A generic diffusion-based approach for 3D human pose prediction in the wild

Saeed Saadatnejad[1] **Ali Rasekh** **Mohammadreza Mofayezi**[2] **Yasamin Medghalchi**[2]

Sara Rajabzadeh[2] **Taylor Mordan**[1] **Alexandre Alahi**[1]

Abstract

3D human pose forecasting, i.e., predicting a sequence of future human 3D poses given a sequence of past observed ones, is a challenging spatio-temporal task. It is more challenging in real-world applications where occlusions will inevitably happen, and estimated 3D coordinates of joints would contain some noise. We provide a unified formulation in which incomplete elements (no matter in the prediction or observation) are treated as noise, and propose a conditional diffusion model that denoises them and forecasts plausible poses. Instead of naively predicting all future frames at once, our model consists of two cascaded sub-models, each specialized for modeling short and long horizon distributions. We also propose a repairing step to improve the performance of any 3D pose forecasting model in the wild, by leveraging our diffusion model to repair the inputs. We investigate our findings on several datasets, and obtain significant improvements over the state of the art.

1 Introduction

Predicting 3D human pose, the task of predicting a sequence of future 3D poses of a person given a sequence of past observed ones, is a challenging task to solve, as it mixes spatial and temporal reasoning, and has multiple modes. While previous models have shown acceptable accurate predictions [15, 13], they perform poorly in imperfect observation settings. In the real world, noise exists in the perceived motion of a person due to sensor errors and occlusions by the same person, other objects in the scene, and other people.

Denoising Diffusion Probabilistic Models (DDPM) [8] are one kind of generative models that denoise input signal iteratively and have shown high-quality image synthesis [21, 20]. Motivated by this property, we propose a diffusion model that explicitly handles noisy data input so not only it predicts accurate and in-distribution poses, but can also be used in the wild. We make a full sequence of observation and future frames by putting noise for the incomplete observation joints and future poses. This sequence is fed to our model for denoising and in T steps, the correct predictions and refinements are achieved by 1) predicting poses for the future frames, and 2) repairing the imperfect observation when there is occlusion, missing whole frame, or noisy observation. Naively predicting all future frames at once leads to inaccurate predictions in later frames, so we break the problem into two simpler tasks and our model consists of two temporally cascaded diffusion blocks. The former predicts the short-term poses and repairs the imperfect observations (if applicable), and the latter takes the output of the former as a condition and predicts the long-term poses. We also introduce a simpler version of our model that could improve the performance of the state-of-the-art models in a black-box manner. We repair the imperfect observation and feed the pseudo-clean data to any prediction model to predict reasonably.

[1]EPFL, Lausanne, Switzerland

[2]Sharif University of Technology, Tehran, Iran

NeurIPS 2022 Workshop on Score-Based Methods.

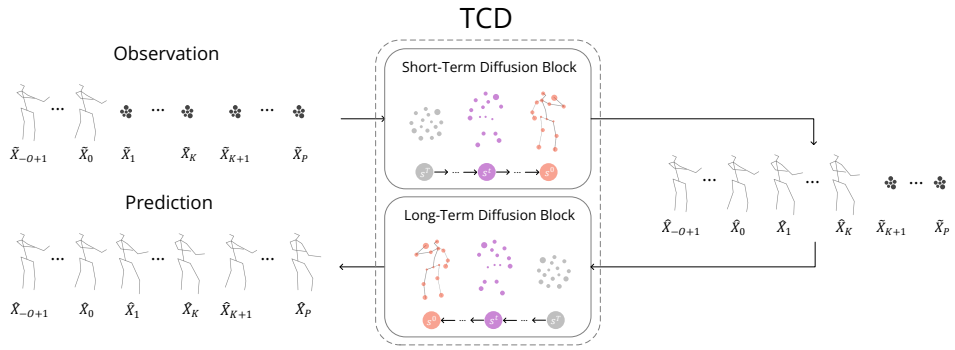

Figure 1: Overview of our Temporal Cascaded Diffusion (TCD). The short-term diffusion block (top) takes the observed sequence padded with random noise and predicts short-term human poses in K frames. The predicted sequence along with the observation padded with random noise is given to the long-term diffusion block (bottom) to predict for all P frames.

To summarize, we propose a two-level diffusion model for long-term 3D human pose prediction in perfect and imperfect input observation settings. We perform extensive experiments and obtain significant improvements over the state of the art. To the best of our knowledge, ours is the first diffusion model for human pose prediction. We also introduce a repairing leveraging our model that can be used with any pose prediction model.

2 Method

2.1 Problem Definition and Notations

Let $X = [X_{-O+1}, X_{-O+2}, \ldots, X_0, X_1, \ldots, X_P] \in \mathbb{R}^{(O+P)\times J\times 3}$ be a clean complete normalized sequence of human body poses with J joints in O frames of observation and P frames of future. Each joint consists of its 3D cartesian coordinates. The availability mask is a binary matrix $M \in \{0,1\}^{(O+P)\times J\times 3}$ where zero determines the parts of the sequence that are not observed due to occlusions or being from future timesteps. Note that the elements of M correspondent to P future frames are always zero. With this notation, the observed sequence $\tilde{X} = [\tilde{X}_{-O+1}, \tilde{X}_{-O+2}, \ldots, \tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_P]$ is derived by applying the element-wise product of M into X and adding a Gaussian noise $\epsilon \sim \mathcal{N}(0,\, I)$ in non-masked area $\tilde{X} = M \odot X + (1 - M)\epsilon$. The model predicts $\hat{X} = [\hat{X}_{-O+1}, \hat{X}_{-O+2}, \ldots, \hat{X}_0, \hat{X}_1, \ldots, \hat{X}_P]$ and the objective is lowering $|\hat{X} - X| \odot (1 - M)$ given $\tilde{X}$.

2.2 Conditional Diffusion Blocks

We propose a conditional diffusion block inspired by [25]. It has multiple residual layers and each layer contains two cascaded transformers with the same input and output shapes. The temporal transformer is responsible for modeling the temporal behavior of data. Its output is fed to the spatial transformer for attending to the body pose inside each frame. For more details, refer to the source code.

At training time, a gaussian noise with zero mean and pre-defined variance is added to the input pose sequence s^0 to make it noisier s^1. This process is repeated for T steps following the markov chain, thus, the output s^T will be close to a pure gaussian noise in the non-masked area:

$$q(s^t|s^{t-1}) = M \odot s^0 + (1 - M) \odot \mathcal{N}(s^t; \sqrt{1-\beta^t}s^{t-1}, \beta^t\mathbf{I}), \tag{1}$$

where β^t the variance of the noise in step t is determined using a scheduler. We utilize the cosine noise scheduler, first introduced in [19]:

$$\beta^t = 1 - \frac{f(t)}{f(t-1)}, \quad f(t) = \cos^2\left(\frac{t/T + c}{1+c} \cdot \frac{\pi}{2}\right), \tag{2}$$

where c is a small offset and is set to 0.008 empirically. Using the above function helps in slowly decreasing the quality of the input compared to commonly used schedulers like quadratic and

Model	80ms	320ms	560ms	720ms	880ms	1000ms
Zero-Vel	23.8	76.0	107.4	121.6	131.6	136.6
Res. Sup. [18]	25.0	77.0	106.3	119.4	130.0	136.6
ConvSeq2Seq [10]	16.6	61.4	90.7	104.7	116.7	124.2
LTD-50-25 [17]	12.2	50.7	79.6	93.6	105.2	112.4
HRI [15]	10.4	47.1	77.3	91.8	104.1	112.1
PGBIG [13]	10.3	**46.6**	76.3	90.9	102.6	110.0
TCD (ours)	**9.9**	48.8	**73.7**	**84.0**	**94.3**	**103.3**

Table 1: Comparison on perfect observation data on Human3.6M [9] in FDE (mm) at different horizons.

linear. Therefore, the input information remains longer and the variances of noises are learned more accurately. The network learns to reverse the diffusion process and retrieve the clean sequence by predicting the cumulative noise that is added to s^t as described in DDPM [8]:

$$Loss = E_{t,s^0,\epsilon}||\epsilon - \epsilon_\theta(s^t, t)||_2^2. \tag{3}$$

At inference time, the model starts from an incomplete noisy observed sequence s^T, where we put a gaussian noise in the non-masked area and observation in the masked area. Then, it predicts the poses $s^{T-1} \dots s^0$ through an iterative process in which we subtract the learned additive noise at each step from the output of the previous iteration until it yields a clean output close to the ground truth.

2.3 Temporal Cascaded Diffusion (TCD)

We illustrate our main model which consists of a short-term and a long-term diffusion blocks in Figure 1. The first block is specialized for pose prediction in short term. The input of this block is $\tilde{X}$ and the model predicts the first K frames of the future $\hat{X}_1 \dots \hat{X}_K$ and completes the observation frames $[\hat{X}_{-O+1} \dots \hat{X}_0]$. The task of the second block is to predict the rest of the future frames $[\hat{X}_{K+1} \dots \hat{X}_P]$ given the observation and the output of the first block. Note that two sub-models are trained separately using clean complete input but in inference time, we give the average of 5 samples of the short-term block to the long-term predictor.

2.4 Sequence Repairing

Since most of the existing pose prediction models cannot handle imperfect observations, we propose a simpler version of our model to repair the observation only. This module takes the imperfect observed sequence $[\tilde{X}_{-O+1}, \tilde{X}_{-O+2}, \dots, \tilde{X}_0]$ as input and repaires it to $[\hat{X}_{-O+1}, \hat{X}_{-O+2}, \dots, \hat{X}_0]$. The architecture is similar to TCD yet predicts in one level and the input and output sequences have O frames. Our accurate repairing enables any pose prediction model trained on complete data to predict reasonably.

3 Experiments

3.1 Experimental Setup

Human3.6M [9] is the largest benchmark dataset for human motion analysis, with 3.6 million body poses. The original 3D pose skeletons in the dataset consist of 32 joints. Previous works reported their performances in different settings. In our experiments, we have 50 observation frames, 25 prediction frames down-sampled to 25 fps, with the subset of 22 joints to represent the human pose.

Our diffusion model has 12 layers of residual blocks and 50 steps. In TCD, the length of short-term prediction K is 20% of the total prediction length P. Each transformer has 64 channels and 8 attention heads.

We measure the Displacement Error (DE), in millimeters (mm), over all joints in a frame and report the Average Displacement Error (ADE) as the average of DE for the whole sequence and Final

Model	80ms	320ms	560ms	720ms	880ms	1000ms
Zero-Vel	84.9	138.2	169.9	184.2	193.7	198.2
HRI [15]	65.2	104.5	130.0	141.6	151.1	157.1
PGBIG [13]	67.0	107.1	132.1	143.5	152.9	158.8
TCD (ours)	**11.2**	**51.3**	**75.4**	**85.4**	**95.4**	**104.5**
Pre(ours) + Zero-Vel	24.1	76.3	107.6	121.7	131.7	136.7
Pre(ours) + HRI [15]	11.4	48.6	78.3	92.7	105.0	112.8
Pre(ours) + PGBIG [13]	11.1	**47.9**	77.2	91.7	103.5	110.8
Pre(ours) + TCD (ours)	**10.8**	49.9	**74.4**	**84.9**	**95.1**	**104.2**

Table 2: Comparison on imperfect observation data and pre-processed observation data (Pre(ours)+) on Human3.6M [9] in FDE (mm) at different horizons.

Displacement Error (FDE) as DE in the final predicted frame. Note that *Zero-Vel* is a simple model that outputs the last observed pose as the predictions for all future poses.

3.2 Results

The results of our model and comparison with previous works in perfect input settings are reported in Table 1. Ours performs better than previous works in the short-term and with a larger margin in the long-term, thanks to our two-level prediction.

We now look at how models perform with imperfect observation data. This is a more realistic scenario than assuming perfect inputs, as occlusions would happen and the estimated 3D coordinates of joints would contain some noise in practice. We randomly remove 40% of the left arm and right leg from the observations of Human3.6M, at both training and evaluation, to simulate occlusions. The state-of-the-art perform poorly on imperfect observation as shown in the top half of Table 2 while ours perform close to the perfect input observation. If our repairing module is added to generate pseudo-perfect observation feeding to the state-of-the-art models and Zero-Vel, the performances significantly improve. In fact, the the results will be close to perfect observation data reported in Table 1.

Extensive experiments can be found in the appendix and complete version of the paper.

4 Conclusion

In this work, we have addressed the task of 3D human pose prediction in imperfect settings. We have proposed a diffusion model suitable to imperfect input data observations happening in the wild. Our model predicts future poses in two levels (short-term and long-term) to better capture human motion dynamics, and yields state-of-the-art results. We have then leveraged it as a repairing step easily applicable to any existing predictor in a black box manner. It rectifies noisy observations before feeding them to the predictor. This paper paves the way for modeling human full-body motions using diffusion models. Further studies in increasing the run-time speed are suggested for real-world deployment.

Acknowledgments and Disclosure of Funding

The authors would like to thank Mohammadhossein Bahari and Bastien Van Delft for their helpful comments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754354.

References

[1] Emad Barsoum, John R. Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–149909, 2018.

[2] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2018.

[3] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4801–4810, June 2021.

[4] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6519–6527, 2020.

[5] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016.

[6] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4941–4949, 2017.

[7] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, jul 2014.

[10] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234, 2018.

[11] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020.

[12] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8161–8171, June 2022.

[13] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6437–6446, June 2022.

[14] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[15] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

[16] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021.

[17] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[18] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017.

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[22] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6457–6466, June 2022.

[23] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010.

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

[25] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

[26] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[27] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 3332–3341, 2017.

[28] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[29] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019.

[30] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

A Appendix

Here, we compare all approaches on four standard 3D human pose forecasting datasets:

Human3.6M [9] is the largest benchmark dataset for human motion analysis, with 3.6 million body poses. The validation set is subject-11, the test set is subject-5, and all the remaining five subjects are training samples. The original 3D pose skeletons in the dataset consist of 32 joints. Previous works reported their performances in different settings. To have a thorough and correct comparison, we define the following settings:

- **Setting-A**: 25 observation frames, 100 prediction frames at 50 fps (frame per second), with the subset of 17 joints to represent the human pose;
- **Setting-B**: 50 observation frames, 25 prediction frames down-sampled to 25 fps, with the subset of 22 joints to represent the human pose;
- **Setting-C**: 25 observation frames, 25 prediction frames down-sampled to 25 fps, with the subset of 17 joints.

AMASS (The Archive of Motion Capture as Surface Shapes) [14] is a recently published human motion dataset that unifies 18 motion capture datasets totaling 13,944 motion sequences from 460 subjects performing a large variety of actions. We use 50 observation frames down-sampled to 25fps with 18 joints similar to previous works.

3DPW (3D Poses in the Wild) [26] is the first dataset with accurate 3D poses in the wild. It contains 60 video sequences taken from a moving phone camera. Each pose is described as a 18-joint skeleton with 3D coordinates similar to AMASS dataset. We use the official instructions to obtain training, validation, and test sets.

HumanEva-I [23] includes 3 subjects that perform different actions captured at 60fps. Each person has 15 body joints. We remove the global translation and use the official train/test split of the dataset. In this dataset, the prediction horizon is 60 frames (1s) given 15 observed frames (0.25s) similar to [16].

We measure the Displacement Error (DE), in millimeters (mm), over all joints in a frame and report the Average Displacement Error (ADE) as the average of DE for the whole sequence and Final Displacement Error (FDE) as DE in the final predicted frame. Similar to [16], we also report the multi-modal versions of ADE (MMADE) and FDE (MMFDE). We follow the same evaluation

Model	Human3.6M [9]					HumanEva-I [23]		
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓
Pose-Knows [27]	6723	461	560	522	569	2308	269	296
MT-VAE [28]	403	457	595	716	883	21	345	403
HP-GAN [1]	7214	858	867	847	858	1139	772	749
BoM [2]	6265	448	533	514	544	2846	271	279
GMVAE [5]	6769	461	555	524	566	2443	305	345
DeLiGAN [6]	6509	483	534	520	545	2177	306	322
DSF [29]	9330	493	592	550	599	4538	273	290
DLow [30]	11741	425	518	495	531	4855	251	268
Motron [22]	7168	375	488	–	–	–	–	–
Multi-Obj [12]	14240	414	516	–	–	5786	228	236
GSPS [16]	14757	389	496	476	525	5825	233	244
TCD (ours)	**19466**	**356**	**396**	**463**	**445**	**6764**	**199**	**215**

Table 3: Comparison with stochastic models on Human3.6M [9] Setting-A and HumanEva-I [23] at a horizon of 2s.

Model	AMASS [14]				3DPW [26]			
	560ms	720ms	880ms	1000ms	560ms	720ms	880ms	1000ms
Zero-Vel	130.1	135.0	127.2	119.4	93.8	100.4	102.0	101.2
convSeq2Seq [10]	79.0	87.0	91.5	93.5	69.4	77.0	83.6	87.8
LTD-10-25 [17]	57.2	65.7	71.3	75.2	57.9	65.8	71.5	75.5
HRI [15]	51.7	58.6	63.4	67.2	56.0	63.6	69.7	73.7
TCD (ours)	**49.8**	**54.5**	**60.1**	**66.7**	**55.4**	**61.6**	**67.9**	**73.4**

Table 4: Comparison with deterministic models on AMASS [14] and 3DPW [26] in FDE (mm) at long horizons.

protocol as in [30] to measure diversity and report the Average Pairwise Distance (APD) between different predictions.

A.1 More Results on Perfect Observation Data

We evaluate our model on two datasets, Human3.6M [9] Setting-A and HumanEva-I [23], and compare it with other approaches in Table 3. Each model is sampled 50 times given each observation sequence. TCD (ours) clearly performs better than previous works in terms of the accuracy of the best sample (ADE and FDE) and multiple samples (MMADE and MMFDE) while generating diverse poses (APD).

The large long-term improvement can be observed in AMASS [14] and 3DPW [26], too. Similar to previous works, we train our model on AMASS and measure FDE on both datasets. The comparison with models that reported in this setting is shown in Table 4.

Qualitative results on Human3.6M are shown in Figure 2. Predictions from our model are displayed along with predictions from several baselines and are superimposed on the ground-truth poses for direct comparison. Our model has correctly learned the data distribution and predicts accurate and plausible poses. For instance, hand movement is natural when the feet move while HRI shows fixed hands and PGBIG has a momentum that avoids large hand movements.

A.2 More Results on Imperfect Observation Data

MT-GCN [3] predicts in incomplete observation settings and reported the performance of some previous models when the input is repaired using their own method and the results on Human3.6M Setting-C is in the first column of Table 5. Ours outperforms MT-GCN by a significant margin of 33.2mm in FDE at 1s horizon (30% improvement). MT-GCN implicitly ignores noise in data [3]. However, we explicitly denoise the input and it leads to a generalizable solution.

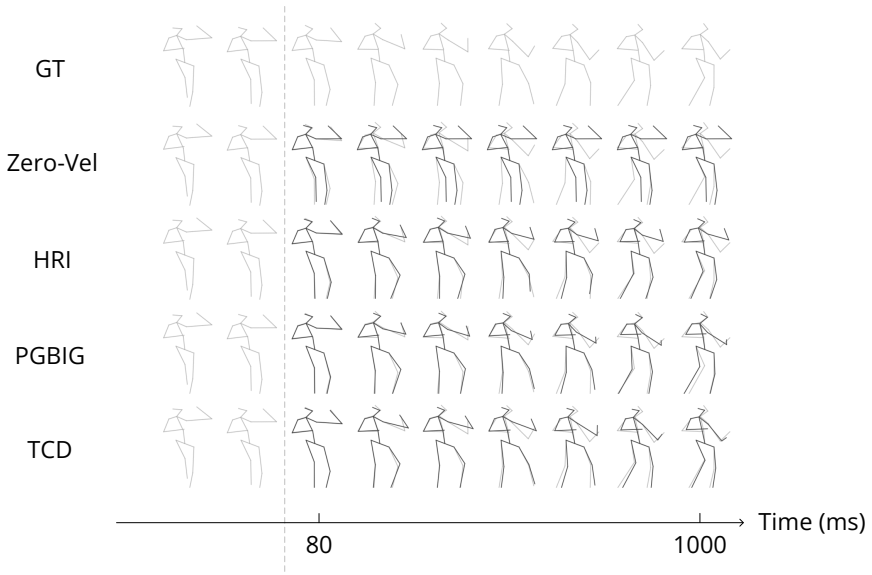

Figure 2: Qualitative results on Human3.6M [9] Setting-B. For each row, the left part is the input observation, and the predicted poses superimposed on the ground truth are displayed on the right.

Model	Random Leg, Arm Occlusions	Structured Joint Occlusions	Missing Frames	Noisy Inputs $\sigma = 25$	Noisy Inputs $\sigma = 50$
R+TrajGCN [17]	121.1	131.5	–	127.1	135.0
R+LDRGCN [4]	118.7	127.1	–	126.4	133.6
R+DMGCN [11]	117.6	126.5	–	124.4	132.7
R+STMIGAN [7]	129.5	128.2	–	–	–
MT-GCN [3]	110.7	114.5	122.0	114.3	119.7
TCD (ours)	**77.5**	**77.2**	**80.5**	**81.9**	**84.9**

Table 5: Comparison on imperfect observation data on Human3.6M [9] Setting-C in FDE (mm) at a horizon of 1s. Models in the top part of the table receive repaired sequences (R+) while others receive imperfect sequences.

We analyze the performance of our model in several occlusion patterns masks M applied to input data:

- Random Leg, Arm Occlusions: leg and arm joints are randomly occluded with the same probability of 40%;
- Structured Joint Occlusions: 40% of the right leg joints for consecutive frames are missing;
- Missing Frames: 20% of consecutive frames are missing;
- Noisy Inputs: Gaussian noise with a standard deviation of $\sigma = 25$ or $\sigma = 50$ is added to the coordinates of the joints, and 50% of the leg joints are randomly occluded.

Table 5 presents the results when training and evaluating in the above observation patterns, in FDE at a prediction horizon of 1 second on Human3.6M Setting-C. Our model outperforms previous works in different patterns of occlusions and noises in input that may happen in the real world. We observed that missing 5 consecutive frames is harder than missing a part of the body in 10 consecutive frames as the network can recover the former with spatial information.

To have a thorough comparison with MT-GCN, we train 4 models on several percentages of random joints missing in the observation pose sequence and the performance of sequence repairing (ADE of the occluded observation sequence) and motion prediction (FDE at 1-second horizon) is presented in Table 6. Our model provides a negligible error of 2.9mm in repairing with 40% of all joints missing,

	Train and Test Missing Ratio			
Model	10%	20%	30%	40%
MT-GCN [3]	109.4 / 8.6	110.5 / 13.7	112.3 / 18.7	114.4 / 24.5
TCD (ours)	**77.1 / 2.2**	**77.2 / 2.3**	**77.6 / 2.6**	**79.1 / 2.9**

Table 6: Results of motion prediction and sequence repairing on Human3.6M [9] Setting-C with varying amounts of randomly occluded joints in input data in FDE (mm) at a horizon of 1s / ADE (mm) of missing elements.

while MT-GCN gives 24.5mm of error. In forecasting, ours achieves more than 31% lower FDE compared to MT-GCN.

A.3 Ablations Studies

Here, we investigate different design choices of the network and report ADE (mm) on Human3.6M [9]. The full model gives an ADE of 63.3mm. Predicting in one level, i.e., 100% at once without having short and long prediction blocks, can increase ADE to 65.5mm mainly caused by wrong predictions in longer horizons. Predicting in 3 levels, i.e., predicting 20%, 20%, and 60% in a sequential manner, lowers the performance to 66.9 because cascading multiple stochastic processes generates either random results or not diverse. The same effect exists for K where a smaller $K = 2$ fades the benefit of 2-level prediction (ADE of 65.1mm) and a larger $K = 10$ makes the task of short-term prediction harder thus ADE increases to 66.6mm.

We have tested a quadratic scheduler instead of our cosine scheduler and it increased ADE by 1mm. Decreasing the number of residual layers in our diffusion blocks from 12 to 4 lowers the performance by 3mm. More than 12 residual layers has a large negative effect on sampling time. Moreover, we have done several experiments on the architecture of the transformers and observed that spatial transformer and time transformer both help in learning the spatio-temporal features of pose sequence, and eliminating them one at a time degrades ADE to 74.5mm and 261.1mm, respectively.

In terms of run-time speed, our model predicts each sequence of 1 second length in around 0.8 seconds which means real-time in our setup. Diffusion models are inherently slow because of their iterative process, however, recent models such as denoising diffusion implicit models (DDIM) could achieve faster sampling in image synthesis [24]. More studies in this direction are suggested for the future.