FMPose3D: monocular 3D pose estimation via flow matching

Supplementary Material

A. Background on Attention and GCN

Attention. The input tokens $X \in \mathbb{R}^{J \times D}$ are first projected to queries $Q \in \mathbb{R}^{J \times d}$, keys $K \in \mathbb{R}^{J \times d}$, and values $V \in \mathbb{R}^{J \times d}$, and then Q, K, V are fed to a scaled dot-product attention [56]:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V, \qquad \text{(L)}$$

where d is the dimension of Q, K, V. Multi-head self-attention (MSA) [56] splits Q, K, V into multiple heads, each of which applies scaled dot-product attention in parallel. This enables the model to efficiently utilize information from various representation subspaces with different locations.

Graph Convolutional Network. Graph Convolutional Network (GCN) [30] is capable of capturing intricate relationships and structures within graph-structured data. Consider an undirected graph $G=\{V, E\}$, where V is the set of nodes and E is the set of edges. The edges can be encoded in an adjacency matrix $A \in \{0, 1\}^{N \times N}$. For the input X_l of the l^{th} layer, the vanilla graph convolution aggregates the features of the neighboring nodes. The output X_{l+1} of the l^{th} GCN layer can be formulated as:

$$X_{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X_l W\right), \qquad \text{(M)}$$

where σ is the ReLU activation function [14], $W_l \in \mathbb{R}^{d_1 \times d}$ is the layer-specific trainable weight matrix. $\tilde{A}=A+I_N$ is the adjacency matrix of the graph with added self-connections, where I_N is the identity matrix. Additionally, $\tilde{D}$ is the diagonal node degree matrix. By stacking multiple GCN layers, it iteratively transforms and aggregates neighboring nodes, thereby obtaining enhanced feature representations.

B. Additional Implementation Details

Humans. For Human3.6M [24] and MPI-INF-3DHP [45], each sample contains $J = 17$ joints. For the model architecture described in Sec. 3.3, each block takes the embedding as input, feeds it into a parallel structure with a GCN branch and an attention branch, concatenates the resulting features, and then processes them with an MLP layer, as illustrated in Figure 2 of the main paper. This block is repeated $L = 5$ times. The dimensionality of the concatenated feature embedding is set to $D = 512$. The learning rate is initialized at 0.001 and is multiplied by 0.98 at each epoch, with the factor replaced by 0.8 every 5 epochs. During inference, we set the number of ODE integration steps to $S = 3$. We employ horizontal flip augmentation, following prior works [7, 47, 67]. Rather than averaging the predictions from the original and flipped inputs, we treat them as two separate hypotheses and feed both into our RPEA module to obtain the final 3D pose. We refer to this strategy as **Flipped Hypothesis Aggregation (FHA)**. Following common practice [7, 47, 64, 67], we use 2D poses detected by the cascaded pyramid network (CPN) [8] for Human3.6M, and the dataset-provided 2D poses for MPI-INF-3DHP.

Animals. For Animal3D [63] and CtrlAni3D [43], each sample contains $J = 26$ joints. We train a single model jointly on the two datasets and evaluate it on each dataset individually. The network architecture follows the same design as in the human setting, with the block repeated for $L = 5$ layers. The dimensionality of the concatenated feature embedding is set to $D = 512$. The model is trained for 300 epochs with a batch size of 13. The learning rate is initialized at 0.001 and is multiplied by 0.95 at each epoch, with the factor replaced by 0.75 every 15 epochs. During inference, we set the number of ODE integration steps to $S = 3$. For the results reported in Table 3 of the main paper, we generate a single prediction per input and do not use flip augmentation or any multi-hypothesis strategy.

C. Additional Quantitative Results

Human3.6M. Table 6 reports the P-MPJPE results on Human3.6M, comparing our FMPose3D with prior state-of-the-art deterministic methods (top) and probabilistic methods (bottom). Our baseline model achieves 38.7 mm. When we increase the number of hypotheses and applying our RPEA module, the error is further reduced to 38.3 mm, demonstrating the effectiveness of multi-hypothesis modeling.

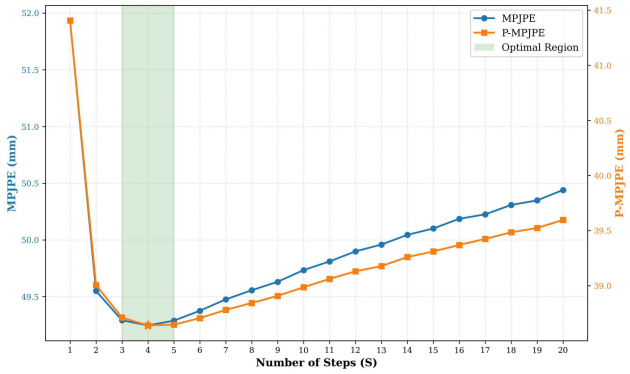

Figure 6. Effect of the number of integration steps S on inference accuracy. The blue curve shows MPJPE (read from the left vertical axis), and the orange curve shows P-MPJPE (read from the right vertical axis); the shaded region marks the range $S \in \{3, 4, 5\}$ where both metrics attain their optimal or near-optimal values.

Table 6. Quantitative comparison with the state-of-the-art methods on Human3.6M under P-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	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
VideoPose3D [47]	CVPR'19	36.0	38.7	38.0	41.7	40.1	45.9	37.1	35.4	46.8	53.4	41.4	36.9	43.1	30.3	34.8	40.0
STGCN [7]	ICCV'19	36.8	38.7	38.2	41.7	40.7	46.8	37.9	35.6	47.6	51.7	41.3	36.8	42.7	31.0	34.7	40.2
SRNet [66]	ECCV'20	35.8	39.2	36.6	36.9	39.8	45.1	38.4	36.9	47.7	54.4	38.6	36.3	39.4	30.3	35.4	39.4
LCN [9]	ICCV'21	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
MLP-JCG [55]	TMM'23	33.7	37.4	37.3	39.6	39.8	47.1	33.7	33.8	45.7	60.5	39.7	37.7	40.1	30.1	33.8	39.3
GKONet [22]	TCSVT'23	35.4	38.8	35.9	40.4	39.6	44.0	36.7	35.4	46.8	53.7	40.9	36.6	42.0	30.6	33.9	39.4
ZEDO [27]	WACV'24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	42.1
Probabilistic Method		Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg ↓
CVAE (N=200) [50]	ICCV'19	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	29.4	34.3	40.9
GAN (N=10) [33]	BMVC'20	41.4	44.3	44.6	50.2	49.3	51.8	40.1	46.2	57.7	72.7	48.7	45.4	49.6	43.8	43.3	48.7
GraphMDN (N=5) [46]	IJCNN'21	39.7	43.4	44.0	46.2	48.8	54.5	39.4	41.1	55.0	69.0	48.0	43.7	49.6	38.4	42.4	46.9
NF (N=1) [62]	ICCV'21	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
DiffPose (N=5) [16]	CVPR'23	33.9	38.2	36.0	39.2	40.2	46.5	35.8	34.8	48.0	52.5	41.2	36.5	40.9	30.3	33.8	39.2
ProPose (N=1) [19]	AAAI'25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.4
FMPose3D (N=2) (Ours)		35.4	38.3	36.0	39.8	39.2	43.5	36.5	34.7	46.3	48.4	40.4	35.9	41.0	31.0	34.2	38.7
FMPose3D (N=40) (Ours)		35.0	37.7	35.7	39.4	38.8	43.0	36.1	34.2	45.7	48.1	40.1	35.5	40.6	30.6	33.7	38.3

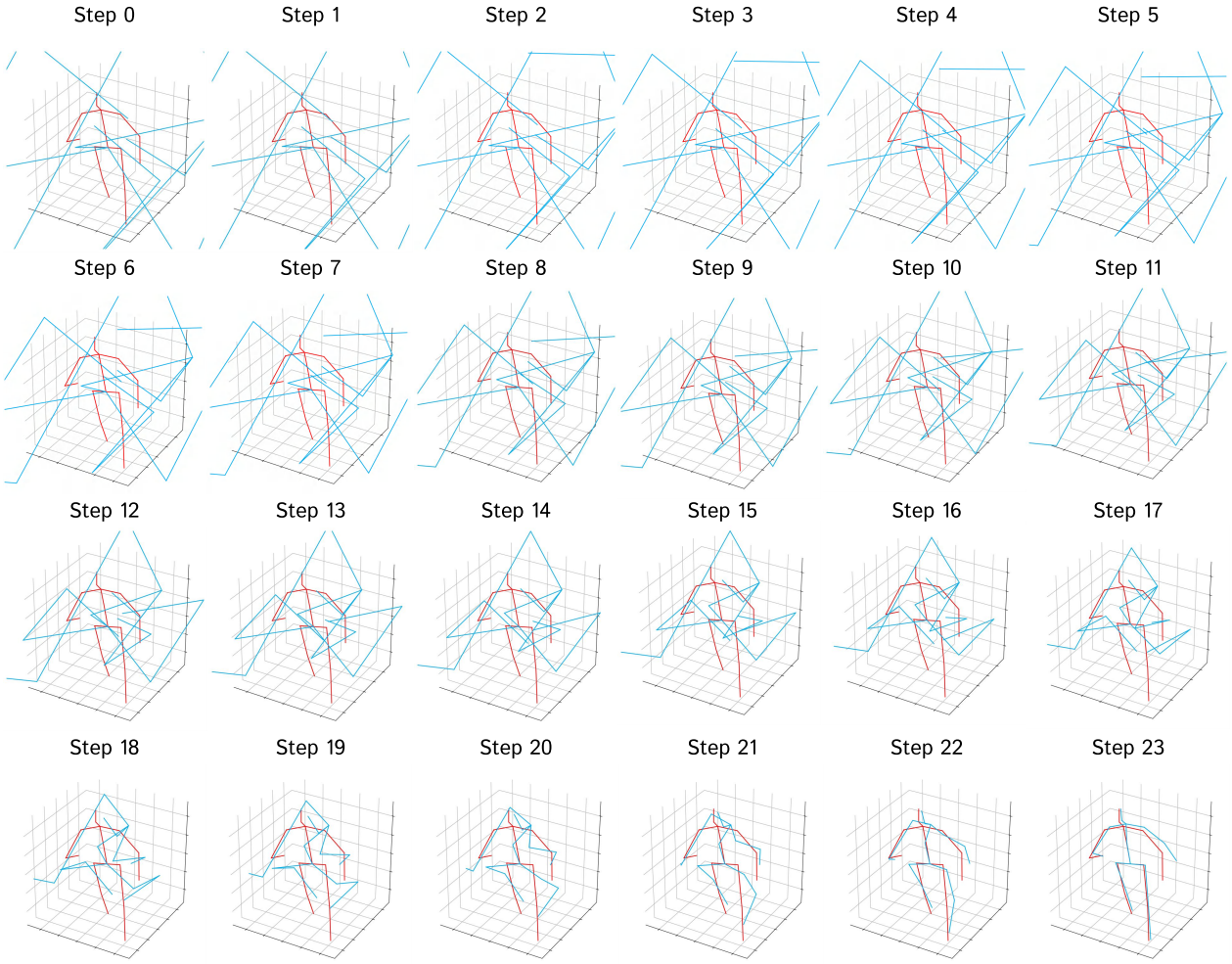

Figure 7. Visualization of intermediate 3D pose predictions during inference with $S = 23$ integration steps. The blue pose represents the predicted results, while the red pose represents the ground truth.

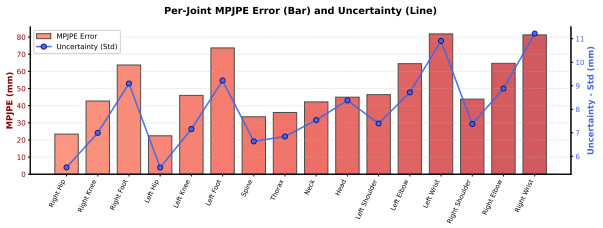

Figure 8. Per-joint uncertainty versus per-joint error on Human3.6M. Uncertainty is measured as the standard deviation (Std) across hypotheses, and error is measured by MPJPE (mm).

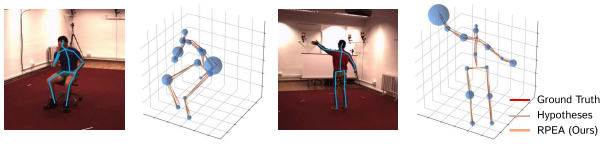

Figure 9. Uncertainty visualization. Left: highest uncertainty at the left elbow. Right: highest uncertainty at the left wrist.

D. Intermediate Integration States

To better understand how the learned velocity field transports a noise sample toward the target 3D pose, we select one example, set the number of integration steps to $S = 23$, and visualize the intermediate predictions along the trajectory. As illustrated in Figure 7, the poses start from random noise and progressively become more structured, gradually converging to a plausible human configuration that closely matches the final target pose.

E. Impact of Integration Steps

During inference, our FMPose3D generates 3D poses by solving the underlying ODE from noise, conditioned on 2D inputs, using S integration steps. Figure 6 reports MPJPE and P-MPJPE on the Human3.6M dataset with 2D poses detected by CPN [8] for different choices of S. With a single integration step, the errors of both metrics are relatively large. As S increases, the errors first decrease, reach the minimum at $S = 4$, and then gradually increase. The results indicate that steps in the range $S \in \{3, 4, 5\}$ yield comparable accuracy, whereas larger S does not provide additional gains. To strike a balance between estimation accuracy and computational efficiency, we set $S = 3$ in all experiments.

F. Uncertainty Estimation

Our multi-hypothesis predictions also enable uncertainty estimation. On the Human3.6M test set, we generate N=40 hypotheses per input and estimate uncertainty as the per-joint standard deviation across these hypotheses. As shown in Figure 8, the average uncertainty is positively correlated with the per-joint error, supporting the validity of this uncertainty measure. Such single-view uncertainty can serve as a confidence signal for unsupervised multi-view fusion and as an intermediate representation to guide mesh reconstruction.

For visualization, we represent uncertainty using spheres centered at the predicted joint locations, with radii proportional to the per-joint variance. Larger spheres indicate higher uncertainty, as illustrated in Figure 9.

Table 7. Results on 3DPW. Top: methods trained on 3DPW. Bottom: methods without 3DPW training (zero-shot evaluation).

Method	P-MPJPE	MPJPE
HMR2.0a [15] (ICCV23)	44.5	70.0
Multi-HMR [3] (ECCV24)	41.7	61.4
AdaptPose [13] (CVPR22)	46.5	81.2
PMCE [65] (ICCV23)	52.3	81.6
ScoreHMR [53] (CVPR24)	50.5	-
FMPose3D (N=40) (Ours)	42.4	70.9

Table 8. Effect of training set size. For each dataset, the model is trained from scratch on randomly subsampled training subsets of 10%, 20%, 40%, and 80%. We run each setting three times with different random seeds and report the mean and the standard deviation. Full training and test set sizes in frames (train, test): Human3.6M (3,119k, 543k), Animal3D (3k, 0.3k), CtrlAni3D (8k, 1.4k).

Data (%)	Human3.6M		Animal3D	CtrlAni3D
	P-MPJPE	MPJPE	P-MPJPE	P-MPJPE
10	$43.5_{\pm 0.85}$	$53.9_{\pm 0.71}$	$120.7_{\pm 1.88}$	$83.5_{\pm 0.95}$
20	$41.8_{\pm 0.62}$	$52.2_{\pm 0.64}$	$98.3_{\pm 0.89}$	$69.5_{\pm 0.30}$
40	$40.4_{\pm 0.33}$	$50.7_{\pm 0.32}$	$81.6_{\pm 0.28}$	$57.8_{\pm 0.24}$
80	$39.3_{\pm 0.15}$	$49.7_{\pm 0.09}$	$67.2_{\pm 1.33}$	$47.9_{\pm 1.02}$

G. Results on an Additional Benchmark: 3DPW

In the main paper, we follow common practice in monocular 3D pose estimation and primarily report results on Human3.6M and MPI-INF-3DHP for fair comparison with prior work. We now additionally evaluate on 3DPW [57] with our Human3.6M pretrained model, the results are reported in Table 7. Since 3DPW is more commonly used in human mesh reconstruction, the methods listed in Table 7 are primarily mesh-based approaches, and thus the comparison should be interpreted with this difference in mind. Our model achieves the best zero-shot performance, comparable even to models pretrained on this dataset.

H. Effect of training set size

To evaluate the sensitivity to training data size, we train our model from scratch on randomly subsampled subsets comprising 10%, 20%, 40%, and 80% of the training set. For each fraction, we run three independent trainings with different random seeds and report the mean performance on

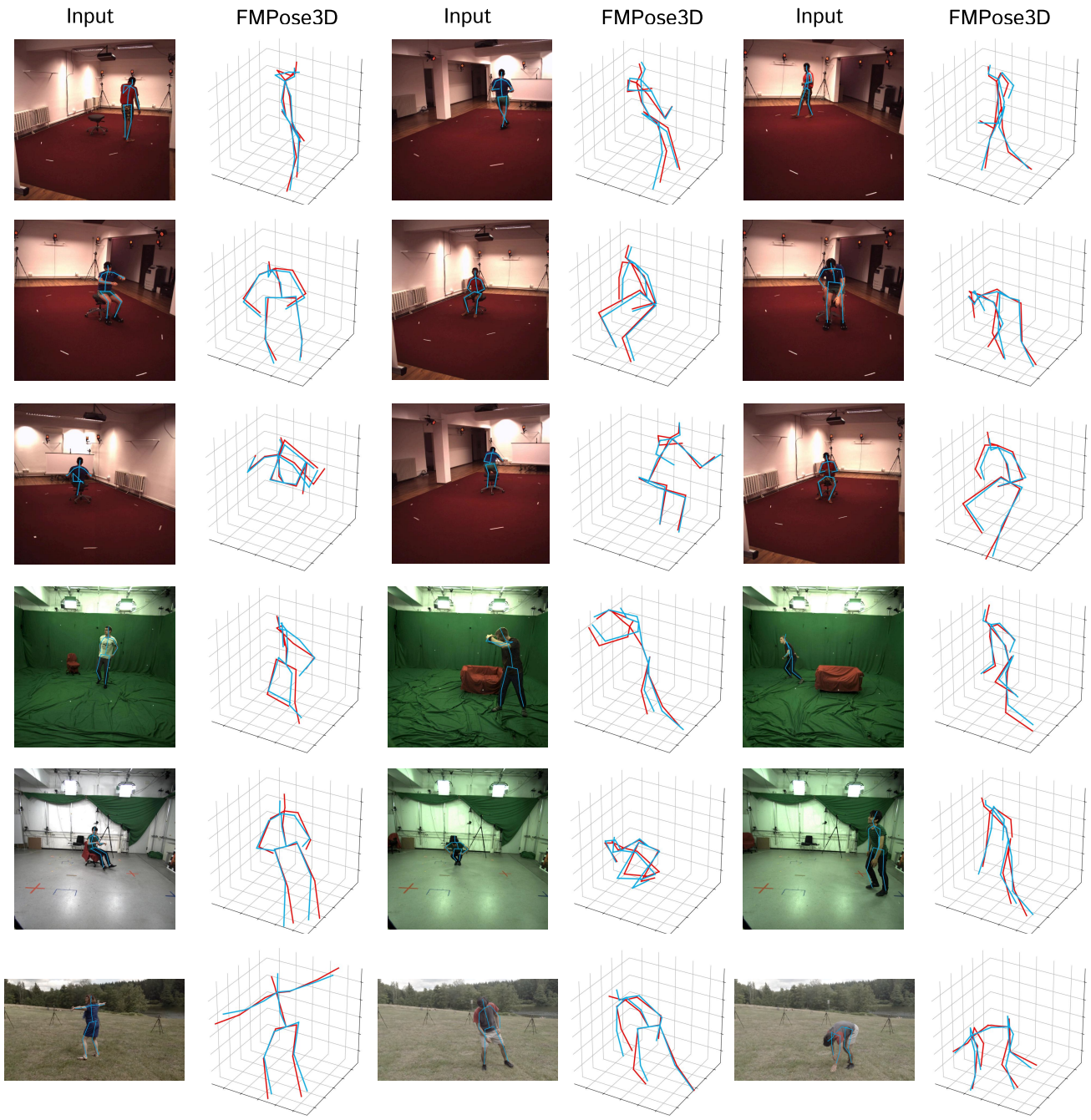

Figure 10. Qualitative results on Human3.6M (top three rows) and MPI-INF-3DHP (bottom three rows). The blue pose represents the predicted results, while the red pose represents the ground truth.

the full test set. Table 8 shows consistent improvements as the training set size increases across all datasets. The gains are modest on Human3.6M, but much larger on Animal3D and CtrlAni3D, as these animal datasets are relatively small and thus more sensitive to reduced training data.

I. Additional Qualitative Results

Figure 10 presents qualitative results of the proposed FMPose3D on the Human3.6M and MPI-INF-3DHP datasets. The model is trained solely on Human3.6M. Human3.6M consists of indoor scenes (top three rows), while the MPI-INF-3DHP test set includes three scenarios: studio with

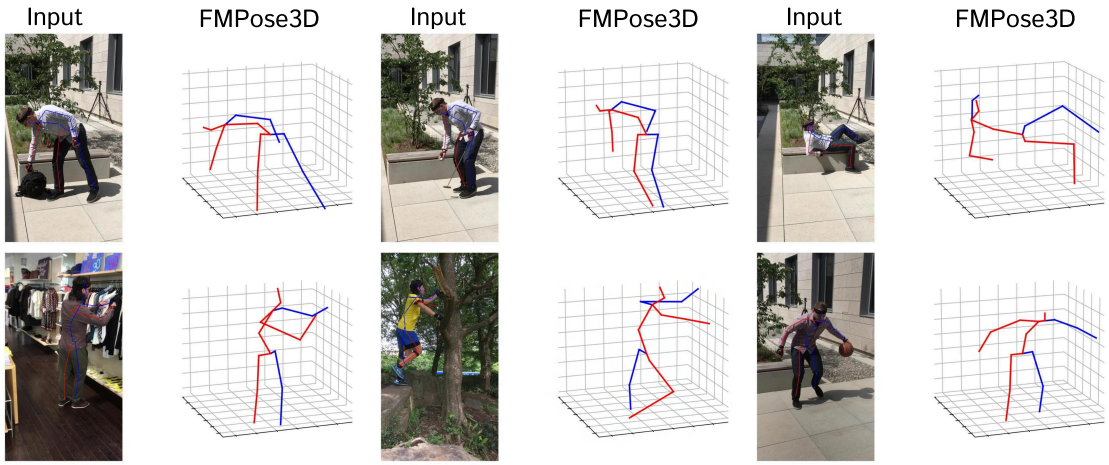

Figure 11. Qualitative results on 3DPW. The 2D pose is detected by HRNet [59].

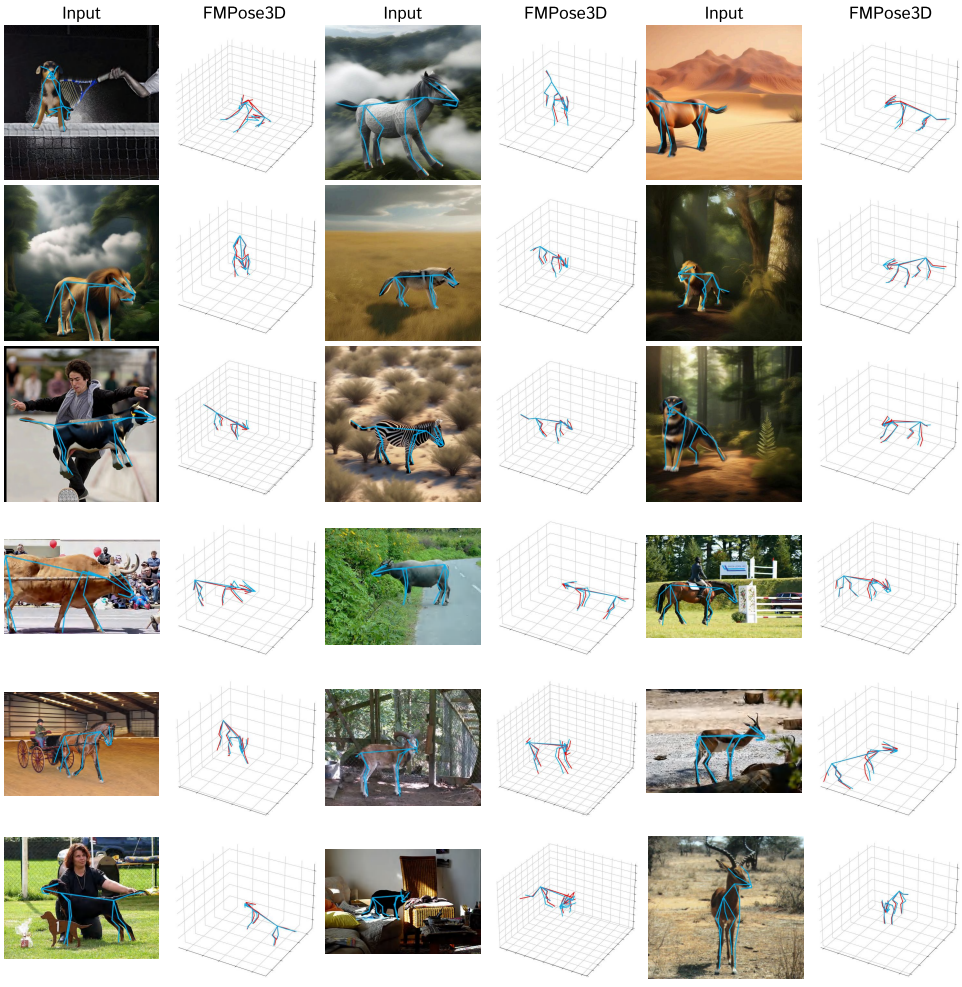

Figure 12. Qualitative results on CtrlAni3D (top three rows) and Animal3D (bottom three rows). The blue pose represents the predicted results, while the red pose represents the ground truth.

green screen (GS, fourth row), studio without green screen (noGS, fifth row), and outdoor scenes (Outdoor, sixth row).

To further assess the generalization ability of our model to outdoor scenarios, we evaluate the model pre-trained on Human3.6M using samples from the 3DPW [57] dataset. Figure 11 presents the qualitative results. The 2D poses are obtained using HRNet [59]. The results indicate that our model generalizes well to unconstrained in-the-wild environments.

Figure 12 further shows qualitative results on animal datasets, including synthetic samples from CtrlAni3D (top three rows) and real-world samples from Animal3D (bottom three rows). Across these challenging cases, our approach consistently produces reliable and anatomically plausible 3D pose predictions.