Noname manuscript No.

Diffusion model with temporal constraint for 3D human pose estimation

Zhangmeng Chen · Ju Dai✉ · Junjun Pan✉ · Feng Zhou

Abstract 3D human pose estimation has received increasing attention as it is the foundation for many downstream tasks. However, this task is challenging due to inherent depth ambiguity and occlusion issues. Thanks to the ability of diffusion models to generate multiple hypotheses, they are promising in reducing uncertainty in results. Inspired by this, we propose a diffusion-based temporal constraint framework for 3D human pose estimation, called DTCPose, which generates multiple 3D candidate poses with 2D poses as conditions to synthesize the final pose to improve estimation accuracy. Simultaneously, to ensure the temporal stability of the 3D output sequences, we introduce temporal constraints into the model to reduce the jitter of the results. Extensive experiments on Human3.6M and MPI-INF3DHP datasets demonstrate that our approach performs predominantly in both single-hypothesis and multi-hypothesis 3D human pose estimation. Code will be available at: https://github.com/czmmmm/DCTPose.

Keywords Diffusion Model · Temporal Constraint · 3D Human Pose Estimation

Zhangmeng Chen
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: zhmchen@buaa.edu.cn

Ju Dai
Peng Cheng Laboratory, Shenzhen, China
E-mail: daij@pcl.ac.cn

Junjun Pan
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: pan_junjun@buaa.edu.cn

Feng Zhou
North China University of Technology, Beijing, China
E-mail: zhoufeng@ncut.edu.cn

1 Introduction

In recent years, there has been growing attention [20] on 3D human pose estimation due to its significant potential applications in many fields, such as action recognition [35], surveillance tracking [2], motion analysis [19], trajectory prediction [3], sports training[18], *etc.* The task aims to estimate the three-dimensional coordinates of human joints from images or videos. Due to the development of 2D human pose estimation [31,6], the two-stage solutions have become the current mainstream approaches [29], which first extracting 2D keypoints by 2D detector and then lifting 2D poses to 3D poses. Nevertheless, despite notable advancements, monocular 3D human pose estimation remains challenging. Especially in the second stage, a single 2D keypoint corresponds to multiple 3D poses, making it an ill-posed problem due to the lack of depth information.

To address this issue, several approaches [34,4] extract temporal information between consecutive video frames based on convolutional neural networks (CNN) or graph convolutional networks (GCN). In recent years, with the notable success of Transformer structures in natural language processing (NLP) [43] and computer vision (CV) [10] fields, transformer-based methods are introduced to 3D Human Pose Estimation. These methods [23,51] leverage attention mechanisms to establish spatio-temporal relationships within action sequences.

Furthermore, due to depth ambiguity, estimating 3D coordinates solely from a 2D single frame image can result in multiple results. Some probabilistic-based methods [24,44] aim to enhance estimation accuracy by generating multiple 3D poses for a single 2D pose. Multiple hypothesis solutions can encompass various scenarios. The currently popular probabilistic approach is denoising diffusion probabilistic models (DDPMs) [14], which is a generative model. In general, diffusion mod-

els have the ability to generate samples that align with a specified data distribution, such as natural images. This is achieved by starting with random noise and progressively removing the noise through multiple steps. Moreover, diffusion models have been introduced in 3D human pose estimation, demonstrating outstanding performance [9,37]. Although these methods have advantages in handling uncertainty, they still face two main problems: 1) the diversity of multiple hypotheses is limited in a narrow range of solutions; 2) the generated 3D sequences exhibit temporal jitter owing to the overlook of temporal constraint. Precisely, motion jitter refers to the phenomenon of discontinuity or abrupt changes in the pose estimation results between consecutive frames. This results in the generated 3D pose sequences being unsmooth and having poor visual effects.

Towards these challenges, we propose a novel architecture, DTCPose, to reformulate 3D human pose estimation as a task of generating 3D poses conditioned on 2D keypoints. Instead of directly concatenating noise and 2D keypoints, we inject 2D keypoints embedding features into the model as conditional information, ensuring the richness of the generated results. Additionally, we introduce time constraints, which can effectively alleviate the jitter in the output 3D sequence.

The main contributions of this work can be summarized as follows:

- We propose a novel diffusion-based temporal constraint for 3D human pose estimation, which reduces estimation errors and enhances smoothness.
- We inject 2D keypoint embedding features as a condition, which increases the diversity in generating multiple hypotheses and involving a broader range of potential poses.
- The introduction of temporal constraint to the model can mitigate jitters in the output 3D pose sequence.
- Our method achieves promising performance on Human3.6M and MPI-INF-3DHP benchmarks.

2 Related Work

2.1 3D human pose estimation

3D human pose estimation involves to repositioning human body joints in 3D space based on the input single view 2D data, *i.e.*, image or 2D keypoints. The early works [1,2] develop various pictorial structures methods to explore the dependencies of the human skeleton and perspective relationships across different spaces. With the development of deep learning, deep neural networks [11,29] are introduced into 3D human pose estimation, which can be categorized into one-stage and two-stage. The one-stage approaches directly regress the 3D pose from the input image and require a large amount of image-pose paired data and powerful computing resources [41]. The two-stage methods first utilize off-the-shelf 2D pose detectors [6,31]to estimate 2D joint coordinates and then lift the 2D coordinates into 3D space by deep neural networks, such as the fully connected network [29], recurrent neural network [11], or graph convolutional network [49]. Although the two-stage methods alleviate the requirement of image-pose pairs, they still heavily suffer from the depth ambiguities problem, which is intrinsically ill-posed due to the lack of depth information. To alleviate depth ambiguity issue, some methods [34,26,51] leverage video as input to fully exploit informative temporal information, while others [13,37] adopt probability-based modeling strategy to generate multiple 3D pose hypotheses.

2.2 3D human pose estimation from video sequence

To address the challenge of depth ambiguities, recent advancements leverage temporal context from neighboring frames to enhance the regression of 3D coordinates. For instance, Pavllo *et al.* [34] introduce a temporal fully convolutional network (TCN) to capture local context by convolving neighboring frames. Liu *et al.* [26] further enhance TCN by incorporating an attention mechanism to dynamically identify significant frames/poses within a sequence. Chen *et al.* [5] decompose pose estimation into predictions of bone length and direction. In contrast to methods relying solely on temporal aggregation, subsequent works [4,15] utilize spatio-temporal graph convolutional networks to model spatial and temporal correlations among joints simultaneously. In particular, Poseformer[51] design a concatenation architecture of several spatial transformer encoders and temporal transformer encoders in PoseFormer. PoseformerV2[50] leverages a compact representation of extensive skeletal sequences within the frequency spectrum to broaden the perceptive scope and enhance resilience against imprecise 2D joint identification. MHFormer [24] attempts to generate multiple hypothesis representations for a pose with the spatial transformer encoder and then models multi-level global correlations with different temporal transformer blocks. The MPA-GNet[17] utilizes a multi-scale parallel adaptive graph network comprising three parallel networks to effectively capture human joint features across multiple scales. AFCSTPA [46] propose a fused convolutional spatio-temporal progressive to obtain rich representations of human pose. Our DTCPose also utilizes 2D pose sequences as conditional inputs and leverages the powerful MixSTE [47] as the denoiser.

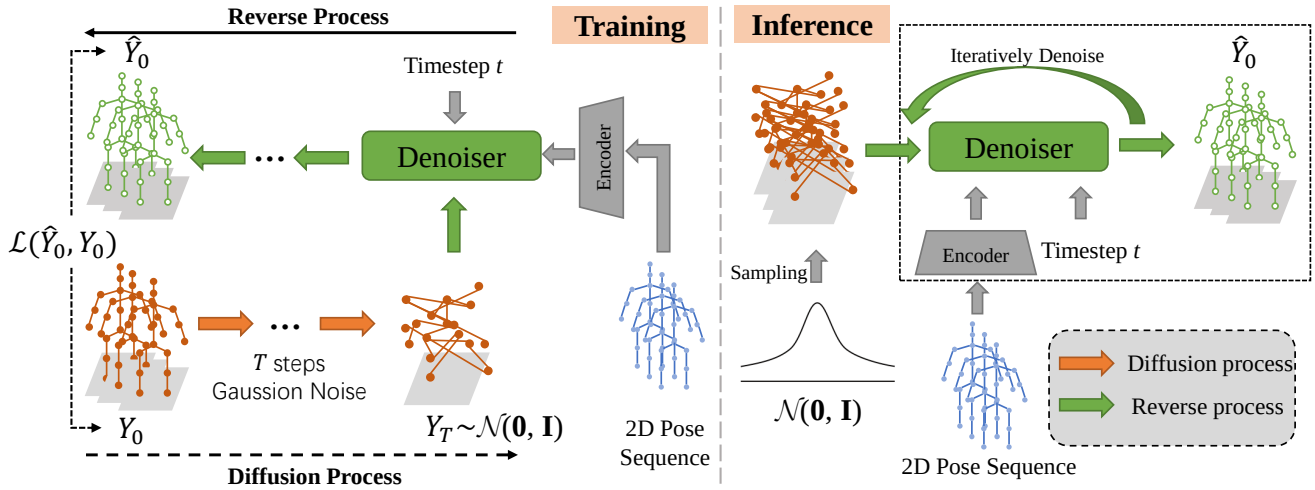

Fig. 1 Overview of the proposed DTCPose framework. Left: Training. In the diffusion process (orange arrows), we progressively diffuse the 3D pose sequence $\mathbf{Y}_0$ to $\mathbf{Y}_T$ by adding Gaussian noise at each step. When T is sufficiently large, $\mathbf{Y}_T$ follows a Gaussian distribution $\mathcal{N} \sim (0, I)$. During the reverse process (green arrows), the 2D pose sequence is encoded to obtain conditional information, which is then fed along with $\mathbf{Y}_t$ and the time step t into the denoiser for training, yielding the predicted $\hat{Y}_0$. Right: Inference. The timestep t, $\mathbf{Y}_t$ obtained through Gaussian sampling, and the conditional information corresponding to the 2D pose sequence are fed to the denoiser. After iterative denoising, the 3D pose sequence $\hat{\mathbf{Y}}_0$ is obtained.

2.3 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs have emerged as a promising approach to learning the data distribution that is straightforward to sample. SohlDickstein *et al.* [39] introduce DDPMs into the field of image generation. In recent years, DDPMs have been further simplified and accelerated [14,40]. In addition to image generation, some studies have extended DDPMs to various tasks such as image inpainting [28] and text generation [25]. Since the generated contents by DDPMs are uniquely diverse, diffusion-based methods [13,37] become the prevalence preference in 3D human pose estimation to address deep ambiguity. These methods explore multi-hypothesis human poses by introducing randomness and gradually refine to approach the true pose. Diffpose[13] proposes various designs, including the initialization of 3D pose distribution, a GMM-based forward diffusion process, and a conditional reverse diffusion process, to achieve advancing results. D3DP[37] employs a novel aggregation strategy that is Joint-wise reProjection-based Multi-hypothesis Aggregation (JPMA) to enhance performance. The proposed DTCPose is also built on the diffusion model, which fully leverages a multi-hypothesis mechanism to effectively enhance the estimation accuracy. Different from existing diffusion-based methods, our model presents the conditional injection strategy of 2D pose to condition the network and utilizes temporal constraint to address sequence jitter.

3 Method

In this section, we elaborate our DTCPose model. Regarding conditional information, we adopt the injection strategy, which can make the generated results more diverse. To ensure the smoothness of the output sequence, we add a temporal constraint. In the following subsections, we provide a detailed description of each item.

3.1 Problem Formulation

The overall framework of the proposed method DTCPose is illustrated in Fig. 1. Given the input 2D keypoints sequence $\mathbf{X} \in R^{L\times N\times 2}$, where L represents the length of the input sequence and N is the number of joints in the human skeleton. Another input is the initial 3D pose with noise $\mathbf{Y}_t \in R^{L\times N\times 3}(t \sim \mathcal{U}(0,T))$, which is obtained through Gaussian sampling. $\mathcal{U}$ represents uniform distribution. Our goal is to use 2D keypoint sequences as conditional information, iteratively denoise on $\mathbf{Y}_t$, and ultimately obtain the 3D coordinates corresponding to the 2D keypoints.

3.2 Background on Diffusion Models

The proposed method is inspired by DDPMs [14], which typically involves two basic processes: a forward process and a reverse process. In the forward process, Gaussian noise is gradually added to the data $\mathbf{Y}_0$ until the data becomes completely noisy, where $\mathbf{Y}_0 \in R^{L\times N\times 3}$ is the

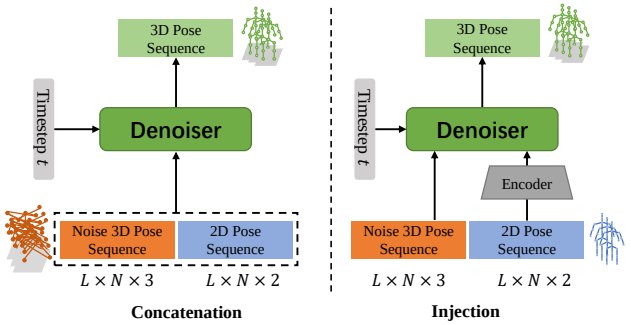

Fig. 2 The conditional strategies. Left: Concatenate strategy. The 2D pose sequence and the noisy 3D pose sequence are first concatenated and then fed into the denoiser along with the timestep t. Right: Injection strategy. The 2D pose sequence is first encoded by an encoder and then fed into the denoiser. A detailed description of the injection strategy can be found in Section 3.3.

groundtruth of 3D pose. In the backward process, a denoiser is trained to denoise the Gaussion noisy data and restore the original data $\mathbf{Y}_0$.

To be more specific, given the groundtruth 3D poses sequence $\mathbf{Y}_0 \sim q(\mathbf{Y}_0)$, $\mathbf{Y}_t$ is obtained by adding Gaussian noise for t-steps, where $t \sim \mathcal{U}(0, T)$ is the timestep sampling and T is the maximum number of timesteps. Following the DDPMs [14], this process is based on a Markov chain. Therefore, the forward process can be formulated as:

$$\begin{aligned} q(\mathbf{Y}_t|\mathbf{Y}_{t-1}) &:= \mathcal{N}(\mathbf{Y}_t; \sqrt{1-\beta_t}\mathbf{Y}_{t-1}, \beta_t\mathbf{I}), \\ \mathbf{Y}_t &= \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{1}$$

where $t = 1, 2, ..., T$ and β_t are referred to as variance schedules to control how much noise should be added at the t step. As t increases, $\mathbf{Y}_t$ becomes closer to Gaussian noise. After applying the re-parameterization trick, the joint probability distribution of $\mathbf{Y}_0$ given $\mathbf{Y}_t$ can be defined as:

$$q(\mathbf{Y}_t|\mathbf{Y}_0) := \mathcal{N}(\mathbf{Y}_t; \sqrt{\bar{\alpha}_t\mathbf{Y}_0}, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \Pi_{i=1}^{t}\alpha_i$. The reverse process is a denoising procedure, where given $\mathbf{Y}_t$, the goal is to predict $\mathbf{Y}_{t-1}$. The process can be expressed as:

$$p(\mathbf{Y}_{t-1}|\mathbf{Y}_t) := \mathcal{N}(\mathbf{Y}_{t-1}; \mu_\theta(\mathbf{Y}_t, t), \Sigma_\theta(\mathbf{Y}_t, t), \tag{3}$$

where $\mu_\theta(\mathbf{Y}_t, t)$ and $\Sigma_\theta(\mathbf{Y}_t, t)$ represent the mean and variance of the sampling at timestep t. A denoiser (a neural network model) can be used for predictions, either by predicting noise ϵ or directly predicting $\mathbf{Y}_0$.

3.3 Conditional Strategy

In this paper, we use 2D keypoints as conditional information. There are various strategies for incorporating conditional information into the model. Here, we provide a detailed explanation of concatenation and injection strategies as shown in Fig. 2.

Table 1 The multi-hypothesis experiments with the Concatenation and Injection strategies. H represents the number of hypotheses. ↓ means the smaller the better.

Method	H	MPJPE (mm)(↓)
Baseline (MixSTE [47])	1	40.9
+Concatenation	1	40.6
+Concatenation	10	40.6
+Concatenation	20	40.5
+Injection	1	39.6
+Injection	10	39.2
+Injection	20	38.8

Concatenation. We concatenate uncertain 3D noisy poses and 2D poses as inputs, and use the timestep t as condition, as shown in Fig. 2 (left). The benefit of this approach is that the merging of noise with 2D conditions can be intuitively observed. Additionally, since the input includes a deterministic 2D pose, it can provide a better initialization method, facilitating the convergence of the model. However, through observation, we find that this configuration does not necessarily lead to a better fusion of multiple hypotheses results. As illustrated in Table 1, the baseline model is MixSTE[99], and a Concatenation strategy is implemented, with various quantities of hypotheses being generated. From the table, it can be observed that compared to a single hypothesis, the final output of multi-hypothesis fusion does not show a significant advantage in reducing estimation errors. Furthermore, when selecting the best result from multiple hypotheses, this strategy does not necessarily choose the pose with the smallest error compared to the injection strategy.

Injection. Unlike the concatenate strategy, we input the uncertain 3D noisy poses to the Denoiser. The embeddings of the 2D pose and timestep t are combined and injected as conditions into the Denoiser, as shown in Fig.2 (right). This process is employed to scale and shift the Denoiser, enabling the 2D pose to better guide the prediction of the 3D pose. The results in Table 1 also demonstrate the effectiveness of our proposed Injection strategy. As the number of hypotheses increases, the estimation error decreases. Specifically, the injection process requires the integration of two terms: the output from other modules and conditional information, inspired by [48]. This process is applied after each Self-attention and feedforward neural network (FFN), as shown in Fig. 3.

As illustrated in Fig. 4, the injection module receives the embeddings of 2D pose sequences and timestep as inputs, merging these embeddings through an addition operation. Subsequently, those embedding vectors are

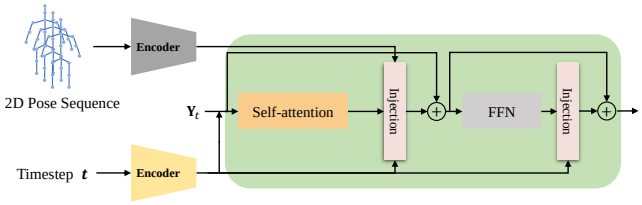

Fig. 3 Pipeline of the injection process.

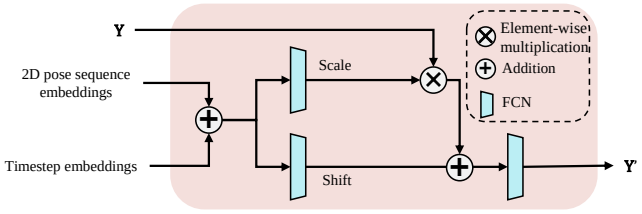

Fig. 4 Details of the injection module.

processed with the fully connection network (FCN), consisting of two fully connected layers, to generate scaling parameters and shifting parameters. Meanwhile, the condition injection module receives another input, namely the output $\mathbf{Y}$ of the upper module. This output $\mathbf{Y}^{'}$ is multiplied element-wise with the previously obtained scaling parameters and then added to the shifting parameters. Finally, this result is processed through another FCN to obtain the final output $\mathbf{Y}^{'}$. Such a process effectively integrates the conditional information and provides accurate guidance for the denoising network.

3.4 Temporal Constraint

Previous probabilistic-based methods [12,37,13] for 3D human pose estimation mainly focused on the error of the 3D pose, whether based on single frames or videos, with little attention given to the smoothness of the generated 3D pose sequences. These methods typically only report a sufficiently small error in the estimated 3D pose to emphasize the superiority of their approaches. However, existing methods tend to produce noticeable jitters in the generated 3D sequences, which is clearly not desirable. This jitter, deviating from the ground truth, is also a form of error. The smoothness of the output 3D skeleton sequence has a significant impact on downstream tasks, such as action recognition, motion analysis, and virtual reality interaction. Smooth sequences can provide more accurate temporal dynamic information, thereby improving the performance and reliability of these applications. To address the jitter problem in pose sequences, existing research such as SmoothNet [45] has proposed a dedicated smoothing network to reduce the jitter in motion sequences. Building on this, we focus on the temporal coherence of motion sequences, introducing temporal constraints to reduce incoherent errors, thereby generating smoother and more realistic 3D human pose sequences.

Our goal is to capture the temporal relationships of joints and impose temporal constraints to alleviate jitter. Unlike [45], which specifically designs additional fully connected networks (FCNs) for motion sequence smoothing, we optimize the objective by adding a temporal loss. This is primarily attributed to our use of a spatio-temporal denoiser, where the temporal module is equivalent to a fully connected network for handling temporal aspects. Jitters can be considered as acceleration errors between adjacent joints, and optimizing for this error contributes to a smoother sequence. The specific definition of the temporal loss is as follows:

$$\mathcal{L}_{temporal} = \frac{1}{(L-2)\times N}\sum_{j=0}^{L}\sum_{i-0}^{N}|\hat{Y}_{j,i}^{''} - Y_{j,i}^{''}|, \tag{4}$$

where $\hat{Y}_{j,i}^{''}$ and $Y_{j,i}^{''}$ represent the acceleration of the predicted 3D sequence and the ground truth 3D sequence, respectively. They can be computed by calculating velocity through the difference in joint positions and then obtaining acceleration through the difference in velocity. The temporal loss $\mathcal{L}_{temporal}$ is one component of the final loss.

3.5 Architecture

Unlike previous methods [24,44] that require designing complex networks as denoisers, our approach focuses on conditional strategies and temporal constraint, making it compatible with various spatio-temporal denoisers. In this paper, we choose MixSTE [47] as the denoiser. For 2D keypoints, we employ fully connected layers (FCN) as the encoder to extract spatio-temporal information. In the case of the concatenate strategy, the 2D keypionts are directly concatenated with 3D noisy poses without any processing. However, for the injection strategy, a simple linear layer and a spatio-temporal fusion model are used for encoding to obtain 2D pose embeddings. The final loss function is composed of the 3D keypoint position estimation loss and the temporal smooth loss:

$$\mathcal{L} = \mathcal{L}_{position} + \mathcal{L}_{temporal}, \tag{5}$$

where $\mathcal{L}_{position}$ is the mean per joint position error, as known as MPJPE, and is described as follows:

$$\mathcal{L}_{position} = \frac{1}{L\times N}\sum_{j=0}^{L}\sum_{i=0}^{N}||\hat{Y}_{j,i} - Y_{j,i}||_2. \tag{6}$$

3.6 Training and Inference

Training. As shown in Fig. 1, the training process of our diffusion model can be viewed as a procedure where inputting 2D poses and 3D noisy poses results in the output of corresponding 3D poses for the given 2D inputs. Specifically, we sample Gaussian noise α_t according to Eq. 1 and add it to the original 3D pose to obtain a 3D pose with noise. The α_t for each sampling timestep t can be obtained through the noise scheduler [14]. Then, we minimize the final loss $\mathcal{L}$ (Eq. 5) to supervise the model training. The training process of DTCPose is detailed in Algorithm 1.

Algorithm 1 Training

Input: 2D keypoints $\mathbf{X}$, the groundtruth 3D pose $\mathbf{Y}_0$

repeat
  $\mathbf{Y}_0 \sim q(\mathbf{Y}_0)$
  $t \sim \mathcal{U}(1, ..., T)$
  $\mathbf{Y}_t = \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
  Take gradient descent step on
$\nabla_\theta ||\mathbf{Y}_0 - Denoiser(\mathbf{X}, \mathbf{Y}_t, t)||^2$
until converged

Algorithm 2 Inference

Input: 2D keypoints $\mathbf{X}$, timestep t
Output: 3D pose Y_0

$\mathbf{Y}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{Y}_{t-1} = Denoiser(\mathbf{X}, \mathbf{Y}_t, t)$
end for
return $\mathbf{Y}_0$

Inference. The proposed DTCPose utilizes 2D keypoints as conditions to denoise 3D noisy poses generated through Gaussian sampling. It progressively refines the predicted results over multiple sampling steps. For each sampling step, the denoiser takes 3D noisy poses and conditional information as input and outputs the 3D pose estimated at the current step. This output then serves as the input for the next step. In this paper, we adopt the DDIM [40] scheme to accelerate the inference process. The inference procedure detail is provided in Algorithm 2.

4 Experiments

4.1 Datasets and Metrics

Human3.6M dataset [16]. Human3.6M is a popular and challenging dataset for monocular 3D human pose estimation. This dataset has 3.6 million human images captured in an indoor controllable environment. In this dataset, 11 subjects execute 15 actions from 4 cameras. Heeding prior endeavors [34,8], 5 subjects (S1, S5, S6, S7, S8) are utilized for training, 2 subjects (S9, S11) are designated for testing. We employ two assessment protocols: Protocol 1 is MPJPE between the estimated 3D pose and the ground truth. Protocol 2 is P-MPJPE, which is MPJPE after rigid alignment 3D pose. To measure jitter errors, we follow related work [19,7,45] and adopt the acceleration error (mm/s^2), which measures the average difference in acceleration between the predicted and the ground truth 3D joints.

MPI-INF-3DHP dataset [30]. MPI-INF-3DHP involves more complex scenarios, consisting of green screen indoor, non-green screen indoor, and outdoor scenes. Utilizing 14 cameras, the dataset captures the performances of 8 actors engaged in 8 activities for training and 7 activities for evaluation purposes. Following the previous works [51,5,36], we report MPJPE, percentage of correct keypoint (PCK) within 150mm range, and area under curve (AUC).

4.2 Implementation Details

The proposed method is implemented with PyTorch [33]. We adopt AdamW [27] optimizer with initial learning rate $1e-4$. We train the proposed model for 1000 epochs. The batch size is 4, and each sample contains 243 frames. We employ a fully connected layer as the encoder. We set the hidden dimension to 512. We set the total sampling steps $T = 1000$. Our experiments are conducted on a single NVIDIA RTX 3090 Ti GPU. Unless otherwise stated, the results reported by our method use the injection strategy and the sequence length is set to 243 on Human3.6M dataset and 27 on MPI-INF-3DHP dataset.

4.3 Comparison with State-of-the-art Methods

Results on the Human3.6M dataset. We compare our proposed DTCPose with state-of-the-art 3D human pose estimation methods on the Human3.6M dataset for all 15 actions, as shown in Table 2. Following [37], we compare deterministic methods (predicting a single result) and probabilistic methods (predicting multiple results) separately. For fairness in comparison with deterministic methods, we set H=1, meaning only a single hypothesis result is generated. As shown in the top of Table 2, our method (L=243) achieves the average MPJPE of 39.6mm, outperforming other deterministic methods with the same sequence length (L=243). Additionally, using MixSTE [47] as the denoiser, our method

Table 2 Results on the Human3.6M dataset under Protocol 1 (MPJPE), using 2D poses detected by CPN [6] as the condition input. Top: results for deterministic methods. Bottom: results for probabilistic methods. L and H represent the sequence length and number of hypotheses, respectively. ↓ means the smaller the better.

Deterministic Methods															
MPJPE (mm)(↓)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
MPA-GNet [17] TVC'2023 (L=81)	43.3	45.7	42.1	45.6	47.5	56.2	44.2	43.5	57.3	63.1	46.0	44.4	45.0	31.8	32.0	45.9
PoseFormer [51] ICCV'2021 (L=81)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
AFCSTPA [46] TVC'2024 (L=81)	41.5	43.3	38.5	43.0	43.9	52.0	41.9	42.3	54.3	61.0	44.7	42.6	44.5	30.1	30.9	43.6
P-STMO [36] ECCV'2022 (L=243)	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
MixSTE [47] CVPR'2022 (L=243)	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
STCFormer [42] CVPR'2023 (L=243)	39.6	41.6	37.4	38.8	43.1	51.1	39.1	39.7	51.4	57.4	41.8	38.5	40.7	27.1	28.6	41.0
D3DP [37] ICCV'2023 (L=243)	37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
DTCPose (Ours) (L=243)	37.6	39.8	36.2	38.0	40.1	48.1	39.3	37.4	50.1	52.8	40.5	39.6	40.1	27.2	27.5	39.6
Probabilistic Methods																
MPJPE (mm)(↓)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
NF [44] ICCV'2021 (L=1, H=1)	52.4	60.2	57.8	57.4	65.7	74.1	56.2	59.1	69.3	78.0	61.2	63.7	67.0	50.0	54.9	61.8
MHFormer [24] CVPR'2022 (L=351, H=3)	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
DiffPose [13] CVPR'2023 (L=243, H=5)	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	24.4	24.1	36.9
D3DP [37] ICCV'2023 (L=243, H=1)	37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
D3DP [37] ICCV'2023 (L=243, H=20)	37.6	39.7	35.8	38.0	41.7	48.7	39.4	38.2	50.3	53.3	41.4	39.4	39.7	27.4	27.3	39.9
DTCPose (Ours) (L=243, H=1)	37.6	39.8	36.2	38.0	40.1	48.1	39.3	37.4	50.1	52.8	40.5	39.6	40.1	27.2	27.5	39.6
DTCPose (Ours) (L=243, H=20)	36.7	38.2	35.5	37.1	39.2	46.9	38.1	36.5	49.1	51.9	40.2	39.1	40.2	26.7	26.6	38.8

Table 3 Results on the Human3.6M under Protocol 2 (P-MPJPE), using 2D poses detected by CPN [6] as the condition information. L and H represent the sequence length and number of hypotheses, respectively. ↓ means the smaller the better.

P-MPJPE (mm)(↓)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
MPA-GNet [17] TVC'2023 (L=81)	33.5	36.8	33.8	37.7	36.6	43.5	34.1	33.1	45.7	49.7	36.9	34.5	34.9	24.7	25.7	36.1
PoseFormer [51] ICCV'2021 (L=81)	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
P-STMO [36] ECCV'2022 (L=243)	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
AFCSTPA [46] TVC'2024 (L=81)	31.3	34.0	30.7	34.5	33.9	38.9	32.2	31.4	43.1	48.8	36.1	32.0	34.4	23.9	24.5	34.0
MixSTE [47] CVPR'2022 (L=243)	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
STCFormer [42] CVPR'2023 (L=243)	29.5	33.2	30.6	31.0	33.0	38.0	30.4	29.4	41.8	45.2	33.6	29.5	31.6	21.3	22.6	32.0
NF [44] ICCV'2021 (L=1, H=1)	37.8	41.7	42.1	41.8	46.5	50.2	38.0	39.2	51.7	61.8	45.4	42.6	45.7	33.7	38.5	43.8
MHFormer [24] CVPR'2022 (L=351, H=3)	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
DiffPose [13] CVPR'2023 (L=243, H=5)	26.3	29.0	26.1	27.8	28.4	34.6	26.9	26.5	36.8	39.2	29.4	26.8	28.4	18.6	19.2	28.7
D3DP [37] ICCV'2023 (L=243, H=1)	30.6	32.5	29.1	31.0	31.9	37.6	30.3	29.4	40.6	43.6	33.3	30.5	31.4	21.5	22.4	31.7
D3DP [37] ICCV'2023 (L=243, H=20)	30.6	32.5	29.1	30.9	31.9	37.5	30.2	29.4	40.6	43.4	33.3	30.4	31.4	21.5	22.4	31.7
DTCPose (Ours) (L=243, H=1)	30.2	31.5	28.9	31.1	31.2	37.0	30.5	28.7	40.5	42.3	33.6	31.3	31.2	20.5	21.9	31.4
DTCPose (Ours) (L=243, H=20)	30.1	30.8	28.4	31.8	31.1	37.3	29.5	28.2	40.9	41.5	33.6	31.2	30.7	20.1	21.3	31.1

obtains better results than D3DP [37] by 0.4 mm under MPJPE. Those results indicate that DTCPose can effectively leverage conditional information to improve the accuracy of 3D human pose estimation.

The advantage of diffusion models is the generation of multiple 3D hypotheses, which is also the core of our method. To validate the performance of our method in generating multiple hypotheses, we compared the proposed method with various probabilistic methods, as shown at the bottom of Table 2. We adopt the strategy of averaging multiple hypothesis results. Under this premise, such manners have set various numbers of hypotheses to assess the impact of multiple hypotheses. It is evident that the NF [44] (L=1) estimation results have the most significant error (61.8 mm), which also confirms that video-based inputs can improve performance. Despite DiffPose [13] achieving the best estimation results with only 5 hypotheses (H=5), this accomplishment can be attributed to the DiffPose's capability to integrate not only 2D pose sequence but also additional inputs, such as statistical insights derived from the dataset and the heat maps generated by the 2D Detector. Compared to methods under the same input conditions (2D keypoints), our method (L=243, H=20) achieved the best result (38.8 mm). Moreover, it is noteworthy that as the number of hypotheses increases (H: 1→20), the performance of the D3DP [37] does not show significant improvement (0.1 mm), while our method exhibits a notable performance enhancement by 0.8 mm. This observation strongly indicates that our method, with the injection conditional strategy, can cover a broader range of hypothesis results, thereby enhancing the accuracy of the final aggregation results.

Additionally, we report the comparison results with other methods under Protocol 2 (P-MPJPE), as shown in Table 3 with the CPN [6] 2D keypoints as inputs. Table 4 shows the MPJPE results under Protocol 1 using the ground truth 2D keypoints as inputs. From both tables, it can be observed that our method achieves significant performance improvements in both single-hypothesis and multi-hypothesis settings. The experimental results powerfully demonstrate the excellent performance of our method.

To further explore the performance upper bound of our method, we calculate the MPJPE between each hypothesis pose in the multiple hypothesis and the ground truth pose, then report the minimum MPJPE (min-MPJPE), following previous works [9,44]. As shown in Table 5, when generating 5 hypothesis poses, our method significantly outperforms MDN [9] (H=200), D3DP [37] (H=20), and NF [44] (H=200) methods. With the increase in the number of hypotheses (H:5→30), the performance of our method continues to improve

Table 4 Results on the Human3.6M dataset under Protocol 1 (MPJPE), using the ground truth 2D poses as the condition information. L and H represent the sequence length and number of hypotheses, respectively. ↓ means the smaller the better.

P-MPJPE (mm)(↓)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
PoseFormer [51] ICCV'2021 (L=81)	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
P-STMO [36] ECCV'2022 (L=243)	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
AFCSTPA [46] TVC'2024 (L=81)	24.3	29.3	25.5	26.8	30.5	29.5	28.3	28.2	33.0	35.1	29.7	29.2	32.2	28.3	26.8	29.1
MixSTE [47] CVPR'2022 (L=243)	21.6	22.0	20.4	21.0	20.8	26.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
STCFormer [42] CVPR'2023 (L=243)	21.4	22.6	21.0	21.3	23.8	26.0	24.2	20.0	28.9	28.0	22.3	21.4	20.1	14.2	15.0	22.0
MHFormer [24] CVPR'2022 (L=351, H=3)	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.3	31.0	29.4	22.2	23.0	30.5
DiffPose [13] CVPR'2023 (L=243, H=5)	18.6	19.3	18.0	18.4	18.3	21.5	21.5	19.1	23.6	22.3	18.6	18.8	18.3	12.8	13.9	18.9
D3DP [37] ICCV'2023 (L=243, H=1)	19.9	19.6	19.7	19.3	20.2	22.7	21.5	19.2	25.5	24.0	20.1	18.9	19.0	14.0	14.5	19.9
D3DP [37] ICCV'2023 (L=243, H=20)	19.9	19.5	19.6	19.2	20.1	22.4	21.5	19.1	25.4	23.7	20.0	18.9	18.8	14.0	14.5	19.8
DTCPose (Ours) (L=243, H=1)	19.7	19.2	19.4	18.9	19.7	22.1	21.3	19.2	24.6	23.9	20.5	18.5	19.1	14.3	14.6	19.7
DTCPose (Ours) (L=243, H=20)	19.4	19.1	19.4	18.5	19.3	21.7	21.1	18.6	24.2	23.4	20.5	18.1	19.0	14.4	14.2	19.4

Table 5 Results on the Human3.6M under Protocol 1 (minMPJPE), using 2D poses detected by CPN [6] as the condition information. H denotes the number of hypotheses. ↓ means the smaller the better.

minMPJPE (mm)(↓)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
MDN [21] CVPR'2019 (H=5)	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
MultiPoseNet [38] ICCV'2019 (H=200)	37.8	43.2	43.0	44.3	51.1	57.0	39.7	43.0	56.3	64.0	48.1	45.4	50.4	37.9	39.9	46.8
Li et al. [22] BMVC'2020 (H=10)	54.8	61.9	48.6	63.6	55.8	73.7	59.0	61.3	62.2	85.7	52.8	60.2	57.5	51.3	56.8	60.0
GraphMDN [32] IJCNN'2021 (H=200)	40.0	43.2	41.0	43.4	50.0	53.6	40.1	41.4	52.6	67.3	48.1	44.2	44.9	39.5	40.2	46.2
NF [44] ICCV'2021 (H=200)	38.5	42.5	39.9	41.7	46.5	51.6	39.9	40.8	49.5	56.8	45.3	46.4	46.8	37.8	40.4	44.3
GFPose [9] CVPR'2023 (H=10)	39.9	44.6	40.2	41.3	46.7	53.6	41.9	40.4	52.1	67.1	45.7	42.9	46.1	36.5	38.0	45.1
GFPose [9] CVPR'2023 (H=200)	31.7	35.4	31.7	32.3	36.4	42.4	32.7	31.5	41.2	52.7	36.5	34.0	36.2	29.5	30.2	35.6
D3DP [37] ICCV'2023 (H=20)	37.3	39.4	35.4	37.8	41.3	48.1	39.0	37.9	49.8	52.8	41.1	39.0	39.4	27.3	27.2	39.5
DTCPose (Ours) (H=5)	35.7	35.1	34.0	34.6	35.2	37.7	35.5	35.7	40.5	40.0	37.3	34.3	32.3	31.7	31.8	35.4
DTCPose (Ours) (H=10)	32.9	32.2	30.8	31.0	31.1	34.3	31.7	32.9	35.0	33.3	33.0	31.0	30.3	29.7	29.6	31.9
DTCPose (Ours) (H=20)	30.8	30.2	28.9	28.9	29.1	31.4	29.8	30.6	32.3	31.0	31.0	29.1	28.7	28.4	28.2	29.9
DTCPose (Ours) (H=30)	29.6	29.2	27.9	28.1	28.3	30.3	28.9	29.6	31.2	30.0	30.0	28.2	28.2	27.7	27.4	29.0

significantly (35.4mm→29.0mm). Due to device limitations, the maximum value for H is set to 30. Despite the device limitations with a maximum value of 30 for H, experimental results indicate that our diffusion-based method, DTCPose, has an exceptionally high-performance ceiling.

Furthermore, to verify that our method can alleviate jitter in the generated 3D pose sequences, we conduct relevant experiments and report the acceleration error. In the comparison methods, we not only selected deterministic methods [51, 47] but also probabilistic methods [37]. For fairness, we employ single-hypothesis generation settings in our experiments. From Table 6, it can be observed that our method achieves a low acceleration error (1.04 mm/s^2) while maintaining the low MPJPE (39.6mm). Additionally, we can observe that MixSTE [47] also exhibits relatively low acceleration errors. This is mainly because MixSTE employs a seq2seq architecture, which allows for better modeling of the temporal dimension of the input sequence. Although D3DP [37] utilizes MixSTE as its backbone, the diversity of the diffusion model leads to decreased temporal stability in the output 3D sequences. With the same backbone (MixSTE) architecture, the incorporation of temporal constraint in our model enables us to achieve the lowest acceleration error. This highlights the importance of time constraints in our approach.

Results on the MPI-INF-3DHP dataset. To evaluate the cross-dataset performance of our method, we also conducted experiments on MPI-INF-3DHP dataset. Table 7 reports three evaluation metrics between our method and other approaches on the dataset. The model

Table 6 Results on the Human3.6M dataset regarding MPJPE and acceleration (Accel) error metrics. We use 2D poses detected by CPN [6] as condition inputs. ↓ means the smaller the better. For fairness, we employ single-hypothesis generation settings in all experiments.

Method	MPJPE(mm)(↓)	Accel (mm/s^2)(↓)
PoseFormer [51] ICCV'2021	44.3	3.84
MixSTE [47] CVPR'2022	40.9	1.87
D3DP [37] ICCV'2023	40.0	4.49
DTCPose (Ours)	39.6	1.04

is trained on the Human3.6M dataset without any training or fine-tuning on the MPI-INF-3DHP dataset with the ground truth 2D keypoints as inputs. Following MixSTE [47], we use 27 frames as input due to the relatively short video lengths in this dataset. As shown in Table 7, it can be observed that under the condition of equal sequence length, when H=1, our method outperforms MixSTE in all three metrics. When the hypothesis result is H=20, the model performance improves significantly. However, compared to the optimal results of existing methods, our method does not achieve superior performance, mainly due to differences in input sequence length. Specifically, the input sequence lengths for P-STMO, D3DP, DiffPose, and STCFormer are 81, 243, 81, and 81, respectively, while the input sequence length used in our method for this experiment is 27. Longer sequences can provide richer temporal information, which might explain why our method performs slightly less well in this experiment. Additionally, DiffPose achieves better results by integrating more comprehensive input information, including 2D poses, heatmap information, and the data distribution

Table 7 Results on MPI-INF-3DHP using ground truth 2D poses as condition inputs. ↓ means the smaller the better. ↑ means the bigger the better.

Method	PCK(↑)	AUC(↑)	MPJPE(↓)
Anatomy[5] TCSVT'2021 (L=81)	87.9	54.0	78.8
PoseFormer[51] ICCV'2021 (L=9)	88.6	56.4	77.1
MHFormer[24] CVPR'2022 (L=9, H=3)	93.8	63.3	58.0
MixSTE[47] CVPR'2022 (L=27)	94.4	66.5	54.9
P-STMO[36] ECCV'2022 (L=81)	97.9	75.8	32.2
D3DP[37] ICCV'2023 (L=243, H=20)	97.7	78.0	30.0
DiffPose[13] CVPR'2023 (L=81, H=5)	98.0	75.9	29.1
STCFormer[42] CVPR'2023 (L=81)	98.7	83.9	23.1
DTCPose (Ours) (L=27, H=1)	95.6	71.4	45.1
DTCPose (Ours) (L=27, H=20)	97.5	76.9	30.2

in the training set. In contrast, our method only leverages the 2D pose sequence as conditional input, which may limit the richness of input information and result in lower performance in this experiment compared to DiffPose. Despite not achieving the best results in comparisons with different input lengths, our method has gained relatively favorable performance. This finding confirms the robustness and generalization capability of our method when applied across different datasets.

4.4 Ablation Study

To verify the impact and performance of critical components of the proposed DTCPose, we conduct extensive ablation experiments on the Human3.6M dataset using 2D poses detected by CPN [6] estimator as inputs.

Impact of Each Component. DTCPose consists of two main components: the conditional strategy (Concatenate and Injection) and the temporal constraint. As shown in Table 8, we use the basic MixSTE [47] as the baseline model and then employ it as the denoiser for our diffusion model. Subsequently, we conduct experiments by adding different components to it. All experiments of Table 8 have been conducted with the single hypothesis case (H=1). The results indicate that the diffusion model can indeed enhance model performance. We also observe that adopting the injection strategy for conditional information leads to a decrease in MPJPE of 0.6mm compared to the concatenate strategy. With the injection strategy, the error further decreases to 39.6mm after adding temporal constraint. The beneficial results demonstrate that the conditional strategy and temporal constraint proposed in our model are crucial for improving model performance.

Impact of the Number of Hypotheses and Multi-Hypotheses Aggregation. Generating multiple hypothesis results is fundamental to diffusion model-based methods, and how to aggregate multiple hypothesis results is equally important. Table 9 illustrates how the performance of our DTCPose is influenced by the number of hypotheses and multiple hypothesis aggregation strategies. As shown in Tabl 9, we employ two aggregation strategies: averaging and minimum. As expected, regardless of the aggregation scheme, as the number of hypotheses increases, the estimation error decreases accordingly. This indicates that leveraging the diffusion model for generating multiple hypotheses helps improve the performance of the model, validating our motivation for using the diffusion model. On the whole, the error with the averaging strategy is higher than with the minimum strategy. This observation suggests that hypothesis diversity has a dual nature. The stronger the diversity, the larger the variance among hypotheses, which is unfavorable for the averaging strategy. Conversely, in such cases, the minimum strategy selects poses with lower errors because multiple hypothesis results can cover a broader range of potential solutions. As shown in Table 9 (Bottom), it is evident that as the number of hypotheses increases, the MPJPE decreases significantly. However, as the num-

Table 8 Ablation study of each component, using 2D poses detected by CPN [6] as the condition input. ✓ denotes having the component. ↓ means the smaller the better.

Models	Concate	Inject	Temporal Constraint	MPJPE (mm)(↓)
Baseline				40.9
	✓			40.6
	✓		✓	40.4
		✓		40.0
Ours		✓	✓	39.6

Table 9 Analysis for the number of hypotheses in our method, using ground truth 2D poses as inputs. H indicates the number of hypotheses. ↓ means the smaller the better.

Method	Aggregation	H	MPJPE (mm)(↓)
Baseline	-	1	40.9
1	Averaging	1	39.6
2	Averaging	20	38.8
3	Minimum	1	39.6
4	Minimum	5	35.4
5	Minimum	10	31.9
6	Minimum	20	29.9
7	Minimum	30	29.0

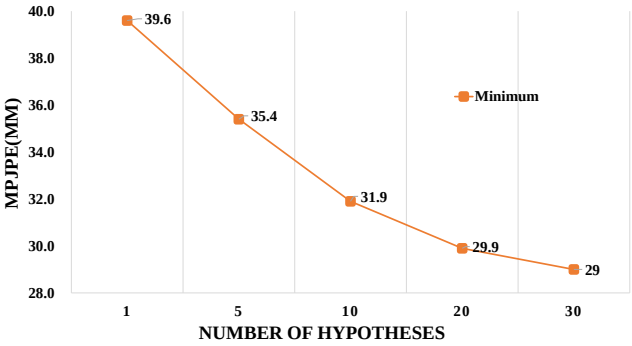

Fig. 5 Ablation study on the number of hypotheses with minimum aggregation strategy.

Table 10 Ablation study on inference speed. The evaluation is conducted on Human3.6M using 2D poses detected by CPN [6] estimator as input. H: the number of hypotheses. ↓ means the smaller the better.

Method	MPJPE(mm)(↓)	Inference FPS
PoseFormer[51] ICCV'2021 (H=1)	44.3	1080
MHFormer[24] CVPR'2022 (H=3)	43.0	347
MixSTE[47] CVPR'2022 (H=1)	40.9	5031
STCFormer[42] CVPR'2023 (H=1)	41.7	292
DiffPose[13] CVPR'2023 (H=5)	36.9	673
D3DP[37] ICCV'2023 (H=20)	39.9	4529
Ours (H=1)	39.6	4895
Ours (H=5)	39.3	979

ber of hypotheses increases, the rate of error reduction becomes slower, as shown in Fig. 5. Although the minimum strategy is not available in real-world applications, it can still be used to explore the upper bound on model performance.

Inference Speed. As shown in Table 10, we compare the proposed method with existing state-of-the-art methods in terms of the estimation error (MPJPE) and Frames Per Second (FPS) during the inference process. The FPS calculation for all methods is performed on a single NVIDIA GeForce RTX 3090 Ti GPU. It can be observed that in the single-hypothesis scenario (H=1), the proposed method achieves a very fast inference speed, with an FPS of 4895. However, when the number of hypotheses is 5 (H=5), the inference speed decreases significantly, with FPS dropping to 979. Although our method does not exhibit superior performance compared to DiffPose, it demonstrates a higher inference speed under the same number of hypotheses. Therefore, it is worth further research on how to balance the number of hypotheses with the performance of the method.

4.5 Qualitative Results

To better evaluate the effectiveness of our method, Fig. 6 shows the visualization comparisons between our method with the MixSTE [47] and D3DP [37] on Human3.6M dataset (S9 and S11). As observed through the magnified effect in Fig. 6, the proposed method produces better 3D poses across different activities, such as *SittingDown, Posing, Sitting*.

Furthermore, we randomly selected some sample persons in Huma3.6M and illustrate the generated multiple hypotheses in Fig. 7. It can be observed from Fig. 7 that our method is capable of generating multiple 3D hypotheses results that are distributed around the ground truth 3D pose. By utilizing a multi-hypothesis integration strategy, the final estimated 3D pose is more closer to the ground truth. Nevertheless, our method still has its limitations. Specifically, the multiple hypotheses generated for the pose display significant variation predomainantly in specific parts, such as the right arm, rather than uniformly across the entire body, as illustrated in the third case of Fig. 7. The failure case indicates that the estimation performance of our method for terminal joints still needs to be improved.

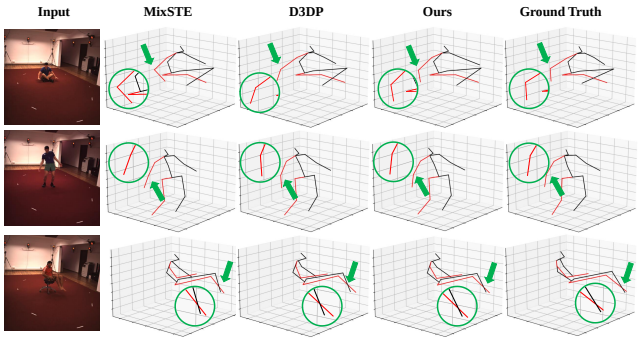

Fig. 6 Qualitative comparison between our method, MixSTE and D3DP on Human3.6M dataset. The green arrows and circles emphasize the superior performance of our approach.

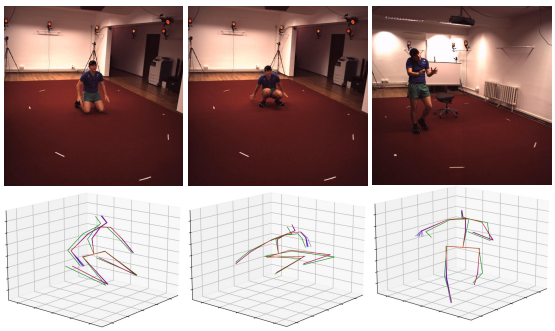

Fig. 7 Qualitative visual results of our method on the Human3.6M test dataset. Blue pose: multiple hypotheses pose; Red pose: the averaging integration pose; Green pose: ground truth 3D pose.

5 Conclusion

In this paper, we propose DTCPose, a novel diffusion model-based framework with temporal constraint. DTCPose transforms the 3D human pose estimation task into a 3D human pose generation problem conditioned on 2D keypoints using a conditional injection strategy. With the proposed injection strategy, the diversity of generated hypotheses is appropriately enhanced, allowing the multiple hypothesis results to cover a broader range of potential poses, thereby improving model performance. Additionally, we introduce the temporal constraint to alleviate the generated 3D pose sequence jit-

ter. Experimental results on human3.6M and MPI-INF-3DHP benchmarks demonstrate that our method can improve estimation accuracy while maintaining the temporal stability of the generated 3D pose sequence. In the future, we will continue to explore the performance upper bound of the model, striving to make it feasible for real-world applications.

Acknowledgments. This research is supported by National Key R&D Program of China (No. 2022ZD0115902), National Natural Science Foundation of China (No. 62102208), Beijing Natural Science Foundation (No. 4232023).

Data availability. Data is available on reasonable request from the corresponding author.

Conflict of interest. The authors declare no potential conflict of interests.

References

1. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(1), 44–58 (2005)
2. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1014–1021 (2009)
3. Aouaidjia, K., Sheng, B., Li, P., Kim, J., Feng, D.D.: Efficient body motion quantification and similarity evaluation using 3-d joints skeleton coordinates. IEEE Transactions on Systems, Man, and Cybernetics: Systems **51**(5), 2774–2788 (2019)
4. Cai, Y., Ge, L., Liu, J., Cai, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2272–2281 (2019)
5. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 198–209 (2021)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
7. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1964–1973 (2021)
8. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: IEEE International Conference on Computer Vision, pp. 2262–2271 (2019)
9. Ci, H., Wu, M., Zhu, W., Ma, X., Dong, H., Zhong, F., Wang, Y.: Gfpose: Learning 3d human pose prior with gradient fields. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4800–4810 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
11. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: AAAI Conference on Artificial Intelligence, vol. 32 (2018)
12. Feng, R., Gao, Y., Tse, T.H.E., Ma, X., Chang, H.J.: Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In: IEEE International Conference on Computer Vision, pp. 14,815–14,826 (2023)
13. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 13,041–13,051 (2023)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
15. Hu, W., Zhang, C., Zhan, F., Zhang, L., Wong, T.T.: Conditional directed graph convolution for 3d human pose estimation. In: ACM International Conference on Multimedia, pp. 602–611 (2021)
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2013)
17. Jia, R., Yang, H., Zhao, L., Wu, X., Zhang, Y.: Mpa-gnet: multi-scale parallel adaptive graph network for 3d human pose estimation. The Visual Computer pp. 1–17 (2023)
18. Kamel, A., Liu, B., Li, P., Sheng, B.: An investigation of 3d human pose estimation for learning tai chi: A human factor perspective. International Journal of Human–Computer Interaction **35**(4-5), 427–439 (2019)
19. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5614–5623 (2019)
20. Lan, G., Wu, Y., Hu, F., Hao, Q.: Vision-based human pose estimation via deep learning: A survey. IEEE Transactions on Human-Machine Systems **53**(1), 253–268 (2023)
21. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9887–9895 (2019)
22. Li, C., Lee, G.H.: Weakly supervised generative network for multiple 3d human pose hypotheses. In: British Machine Vision Conference (2020)
23. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia **25**, 1282–1293 (2022)
24. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 13,147–13,156 (2022)
25. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-lm improves controllable text generation. Advances in Neural Information Processing Systems **35**, 4328–4343 (2022)

26. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5064–5073 (2020)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
28. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 11,461–11,471 (2022)
29. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: IEEE International Conference on Computer Vision, pp. 2640–2649 (2017)
30. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: International Conference on 3D Vision, pp. 506–516 (2017)
31. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499 (2016)
32. Oikarinen, T., Hannah, D., Kazerounian, S.: Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In: International Joint Conference on Neural Networks, pp. 1–9 (2021)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)
34. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2020)
35. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 640–649 (2023)
36. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision, pp. 461–478. Springer (2022)
37. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: IEEE International Conference on Computer Vision, pp. 14,715–14,725 (2023)
38. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: IEEE International Conference on Computer Vision, pp. 2325–2334 (2019)
39. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
41. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: European Conference on Computer Vision, pp. 529–545 (2018)
42. Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4790–4799 (2023)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems pp. 5998–6008 (2017)
44. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. In: IEEE International Conference on Computer Vision, pp. 11,179–11,188 (2021)
45. Zeng, A., Yang, L., Ju, X., Li, J., Wang, J., Xu, Q.: Smoothnet: A plug-and-play network for refining human poses in videos. In: European Conference on Computer Vision, pp. 625–642. Springer (2022)
46. Zhang, H., Hu, Z., Sun, Z., Zhao, M., Bi, S., Di, J.: A fused convolutional spatio-temporal progressive approach for 3d human pose estimation. The Visual Computer **40**(6), 4387–4399 (2024)
47. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 13,232–13,242 (2022)
48. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
49. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3425–3435 (2019)
50. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8877–8886 (2023)
51. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: IEEE International Conference on Computer Vision, pp. 11,656–11,665 (2021)