DiffusionPose: Markov-Optimized Diffusion Model for Human Pose Estimation

Zhigang Wang[1], Zhenguang Liu[1,2,3], Shaojing Fan[4], Sifan Wu[5,6], Yingying Jiao[7]*,

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Shandong Rendui Network Co., Ltd.
[3]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[4]Department of Electrical and Computer Engineering, National University of Singapore
[5]College of Computer Science and Technology, Jilin University
[6]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
[7]College of Computer Science and Technology, Zhejiang University of Technology
{wangzhigang2024, liuzhenguang2008, fanshaojing, wusifan2021, yingyingjiao21}@gmail.com

Abstract

Video-based human pose estimation has long been a non-trivial task due to its dynamic nature and challenging detection scenarios such as occlusion and defocus. Inspired by the success of diffusion models, researchers have applied them to video pose estimation, outperforming traditional joint detection methods. However, existing diffusion model-based methods still face challenges like slow convergence and unstable pose generation. To tackle these issues, we propose DiffusionPose, a novel framework for video pose estimation that integrates diffusion models with optimization strategies: (1) We combine the emerging Mamba with Transformers to balance global and local spatio-temporal modeling. (2) We integrate Markov Random Fields into the reverse diffusion process to enhance the denoising of pose heatmaps, particularly addressing the issue of erroneous generation of occluded joints. (3) We mathematically formulate a Markov objective to supervise the heatmap denoising process, enabling the model to generate anatomically plausible skeletons. Our method achieves state-of-the-art performance on three large-scale benchmark datasets. Interestingly, it shows notable robustness in challenging video scenarios, improving the accuracy of the most difficult ankle joint by 16.9% compared to the previous best diffusion model-based method on the Challenging-PoseTrack dataset.

Introduction

Human pose estimation, which involves accurately locating the anatomical keypoints of the human body, has become a fundamental field in the realm of computer vision and artificial intelligence (Geng et al. 2023). With the growing availability of video data, video-based human pose estimation has gained increasing attention for its effectiveness in dynamic motion understanding and temporal consistency (Dubey and Dixit 2023). This advancement enables a wide range of applications, including *action recognition*, *human-computer interaction*, *sports analytics*, and *healthcare monitoring* (Liu et al. 2020; Su et al. 2021; Wu et al. 2024b).

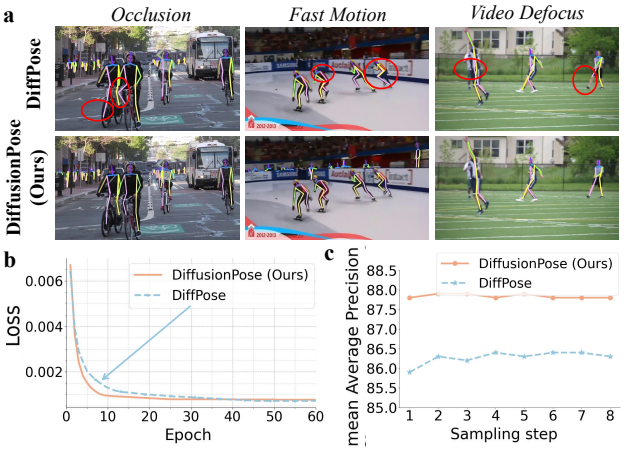

Figure 1: We propose DiffusionPose, a novel video-based pose estimation framework that leverages diffusion models with joint-aware optimization. Sub-figure (a) compares DiffPose (Feng et al. 2023b) and our method in challenging scenes, where DiffusionPose shows greater accuracy and robustness by modeling joint relationships via Markov random fields. Red circles highlight erroneous predictions. Sub-figures (b) and (c) show that DiffusionPose converges within 10 epochs, compared to 30 for DiffPose, and achieves high accuracy in a single step through optimized denoising and limb-aware supervision.

Despite the growing interest in video-based pose estimation, the majority of existing research (Sun et al. 2019; Yuan et al. 2021) focuses on estimating human pose from static images, lacking the robustness and capability to address the inevitable blurriness found in video frames. Naively applying image-based methods to video data, without accounting for temporal continuity, readily degrades performance due to motion blur and occlusion.

To overcome the limitations of image-based pose estimation, many video-based methods (Liu et al. 2021; Feng et al. 2023a) leverage temporal cues from adjacent frames. Early works (Bertasius et al. 2019; Liu et al. 2021) use de-

*Corresponding author

formable convolutions to align poses across frames, while recent Transformer-based models (He and Yang 2024; Feng et al. 2023b) employ attention to model spatio-temporal dependencies. Emerging research has explored State Space Models (Gupta, Gu, and Berant 2022) (SSMs) to address the drawbacks of Transformers by leveraging the linear complexity and long-range modeling capabilities of SSMs, yet their potential in video-based human pose estimation remains largely unexplored.

Recently, denoising diffusion probabilistic models (DDPMs) (Ho, Jain, and Abbeel 2020) have garnered significant attention due to their superior generative capabilities, facilitating their expansion into diverse applications. Motivated by advancements in other visual tasks (Chen et al. 2023a,b), DiffPose (Feng et al. 2023b) leads the way in applying DDPMs to human pose estimation, achieving impressive success. This approach transforms the general workflow of human pose estimation from pose detection in video frames to the conditional denoising generation of pose heatmaps. However, as illustrated in Fig. 1, DiffPose demonstrates slow convergence during training and requires multiple iterative refinement steps during inference to achieve optimal performance. We speculate that the reasons are twofold. (i) The encoder, which performs multi-frame full-sequence modeling using Transformers, lacks fine-grained spatio-temporal modeling conducted from the perspective of global-local separation. (ii) Its lookup-based strategy obtains only minimal pose priors from noisy heatmaps, leading to insufficient positional guidance during the denoising process. Additionally, when joints are occluded or blurred, existing methods like DiffPose and JMPose (Wu et al. 2024a) tend to produce unreasonable poses, as they fail to extract features of those joints and neglect spatial relationships among joints as well as the understanding of human anatomy.

Inspired by these insights, we propose a novel architecture based on the diffusion model, called **DiffusionPose**, which is designed to enhance the robustness and precision of human pose estimation in videos. Specifically, we embrace three innovative designs. (1) We design a hybrid encoder that combines Mamba and Transformer to perform spatio-temporal modeling and extract position and feature conditions from consecutive video frames. These two types of conditions provide explicit location guidance and rich informational cues during the reverse diffusion process, thereby facilitating the denoising process. (2) We theoretically explore the potential of Markov Random Fields (MRFs) in the context of the human pose estimation task and innovatively integrate MRFs into the Pose Denoising Decoder. A Markov optimization layer is designed and embedded to dynamically optimize the heatmap of each joint during pose generation based on the relationships among human joints. (3) Additionally, we mathematically formulate a Markov loss as a penalty term to constrain the length and angle of the generated human limbs, ensuring they conform to human anatomy and kinematics. Empirically, our approach significantly and consistently outperforms all state-of-the-art methods across three benchmarks.

Overall, the main contributions can be summarized as:

- We propose a novel hybrid Mamba-Transformer diffusion model framework that effectively integrates global temporal modeling with local spatial cues for video-based human pose estimation.
- To our knowledge, we are the first to optimize the denoising generation process of diffusion models from the perspective of Markov Random Fields. Our Markov optimization layers and loss penalty terms guide pose generation toward anatomically plausible human structures, effectively resolving occluded joint ambiguities.
- Our method establishes new benchmarks on three large-scale datasets and provides novel insights: robust condition guidance and task-specific optimizations during the reverse diffusion process enable top-tier results with minimal sampling steps.

Related Work

Human Pose Estimation. The field of human pose estimation in static images has seen a multitude of research efforts, transitioning from initial approaches that relied on tree-based and random forest models (Zhang et al. 2009) to contemporary techniques that leverage convolutional neural networks (Sun et al. 2019; Xiao, Wu, and Wei 2018) and Transformers (Xu et al. 2022; Yuan et al. 2021). However, while these methods are still effective as backbones for processing single frames of images, their performance inevitably deteriorates when applied to video-based human pose estimation. To capitalize on the additional temporal dimension, a multitude of approaches (Jin et al. 2023; Wu et al. 2024a; Bertasius et al. 2019; Wu et al. 2025) have begun to explore the spatio-temporal relationships between consecutive video frames, supplementing the temporal cues that static image methods are incapable of capturing.

State Space Models. Recently, State Space Models (SSMs) (Gupta, Gu, and Berant 2022) have shown strong efficiency in modeling long-range dependencies. Mamba (Gu and Dao 2023), with its hardware-efficient design and parallel scan–based S6 mechanism, surpasses Transformers on large-scale sequence modeling. Its effectiveness has motivated applications in vision tasks such as object detection (Huang et al. 2024), image segmentation (Xing et al. 2024), and video understanding (Li et al. 2025). Despite its absence in video-based human pose estimation, our work introduces the first SSM-based framework for this problem, designing a tailored architecture for effective spatio-temporal modeling.

Diffusion Model. Diffusion-based models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Zhang et al. 2025a) are powerful deep generative frameworks that model natural image distributions by reversing a Markov chain initialized from a standard Gaussian. They first demonstrated remarkable success in image synthesis (Saharia et al. 2022; Zhang et al. 2025b) and were later extended to discriminative tasks, including object detection (Chen et al. 2023a) and semantic segmentation (Chen et al. 2023b; Amit et al. 2021). DiffusionDet (Chen et al. 2023a) formulates object detection as a generative denoising process from noisy to precise bounding boxes. DiffPose (Feng et al. 2023b) further introduces

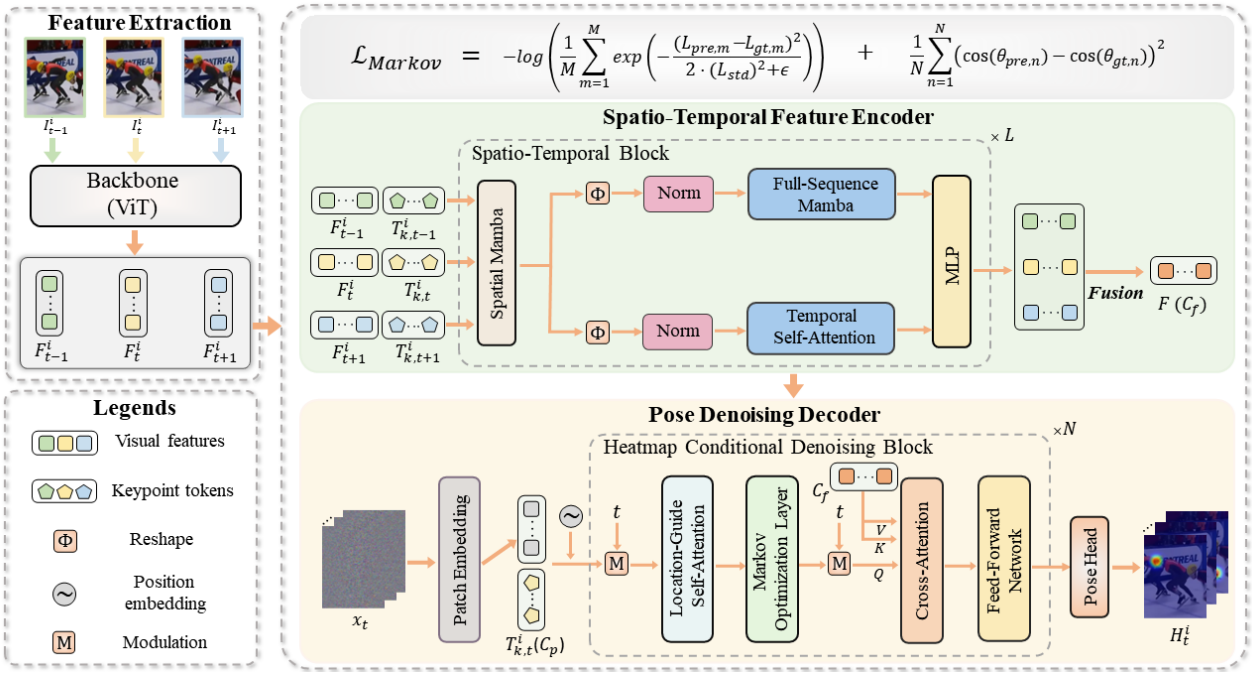

Figure 2: The overall pipeline of the proposed DiffusionPose framework. DiffusionPose leverages diffusion models for video-based human pose estimation with task-specific optimizations based on Markov Random Fields (MRFs).

a lookup-based interaction mechanism that combines noisy heatmaps with visual features to enable inductive denoising for human pose estimation. However, despite adopting diffusion models, DiffPose still suffers from slow convergence and inefficient multi-step denoising. To further exploit the potential of diffusion models for pose estimation, our framework introduces: (i) precise positional conditions and information-rich feature conditions, and (ii) Markov optimization with limb constraints.

Method

Preliminaries

Diffusion Model. Diffusion model-based approaches (Ho, Jain, and Abbeel 2020) generally encompass two key processes: a forward diffusion phase that introduces noise, and a reverse diffusion phase that denoises the data. The forward diffusion process involves continuously adding Gaussian noise to an original data sample until it eventually reaches a random noise distribution. For the task of human pose estimation, we consider the ground-truth pose heatmaps as the original data $\mathbf{x}_0$ which is transformed into a noisy sample $\mathbf{x}_t$ at a time step $t \in \{0, 1, \cdots, T\}$. After obtaining the noisy data $\mathbf{x}_t$, it is necessary to train a denoising network f_θ to perform the denoising process. Unlike conventional Diffusion models that output predicted noise maps, we are inspired by DiffusionDet (Chen et al. 2023a) and directly predict $\mathbf{x}_0$ from $\mathbf{x}_t$ through the denoising network.

Problem Formulation. In this approach, we adopt a top-down paradigm that starts by detecting the bounding boxes of all individuals within a video frame using a generic object detection model. To ensure consistency in tracking the same individual across subsequent frames, we expand each bounding box by 25%. This enlargement creates a temporal window that includes the key frame I_t, along with its two neighboring frames I_{t-1} and I_{t+1}. By doing this, we obtain the cropped video segment $\boldsymbol{\mathcal{I}}_{\boldsymbol{t}}^{\boldsymbol{i}} = \{I_{t-1}^i, I_t^i, I_{t+1}^i\}$ for person i, and the goal is to estimate the pose heatmaps for I_t^i. The overview framework of the DiffusionPose is depicted in Fig. 2. Given a frame sequence $\boldsymbol{\mathcal{I}}_{\boldsymbol{t}}^{\boldsymbol{i}}$, we first utilize a Vision Transformer (ViT) backbone to extract visual features $\{\boldsymbol{F}_{t-1}^i, \boldsymbol{F}_t^i, \boldsymbol{F}_{t+1}^i\}$, which are then fed into a Spatio-Temporal Feature Encoder (STFE). The STFE constructs a Mamba-Transformer hybrid architecture to perform decoupled spatio-temporal modeling to obtain two outputs: position conditions and feature conditions. These two conditions are then fed into a Pose Denoising Decoder (PDD) to assist in the denoising of poses during the reverse diffusion process. The PDD takes the $\mathbf{x}_t$ (*i.e.*, the noisy pose heatmaps) obtained from the forward diffusion process as the initial input and outputs the denoised generated heatmaps $\boldsymbol{H}_t^i$ for I_t^i by utilizing a denoising network.

Spatio-Temporal Feature Encoder

Inspired by the success of the Mamba model (Gu and Dao 2023; Zhu et al. 2024) in computer vision tasks, we explore its potential for efficient spatio-temporal modeling in video-based human pose estimation. However, after designing various architectures and conducting validation, we observed that while Mamba has a clear advantage in long sequence modeling, it lacks the ability to capture local details. Specifically, in recognizing the differences in the same spatial position across different frames, Mamba tends to be less effec-

tive than self-attention. Considering this, we design a parallel temporal modeling mechanism and dynamic fusion strategy to adaptively aggregate features extracted by Mamba and self-attention. To provide clear and semantically rich guidance for the subsequent reverse diffusion process, this module aims not only to output feature conditions like DiffPose (Feng et al. 2023b) but also position conditions.

Spatio-Temporal Block (STB). Given a visual feature sequence $\{\boldsymbol{F}_{t-1}^i, \boldsymbol{F}_t^i, \boldsymbol{F}_{t+1}^i\} \in \mathbb{R}^{3\times N\times D}$ output by the ViT backbone, we concatenate a learnable keypoint token sequence $\{\boldsymbol{T}_{k,t-1}^i, \boldsymbol{T}_{k,t}^i, \boldsymbol{T}_{k,t+1}^i\} \in \mathbb{R}^{3\times J\times D}$ to each feature, where J represents the number of human joints. The concatenated feature sequence then enters the Spatio-Temporal Block for spatio-temporal learning.

First, we utilize Spatial Mamba (SM) to undertake frame-level spatial modulation to model spatial relationships. Subsequently, the modulated features are reshaped into $\boldsymbol{F}_{full} \in \mathbb{R}^{(3\cdot(N+J))\times D}$ and $\boldsymbol{F}_{temp} \in \mathbb{R}^{(N+J)\times 3\times D}$, which, after normalization, are fed into Full-Sequence Mamba (FSM) and Temporal Self-Attention (TSA) for global-local temporal modeling, respectively. The Mamba module adheres to the standard computational paradigm of the original Mamba paper (Gu and Dao 2023) and, respecting the temporal unidirectionality, performs only forward (causal) scanning. This normalization is implemented via LayerNorm. Finally, a multilayer perceptron (MLP) restores the token count and performs a nonlinear mapping. After multiple stacked STB computations, we fuse multi-frame visual features into $\mathbf{F}$ as feature conditions $\mathbf{C}_f$ and extract the learned position information of keypoint tokens corresponding to the key frame I_t^i as position conditions $\mathbf{C}_p$.

Pose Denoising Decoder

A naive approach to performing the reverse diffusion process is to utilize a U-Net (Ronneberger, Fischer, and Brox 2015) as a denoising neural network, which takes the noisy maps and concatenated feature maps as input and outputs the predicted noise. DiffPose (Feng et al. 2023b) deviates from this paradigm by combining feature and noise maps for lookup-based interaction and initializing multiple sets of noise maps to achieve pose ensemble, thereby consolidating multiple predictions toward the final pose. However, this random independent generation of pose can easily lead to instability in the reverse diffusion process, especially in video scenes with severe occlusion and blur. The proposed DiffusionPose explores the relationships between human joints through Markov Random Fields, dynamically optimizing the state of each joint heatmap during the generation of pose. Furthermore, the decoder leverages the dual conditions emitted by the encoder to navigate the denoising of the noise map, bolstering the efficiency of the reverse diffusion process.

Heatmap Conditional Denoising Block. Our objective is to leverage the position conditions $\boldsymbol{C}_p$, feature conditions $\boldsymbol{C}_f$, and sampling step t as conditional inputs to guide the Pose Denoising Decoder in denoising the noisy pose heatmap $\mathbf{x}_t$ and producing the predicted heatmaps $\boldsymbol{H}_t^i$. First, we perform patch embedding and add a position embedding to $\mathbf{x}_t$, then feed the result into the Heatmap Conditional Denoising Block (HCDB). This block consists of four cascaded main layers: (i) A location-guide self-attention layer utilizes positional information from the position conditions $\boldsymbol{C}_p$ to guide the initial denoising of the pose heatmaps. (ii) A Markov optimization layer optimizes the pose location state by exploiting the kinematic relationships inherent in human joint connections. (iii) A cross-attention layer aggregates useful spatio-temporal cues from the feature conditions $\boldsymbol{C}_f$ to enrich the missing heatmaps. (iv) A feed-forward network for the final transformation. The sampling step t, after being projected into an embedding, modulates the location-guide self-attention layer and the cross-attention layer.

Markov Optimization Layer. Markov Random Fields (MRFs) (Cross and Jain 1983), a probabilistic graphical model, describes the interdependencies among a set of random variables that exhibit a particular spatial configuration. Specifically for human pose estimation, the nodes in this graph correspond to the joints of the human body, and the edges represent the connections between joints, *i.e.*, the limbs of the human body. The core characteristic of Markov Random Fields is the local Markov property. This property asserts that the distribution of a variable is contingent solely upon its own state and the states of its immediate neighbors, and is unaffected by the states of non-adjacent nodes. This relationship is typically represented through graph models. Leveraging this property, we optimize the state of each joint using only the anatomically adjacent joints. To this end, we design a node potential function ψ and an edge potential function φ that update the heatmap of each joint based on the current state of the joint and the states of the edges.

Given a node i, we first propose a message passing formula msg to quantify the influence of the state of node i and the states of its neighboring nodes:

$$\mathrm{msg}_{j\to i}\left(N_i\right) \propto \psi_i\left(N_i\right)\varphi_{ij}\left(N_i, N_j\right) \prod_{k\in\mathcal{N}g(j)\setminus i} \mathrm{msg}_{k\to j}\left(N_j\right), \quad (1)$$

$$\begin{aligned} \psi_i\left(N_i\right) &= N_i \otimes W_n, \\ \varphi_{ij}\left(N_i, N_j\right) &= \left(N_i \oplus N_j\right) \otimes W_e, \end{aligned} \quad (2)$$

where N_i represents the current state of node i. $\mathcal{N}g(j)$ denotes the set of neighboring nodes of node j. $\otimes$ stands for matrix multiplication. $\oplus$ signifies concatenation. W_n and W_e are learnable mapping matrices. After receiving messages from all neighboring nodes, node i updates its state to obtain $\overline{N}_i$.

$$\overline{N}_i = log(\sum_{j\in\mathcal{N}g(i)} exp(\mathrm{msg}_{j\to i}\left(N_i\right))) + N_i. \quad (3)$$

Markov Loss. We also design a Markov loss to control the length and angle of generated limbs so that the skeletal structure conforms to human anatomy in complex interaction scenarios. First, we solve for the coordinates of each joint $\boldsymbol{C}_t^i$ from the decoder output $\boldsymbol{H}_t^i$, and then calculate the predicted limb length and angle differences from the ground-truth based on the connections between joints. The specific

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseTracker (Girdhar et al. 2018)	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow (Xiu et al. 2018)	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
HRNet (Sun et al. 2019)	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
PoseWarper (Bertasius et al. 2019)	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose (Liu et al. 2021)	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
SLT-Pose (Gai et al. 2023)	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
HANet (Jin et al. 2023)	90.0	90.0	85.0	78.8	83.1	82.1	77.1	84.2
KPM (Fu et al. 2023)	89.5	90.0	87.6	81.8	81.1	82.6	76.1	84.6
FAMI-Pose (Liu et al. 2022)	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
DSTA (He and Yang 2024)	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
TDMI-ST (Feng et al. 2023a)	90.6	91.0	87.2	81.5	85.2	84.5	78.7	85.9
JM-Pose (Wu et al. 2024a)	90.7	91.6	87.8	82.1	85.9	85.3	79.2	86.4
TPSD-ViT (Zhang et al. 2025c)	90.7	91.1	87.9	83.6	85.3	86.3	80.0	86.7
DiffPose (Feng et al. 2023b)	89.0	91.2	87.4	83.5	85.5	87.2	80.2	86.4
DiffusionPose (Ours)	**90.8**	**91.8**	**88.7**	**84.9**	**88.1**	**88.6**	**81.8**	**87.8**

Table 1: Comparisons with the state-of-the-art methods for video-based human pose estimation on the validation set of the **PoseTrack2017** dataset (Iqbal, Milan, and Gall 2017).

formula is expressed as follows:

$$\mathcal{L}_{\text{Markov}} = \underbrace{-log\left(\frac{1}{M}\sum_{m=1}^{M} exp\left(-\frac{(L_{pre,m}-L_{gt,m})^2}{2\cdot(L_{std})^2+\epsilon}\right)\right)}_{\boldsymbol{length}} + \underbrace{\frac{1}{N}\sum_{n=1}^{N}\left(cos(\theta_{pre,n})-cos(\theta_{gt,n})\right)^2}_{\boldsymbol{angle}}, \tag{4}$$

where M and N represent the number of limb segments with length constraints and the number of connected limb pairs with angle constraints, respectively. $L_{\text{pre},m}$ and $L_{\text{gt},m}$ denote the predicted length and ground-truth length of the m-th limb segment, respectively. L_{std} is the standard deviation of the ground-truth lengths. ϵ is an extremely small positive number. $cos(\theta_{pre,n})$ and $cos(\theta_{gt,n})$ are the predicted cosine of the angle and the ground-truth cosine of the angle for the n-th pair of connected limbs, respectively.

Training and Inference Process

Training. We first perform a forward diffusion process to corrupt the ground-truth pose heatmaps into noisy heatmaps. At each step t, we pre-determined a noise parameter α_t according to a monotonically decreasing cosine scheme (Ho, Jain, and Abbeel 2020). Then, we train the Pose Denoising Decoder to denoise the heatmaps, thereby reversing the diffusion process. We use a standard pose estimation loss to supervise the final heatmaps, while also incorporating a Markov loss to regularize representation learning and the skeletal constraints.

$$\mathcal{L}_{\text{total}} = \left\|\boldsymbol{H}_t^i - \boldsymbol{G}_t^i\right\|_2^2 + \alpha\cdot\mathcal{L}_{\text{Markov}}, \tag{5}$$

where α is an adjustment factor.

Inference. At the beginning of the inference phase, we randomly sample Gaussian distribution heatmaps with an equal number of human joints as the initial input for the decoder. We employ DDIM (Song, Meng, and Ermon 2020) sampling, which allows for a variable number of denoising steps, and in this case, we only perform one denoising step, running the decoder once to obtain the generated final pose heatmaps. This minimal initialization and single-step denoising are made possible by the various design components of DiffusionPose, significantly enhancing its stability and efficiency.

Experiments

Datasets and Experimental Settings

We have conducted extensive experiments on three widely recognized large-scale benchmarks for video-based human pose estimation: PoseTrack2017 (Iqbal, Milan, and Gall 2017), PoseTrack2018 (Andriluka et al. 2018), and PoseTrack21 (Doering et al. 2022). These datasets include video sequences that capture complex scenarios featuring individuals with rapid movements and significant occlusion in crowded environments. To evaluate the performance of our model, we first calculate the Average Precision (AP) for each joint, and then compute the mAP across all joints. The ViT backbone has been pre-trained on the COCO dataset (Lin et al. 2014). We train it on a single RTX 4090 GPU for 20 epochs, rather than the 60 epochs like DiffPose (Feng et al. 2023b), utilizing PyTorch. The initial learning rate is set to 5e-4 and is reduced to one-tenth at the 16th epoch.

Comparison with State-of-the-art Approaches

Results on the PoseTrack2017 Dataset. We first benchmark our method on the PoseTrack2017 (Iqbal, Milan, and Gall 2017) dataset. A total of 14 methods are compared and their performances on the PoseTrack2017 validation set are summarized in Table 1. It can be observed that our DiffusionPose consistently outperforms all state-of-the-art methods. The final performance achieved an mAP of 87.8, surpassing the most advanced DiffPose (Feng et al. 2023b), which is also based on DDPM, by 1.4 mAP. Furthermore, for challenging joints like the wrist and ankle, DiffusionPose has demonstrated impressive results, with mAP of 84.9 (↑ 1.4) and 81.8 (↑ 1.6) respectively. These consistent and promising improvements underscore the effectiveness of condition guidance and specific optimizations in the denoising process for DiffusionPose.

Results on the PoseTrack2018 Dataset. We further evaluate our DiffusionPose on the PoseTrack2018 dataset (Andriluka et al. 2018), comparing it with 10 existing methods.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
AlphaPose (Fang et al. 2017)	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
PoseWarper (Bertasius et al. 2019)	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
DCPose (Liu et al. 2021)	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose (Liu et al. 2022)	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
HANet (Jin et al. 2023)	86.1	88.5	84.1	78.7	79.0	80.3	77.4	82.3
DSTA (He and Yang 2024)	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
TDMI-ST (Feng et al. 2023a)	86.7	88.9	85.4	80.6	82.4	82.1	77.6	83.6
JM-Pose (Wu et al. 2024a)	86.6	88.7	86.0	81.6	83.3	83.2	78.2	84.1
TPSD-ViT (Zhang et al. 2025c)	86.9	89.0	86.0	81.6	83.3	83.3	78.0	84.2
DiffPose (Feng et al. 2023b)	85.0	87.7	84.3	81.5	81.4	82.9	77.6	83.0
DiffusionPose (Ours)	**87.1**	**90.2**	**86.8**	**83.7**	**83.8**	**84.7**	**80.5**	**85.4**

Table 2: Comparisons with the state-of-the-art methods for video-based human pose estimation on the validation set of the **PoseTrack2018** dataset (Andriluka et al. 2018).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
DCPose (Liu et al. 2021)	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose (Liu et al. 2022)	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
DSTA (He and Yang 2024)	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5
TDMI-ST (Feng et al. 2023a)	86.8	87.4	85.1	81.4	83.8	82.7	78.0	83.8
JM-Pose (Wu et al. 2024a)	85.8	88.1	85.7	82.5	84.1	83.1	78.5	84.0
TPSD-ViT (Zhang et al. 2025c)	87.7	88.0	85.0	81.7	83.4	82.8	78.3	84.1
DiffPose (Feng et al. 2023b)	84.7	85.6	83.6	80.8	81.4	83.5	80.0	82.9
DiffusionPose (Ours)	**88.1**	**89.2**	**86.1**	**83.7**	**84.9**	**84.8**	**81.2**	**85.1**

Table 3: Comparisons with the state-of-the-art methods for video-based human pose estimation on the validation set of the **PoseTrack2021** dataset (Doering et al. 2022).

As shown in Table 2, DiffusionPose surpasses all state-of-the-art methods, achieving the top-performing result of 85.4 mAP. The results for wrist and ankle joints once again illustrate a significant improvement over DiffPose (Feng et al. 2023b), with 83.7 (↑ 2.2) and 80.5 (↑ 2.9) respectively.

Results on the PoseTrack2021 Dataset. The performance of our model, when compared to previous state-of-the-art methods on the PoseTrack21 dataset, is detailed in Table 3. Our model continues to push the boundaries of performance, achieving an overall mAP of 85.1 on the PoseTrack2021 validation dataset (Doering et al. 2022), outperforming DiffPose by 2.2 mAP. Notably, our approach maintains its dominance across all joints, improving the previous best results by 1.2 mAP at the wrist, attaining 83.7, and by 1.2 mAP at the ankle, reaching 81.2.

Results on the Challenging-PoseTrack Dataset. We also evaluate the robustness of our model on the Challenging-PoseTrack dataset collected and set up by JM-Pose (Wu et al. 2024a). This dataset only includes videos involving complex spatio-temporal interactions, such as rapid motion and severe occlusion. It is evident from the Table 4 that the proposed DiffusionPose significantly outperforms other leading methods, obtaining a gain of **4.3 mAP** over DiffPose, which achieves a total of 59.7 mAP. Additionally, remarkable progress has been made on the wrist and ankle joints, with improvements of **4.7** and **7.0 mAP**, respectively. These results strongly demonstrate that the condition guidance and Markov processes contribute to the stability of pose denoising generation while enhancing the robustness and precision for human pose estimation. The comparison between our method and DiffPose (Feng et al. 2023b) in challenging scenarios can refer to Fig. 1. In Fig. 3, we further demonstrate the superiority of our approach on complex interactive scenarios.

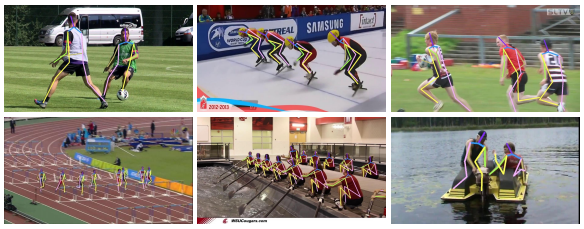

Figure 3: Visual results of our DiffusionPose in challenging scenes, such as those involving fast motion and occlusions.

Ablation Study

In this section, we first carry out extensive ablation studies centered on evaluating the effect of individual components within the DiffusionPose framework, encompassing the Spatio-Temporal Feature Encoder (STFE) and the Pose Denoising Decoder (PDD). We additionally investigate the efficacy of each sub-design incorporated in these components. All experiments are performed on the PoseTrack2017 dataset.

Study on components of DiffusionPose. We experimentally evaluate the effectiveness of each component in our DiffusionPose, depicting the results in Table 5. Firstly, we establish a baseline for this experiment by constructing a Vision Transformer (ViT) as the backbone for feature extraction and applying a standard U-Net (Ronneberger, Fischer, and Brox 2015) as the denoising network. (a) Integrating the Spatio-Temporal Feature Encoder after the backbone to extract spatio-temporal cues for generating two conditions yields a substantial gain of 4.5 mAP. This substantial progress indicates that this encoder can effectively model temporal continuity and spatial correlation, thereby facilitating human pose estimation. (b) In the next setup, we exclusively replace the U-Net with the Pose Denoising De-

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
HRNet-W48 (Sun et al. 2019)	61.1	56.2	48.0	39.5	46.4	34.2	32.0	46.4
PoseWarper (Bertasius et al. 2019)	65.2	58.7	49.6	40.9	47.8	36.8	30.4	48.0
DCPose (Liu et al. 2021)	67.5	60.4	49.7	40.0	50.1	37.3	30.0	48.7
FAMI-Pose (Liu et al. 2022)	69.3	62.2	50.0	43.7	49.4	40.2	38.0	51.6
TDMI (Feng et al. 2023a)	71.8	63.4	53.7	46.0	53.5	44.3	39.6	54.4
JM-Pose (Wu et al. 2024a)	71.9	65.7	56.4	47.4	56.0	45.6	42.0	56.1
DiffPose (Feng et al. 2023b)	71.3	64.2	56.8	47.9	54.9	46.6	41.5	55.3
DiffusionPose (Ours)	**77.9**	**68.8**	**58.3**	**52.6**	**60.7**	**51.2**	**48.5**	**59.7**

Table 4: Comparisons with the state-of-the-art methods on the **Challenging-PoseTrack** dataset.

Method	Spatio-Temporal Feature Encoder	Pose Denoising Decoder	Mean
Baseline	✗	✗	81.2
(a)	✓	✗	85.7 (↑ 4.5)
(b)	✗	✓	86.1 (↑ 4.9)
(c)	✓	✓	**87.8** (↑ 6.6)

Table 5: Ablation of different components in our **DiffusionPose**.

Method	Transformer Block	Spatio-Temporal Block	Mean
(a)	✗	✗	86.1
(b)	✓	✗	86.3 (↑ 0.2)
(c)	✗	✓	**87.8** (↑ 1.7)

Table 6: Ablation of various designs in the Spatio-Temporal Feature Encoder.

coder as the denoising network. The improved architecture achieves an mAP of 86.1, marking an increase of 4.9 mAP over the baseline. Such a significant boost in performance unequivocally proves the proficiency of the decoder, which is designed based on Markov Random Fields, in precisely optimizing the state of joint heatmaps and effectively controlling the generation of human pose. (c) Finally, by integrating both the customized encoder and decoder into our framework, we achieve a peak performance of 87.8 mAP. This result highlights the synergy of these components in enhancing performance.

Study on Spatio-Temporal Feature Encoder. Additionally, we explore the influence of sub-designs in the Spatio-Temporal Feature Encoder (STFE). Three experiments are performed and displayed in Table 6. (a) First, we remove the Spatio-Temporal Feature Encoder. (b) We replace our Spatio-Temporal Block with the vanilla Transformer block used by DiffPose (Feng et al. 2023b) for the extraction of spatio-temporal features. We observe a slight performance improvement, that is, an increase of 0.2 mAP, indicating that utilizing spatio-temporal modeling is beneficial. (c) We further introduce our novel hybrid block, which combines the strengths of long-range modeling and local focus from Mamba and self-attention, achieving a 1.5 mAP improvement. These results demonstrate that our method can effectively extract semantically rich spatio-temporal cues, enhancing subsequent pose generation.

Study on Pose Denoising Decoder. We then investigate the impact of various designs in the Pose Denoising Decoder on overall performance, as shown in Table 7. (a) The Spatio-Temporal Feature Encoder is adopted, but the decoder still uses only U-Net. (b) The U-Net is replaced with the Heatmap Conditional Denoising Block (HCDB), which currently includes only one cross-attention layer to mine the spatio-temporal information contained in the feature conditions $\boldsymbol{C}_f$ and a feed-forward network. (c) The location-guide self-attention layer is injected into the HCDB to learn explicit location guidance from the position conditions $\boldsymbol{C}_p$. (d) We then insert the Markov Optimization Layer (MOL) to optimize the pose heatmap features. (e) Finally, the Markov loss is added to the total loss as a penalty term to constrain the generation of reasonable human limbs. We can observe from this table that integrating only the feature conditions (b) results in a slight improvement of 0.3 mAP. When both the feature conditions and position conditions are applied simultaneously (c), better progress is achieved, reaching 86.5 mAP. Subsequently, applying the Markov Optimization Layer (d) further brings a performance improvement of 0.7 mAP, indicating that this design's profound understanding of the spatial relationships between joints is conducive to promoting more accurate human pose estimation. Finally, the combination of Markov loss (e) pushes performance to its peak, reaching a final 87.8 mAP for our DiffusionPose. These outcomes attest to the superiority of our condition guidance and the application of Markov theory, which constructs a stable and controllable heatmap generation process, enabling more accurate human pose estimation.

Method	$\boldsymbol{C}_f$	$\boldsymbol{C}_p$	MOL	Markov Loss	Mean
(a)	✗	✗	✗	✗	85.7
(b)	✓	✗	✗	✗	86.0 (↑ 0.3)
(c)	✓	✓	✗	✗	86.5 (↑ 0.8)
(d)	✓	✓	✓	✗	87.2 (↑ 1.5)
(e)	✓	✓	✓	✓	**87.8** (↑ 2.1)

Table 7: Ablation of designs in the Pose Denoising Decoder.

Conclusion

In this paper, we demonstrate the potential of diffusion models for video-based human pose estimation through novel task-specific optimizations using Markov Random Fields (MRFs). We introduce the Spatio-Temporal Feature Encoder (STFE), which balances local and global spatio-temporal modeling to capture richer complementary cues. Furthermore, we enhance the control guidance of MRFs for the reverse diffusion process by leveraging joint relationships for pose optimization during generation. To supervise plausible skeleton generation and integrate kinematic priors, we incorporate Markov loss items into the framework. Our method achieves state-of-the-art performance and demonstrates exceptional robustness across three benchmark datasets, offering valuable insights into leveraging diffusion models for challenging video tasks like human pose estimation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62372402) and the Key R&D Program of Zhejiang Province (No. 2025C01084).

References

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5167–5176.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning temporal pose estimation from sparsely-labeled videos. *Advances in neural information processing systems*, 32.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2023b. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 909–919.

Cross, G. R.; and Jain, A. K. 1983. Markov random field texture models. *IEEE Transactions on pattern analysis and machine intelligence*, (1): 25–39.

Doering, A.; Chen, D.; Zhang, S.; Schiele, B.; and Gall, J. 2022. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20963–20972.

Dubey, S.; and Dixit, M. 2023. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29(1): 167–195.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2334–2343.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023a. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17131–17141.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023b. DiffPose: SpatioTemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14861–14872.

Fu, Z.; Zuo, W.; Hu, Z.; Liu, Q.; and Wang, Y. 2023. Improving Multi-Person Pose Tracking with A Confidence Network. *IEEE Transactions on Multimedia*.

Gai, D.; Feng, R.; Min, W.; Yang, X.; Su, P.; Wang, Q.; and Han, Q. 2023. Spatiotemporal learning transformer for video-based human pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4564–4576.

Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; and Hu, H. 2023. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 660–671.

Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; and Tran, D. 2018. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 350–359.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gupta, A.; Gu, A.; and Berant, J. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35: 22982–22994.

He, J.; and Yang, W. 2024. Video-Based Human Pose Regression via Decoupled Space-Time Aggregation. *arXiv preprint arXiv:2403.19926*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*.

Iqbal, U.; Milan, A.; and Gall, J. 2017. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2020.

Jin, K.-M.; Lim, B.-S.; Lee, G.-H.; Kang, T.-K.; and Lee, S.-W. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5725–5734.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2025. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 237–255. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 525–534.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11006–11016.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Su, P.; Liu, Z.; Wu, S.; Zhu, L.; Yin, Y.; and Shen, X. 2021. Motion Prediction via Joint Dependency Modeling in Phase Space. In *ACM Multimedia*, 713–721.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 8962–8971. ACM.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 466–481.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 578–588. Springer.

Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; and Lu, C. 2018. Pose Flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35: 38571–38584.

Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; and Wang, J. 2021. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34: 7281–7293.

Zhang, R.; Huang, Y.; Cao, Y.; and Wang, H. 2025a. MoleBridge: Synthetic Space Projecting with Discrete Markov Bridges. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, R.; Huang, Y.; Lou, Y.; Xin, Y.; Chen, H.; Cao, Y.; and Wang, H. 2025b. Exploit Your Latents: Coarse-Grained Protein Backmapping with Latent Diffusion Models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 1111–1119. AAAI Press.

Zhang, R.; Lin, D.; Wang, X.; Liu, R.; Sheng, B.; Baciu, G.; Chen, C. P.; and Li, P. 2025c. Temporal-Interim Pose Synthesis and Distillation for Dynamic Human Pose Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, X.; Li, C.; Tong, X.; Hu, W.; Maybank, S.; and Zhang, Y. 2009. Efficient human pose estimation via parsing a tree structure based human model. In *2009 IEEE 12th International Conference on Computer Vision*, 1349–1356. IEEE.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.