435, 2019.

[22] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 11656–11665, 2021.

[23] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 13147–13156, 2022.

[24] R. Liu, J. Shen, H. Wang, C. Chen, S. Cheung, and V. K. Asari, "Enhanced 3D human pose estimation from videos by using attention-based neural network with dilated convolutions," *International Journal of Computer Vision,* vol. 129, no. 5, pp. 1596–1615, 2021.

[25] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 13232–13242, 2022.

[26] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision,* pp. 2640–2649, 2017.

[27] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proceedings of the European Conference on Computer Vision,* pp. 561–578, 2016.

[28] I. T. Young and L. J. Van Vliet, "Recursive implementation of the Gaussian filter," *Signal Processing,* vol. 44, no. 2, pp. 139–151, 1995.

[29] W. H. Press and S. A. Teukolsky, "Savitzky-Golay smoothing filters," *Computer Physics,* vol. 4, no. 6, pp. 669–672, 1990.

[30] G. Casiez, N. Roussel, and D. Vogel, "1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 2527–2530, 2012.

[31] D. Mehta *et al.*, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *Proceedings of the 2017 International Conference on 3D Vision (3DV),* pp. 506–516, 2017.

[32] I. J. Goodfellow *et al.*, "Generative adversarial nets," *Neural Information Processing Systems,* vol. 27, 2014.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics,* vol. 1, pp. 4171–4186, 2019.

[34] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” *arXiv preprint arXiv:1803.02155,* 2018.

[35] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing,* vol. 568, p. 127063, 2024.

[36] O. Press, N. A. Smith, and M. Lewis, “Train short, test long: Attention with linear biases enables input length extrapolation,” *arXiv preprint arXiv: 2108.12409,* 2021.