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

北京大学 PEKING UNIVERSITY

北京科技大学 University of Science and Technology Beijing

Jinglin Xu, Yijie Guo and Yuxin Peng*

xujinglinlove@gmail.com; 2000012936@stu.pku.edu.cn; pengyuxin@pku.edu.cn

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation

- 3D Human Pose Estimation (3DHPE) is a task to estimate the 3D body joints and bones from 2D images or videos.
- 3DHPE is challenging since its uncertainty (depth ambiguity), complexity (complex human body structure).
- Most methods 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.
- Previous efforts often neglect fine-grained guidance hidden in different body parts.

(1) depth ambiguity
(2) complex human body structure

Contribution

- We propose **FinePOSE**, a new fine-grained part-aware prompt learning mechanism coupled with diffusion models.
- Our FinePOSE encodes multi-granularity information and establishes fine-grained communications between learnable part-aware prompts and poses.
- Extensive experiments illustrate that our approach obtains substantial improvements and achieves the state-of-the-art.

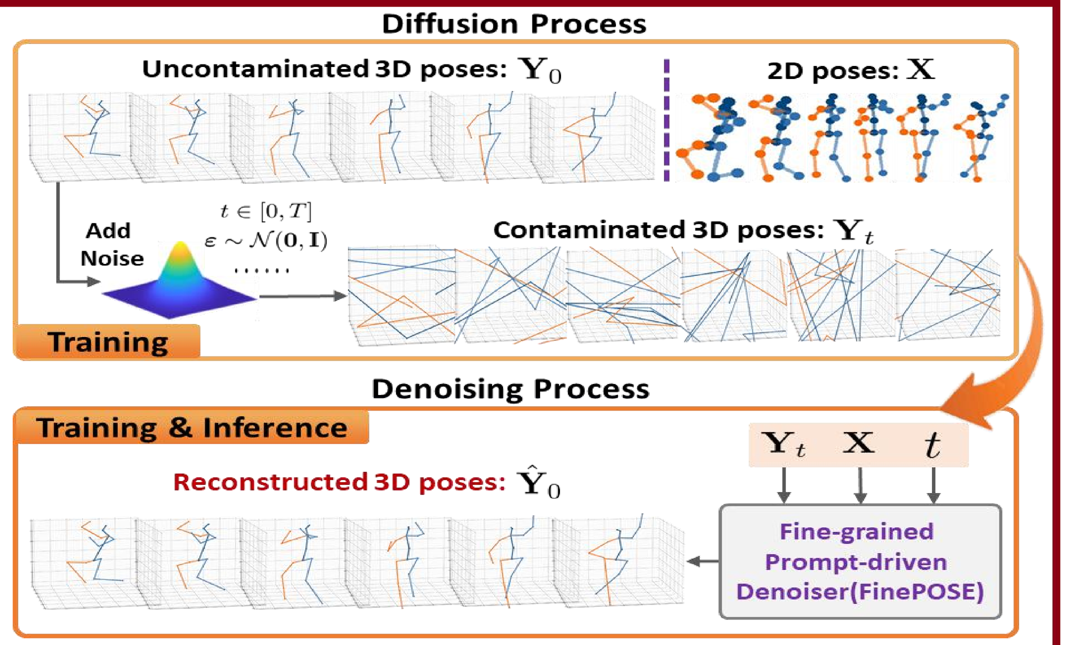
Diffusion Model-based 3D HPE

- Add Gaussian noise into Y_0

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

- Passing Y_t to D to get $\hat{Y}_0$
- Obtain Y_{t-1} at timestamp $t-1$

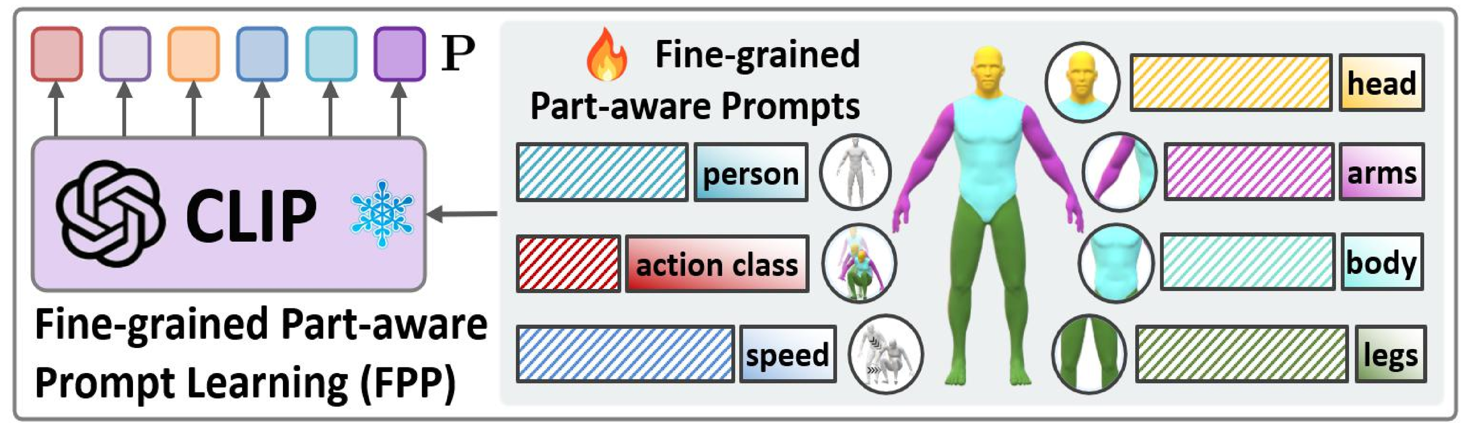
$$\mathbf{Y}_{t\text{-}1} = \sqrt{\bar{\alpha}_{t\text{-}1}}\hat{\mathbf{Y}}_0 + \epsilon_t\sqrt{1-\bar{\alpha}_{t\text{-}1}-\sigma_t^2} + \sigma_t\epsilon,$$



Method: FinePOSE

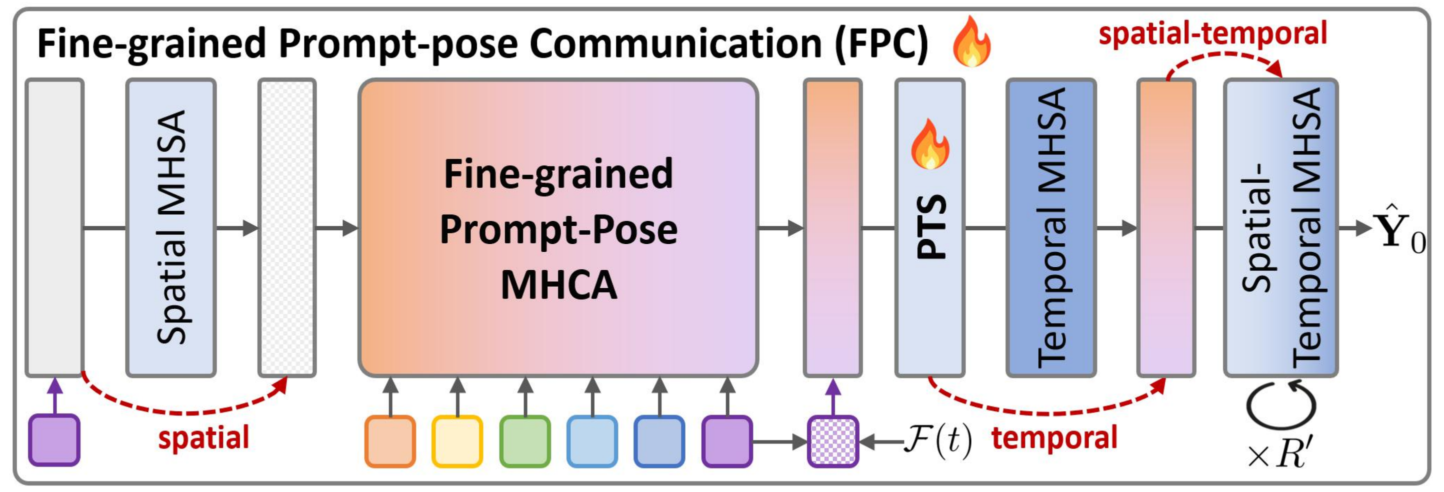
Fine-grained Part-aware Prompt Learning (FPP)

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 .



Fine-grained Prompt-pose Communication (FPC)

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.



Prompt-driven timestamp Stylization (PTS)

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.

Project Page Website GitHub Homepage

[1] D3DP: Diffusion based 3d human pose estimation with multi-hypothesis aggregation. ICCV 2023

[2] Motionbert: A unified perspective on learning human motion representations. ICCV 2023

Experiments

Method	N	Human3.6M (DET)			Human3.6M (GT)		
		Detector	MPJPE ↓	P-MPJPE ↓	Detector	MPJPE ↓	P-MPJPE ↓
TCN [29]	243	CPN	46.8	36.5	GT	37.8	/
Anatomy [6]	243	CPN	44.1	35.0	GT	32.3	/
P-STMO [33]	243	CPN	42.8	34.4	GT	29.3	/
MixSTE [52]	243	HRNet	39.8	30.6	GT	21.6	/
PoseFormerV2 [54]	243	CPN	45.2	35.6	GT	35.5	/
MHFormer [19]	351	CPN	43.0	34.4	GT	30.5	/
Diffpose [10]	243	CPN	36.9	<u>28.7</u>	GT	18.9	/
GLA-GCN [48]	243	CPN	44.4	34.8	GT	21.0	17.6
ActionPrompt [55]	243	CPN	41.8	29.5	GT	22.7	/
MotionBERT [59]	243	SH	37.5	/	GT	<u>16.9</u>	/
D3DP [34]	243	CPN	<u>35.4</u>	<u>28.7</u>	GT	18.4	/
FinePOSE (Ours)	243	CPN	**32.2** (-3.2)	**25.0** (-3.7)	GT	**16.7** (-0.2)	**12.7** (-4.9)

Method	N	MPI-INF-3DHP		
		PCK↑	AUC↑	MPJPE ↓
TCN [29]	81	86.0	51.9	84.0
Anatomy [6]	81	87.9	54.0	78.8
P-STMO [33]	81	97.9	75.8	32.2
MixSTE [52]	27	94.4	66.5	54.9
PoseFormerV2 [54]	81	97.9	78.8	<u>27.8</u>
MHFormer [19]	9	93.8	63.3	58.0
Diffpose [10]	81	98.0	75.9	29.1
GLA-GCN [48]	81	<u>98.5</u>	<u>79.1</u>	<u>27.8</u>
D3DP [34]	243	98.0	<u>79.1</u>	28.1
FinePOSE (Ours)	243	**98.7** (+0.2)	**79.7** (+0.6)	**26.8** (-1.0)

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**

AUC: area under curve
MPJPE: mean per joint position error
PCK: percentage of correct keypoint

Visualization

The 3D poses predicted by FinePOSE match better with ground-truth 3D poses than other methods.