FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Models

Jinglin Xu[1] Yijie Guo[2] Yuxin Peng[2*]
[1] School of Intelligence Science and Technology, University of Science and Technology Beijing
[2] Wangxuan Institute of Computer Technology, Peking University
xujinglinlove@gmail.com; 2000012936@stu.pku.edu.cn; pengyuxin@pku.edu.cn

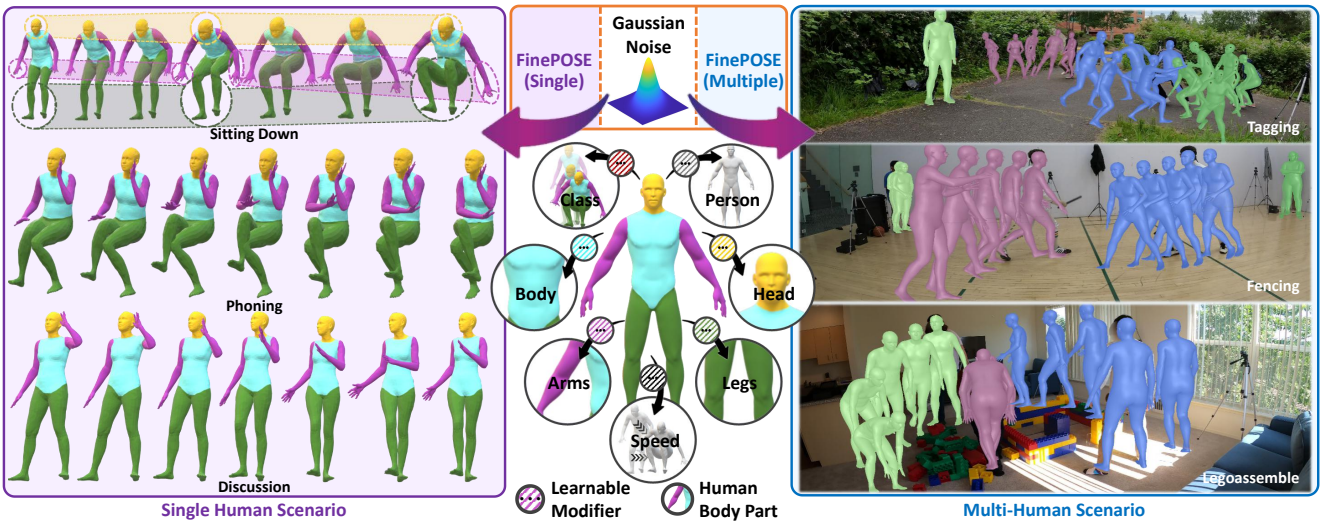

Figure 1. **Illustration of Fine-grained Prompt-driven Denoiser (FinePOSE).** FinePOSE, the proposed diffusion model-based 3D human pose estimation approach, enables multi-granularity manipulation controlled by learnable modifiers (e.g., "action class", coarse- and fine-grained human body parts including "person, head, body, arms, legs", and kinematic information "speed"), boosting motion reconstruction for single human and multi-human scenarios.

Abstract

The 3D Human Pose Estimation (3D HPE) task uses 2D images or videos to predict human joint coordinates in 3D space. Despite recent advancements in deep learning-based methods, they mostly ignore the capability of coupling accessible texts and naturally feasible knowledge of humans, missing out on valuable implicit supervision to guide the 3D HPE task. Moreover, previous efforts often study this task from the perspective of the whole human body, neglecting fine-grained guidance hidden in different body parts. To this end, we present a new Fine-Grained Prompt-Driven Denoiser based on a diffusion model for 3D HPE, named ***FinePOSE****. It consists of three core blocks enhancing the reverse process of the diffusion model: (1) Fine-grained Part-aware Prompt learning (FPP) block constructs fine-grained part-aware prompts via coupling accessible texts and naturally feasible knowledge of body parts with learnable prompts to model implicit guidance. (2) Fine-grained Prompt-pose Communication (FPC) block establishes fine-grained communications between learned part-aware prompts and poses to improve the denoising quality. (3) Prompt-driven Timestamp Stylization (PTS) block integrates learned prompt embedding and temporal information related to the noise level to enable adaptive adjustment at each denoising step. Extensive experiments on public single-human pose estimation datasets show that FinePOSE outperforms state-of-the-art methods. We further extend FinePOSE to multi-human pose estimation. Achieving 34.3mm average MPJPE on the EgoHumans dataset demonstrates the potential of FinePOSE to deal with complex multi-human scenarios. Code is available at* https://github.com/PKU-ICST-MIPL/FinePOSE_CVPR2024.

1. Introduction

Given monocular 2D images or videos, 3D Human Pose Estimation (3D HPE) aims to predict the positions of human body joints in 3D space. It is vital in various applications, including self-driving [50, 56], sports analysis [13, 31, 46],

*Corresponding author.

abnormal detection [9, 45], and human-computer interaction [11, 25, 42]. Considering the expensive computational costs of directly obtaining 3D human poses from 2D contents, 3D HPE is usually decomposed into two stages: 1) detecting 2D keypoints in images or videos [5, 7, 24, 39], and 2) mapping 2D keypoints to 3D human poses [6, 10, 35, 48, 52]. In this work, we mainly focus on the second stage, estimating 3D human poses given 2D keypoints.

Existing monocular 3D HPE methods [4, 6, 10, 17–19, 27, 28, 35, 36, 43, 44, 47, 48, 52, 54, 59, 61] usually have three challenges as follows: 1) Uncertainty: the depth ambiguity inherently exists in the mapping from 2D skeletons to 3D ones (one-to-many); 2) Complexity: flexible human body structure, complex inter-joint relationships, and a high limb freedom degree lead to self-occlusion or rare and complicated poses; 3) Generalizability: current publicly available 3D HPE datasets have limited action classes, and thus, the models trained on such data are prone to overfitting and difficult to generalize to more diverse action classes.

To address these issues, we consider improving the 3D HPE model performance by enhancing the input information. We found that existing methods ignore accessible texts and naturally feasible knowledge of humans while they promise to provide the model with more guidance. We explicitly utilize (1) the action class of human poses, (2) kinematic information "speed", and (3) the way that different human body parts (e.g., person, head, body, arms, and legs) move in human activities to build ***fine-grained part-aware prompts*** for the reconstruction task. Specifically, we incorporate a fine-grained part-aware prompt learning mechanism into our framework to drive 3D human pose estimation via vision-language pre-trained models. It is well known that text prompts play a crucial role in various downstream tasks for vision-language pre-training models (e.g., CLIP [30]). However, manually designing prompt templates is expensive and cannot ensure that the final prompt is optimal for the 3D HPE task. Thus, we create a new fine-grained part-aware prompt learning mechanism that adaptively learns modifiers for different human body parts to precisely describe their movements from multiple granularities, including action class, speed, the whole person, and fine-grained human body parts. This new mechanism, coupled with diffusion models, possesses controllable high-quality generation capability, which is beneficial in addressing the challenges of the 3D human pose estimation task.

In this work, we propose a Fine-grained Prompt-driven Denoiser (***FinePOSE***) based on diffusion models for 3D human pose estimation, in Fig. 1, which is composed of a fine-grained part-aware prompt learning (***FPP***) block, fine-grained prompt-pose communication (***FPC***) block, and prompt-driven timestamp stylization (***PTS***) block. Concretely, the FPP block encodes three kinds of information about the human pose, including action class, coarse- and fine-grained parts of humans like "person, head, body, arms, legs", and kinematic information "speed", and integrates them with pose features for serving subsequent processes. Then, the FPC block injects fine-grained part-aware prompt embedding into noise 3D poses to establish fine-grained communications between learnable part-aware prompts and poses for enhancing the denoising capability. To handle 3D poses with different noise levels, the PTS block introduces the timestamp coupled with fine-grained part-aware prompt embedding into the denoising process to enhance its adaptability and refine the prediction at each noise level.

Our contributions can be summarized as follows:

- We propose a new fine-grained part-aware prompt learning mechanism coupled with diffusion models that possesses human body part controllable high-quality generation capability, beneficial to the 3D human pose estimation task.
- Our FinePOSE encodes multi-granularity information about action class, coarse- and fine-grained human parts, and kinematic information, and establishes fine-grained communications between learnable part-aware prompts and poses for enhancing the denoising capability.
- Extensive experiments illustrate that our FinePOSE obtains substantial improvements on Human3.6M and MPI-INF-3DHP datasets and achieves state-of-the-art. More experiments on EgoHumans demonstrate the potential of FinePOSE to deal with complex multi-human scenarios.

2. Related Work

Diffusion Models. Diffusion models [12, 26, 37, 38] are a kind of generative models that sequentially add a series of noise with different levels to the raw data, gradually transforming it from an original data distribution to a noisy distribution, and subsequently reconstructing the original data by denoising. Diffusion models have strong capabilities in many applications, from 2D image or video generation/editing [1–3, 16, 49] to 3D human pose estimation/generation [10, 17, 19, 27, 35, 47, 48, 52, 54, 59]. The 3D HPE task, for example, encounters various difficulties, including occlusions, limited training data, and inherent ambiguity in pose representations. Therefore, diffusion models' ability to generate high-fidelity 3D human poses makes them more suitable for 3D HPE.

3D Human Pose Estimation. Considering that extracting 2D human skeletons from videos or images requires expensive costs, the 3D human pose estimation task is usually divided into two phases: (1) estimating 2D positions of human joints from images or videos [5, 7, 22, 41], and (2) mapping 2D positions to the 3D space to estimate the 3D positions of human joints [4, 6, 10, 17–19, 27, 28, 35, 36, 43, 47, 48, 52, 54, 59, 61]. In this work, we focus on the second phase. Early, TCN [29] used a fully convolutional network based on dilated temporal convolutions over 2D keypoints to estimate 3D poses in video.

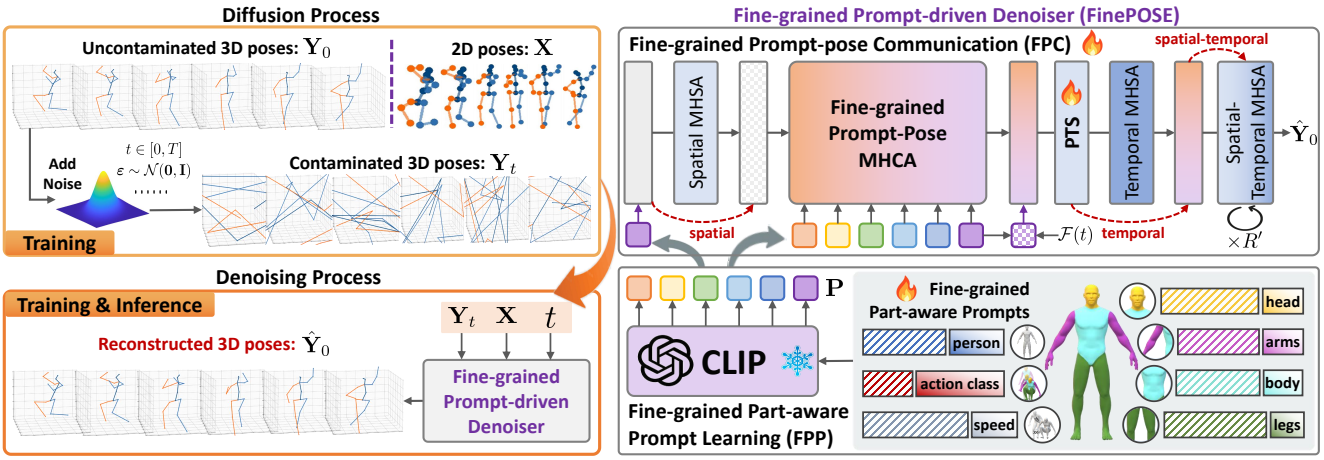

Figure 2. **The architecture of the proposed FinePOSE.** In the diffusion process, Gaussian noise is gradually added to the ground-truth 3D poses $\mathbf{Y}_0$, generating the noisy 3D poses $\mathbf{Y}_t$ for the timestamp t. In the denoising process, $\mathbf{Y}_t$, $\mathbf{X}$ and t are fed to fine-grained prompt-driven denoiser $\mathcal{D}$ to reconstruct pure 3D poses $\hat{\mathbf{Y}}_0$. $\mathcal{D}$ is composed of a Fine-grained Part-aware Prompt learning (FPP) block, a Fine-grained Prompt-pose Communication (FPC) block, and a Prompt-driven Timestamp Stylization (PTS) block, where FPP provides more precise guidance for all human part movements, FPC establishes fine-grained communications between learnable prompts and poses for enhancing the denoising capability, and PTS integrates learned prompt embedding and current timestamp for refining the prediction at each noise level.

SRNet [51] proposed a split-and-recombine approach, leading to appreciable improvements in predicting rare and unseen poses. Anatomy [6] decomposed the task into bone direction prediction and bone length prediction, from which the 3D joint locations can be derived entirely. Recently, MixSTE [52] used temporal and spatial transformers alternately to obtain better spatio-temporal features. MotionBERT [59] proposed a pretraining stage to recover the underlying 3D motion from noisy partial 2D observations. GLA-GCN [48] globally modeled the spatio-temporal structure for 3D human pose estimation. D3DP [35] proposed the joint-level aggregation strategy to benefit from all generated poses. Unlike previous methods, our approach proposes a new fine-grained part-aware prompt learning mechanism coupled with diffusion models that possess controllable, high-quality generation capability of human body parts, which benefits the 3D human pose estimation task.

Prompt Learning. Prompt learning has been widely used in the computer vision community [8, 21, 57, 58]. Typically, CoOp [58] utilized a continuous prompt optimization from downstream data instead of hand-craft design, the pioneering work that brings prompt learning to adapt pre-trained vision language models. CoCoOp [57] extended CoOp by learning image conditional prompts to improve generalization. ProDA [21] learned a prompt distribution over the output embedding space. VPT [8] introduced variational prompt tuning by combining a base learned prompt with a residual vector sampled from an instance-specific underlying distribution. PointCLIPV2 [60] combined CLIP [30] with GPT [20] to be a unified 3D open-world learner. Unlike the above methods, we propose a new fine-grained part-aware prompt learning mechanism, which encodes multi-granularity information about action class, coarse- and fine-grained human parts, and kinematic data, and establishes fine-grained communications between learnable part-aware prompts and poses for enhancing the denoising capability.

3. The Proposed Approach: FinePOSE

Given a 2D keypoints sequence $\mathbf{X} \in \mathbb{R}^{N \times J \times 2}$, constructed by N frames with J joints in each, the proposed approach is formulated to predict the 3D pose sequence $\mathbf{Y} \in \mathbb{R}^{N \times J \times 3}$. Considering the high-quality generation capability of the text-controllable denoising process of diffusion models, we develop a Fine-grained Prompt-driven Denoiser (FinePOSE) $\mathcal{D}$ for 3D human pose estimation. FinePOSE generates accurate 3D human poses enhanced by three core blocks: Fine-grained Part-aware Prompt learning (FPP), Fine-grained Prompt-pose Communication (FPC), and Prompt-driven Timestamp Stylization (PTS) blocks.

3.1. Diffusion-Based 3D Human Pose Estimation

Diffusion models are generative models that model the data distribution in the form of $p_\theta(\mathbf{Y}_0) := \int p_\theta(\mathbf{Y}_{0:T}) d\mathbf{Y}_{1:T}$ through chained diffusion and reverse (denoising) processes. The diffusion process gradually adds Gaussian noise into the ground truth 3D pose sequence $\mathbf{Y}_0$ to corrupt it into an approximately Gaussian noise $\mathbf{Y}_t (t \to T)$ using a variance schedule $\{\beta_t\}_{t=1}^T$, which can be formulated as

$$q(\mathbf{Y}_t \mid \mathbf{Y}_0) := \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \tag{1}$$

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$. Afterward, the denoising process reconstructs the uncontaminated 3D poses by a denoiser $\mathcal{D}$. Since the degraded data is well approximated

by a Gaussian distribution after the diffusion process, we can obtain initial 3D poses $\mathbf{Y}_T$ by sampling noise from a unit Gaussian. Passing $\mathbf{Y}_T(t=T)$ to the denoiser $\mathcal{D}$, we obtain $\hat{\mathbf{Y}}_0$ that is thereafter used to generate the noisy 3D poses $\hat{\mathbf{Y}}_{t-1}$ as inputs to the denoiser $\mathcal{D}$ at timestamp $t-1$ via DDIM [37], which can be formulated as

$$\mathbf{Y}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{Y}}_0 + \epsilon_t\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} + \sigma_t\epsilon, \tag{2}$$

where t is from T to 1, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise independent of $\mathbf{Y}_t$, and

$$\epsilon_t = \left(\mathbf{Y}_t - \sqrt{\bar{\alpha}_t}\cdot\hat{\mathbf{Y}}_0\right)/\sqrt{1-\bar{\alpha}_t}, \tag{3a}$$

$$\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\cdot\sqrt{1-(\bar{\alpha}_t/\bar{\alpha}_{t-1})}, \tag{3b}$$

where ϵ_t is the noise at timestamp t, and σ_t controls how stochastic the diffusion process is.

3.2. Fine-grained Prompt-driven Denoiser

Fine-grained Part-aware Prompt Learning (FPP). To assist the reconstruction of pure 3D poses $\hat{\mathbf{Y}}_0$ from contaminated 3D poses $\mathbf{Y}_t$ with additional information, FinePOSE guides the denoising process with regular 2D keypoints $\mathbf{X}$, timestamp t, and fine-grained part-aware prompt embedding $\mathbf{P}$. We design the FPP block to learn $\mathbf{P}$. It encodes three pose-related information in the prompt embedding space, including its action class, coarse- and fine-grained parts of humans like "person, head, body, arms, legs", and kinematic information "speed". Afterward, $\mathbf{P}$ is integrated with pose features for subsequent processes.

A learnable prompt embedding $\mathbf{P} = \{\boldsymbol{p}\}_{k=1}^{K}$ is with the shape of $K \times L \times D$, where K denotes the number of text prompts, L indicates the number of tokens in each text prompt, and D is the dimension of token embedding. Since the number of valid tokens is found to be three to four through the text encoder $\mathcal{E}_{\text{tx}}$, the first four tokens are taken as representations $\tilde{\boldsymbol{p}}_k$ for each text. Moreover, since modifiers help precisely describe the movements of human body parts, we design a learnable vector $\boldsymbol{r}_k \in \mathbb{R}^{(L_k-4)\times D}$ to wrap the representations as $\boldsymbol{p}_k$. The above can be formulated as

$$\tilde{\boldsymbol{p}}_k = \mathcal{E}_{\text{tx}}(\text{text}_k)[:4],\ k \in [1, K], \tag{4a}$$

$$\boldsymbol{p}_k = \text{Concat}(\boldsymbol{r}_k, \tilde{\boldsymbol{p}}_k), \tag{4b}$$

where $K = 7$ and $\{\text{text}_k\}_{k=1}^{7}$ indicate {person, [Action Class], speed, head, body, arms, legs}. $\boldsymbol{r}_k$ is initialized with Gaussian distribution of $\mu = 0$ and $\sigma = 0.02$, and $\{L_k\}_{k=1}^{7} = \{7, 12, 10, 10, 10, 14, 14\}$, which sums to 77 regarding the text embedding dimension of CLIP [30]. In short, the FPP block builds multi-granularity text prompts and learnable modifiers, providing precise guidance for each human body part, as shown in Fig. 2.

Fine-grained Prompt-pose Communication (FPC). After obtaining fine-grained part-aware prompt embedding $\mathbf{P}$, we establish fine-grained communications between learned part-aware prompts and poses using the FPC block to improve the denoising quality. Specifically, when processing the noised 3D poses $\mathbf{Y}_t$, it injects prompt embedding $\mathbf{P}$, 2D keypoints $\mathbf{X}$, and timestamp t within.

First, FPC integrates $\mathbf{Y}_t$ and guidance information (i.e., $\mathbf{X}$, t, and $\mathbf{P}$) by a series of concatenation and addition operations, as $\mathbf{Z}_t = \text{Concat}(\mathbf{Y}_t, \mathbf{X}) + \mathbf{P}[L] + \mathcal{F}(t)$. $\mathcal{F}$ is the timestamp embedding network containing a sinusoidal function followed by two Linear layers connected by a GELU non-linearity. The timestep embedding adaptively adjusts the quantity of Gaussian noise additions. Since the denoiser $\mathcal{D}$ works iteratively, providing detailed information about the current timestamp t is crucial for $\mathcal{D}$ to handle 3D poses containing different noise levels effectively. Then, $\mathbf{Z}_t$ is encoded by a spatial transformer, where the multi-head self-attention (MHSA) mechanism helps to focus on the fine-grained relationships between joints within each frame, obtaining $\mathbf{Z}_t^s$.

To completely inject prompt embedding $\mathbf{P}$ into $\mathbf{Z}_t^s$, we implement a multi-head cross-attention model, where the *query*, *key*, and *value* are as $\mathbf{Q} = \mathbf{W}_Q\mathbf{Z}_t^s$, $\mathbf{K} = \mathbf{W}_K\mathbf{P}$, $\mathbf{V} = \mathbf{W}_V\mathbf{P}$. The *value* is aggregated with cross-attention $\mathbf{A}$ to generate fine-grained prompt-driven pose features $\mathbf{Z}_t^{sp}$, achieving fine-grained prompt-pose communication. The mechanism can be formulated as

$$\mathbf{A} = \text{softmax}(\mathbf{Q} \otimes \mathbf{K}^\top/\sqrt{d}), \tag{5a}$$

$$\mathbf{Z}_t^{sp} = \mathbf{A} \otimes \mathbf{V},\ \tilde{\mathbf{Z}}_t^{sp} = \mathcal{P}(\mathbf{Z}_t^{sp}), \tag{5b}$$

where $d = D/H$ and H is the number of attention heads. $\mathcal{P}$ indicates the PTS block that bring timestamp t into the generation process to obtain timestamp stylized output $\tilde{\mathbf{Z}}_t^{sp}$. On the other hand, to model inter-frame relationships between poses, $\tilde{\mathbf{Z}}_t^{sp}$ is encoded using a temporal transformer via MHSA to obtain $\tilde{\mathbf{Z}}_t^{spf}$. Finally, we utilize a spatial-temporal transformer accompanied by permutation operations between spatial and temporal dimensions to extract more compact fine-grained prompt-driven pose features from $\tilde{\mathbf{Z}}_t^{spf}$, which are decoded as the predicted 3D poses $\hat{\mathbf{Y}}_0$.

Prompt-driven timestamp Stylization (PTS). As mentioned, providing timestamp embedding to the denoising process is critical for handling 3D poses with different noise levels. Therefore, inspired by Motiondiffuse [53], we introduce the PTS block that explicitly embeds timestamp t by positional embedding [40] and sums it with the learnable prompt embedding $\mathbf{P}$ obtained by the FPP block, as $\boldsymbol{v} = \mathbf{P}[L] + \mathcal{F}(t)$. Given the intermediate output $\mathbf{Z}_t^{sp}$ of the FPC block, the PTS block calculates $\tilde{\mathbf{Z}}_t^{sp} = \mathbf{Z}_t^{sp}\cdot\psi_w(\phi(\boldsymbol{v})) + \psi_b(\phi(\boldsymbol{v}))$, where ψ_b, ψ_w, ϕ are three different linear projections, and $(\cdot)$ is the Hadamard product.

3.3. Training & Inference

Training. The contaminated 3D poses $\mathbf{Y}_t$ is sent to a fine-grained prompt-driven denoiser $\mathcal{D}$ to reconstruct the 3D

Method	N	Human3.6M (DET)			Human3.6M (GT)			Year
		Detector	MPJPE ↓	P-MPJPE ↓	Detector	MPJPE ↓	P-MPJPE ↓	
TCN [29]	243	CPN	46.8	36.5	GT	37.8	/	CVPR'19
Anatomy [6]	243	CPN	44.1	35.0	GT	32.3	/	CSVT'21
P-STMO [33]	243	CPN	42.8	34.4	GT	29.3	/	ECCV'22
MixSTE [52]	243	HRNet	39.8	30.6	GT	21.6	/	CVPR'22
PoseFormerV2 [54]	243	CPN	45.2	35.6	GT	35.5	/	CVPR'23
MHFormer [19]	351	CPN	43.0	34.4	GT	30.5	/	CVPR'22
Diffpose [10]	243	CPN	36.9	<u>28.7</u>	GT	18.9	/	CVPR'23
GLA-GCN [48]	243	CPN	44.4	34.8	GT	21.0	17.6	ICCV'23
ActionPrompt [55]	243	CPN	41.8	29.5	GT	22.7	/	ICME'23
MotionBERT [59]	243	SH	37.5	/	GT	<u>16.9</u>	/	ICCV'23
D3DP [34]	243	CPN	<u>35.4</u>	<u>28.7</u>	GT	18.4	/	ICCV'23
FinePOSE (Ours)	243	CPN	**31.9** (-3.5)	**25.0** (-3.7)	GT	**16.7** (-0.2)	**12.7** (-4.9)	

Table 1. **Quantitative comparison with the state-of-the-art 3D human pose estimation methods on the Human3.6M dataset.** N: the number of input frames. CPN, HRNet, SH: using CPN [7], HRNet [39], and SH [24] as the 2D keypoint detectors to generate the inputs. GT: using the ground truth 2D keypoints as inputs. The best and second-best results are highlighted in **bold** and <u>underlined</u> formats.

poses $\hat{\mathbf{Y}}_0 = \mathcal{D}(\mathbf{Y}_t, \mathbf{X}, t, \mathbf{P})$ without noise. The entire framework is optimized by minimizing the MSE loss $\|\mathbf{Y}_0 - \hat{\mathbf{Y}}_0\|_2$.

Inference. Since the distribution of $\mathbf{Y}_T$ is nearly an isotropic Gaussian distribution, we sample H initial 3D poses $\{\mathbf{Y}_T^h\}_{h=1}^H$ from a unit Gaussian. After passing them to the denoiser $\mathcal{D}$, we obtain H feasible 3D pose hypotheses $\{\hat{\mathbf{Y}}_0^h\}_{h=1}^H$. Each hypothesis $\hat{\mathbf{Y}}_0^h$ is used to generate the noisy 3D poses $\hat{\mathbf{Y}}_{t-1}^h$ as inputs to the denoiser $\mathcal{D}$ for the next timestamp $t-1$. Then, we regenerate $\{\hat{\mathbf{Y}}_0^h\}_{h=1}^H$ using $\{\hat{\mathbf{Y}}_{t-1}^h\}_{h=1}^H$ as inputs to the denoiser $\mathcal{D}$ for the next timestamp $t-2$. Analogously, this process iterates M times starting from the timestamp T, so each iteration $m \in [1, M]$ is with the timestamp $t = T(1 - \frac{m}{M})$. Following Joint-Wise Reprojection-Based Multi-Hypothesis Aggregation (JPMA) in [35], we reproject $\{\hat{\mathbf{Y}}_0^h\}_{h=1}^H$ to the 2D camera plane using known or estimated intrinsic camera parameters and then choose joints with minimum projection errors with the input $\mathbf{X}$, as

$$h' = \underset{h \in [1,H]}{\arg\min} \|\mathcal{P}_R(\hat{\mathbf{Y}}_0^h)[j] - \mathbf{X}[j]\|_2, \tag{6a}$$

$$\hat{\mathbf{Y}}_0[j] = \hat{\mathbf{Y}}_0^{h'}[j],\ j \in [1, J], \tag{6b}$$

where $\mathcal{P}_R$ is the reprojection function, j is the index of joints, and h' indicates the index of selected hypothesis. JPMA enables us to select joints from distinct hypotheses automatically to form the final prediction $\hat{\mathbf{Y}}_0$.

3.4. Extension to 3D Multi-Human Pose Estimation

We append a post-integration to FinePOSE to apply for the multi-human scenario, avoiding incorporating extra computational cost. Specifically, given a multi-human 2D keypoints sequence $\mathbf{X}_{\text{mul}} \in \mathbb{R}^{C \times N \times J \times 2}$, which involves C human characters, FinePOSE first predicts $\hat{\mathbf{Y}}_0^c$ for each character $c \in [1, C]$. Considering that some characters may temporarily leave the camera field of view, their positions in those frames are set as zeros to ensure synchronization of all characters' states in $\mathbf{X}_{\text{mul}}$. Next, we integrate $\{\hat{\mathbf{Y}}_0^c\}_{c=1}^C$ by stacking over the character dimension, obtaining the final prediction $\hat{\mathbf{Y}}_0^C \in \mathbb{R}^{C \times N \times J \times 3}$.

4. Experiments

4.1. Datasets and Metrics

Human3.6M [14] is a widely used benchmark dataset in human pose estimation tasks, which provides a large-scale collection of accurate 3D joint annotations on diverse human activities. Human3.6M consists of 3.6 million RGB images, captured from multiple camera views, of 11 professional actors performing 15 activities, e.g., walking, running, and jumping. Following previous efforts [19, 29, 34], our FinePOSE is trained on five subjects (S1, S5, S6, S7, S8) and evaluated on two subjects (S9, S11). We calculate the mean per joint position error (i.e., MPJPE) to measure the average Euclidean distance in millimeters between the ground truth and estimated 3D joint positions for evaluation. We also report procrustes MPJPE (i.e., P-MPJPE) that calculates MPJPE after aligning the estimated poses to the ground truth using a rigid transformation.

MPI-INF-3DHP [23] provides synchronized RGB video sequences with accurate 3D joint annotations for 3D human pose estimation. It comprises 8 activities conducted by 8 actors in the training set, while the test set encompasses 7 activities. We calculate MPJPE, the percentage of correctly estimated keypoints (i.e., PCK) within a 150mm range, and the area under the curve (i.e., AUC).

EgoHumans [15] collects multi-human ego-exo videos covering 7 sports activities. Recently, a subset of 2D to 3D keypoints annotations has been released covering tagging, lego-assembling, and fencing. It contains 105 RGB videos taken by ego cameras. Between 1 and 3 human characters

Method / MPJPE ↓	Human3.6M (DET)															
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
TCN [29]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
SRNet [51]	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
RIE [32]	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy [6]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
P-STMO [33]	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 [52]	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
PoseFormerV2 [54]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
MHFormer [19]	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 [10]	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
GLA-GCN [48]	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
ActionPrompt [55]	37.7	40.2	39.8	40.6	43.1	48.0	38.8	38.9	50.8	63.2	42.0	40.0	42.0	30.5	31.6	41.8
MotionBERT [59]	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	46.9	53.9	39.5	36.3	35.8	25.1	25.3	37.5
D3DP [34]	33.0	34.8	31.7	33.1	37.5	43.7	34.8	33.6	45.7	47.8	37.0	35.0	35.0	24.3	24.1	35.4
FinePOSE (Ours)	**31.4** (-1.6)	**31.5** (-3.3)	**28.8** (-2.9)	**29.7** (-2.4)	**34.3** (-3.2)	**36.5** (-7.2)	**29.2** (-5.6)	**30.0** (-3.6)	**42.0** (-3.7)	**42.5** (-5.3)	**33.3** (-3.7)	**31.9** (-3.1)	**31.4** (-3.6)	**22.6** (-1.7)	**22.7** (-1.4)	**31.9** (-3.5)

Table 2. **Quantitative comparison with the state-of-the-art 3D human pose estimation methods on the Human3.6M dataset using 2D keypoint detectors to generate the inputs.** *Dir.*, *Disc.*, $\cdots$, and *WalkT.* correspond to 15 action classes. *Avg* indicates the average MPJPE among 15 action classes. The best and second-best results are highlighted in **bold** and underlined formats.

Method	N	MPI-INF-3DHP			Year
		PCK↑	AUC↑	MPJPE ↓	
TCN [29]	81	86.0	51.9	84.0	CVPR'19
Anatomy [6]	81	87.9	54.0	78.8	CSVT'21
P-STMO [33]	81	97.9	75.8	32.2	ECCV'22
MixSTE [52]	27	94.4	66.5	54.9	CVPR'22
PoseFormerV2 [54]	81	97.9	78.8	27.8	CVPR'23
MHFormer [19]	9	93.8	63.3	58.0	CVPR'22
Diffpose [10]	81	98.0	75.9	29.1	CVPR'23
GLA-GCN [48]	81	98.5	79.1	27.8	ICCV'23
D3DP [34]	243	98.0	79.1	28.1	ICCV'23
FinePOSE (Ours)	243	**98.9** (+0.4)	**80.0** (+0.9)	**26.2** (-1.6)	

Table 3. **Quantitative comparison with the state-of-the-art 3D human pose estimation methods on the MPI-INF-3DHP dataset using ground truth 2D keypoints as inputs.** N: the number of input frames. The best and second-best results are highlighted in **bold** and underlined formats.

appear in each video, resulting in a total of 238 subsequences. We report the average MPJPE per video.

4.2. Implementation Details

We take MixSTE [52] as the backbone of the denoiser $\mathcal{D}$ and CLIP as the frozen text encoder $\mathcal{E}_{\text{tx}}$. The numbers of MHSA-MLP-LN building blocks of the spatial, temporal, and spatio-temporal transformer in the FPC block are 1, 1, and 3. The training epoch in all the experiments below is 100, and the batch size is 4. We adopt AdamW optimizer with the momentum parameters of β_1 =0.9, β_2 =0.999, and the weight decay of 0.1. The learning rate starts from $6e^{-5}$ and shrinks after each epoch with a factor of 0.993. For fair comparisons, we set the number of hypotheses $H = 1$ and iterations $M = 1$ during training, and $H = 20$ and $M = 10$ during inference, as in D3DP [34].

Method	Human3.6M (DET)	
	MPJPE ↓	P-MPJPE ↓
w/o Prompt	37.2	29.1
M-Prompt	35.8	28.1
S-Prompt	36.2	28.9
C-Prompt	34.7	27.4
AL-Prompt	34.6	27.4
FinePOSE (Ours)	**31.9**	**25.0**

Table 4. **Ablation study on different designs of prompt learning in the FPP block.** w/o Prompt: without any textual information and learnable prompts. M-Prompt: using the action class to design the prompt manually. S-Prompt: using a learnable prompt combined with the action class. C-Prompt: employing the action class and coarse-grained information to create the prompt. AL-Prompt: only learnable prompts without any manual design.

4.3. Comparison with the State-of-the-Arts

Human3.6M. Tab. 1 reports comparisons between our FinePOSE with state-of-the-art (SOTA) 3D HPE methods on the Human3.6M dataset. FinePOSE significantly achieves new SOTA performance, especially when using detected 2D keypoints as inputs. Compared with existing 3D HPE methods, FinePOSE surpasses the SOTA method D3DP [34] by 3.5mm in MPJPE and 3.7mm in P-MPJPE. When using ground truth 2D keypoints as inputs, FinePOSE also significantly outperforms the SOTA method MotionBERT [59], improving MPJPE by 0.2mm. Tab. 2 provides detailed comparisons between on each action class using 2D keypoint detectors as inputs. For example, our FinePOSE achieves noticeable improvements (43.7mm→36.5mm) for the action class "*Photo*" and decreases average MPJPE by 3.5mm (35.4mm→31.9mm).

MPI-INF-3DHP. Tab. 3 reports comparisons between our FinePOSE and SOTA 3D HPE methods on the MPI-INF-

Method	Configuration			MPJPE ↓	P-MPJPE ↓
	FPP	FPC	PTS		
Baseline				37.2	29.1
w FPP	✓			35.3	28.0
w/o FPP			✓	37.1	29.2
w/o FPC	✓		✓	35.7	27.8
w/o PTS	✓	✓		36.6	29.0
FinePOSE (Ours)	✓	✓	✓	**31.9**	**25.0**

Table 5. **Ablation study on different configurations of FinePOSE on Human3.6M using 2D keypoint detectors as inputs.** Baseline: the method without any textual information via prompt learning. w FPP: the method only contains the FPP block and adds $\mathbf{P}[L]$ to the input. w/o FPP: the method without the FPP block leads to an infeasible FPC block. w/o FPC: the method without the FPC block. w/o PTS: the method without the PTS block.

3DHP dataset, using ground truth 2D keypoints as inputs. Compared with the SOTA existing method GLA-GCN [48], FinePOSE decreases MPJPE by 1.6mm and increases the PCK by 0.4% and AUC by 0.9%. Overall, these experimental results demonstrate that our FinePOSE benefits from fine-grained part-aware prompt learning and pose-prompt communications, resulting in higher denoising quality and estimation accuracy.

4.4. Ablation Study

We conduct a series of analysis experiments of our FinePOSE on the Human3.6M dataset to investigate the effects on the performance of different prompt learning designs in the FPP block and different blocks in FinePOSE.

Effects of Different Designs in FPP. We design various versions of the FPP block for our FinePOSE, including a) w/o Prompt, b) M-Prompt, c) S-Prompt, d) C-Prompt, and e) AL-Prompt. Specifically, w/o Prompt denotes FinePOSE without introducing textual information and learnable prompts. M-Prompt indicates using the action class to design the prompt manually instead of the FPP block. Taking the action class "*Directions*" as an example, the manually designed prompt is "*a person is pointing directions with hands*". There are 15 action classes available in the Human3.6M dataset corresponding to 15 kinds of manually designed prompts. S-Prompt indicates utilizing learnable prompts combined with the action class. C-Prompt indicates employing the action class and coarse-grained information like "*person*" and "*speed*" to create the prompt. Finally, AL-Prompt means only using learnable prompts without any manual design.

We first evaluate the effect of manually designed prompts (i.e., M-Prompt) on Human3.6M. As shown in Tab. 4, compared to w/o Prompt, M-Prompt achieves a decrease of 1.4mm on MPJPE and 1.0mm on P-MPJPE, indicating that manually designing prompts is a practical strategy even though they cannot guarantee the prompt is optimal during the denoising process for the 3D HPE task. To evaluate the effectiveness of S-Prompt, we compare it with w/o

Method / MPJPE ↓	EgoHumans			
	Tag.	Lego	Fenc.	Avg
D3DP [35]	30.7	29.0	46.6	35.4
FinePOSE (Ours)	**30.0** (-0.7)	**26.7** (-2.3)	**46.2** (-0.4)	**34.3** (-1.1)

Table 6. **Quantitative comparison with D3DP on the EgoHumans dataset using 2D keypoints as inputs.** *Tag.*, *Lego*, and *Fenc.* correspond to 3 action classes. *Avg* indicates the average MPJPE among 3 action classes.

Prompt. As shown in Tab. 4, MPJPE and P-MPJPE are reduced by 1.0mm and 0.2mm, respectively, for S-Prompt, which demonstrates that with the help of learnable prompts, integrating textual information can improve the performance on 3D HPE task. While compared to M-Prompt, S-Prompt results in performance degradation, indicating that learnable prompts must be meticulously designed. In addition, we also investigate the impact of manual intervention degrees on 3D HPE performance using two groups of comparative experiments. In the first group, we used only learnable prompts without any textual information and manual intervention, named AL-Prompt, which differs from S-Prompt with the action class. The second group designed a coarse-grained prompt involving action class, "person", "speed", and corresponding learnable prompts, denoted as C-Prompt. We see that both AL-Prompt and C-Prompt outperform S-Prompt since AL-Prompt is without interference from uncomplete textual information and C-Prompt contains some important textual information like action class, "person", and "speed", which provide the action subject and kinematic data. Finally, it is observed that our FinePOSE outperforms various versions of prompt learning on both MPJPE and P-MPJPE, indicating the effectiveness of the fine-grained part-aware prompt learning mechanism in FinePOSE.

Effects of Different Blocks in FinePOSE. In Tab. 5, we provide different settings of our FinePOSE to evaluate the effects of different blocks for the 3D HPE performance, including Baseline, w FPP, w/o FPP, w/o FPC, and w/o PTS. Specifically, Baseline denotes FinePOSE without introducing textual information and learnable prompts, the same as the configuration of w/o Prompt. w FPP indicates FinePOSE only contains the FPP block without introducing the FPC and PTS blocks and only adds textual information $\mathbf{P}[L]$ to the input. w/o FPP denotes FinePOSE without the FPP block, leading to the FPC block being infeasible and only utilizing the PTS block. w/o FPC means FinePOSE without the FPC block but using the FPP and PTS blocks. w/o PTS refers to FinePOSE without the PTS block but using the FPP and FPC blocks to integrate textual information for fine-grained part-aware prompt learning.

Compared w FPP and Baseline, we observe that the former can achieve 1.9mm and 1.1mm improvements on MPJPE and P-MPJPE. This is because our FinePOSE con-

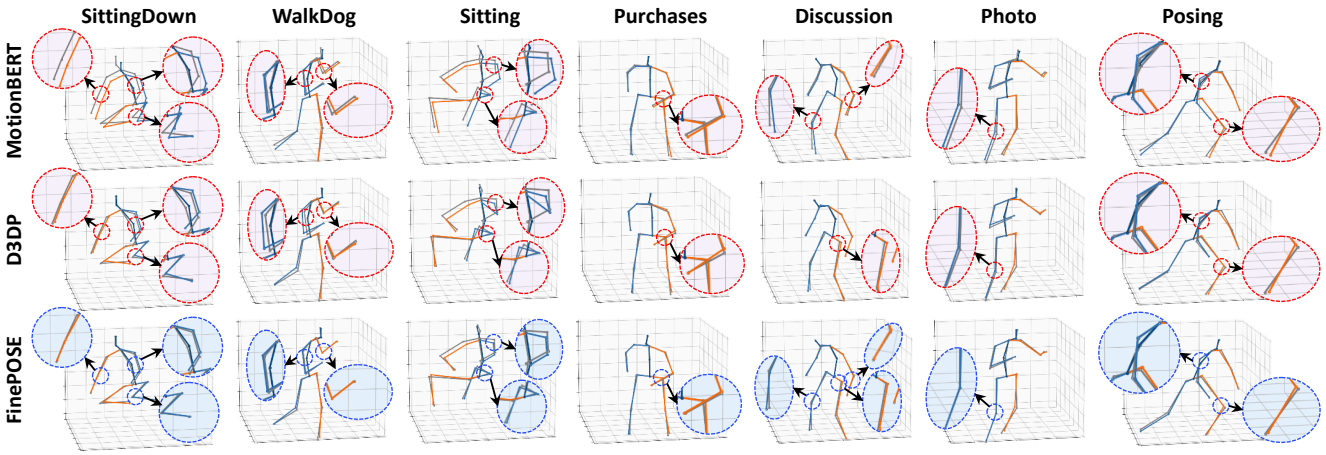

Figure 3. **Qualitative comparisons of our FinePOSE with MotionBERT [59] and D3DP [34] on Human3.6M.** The gray skeleton is the ground-truth 3D pose. The blue skeleton represents the prediction of the human left part, and the orange indicates the human right part. The red dashed line represents the incorrect regions of the compared methods, and the blue dashed line indicates the counterparts of FinePOSE.

tains the FPP block, which adds the prompt embedding $\mathbf{P}[L]$ into the input $\mathbf{Z}_t$ of denoiser $\mathcal{D}$, significantly improving the denoising capability. We observe that the results between w/o FPP and Baseline are almost equivalent. The baseline has already brought timestamp t into the denoising process, while the PTS block refines the prediction at each noise level by reusing the timestamp to the denoising process after the FPP and FPC block. Thus, there is nearly no effect in adding only the PTS block without FPP and FPC blocks to the denoiser. Making a comparison between w/o FPC and w/o FPP, the former achieves a decrease of 1.4mm on both MPJPE and P-MPJPE over w/o FPP, indicating that the FPP block in the denoiser plays a critical role in the fine-grained part-aware prompt learning mechanism. Finally, we observe that FinePOSE achieves a decrease of 4.7mm on MPJPE and 4.0mm on P-MPJPE compared to w/o PTS, indicating the necessity to integrate learned prompt embeddings and timestamps in the PTS block.

4.5. Results on 3D Multi-Human Pose Estimation

In real-world applications, the multi-human scenario is more common than the single-human one. However, its complexity hinders existing work from handling it. In Sec. 3.4, we present a post-integration to extend FinePOSE for the multi-human pose estimation task. We implemented the extension using the SOTA method D3DP for a convincing comparison. The experimental results on EgoHumans are reported in Tab. 6, demonstrating that (1) the integration strategy indeed has potential feasibility and (2) FinePOSE has a dominant performance even in the complex multi-human scenario.

4.6. Visualization

Fig. 3 shows the visualization results of D3DP [35], MotionBERT [59] and our FinePOSE on Human3.6M. These methods have performed well for actions in which the body, legs, and other parts of the person in the scene are relatively clear. For the actions with simple shapes, e.g., "Discussion" and "Photo", the 3D poses predicted by FinePOSE match better with ground-truth 3D poses than those of D3DP and MotionBERT, especially in the left knee, right arm, and right hip of "Discussion" and in the left knee of "Photo". For the actions with complex shapes, e.g., "Sitting" and "SittingDown", FinePOSE is more accurate at various joints, especially for arms and legs, while the 3D poses predicted by D3DP and MotionBERT differ significantly from ground-truth 3D poses.

5. Conclusion and Discussion

This work has presented FinePOSE, a new fine-grained prompt-driven denoiser for 3D human pose estimation. FinePOSE was composed of FPP, FPC, and PTS blocks. FPP learned fine-grained part-aware prompts to provide precise guidance for each human body part. FPC established fine-grained communication between learnable part-aware prompts and poses to enhance denoising capability. PTS brought timestamp information to the denoising process, strengthening the ability to refine the prediction at each noise level. Experimental results on two benchmarks demonstrated that FinePOSE surpasses the state-of-the-art methods. We have also extended FinePOSE from single-human scenarios to multi-human ones, exhibiting that our model performs well in complex multi-human scenarios.

Limitations. FinePOSE is not designed explicitly for the multi-person scenario. The diffusion model-based 3D HPE method is relatively computationally expensive.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (61925201, 62132001, 62373043) and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, 34:17981–17993, 2021. 2

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *TOG*, 42(4):1–11, 2023.

[3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[4] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, pages 2272–2281, 2019. 2

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 2

[6] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *TCSVT*, 32(1): 198–209, 2021. 2, 3, 5, 6

[7] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 2, 5

[8] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022. 3

[9] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, pages 919–929, 2020. 2

[10] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, pages 13041–13051, 2023. 2, 5, 6

[11] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, pages 14708–14718, 2021. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[13] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *CVPR*, pages 5218–5227, 2023. 1

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 5

[15] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *ICCV*, pages 19807–19819, 2023. 5

[16] Gyeongman Kim, Hajin Shim, Hyunsu Kim, Yunjey Choi, Junho Kim, and Eunho Yang. Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. In *CVPR*, pages 6091–6100, 2023. 2

[17] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, pages 9887–9895, 2019. 2

[18] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. In *BMVC*, 2020.

[19] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156, 2022. 2, 5, 6

[20] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023. 3

[21] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022. 3

[22] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *ECCV*, pages 72–88, 2022. 2

[23] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017. 5

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 2, 5

[25] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, pages 9890–9900, 2020. 2

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 2

[27] Qiang Nie, Ziwei Liu, and Yunhui Liu. Lifting 2d human pose to 3d with domain adapted 3d body concept. *IJCV*, 131 (5):1250–1268, 2023. 2

[28] Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *IJCNN*, pages 1–9, 2021. 2

[29] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762, 2019. 2, 5, 6

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3, 4

[31] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *CVPR*, pages 4738–4747, 2018. 1

[32] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *ACM MM*, pages 3446–3454, 2021. 6

[33] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *ECCV*, pages 461–478, 2022. 5, 6

[34] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, pages 14761–14771, 2023. 5, 6, 8

[35] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, pages 14761–14771, 2023. 2, 3, 5, 7, 8

[36] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, pages 2325–2334, 2019. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[39] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 2, 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4

[41] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10):3349–3364, 2020. 2

[42] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, pages 9401–9411, 2021. 2

[43] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *ICCV*, pages 11199–11208, 2021. 2

[44] Cunlin Wu, Yang Xiao, Boshen Zhang, Mingyang Zhang, Zhiguo Cao, and Joey Tianyi Zhou. C3p: Cross-domain pose prior propagation for weakly supervised 3d human pose estimation. In *ECCV*, pages 554–571, 2022. 2

[45] Jinglin Xu, Guangyi Chen, Jiwen Lu, and Jie Zhou. Unintentional action localization via counterfactual examples. *TIP*, 31:3281–3294, 2022. 2

[46] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 1

[47] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *CVPR*, pages 16105–16114, 2021. 2

[48] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *ICCV*, pages 8818–8829, 2023. 2, 3, 5, 6, 7

[49] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, pages 18456–18466, 2023. 2

[50] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *RL*, pages 1114–1124, 2023. 1

[51] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, pages 507–523, 2020. 3, 6

[52] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, pages 13232–13242, 2022. 2, 3, 5, 6

[53] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 4

[54] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *CVPR*, pages 8877–8886, 2023. 2, 5, 6

[55] Hongwei Zheng, Han Li, Bowen Shi, Wenrui Dai, Botao Wang, Yu Sun, Min Guo, and Hongkai Xiong. Actionprompt: Action-guided 3d human pose estimation with text and pose prompting. In *ICME*, pages 2657–2662, 2023. 5, 6

[56] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *CVPR*, pages 4478–4487, 2022. 1

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[59] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, pages 15085–15099, 2023. 2, 3, 5, 6, 8

[60] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, pages 2639–2650, 2023. 3

[61] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *ICCV*, pages 11477–11487, 2021. 2