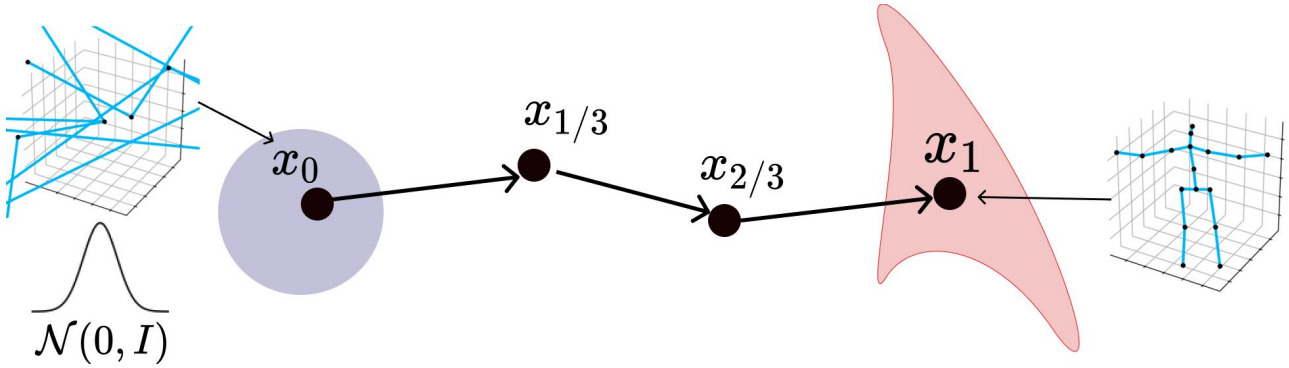
FMPose3D: monocular 3D pose estimation via flow matching

Ti Wang, Xiaohang Yu, Mackenzie Weygandt Mathis



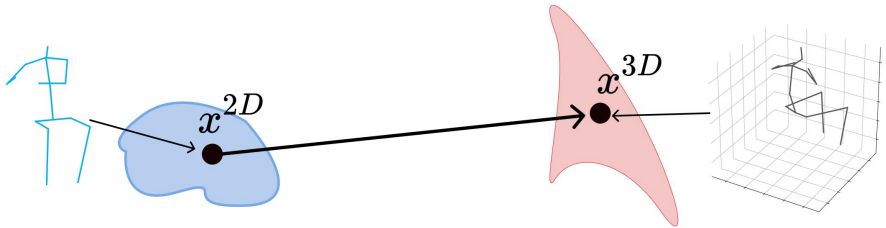
EPFL CVPR JUNE 3-7, 2026 DENVER COLORADO

Monocular 3D pose estimation

- How do humans infer 3D from 2D?

- Given a 2D image of an animal, we can tell the front paws are in front, and the tail is behind.
- Our brain relies on learned 3D priors to interpret ambiguous 2D cues.

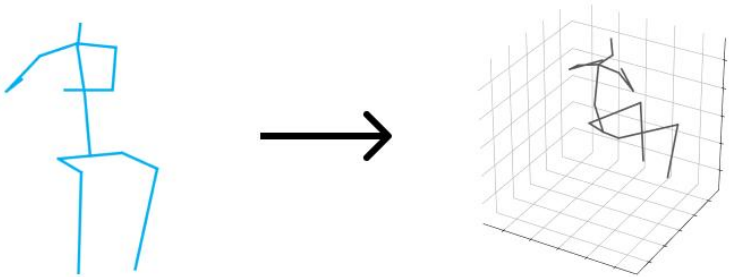
- Learning 3D priors
  - Generative models learn a distribution that captures the pose priors conditioned on the 2D input.



Two paradigms for 2D-to-3D pose lifting

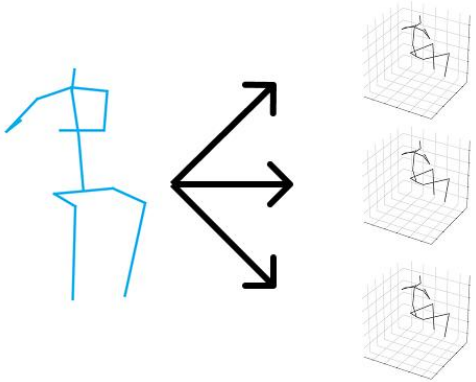
Deterministic approaches

- Learn a single mapping
- 2D pose → one 3D pose



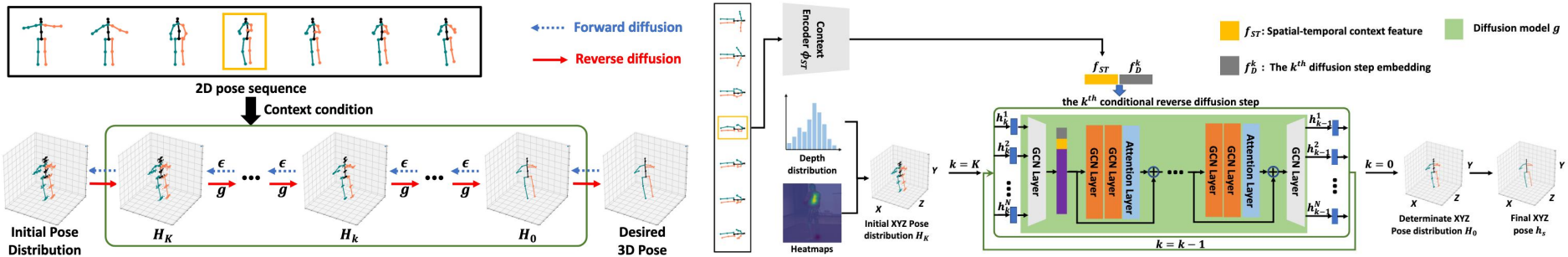
Probabilistic approaches

- Learn a conditional pose distribution
- 2D pose → multiple plausible 3D hypotheses



Generative model in 3D HPE

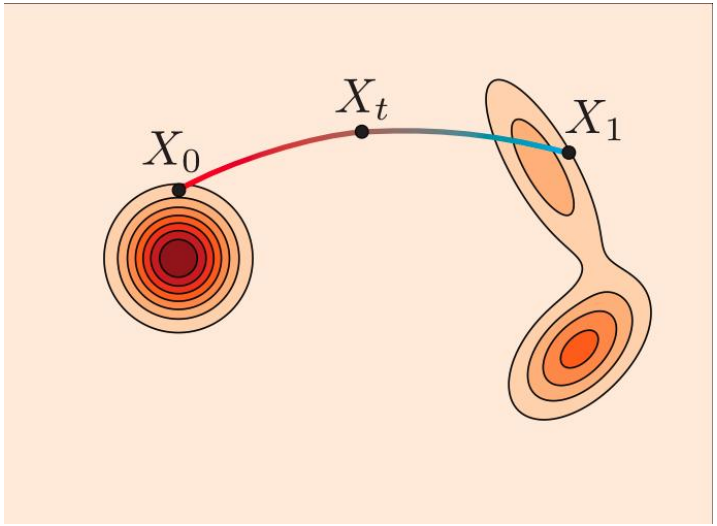
- DiffPose (CVPR'23)



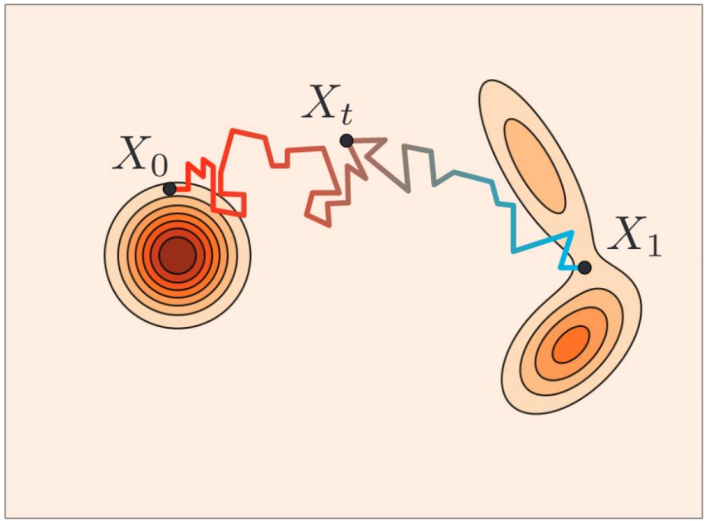
- Strength: the stochastic nature of diffusion promotes more diverse results
- Limitations: The iterative denoising process makes inference very slow

[1] Gong, Jia, et al. "DiffPose: Toward more reliable 3d pose estimation." CVPR. 2023.

Flow Matching: fast and deterministic



(a) Flow

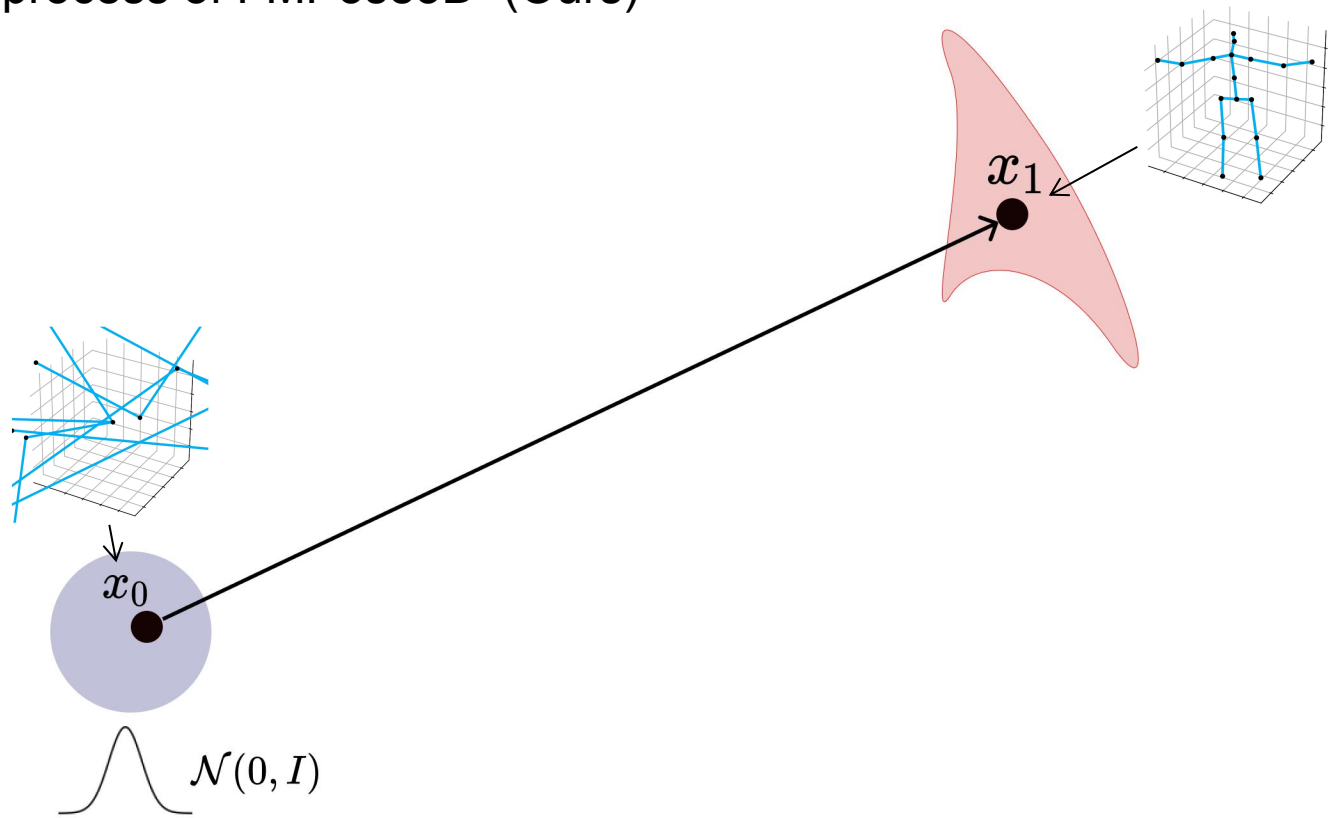


(b) Diffusion

[1] Lipman Y, Havasi M, Holderrieth P, et al. Flow matching guide and code[J]. arXiv preprint arXiv:2412.06264, 2024.

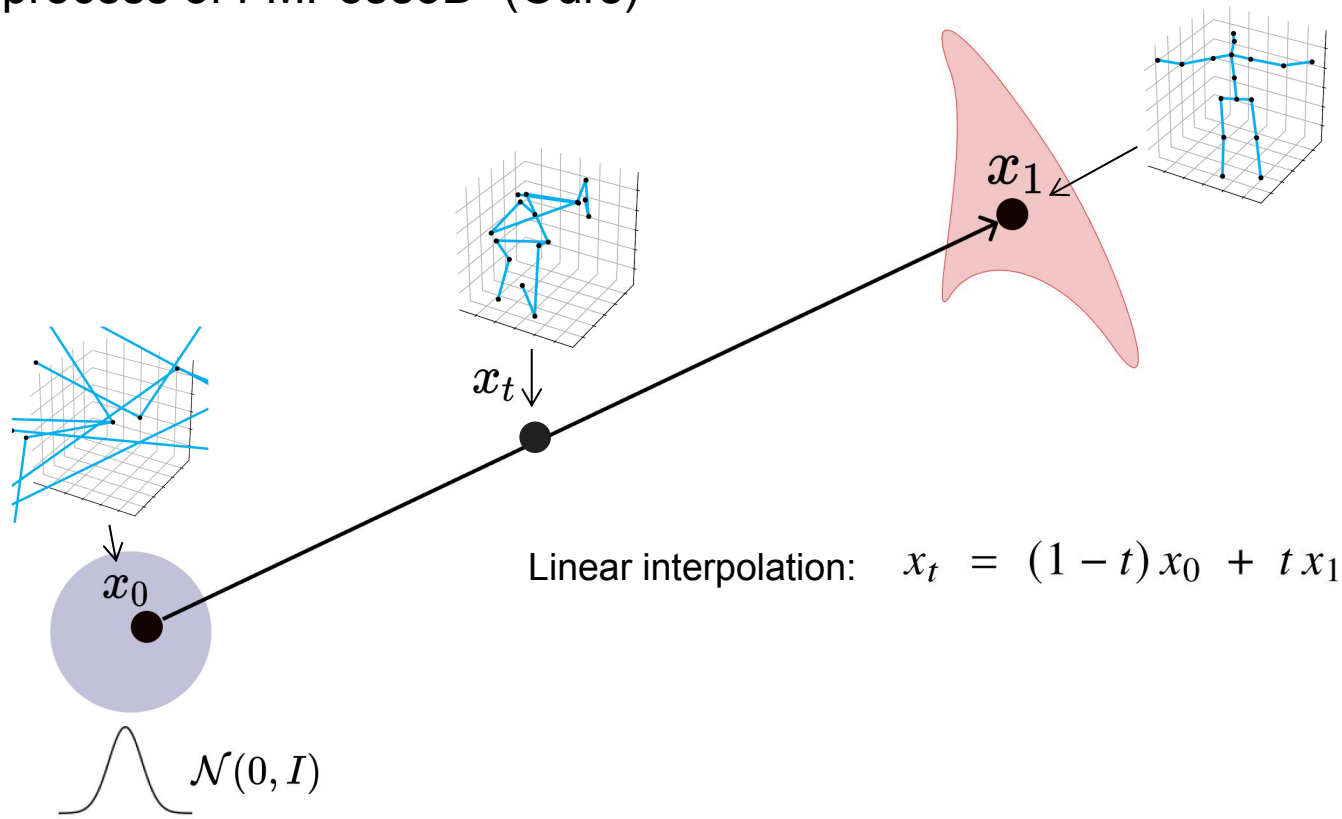
3D pose estimation via flow matching

- Training process of FMPose3D (Ours)



3D pose estimation via flow matching

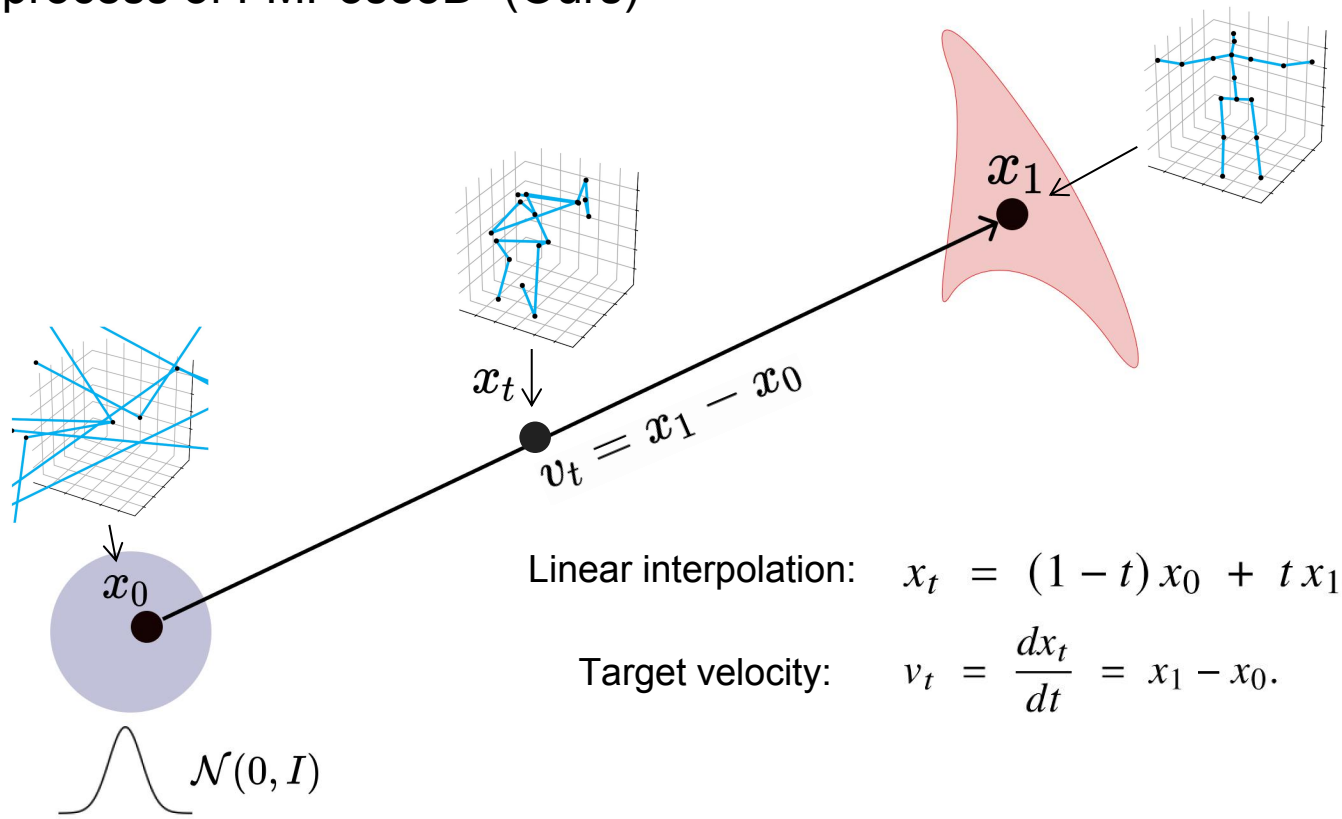
- Training process of FMPose3D (Ours)



Linear interpolation: $x_t = (1-t)\,x_0 + t\,x_1, \quad t \in [0, 1).$

3D pose estimation via flow matching

- Training process of FMPose3D (Ours)

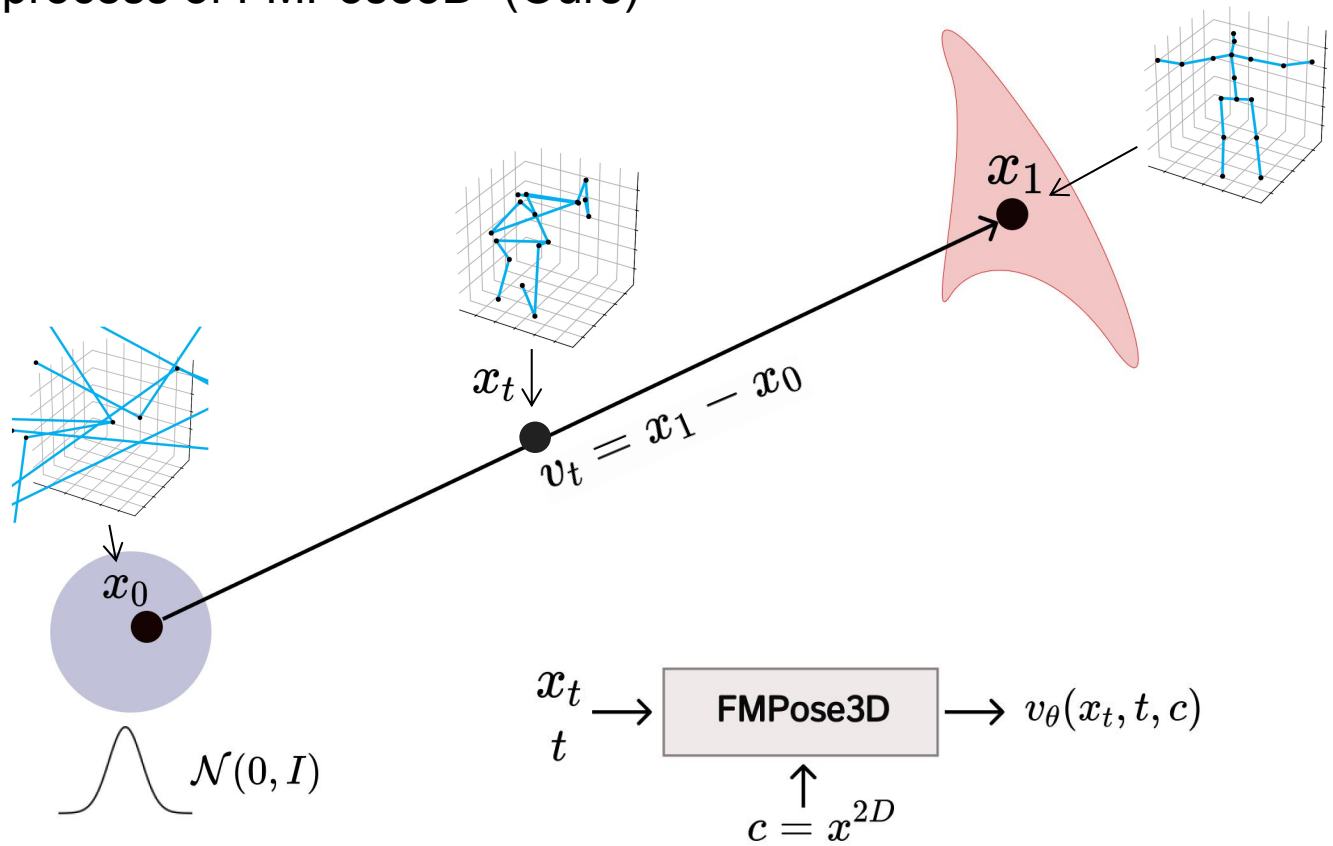


Linear interpolation: $x_t = (1-t)\,x_0 + t\,x_1, \quad t \in [0, 1).$

Target velocity: $v_t = \frac{dx_t}{dt} = x_1 - x_0.$

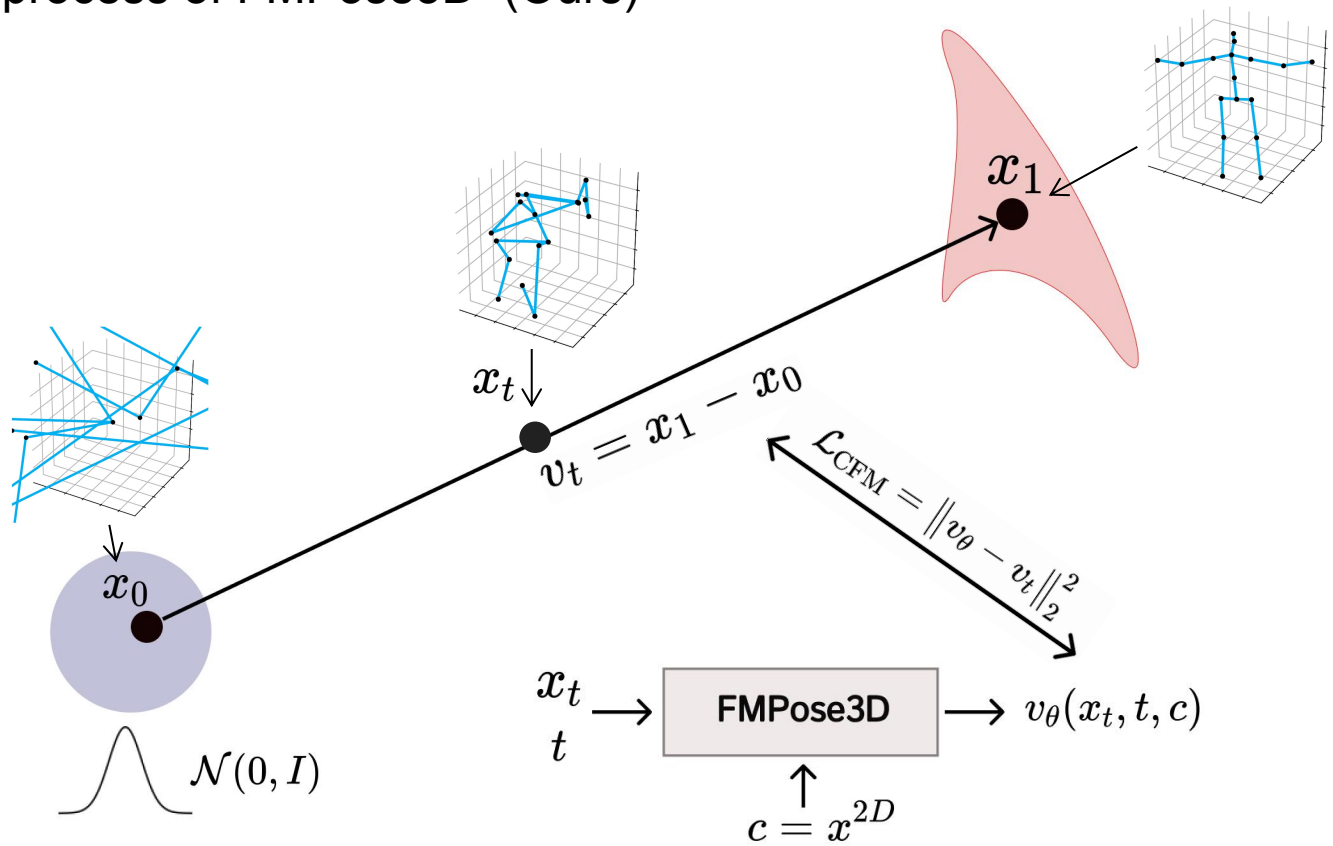
3D pose estimation via flow matching

- Training process of FMPose3D (Ours)



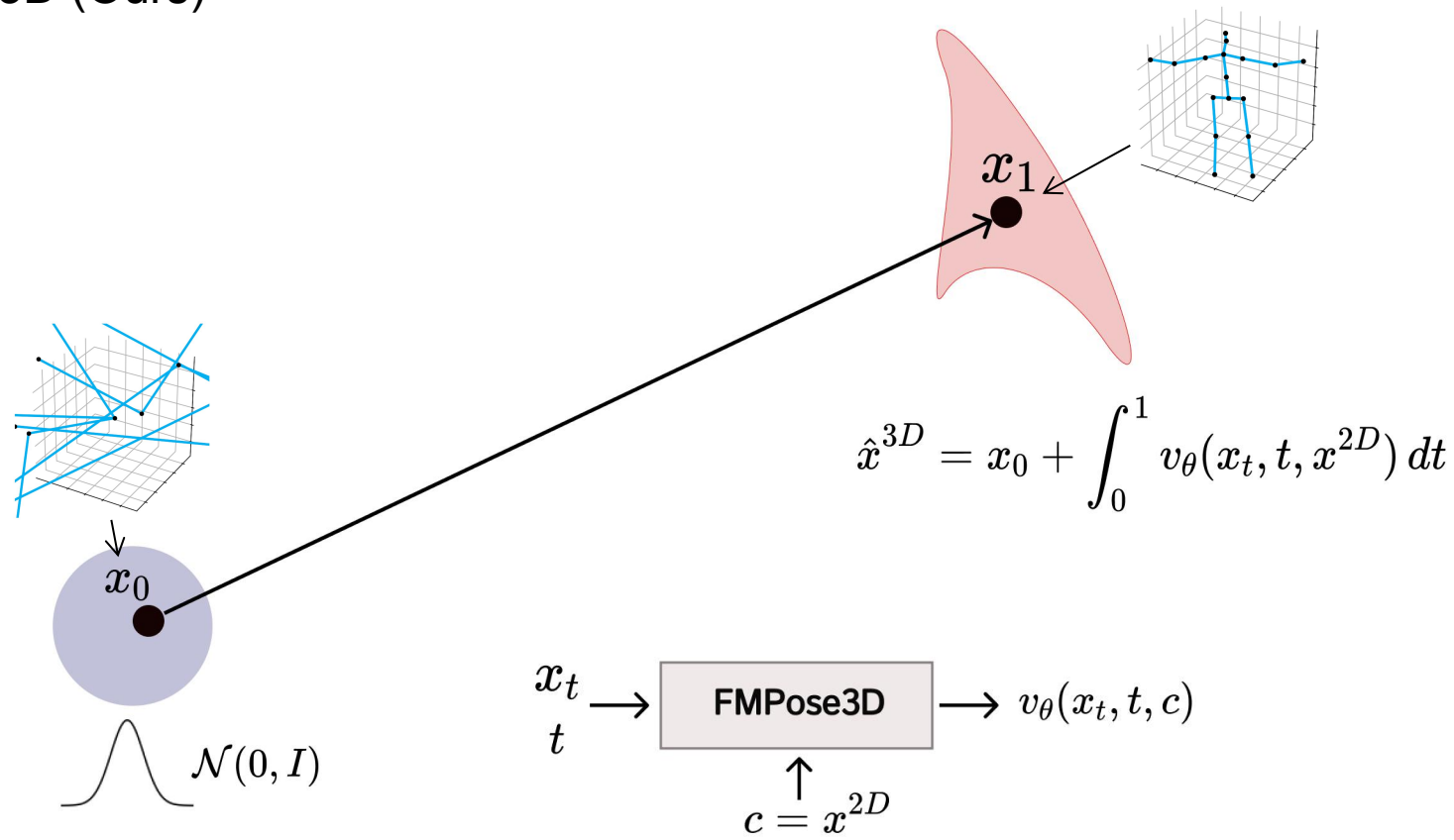
3D pose estimation via flow matching

- Training process of FMPose3D (Ours)



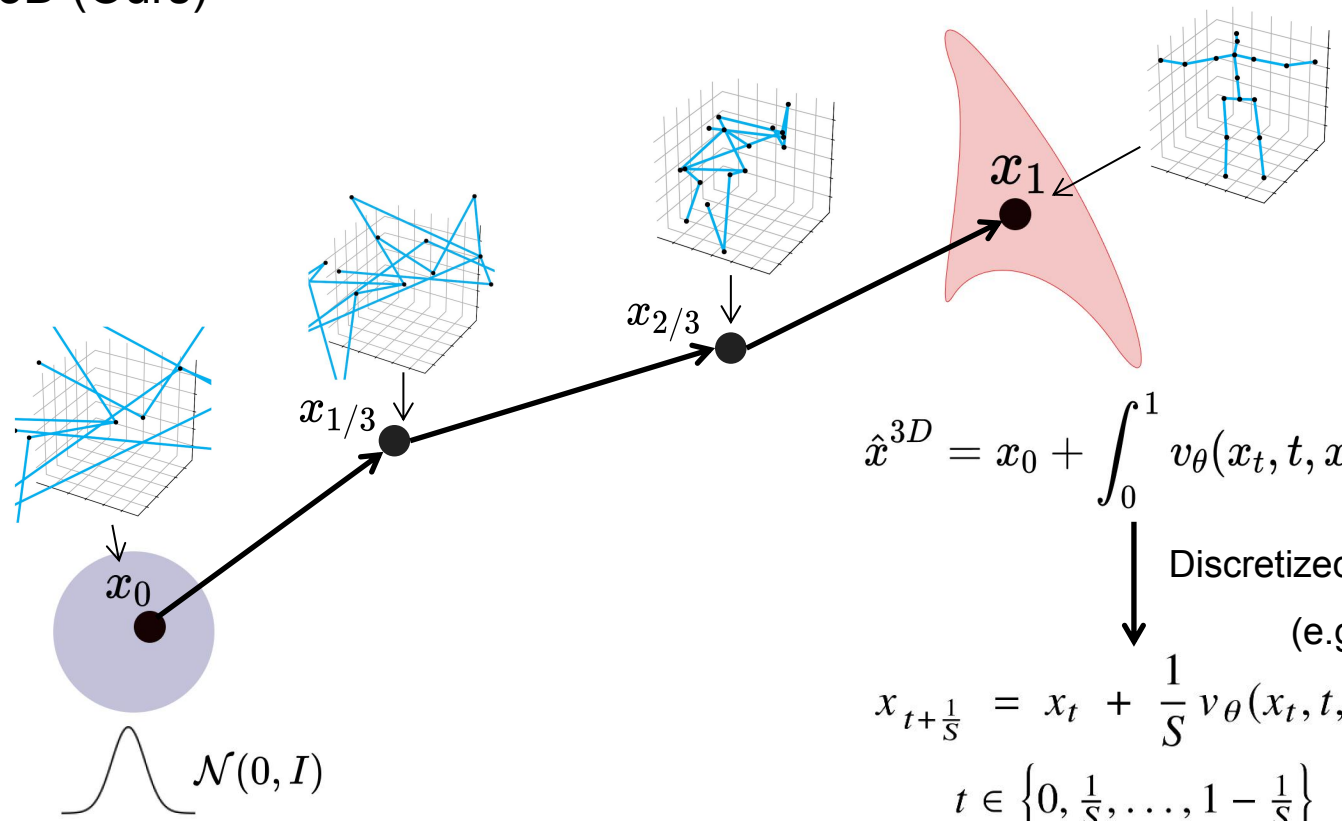
Inference process of FMPose3D

- FMPose3D (Ours)



Inference process of FMPose3D

- FMPose3D (Ours)



$$\hat{x}^{3D} = x_0 + \int_0^1 v_\theta(x_t, t, x^{2D})\, dt$$

Discretized approximation

(e.g. ***S*** = 3)

$$x_{t+\frac{1}{S}} = x_t + \frac{1}{S} v_\theta(x_t, t, c),$$

$$t \in \left\{0, \tfrac{1}{S}, \ldots, 1 - \tfrac{1}{S}\right\}$$

FMPose3D

Algorithm 1: Training and inference of FMPose3D

Input: Training set $\mathcal{D} = \{(x^{2D}, x^{3D})\}$ with $x^{2D} \in \mathbb{R}^{J\times 2}$ and $x^{3D} \in \mathbb{R}^{J\times 3}$; number of training iterations T_{train}; number of Euler steps S used at inference.

Output: Trained FMPose3D parameters θ (velocity field v_θ); at inference time, the predicted 3D pose $\hat{x}^{3D}$ for a given 2D pose.

Training phase:

for $iter = 1$ **to** T_{train} **do**

- Sample a mini-batch of data pairs (x^{2D}, x^{3D}) from $\mathcal{D}$;
- Set $x_1 \leftarrow x^{3D}$ and condition $c \leftarrow x^{2D}$;
- Sample $x_0 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}[0, 1)$;
- Compute interpolated states $x_t \leftarrow (1-t)\,x_0 + t\,x_1$;
- Predict velocities $v^{\text{pred}} \leftarrow v_\theta(x_t, t, c)$;
- Set target velocities $v^{\text{target}} \leftarrow x_1 - x_0$;
- Compute the CFM loss on the batch $\mathcal{L}_{\text{CFM}} \leftarrow \|v^{\text{pred}} - v^{\text{target}}\|_2^2$;
- Update parameters $\theta \leftarrow \text{Adam}(\theta, \nabla_\theta \mathcal{L}_{\text{CFM}})$;

Inference phase (given trained θ):

Given a 2D pose x^{2D} and the number of Euler steps S;

Set condition $c \leftarrow x^{2D}$ and sample $x_0 \sim \mathcal{N}(0, I)$;

Initialize $x \leftarrow x_0$;

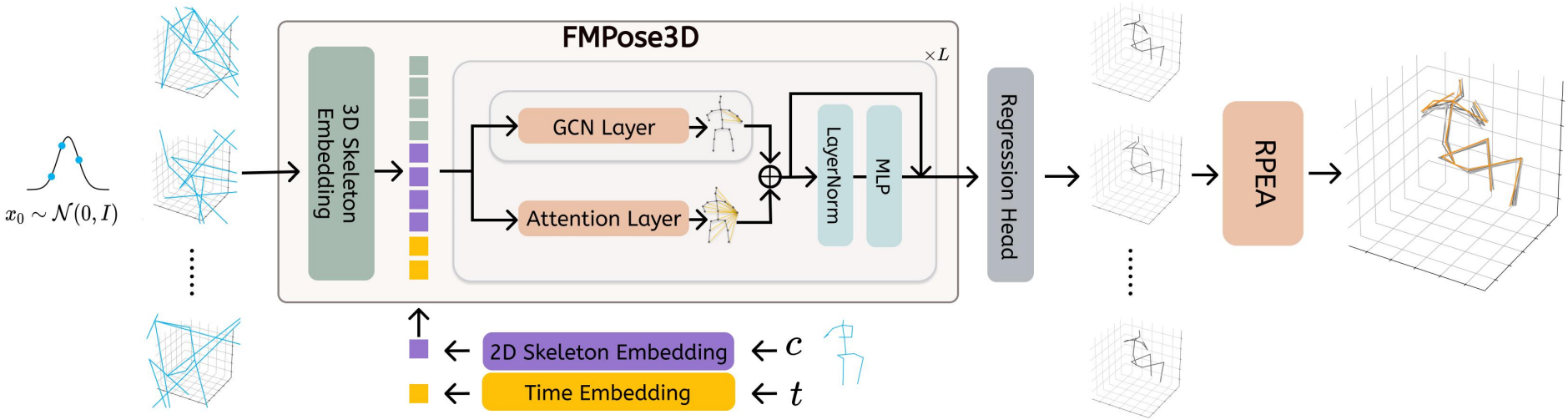
for $k = 0$ **to** $S - 1$ **do**

- $t \leftarrow k/S$;
- $x \leftarrow x + \frac{1}{S}\, v_\theta(x, t, c)$;

return $\hat{x}^{3D} \leftarrow x$;

3D human pose estimation via flow matching

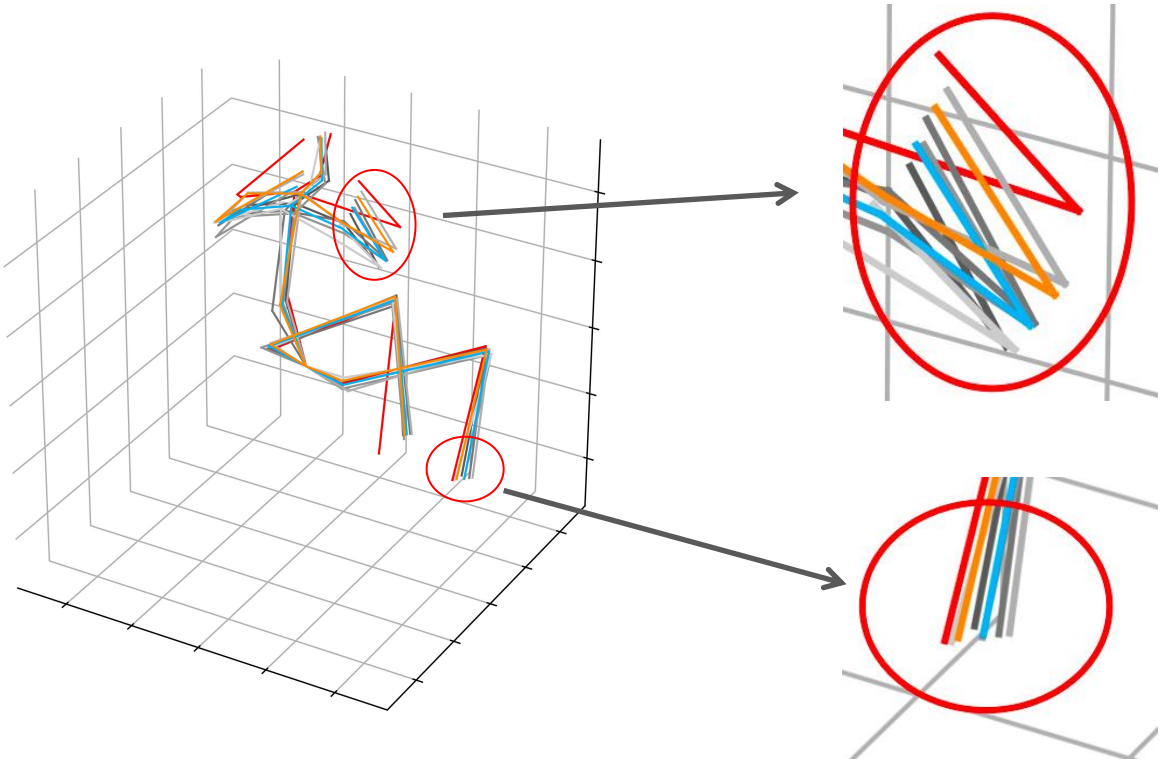
- FMPose3D (Ours)



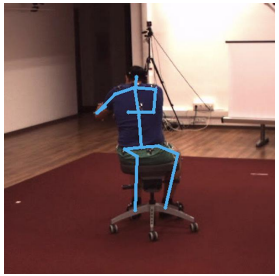
$$\hat{x}^{3D} = x_0 + \int_0^1 v_\theta(x_t, t, x^{2D})\, dt$$

Reprojection-based Posterior Expectation Aggregation (RPEA)

- Visualization comparison (RPEA (ours) vs. Mean[1])



- Red: Ground Truth
- Gray: Hypotheses
- Blue: Mean
- Orange: RPEA



[1] Gong, Jia, et al. "DiffPose: Toward more reliable 3d pose estimation." CVPR. 2023.

Datasets

- **Human3.6M dataset**
  - The largest and most representative dataset for 3D HPE.
  - Constrained indoor scenes, 3.6 million frames for 11 subjects.
  - Each subject performs 15 actions from 4 views.
- **MPI-INF-3DHP dataset**
  - A large-scale dataset for 3D HPE.
  - Contains both indoor and complex in-the-wild outdoor scenes.
  - There are 8 actions from 14 views by 8 actors.
- Evaluation protocols
  - MPJPE (Protocol 1): Mean Per Joint Position Error in millimeter
  - P-MPJPE (Protocol 2): MPJPE after Procrustes alignment
  - PCK: Percentage of Correct Keypoints
  - AUC: Area Under the Curve

Human3.6M

MPI-INF-3DHP

Quantitative comparison on Human3.6M

- Quantitative comparison with state-of-the-art methods on Human3.6M

Table 1. Quantitative comparison with the state-of-the-art methods on Human3.6M under MPJPE. The detected 2D pose is used as input. N denotes the number of hypotheses. Red: Best. Blue: Second Best. Grey : our method.

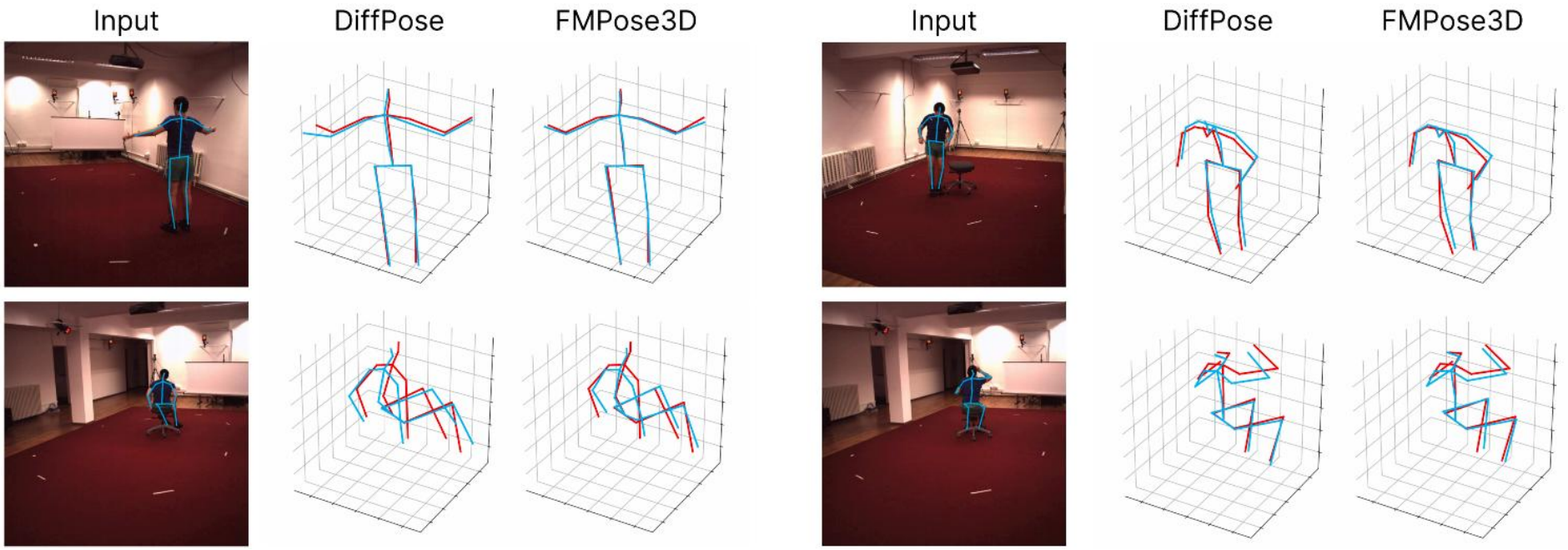
Deterministic Method		Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg ↓
SimpleBaseline [44]	ICCV'17	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
VideoPose3D [47]	CVPR'19	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
LCN [9]	ICCV'21	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
SRNet [66]	ECCV'20	44.5	48.2	47.1	47.8	51.2	56.8	50.1	45.6	59.9	66.4	52.1	45.3	54.2	39.1	40.3	49.9
GraphSH [64]	CVPR'21	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
GraFormer [72]	CVPR'22	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	39.7	43.1	51.8
UGRN [34]	AAAI'23	47.9	50.0	47.1	51.3	51.2	59.5	48.7	46.9	56.0	61.9	51.1	48.9	54.3	40.0	42.9	50.5
MLP-JCG [55]	TMM'23	43.8	46.7	46.9	48.9	50.3	60.1	45.7	43.9	56.0	73.7	48.9	48.2	50.9	39.9	41.5	49.7
PerturbPE [1]	ECCV'24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.8
Probabilistic Method		Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg ↓
CVAE (N=200) [50]	ICCV'19	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
GAN (N=10) [33]	BMVC'20	66.0	74.7	71.1	80.6	81.1	93.0	73.2	83.7	90.0	117.4	75.8	79.3	82.1	74.4	77.8	80.9
GraphMDN (N=5) [46]	IJCNN'21	51.9	56.1	55.3	58.0	63.5	75.1	53.3	56.5	69.4	92.7	60.1	58.0	65.5	49.8	53.6	61.3
NF (N=1) [62]	ICCV'21	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
DiffPose (N=5) [16]	CVPR'23	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52.0	37.6	40.2	49.7
ProPose (N=1) [19]	AAAI'25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51.9
FMPose3D (N=2) (Ours)		45.7	49.2	46.0	49.4	50.9	57.3	47.2	45.0	57.9	61.5	49.8	46.8	52.1	38.9	41.6	49.3
FMPose3D (N=40) (Ours)		43.5	47.2	44.4	47.7	48.9	55.1	45.5	42.7	55.7	59.4	47.9	45.1	49.8	37.1	39.6	47.3

Quantitative comparison on MPI-INF-3DHP

- Generalization performance on MPI-INF-3DHP

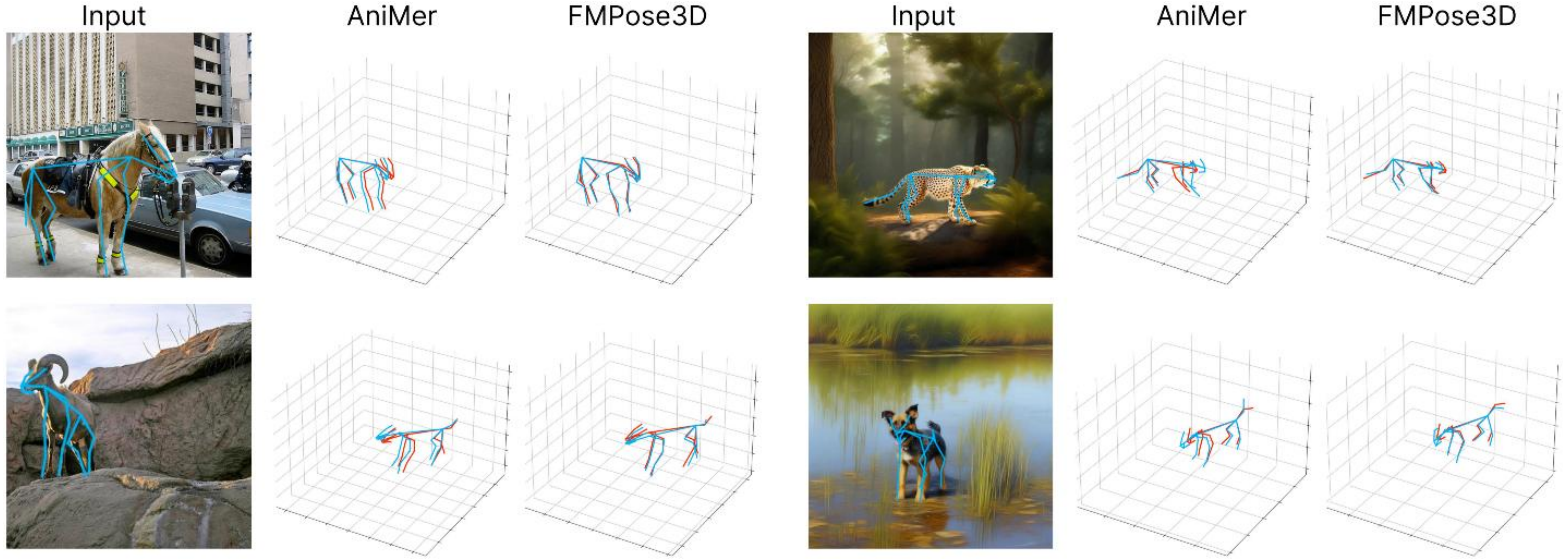
Method	GS ↑	noGS ↑	Outdoor ↑	All PCK ↑	All AUC ↑
SimpleBaseline [44]	49.8	42.5	31.2	42.5	17.0
GraphSH [64]	81.5	81.7	75.2	80.1	45.8
Zeng *et al.* [67]	-	-	84.6	82.1	46.2
GraFormer [72]	80.1	77.9	74.1	79.0	43.8
UGRN [34]	86.2	84.7	81.9	84.1	53.7
PerturbPE [1]	80.0	79.0	84.0	82.0	-
ProPose [19]	83.9	85.5	83.4	84.4	52.1
FMPose3D (N=2) (Ours)	85.7	86.4	85.6	85.9	53.7
FMPose3D (N=20) (Ours)	86.1	87.1	86.5	86.4	54.6

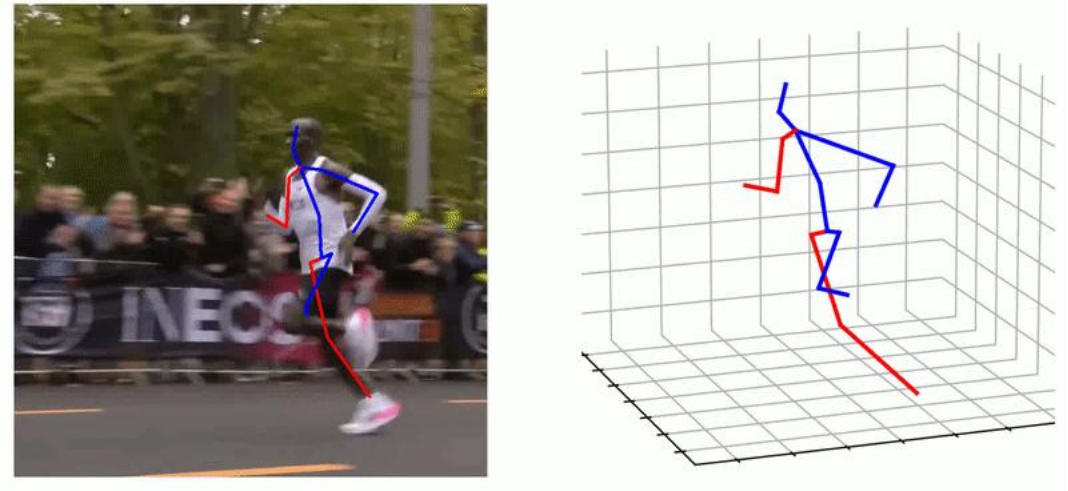
Qualitative comparison on Human3.6M



Results on animal datasets

Dataset	Animal3D	CtrlAni3D
Metric	P-MPJPE ↓	P-MPJPE ↓
HMR [28]	123.5	123.5
WLDO [5]	112.3	71.5
HMR2.0 [15]	94.1	60.9
AniMer [43]	80.4	44.1
Ours	60.5	43.7





Thanks for watching!

Project Page: https://xiu-cs.github.io/FMPose3D/

Paper link: https://arxiv.org/abs/2602.05755

Github Repository: https://github.com/AdaptiveMotorControlLab/FMPose3D